

Sprachkommunikation

– Sprachsignalverarbeitung und Sprachtechnologie –

Prof. Dr.-Ing. Sebastian Möller
Quality and Usability Lab
Technische Universität Berlin
Wintersemester 2022/2023

Organisation

Die Veranstaltung wird als Vorlesung mit integrierten Übungen durchgeführt, wobei einige Übungsaufgaben zu Hause gerechnet bzw. implementiert werden sollten.

- *Vorlesung*: Montags, 10-12 Uhr, Zoom Meeting und MA001
- *Übungen*: Montags, 12-14 Uhr, Zoom Meeting

Übersicht

Sprache ist das wichtigste Kommunikationsmittel des Menschen, und es entwickelt sich zunehmend auch zu einer wichtigen Modalität bei der Mensch-Maschine-Interaktion. So sind bereits heute Systeme verfügbar, die über eine Spracherkennung, eine Interpretation sprachlicher Inhalte, eine Steuerung des Dialogverlaufes, eine Generierung von Antworten, sowie eine Erzeugung von Sprachsignalen verfügen. Darüber hinaus kommt der effizienten Übertragung von Sprache eine überaus große Bedeutung zu, sowohl in herkömmlichen als auch in paketvermittelten Netzen (bspw. *Voice over IP*).

Im Rahmen dieser Vorlesung sollen die Grundlagen für das Verständnis und die Gestaltung kommunikationstechnischer Systeme gelegt werden, die auf Sprache beruhen. Ausgangspunkt dazu ist die Erzeugung und die Wahrnehmung natürlicher Sprache durch den Menschen, da sich hieraus viele wichtige Eigenschaften von Sprachsignalen sowie Anforderungen an ihre Verarbeitung ergeben. Hierzu werden die Grundlagen einer Beschreibung von Sprachsignalen im Zeit- und Frequenzbereich gelegt. Auf Basis dieser Grundlagen wird die Funktion wichtiger Komponenten sprachtechnologischer Systeme erläutert. Neben der effizienten Kodierung von Sprache stehen dabei die Spracherkennung, die Sprachsynthese sowie die Interaktion mittels sprachverarbeitender Systeme (Sprachdialogsysteme, auch *Voice User Interfaces*) im Mittelpunkt. Abschließend wird gezeigt, wie solche Systeme mittels multimodaler Ein- und Ausgabeschnittstellen noch besser an die menschlichen Kommunikationsbedürfnisse angepasst werden können (multimodale Dialogsysteme).

Die Vorlesung richtet sich gleichermaßen an Studierende der Elektrotechnik, der technischen Informatik sowie der Informatik. Darüber hinaus sind auch Studierende aus den Sprach- und Kommunikationswissenschaften, der Technischen Akustik, der Soziologie und dem Bereich Human Factors sowie aus anderen Fachbereichen herzlich willkommen. Die Vorlesung setzt keine Vorkenntnisse im Bereich der Signalverarbeitung oder der Linguistik voraus.

Inhalt

1. Motivation und Zielsetzung.....	5
1.1 Was ist Sprache	5
1.2 Anwendungen	7
1.3 Marktentwicklung	10
1.4 Kommunikationsmodelle	11
1.5 Literatur.....	13
2. Sprachsignaldarstellung und -eigenschaften	14
2.1 Darstellung kontinuierlicher Signale im Zeitbereich	14
2.2 LTI-Systeme	15
2.3 Darstellung kontinuierlicher Signale im Frequenzbereich, Spektrum	16
2.4 Impulsantwort und Übertragungsfunktion	20
2.5 Statistische Beschreibung von Sprachsignalen.....	21
2.6 Energiedichte- und Leistungsdichtespektrum.....	29
2.7 Darstellung diskreter Signale im Zeit- und Frequenzbereich	30
2.8 Langzeit- und Kurzzeit-Signaleigenschaften	36
2.9 Spektrogramm.....	38
2.10 Amplitudenverteilung	39
2.11 Literatur.....	41
3. Grundlagen der menschlichen Spracherzeugung	42
3.1 Anatomie des menschlichen Sprechapparates	42
3.2 Anregung.....	44
3.3 Lautformung	46
3.4 Sprachlaute.....	48
3.5 Modelle der Spracherzeugung	52
3.6 Literatur.....	57
4. Sprachsignalanalyse	58
4.1 Spektralanalyse	58
4.2 Cepstrum	64
4.3 Lineare Prädiktion.....	66
4.4 Literatur.....	69
5. Grundlagen der auditiven Wahrnehmung	70
5.1 Außenohr.....	70
5.2 Mittelohr	72
5.3 Innenoehr und Nervensystem	72
5.4 Frequenzauflösung und Tonhöhenwahrnehmung	76
5.5 Lautheitswahrnehmung	78
5.6 Literatur.....	83
6. Sprachsignalübertragung und -kodierung.....	84
6.1 Klassen von Sprach- und Audiosignalcodierern.....	84
6.2 Quantisierung	86
6.2.1 Lineare Quantisierung, Pulse-Code-Modulation (PCM)	86
6.2.2 Nichtlineare Quantisierung	89
6.2.3 Optimalquantisierung	92
6.2.4 Adaptive Quantisierung.....	93

6.2.5 Vektorquantisierung.....	96
6.3 Differentielle PCM.....	98
6.4 Parametrische Kodierung.....	103
6.5 Sprachkodierung bei mittleren Bitraten, Hybrid-Kodierung	107
6.5.1 RELP-Codierung.....	110
6.5.2 CELP-Codierung.....	114
6.6 Sprachkodierung im Frequenzbereich	115
6.6.1 Transformationskodierung	116
6.6.2 Teilbandkodierung	118
6.7 Kriterien zur Auswahl und Beurteilung von Kodierern.....	119
6.8 Literatur.....	120
7. Sprachtechnologische Systeme.....	121
7.1 Spracherkennung.....	121
7.1.1 Problemstellung.....	121
7.1.2 Aufbau eines Spracherkenners.....	122
7.1.3 Merkmalsextraktion	123
7.1.4 Hidden-Markov-Modelle und neuronale Netze	126
7.1.5 Sprachmodelle	130
7.1.6 Erkennungsleistungen	131
7.1.7 Literatur.....	132
7.2 Sprachsynthese.....	134
7.2.1 Struktur eines Vorleseautomaten	136
7.2.2 Symbolische Verarbeitung	137
7.2.3 Prosodiegenerierung.....	138
7.2.4 Sprachsignalgenerierung	139
7.2.5 Literatur.....	145
7.3 Natürlichsprachliche Dialogsysteme.....	146
7.3.1 Struktur eines Sprachdialogsystems.....	147
7.3.2 Spracherkennung.....	149
7.3.3 Sprachverstehen	150
7.3.4 Dialogmanagement.....	151
7.3.5 Sprachausgabe	154
7.3.6 Beispiele	154
7.3.7 Literatur.....	155
8. Multimodale Dialogsysteme	157
8.1 Eigenschaften von Modalitäten	158
8.2 Allgemeine Architektur eines multimodalen Dialogsystems.....	161
8.3 Multimodale Eingabe-Schnittstellen	163
8.4 Multimodale Verarbeitung	167
8.5 Multimodale Ausgabe-Schnittstellen	168
8.6 Systembeispiele	171
8.7 Literatur.....	171

1. Motivation und Zielsetzung

1.1 Was ist Sprache

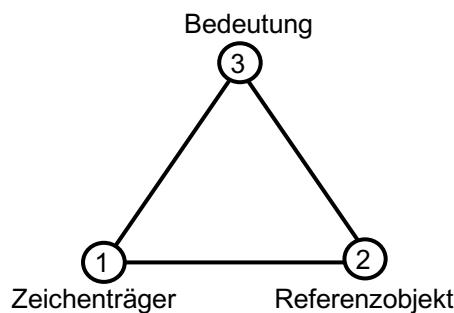
Blitzlicht:

-
-
-
-
-
-

Definition Linguistisches Wörterbuch (Lewandowski, 1994, p. 994):

„Die natürliche Sprache ist eine typisch menschliche und zugleich gesellschaftliche Erscheinung; sie ist das primäre System von Zeichen, ein Werkzeug des Denkens und Handelns und das wichtigste Kommunikationsmittel.“

Sprache als System von Zeichen → *Semiotische Betrachtung* von Sprache:



Semiotisches Dreieck, Darstellung nach Nöth (2000, 140)

Darstellung des Zeichens als Triade mit drei Korrelaten (nach Peirce, Ogden und Richards, u.a.):

- Zeichenträger (Repräsentamen, Symbol)
- Bedeutung (Interpretant, Gedanke oder Referenz)
- Referenzobjekt (Objekt, Referent)

Auch andere (monadische, dyadische) Darstellungen möglich (vgl. Nöth, 2000, 136-141).

Information vs. Bedeutung (Jekosch, 2000):

- Information: konventionalisierter bzw. institutionalisierter Kode; Kode ist für uns alle tatsächlich der gleiche
- Bedeutung: Ein Zeichen ruft weitere Begriffe oder Vorstellungen hervor:
 - Konnotationen (Mit-Gemeintes)
 - Assoziationen (individuelle Erfahrungen)

Formen:

Sprache lässt sich zunächst unterscheiden in

- Sprache als System (franz. *langue*)
- Realisierung dieses Systems als Sprechakte (franz. *parole*)

Daneben ist natürlich auch noch die Unterscheidung in gesprochene und geschriebene Sprache sinnvoll:

- Gesprochene Sprache (engl. *speech*): Aufgabengebiet der Sprachsignalverarbeitung (*speech signal processing*)
- Geschriebene Sprache (engl. *language*): Aufgabengebiet der Sprach-Symbolverarbeitung (*natural language processing*)

Darüber hinaus gibt es auch noch Sonderformen, z.B. Gebärdensprache für Gehörlose.

Sprachsignale und Sprachlaute:

- *Sprachsignale*: Bedeutendste Informationsträger für die zwischenmenschliche Kommunikation, aber in zunehmendem Maße auch für die Mensch-Maschine-Interaktion; dabei:
 - Elektrische Signale: Verlauf der elektrischen Größe (Strom oder Spannung) über der Zeit; werden normalerweise mit Hilfe eines Mikrofons aus dem entsprechenden akustischen Signal gewonnen
 - Akustische Signale: Verlauf der akustischen Größe (Druck oder Schnelle) über der Zeit
- *Sprachlaute*: Sprach-Hörereignisse
- *Hörereignis*: Das akustisch Wahrgenommene, Hörempfindung, Hörgegenstand. Hörereignisse sind wie alles Wahrgenommene räumlich, zeitlich und eigenschaftlich bestimmt (vgl. Blauert, 1994).

Sprachliche Einheiten:

Es stellt sich die Frage, wie viele Laute überhaupt in einer Sprache unterschieden werden können. Hierzu sollen zunächst folgende Begriffe erläutert werden:

(Sprach-) Laut:

„[...] Segment, in das sich eine sprachliche Äußerung auditiv zerlegen lässt, oder auch eine Klasse solcher, miteinander ähnlicher Segmente [...] Der Laut kann unter artikulatorischem, akustischem und auditivem Aspekt untersucht werden.“ (Lewandowski, 1994, 631).

Phonem:

Kleinste bedeutungsunterscheidende, aber nicht selbst bedeutungstragende Einheit einer Sprache, z.B. /r/ und /f/ in rein bzw. fein.

Morphem:

Kleinste selbst bedeutungstragende Einheit einer Sprache.

Diese Begrifflichkeit ist allerdings teilweise etwas vereinfacht; weitere Unterscheidungen bezeichnen die Einheiten als Phone bzw. Morphe, die in Klassen (Phoneme, Morpheme) eingeteilt werden und unterschiedlich realisiert werden können (Allophone, Allomorphe).

Im Gegensatz zum Laut ist das Phonem ein form- und funktionsbezogener Begriff. Das Phoneminventar der bekannten Sprache schwankt je nach Zählung von 25 bis zu 60 Phonemen. Phoneme schließen sich gegenseitig aus, d.h. wenn man ein Phonem gegen ein anderes auswechselt ändert sich die Bedeutung des Gesagten. Laute, die in einer Sprache Phonemcharakter haben, müssen dies nicht notwendigerweise auch in einer anderen Sprache haben (bspw. Tonhöhenänderungen in tonalen Sprachen).

Die Beziehungen zwischen den Elementen der Sprache lassen sich ebenfalls auf unterschiedlichen Ebenen spezifizieren (Bußmann, 1990, nach Vary et al., 1998, 41):

- *Syntax*: Beziehung der Zeichen untereinander (z.B. Regeln, nach denen der Plural gebildet wird oder ein Satz aus Wörtern aufgebaut wird)
- *Semantik*: Beziehung zwischen den Zeichen und dem, was sie bezeichnen, also die Lehre von der Bedeutung der sprachlichen Zeichen
- *Pragmatik*: Beziehung zwischen den Zeichen und ihrem Benutzer; repräsentieren das „Weltwissen“, das hinter einer sprachlichen Äußerung zu finden ist.

Prosodie:

Die Prosodie (aus dem Griechischen „das Hinzugesungene“) kennzeichnet das Klangbild des Gesprochenen. Sie besteht hauptsächlich aus den drei Aspekten

- *Quantität*: Dies betrifft insbes. die zeitliche Struktur der Sprache, z.B. Lautdauern, Pausendauern, Silbendauern, Sprechrythmus, Sprechgeschwindigkeit, etc.
- *Intensität* oder *Akzentuierung*: Alle Aspekte von Akzent und Betonung auf der Wort-, Satz- und Äußerungsebene
- *Intonation*: Der melodische Aspekt von Sprache, der Informationen wie Phrasierung, Satzmodus und Fokus trägt – also Informationen darüber, was für den Hörer wichtig und was unwichtig sein soll. Weiterhin trägt die Intonation auch Informationen über Haltung und Emotionen des Sprechers.

Auf der akustischen Ebene werden diese drei Aspekte durch die akustischen Parameter

- *Dauer*
- *Amplitude* und
- *Grundfrequenz* (vgl. hierzu auch Kapitel 3)

gekennzeichnet. Allerdings besteht – wie wir noch sehen werden – keine eindeutige Zuordnung zwischen den linguistischen Aspekten und den akustischen Realisierungen.

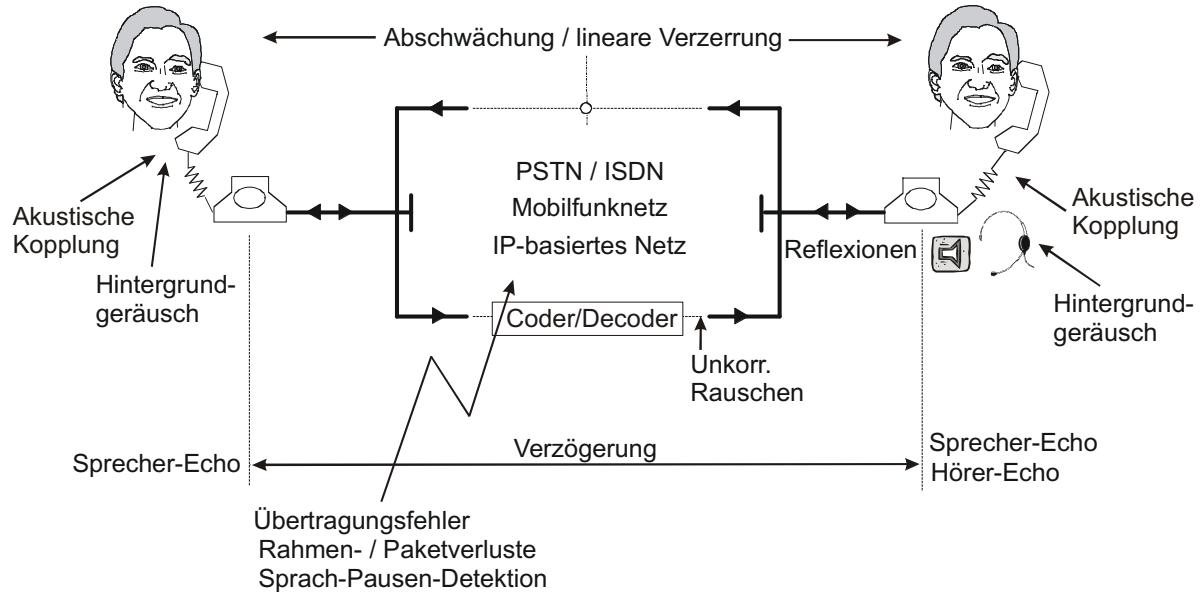
1.2 Anwendungen

Technik-vermittelte zwischenmenschliche Kommunikation:

- Sprachübertragung (Telefon, Voice-over-IP, Rundfunk, Fernsehen, Beschallungsanlagen)

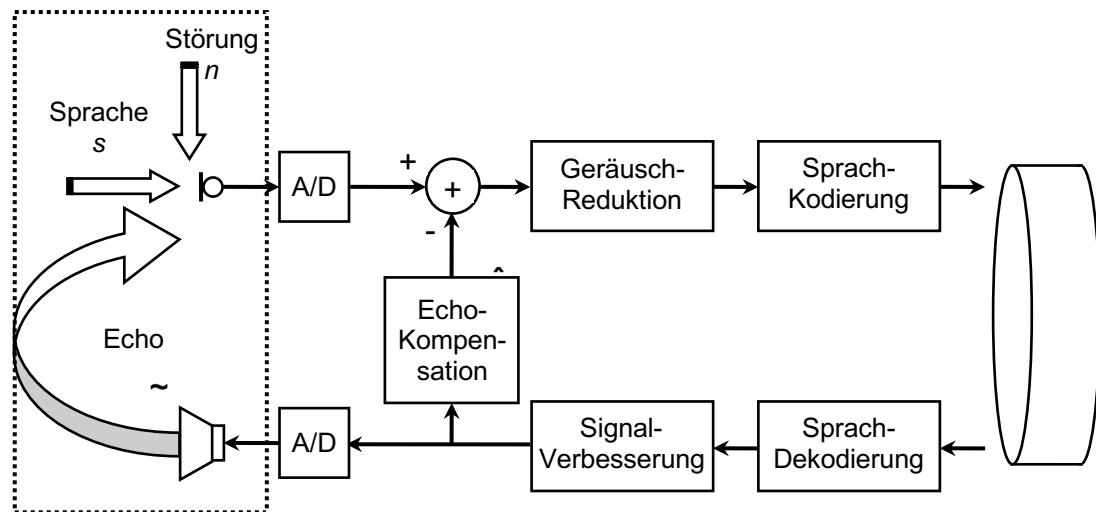
- Sprachaufzeichnung und -wiedergabe (Schallplatte, CD, DVD, MP3, Video, Kino)

Beispiel: Zwischenmenschlichen Kommunikation über einen Telefonkanal



Zwischenmenschliche Kommunikation über einen Telefonkanal (Möller, 2005).

Beispiel: Telekommunikations-Endgerät

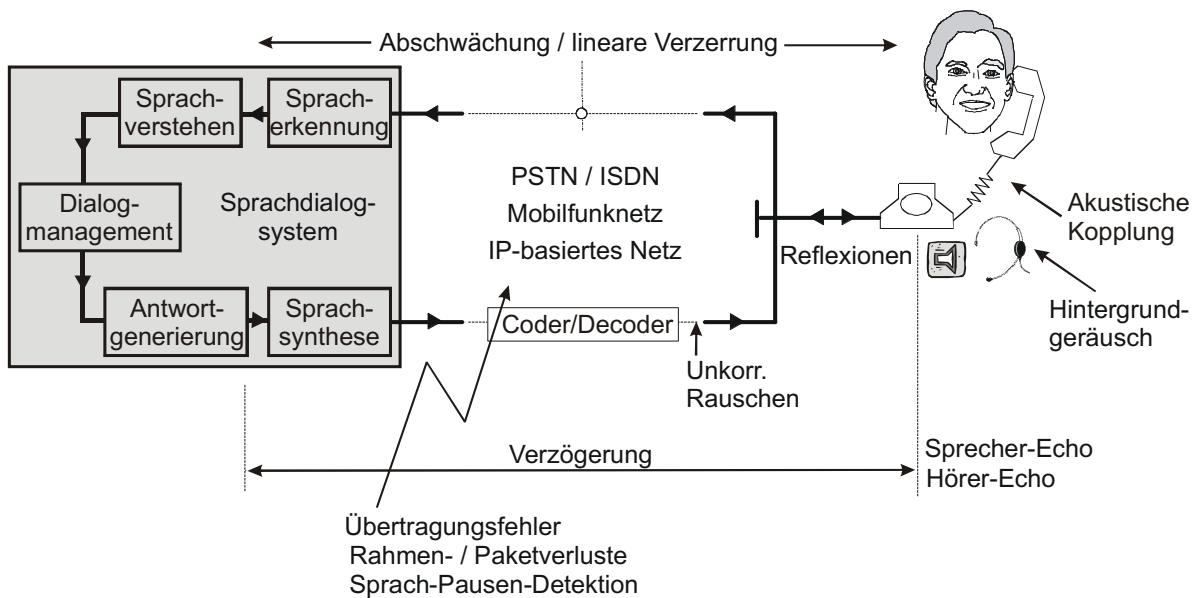


Sprachsignalverarbeitung in einem Telekommunikations-Endgerät
(vgl. Vary et al., 1998, 2).

Sprache zur Mensch-Maschine-Interaktion:

- Vorteile:
 - Das Kommunikationsmedium des Menschen
 - intuitiv und natürlich
 - erfordert keine speziellen Kenntnisse oder Erlernen
 - geeignet in *Hands-Busy-Eyes-Busy*-Situationen
 - speziell geeignet für sehbehinderte / nicht mobile Benutzer
 - von praktisch jedem Ort aus anwendbar (Telefon)
- Beispiele:
 - Maschinelle Spracheingabe (Diktieren, Sprachsteuerung)
 - Maschinelle Sprachausgabe (Vorlesesysteme)
 - Sprecheridentifikation und -verifikation (IT-Sicherheit, Forensik); wird hier nicht im Detail behandelt
 - Natürlichsprachliche Dialogsysteme

Schema der Mensch-Maschine-Interaktion über einen Telefonkanal:



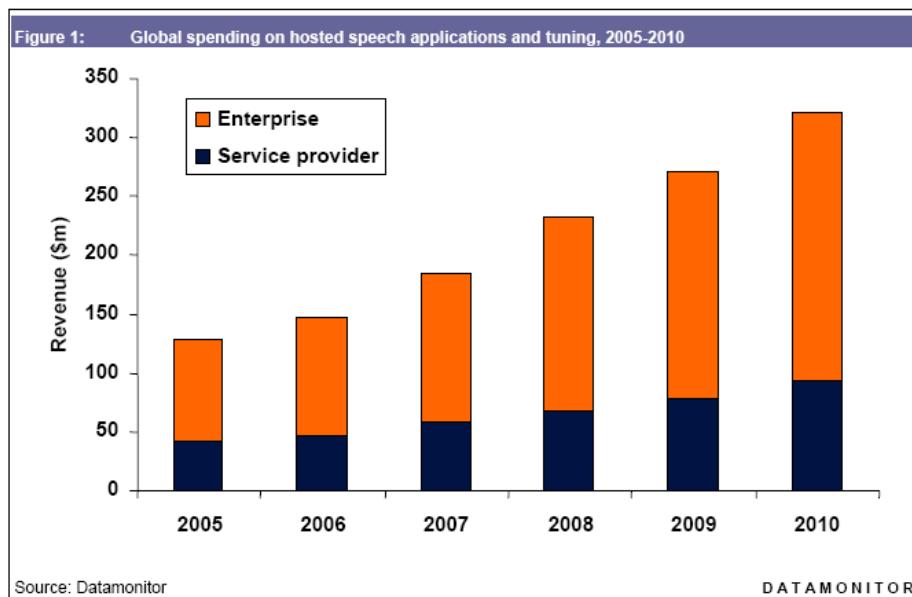
Mensch-Maschine-Interaktion über einen Telefonkanal (Möller, 2005).

Hier betrachten wir vorwiegend:

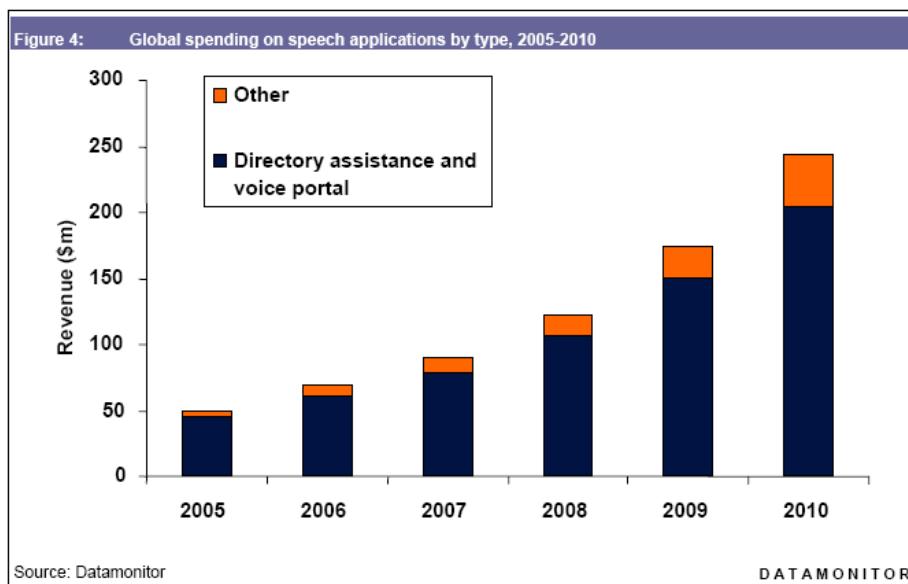
- Funktionsweise der menschlichen Spracherzeugung und -wahrnehmung
- Funktionsweise der Komponenten des maschinellen Interaktionspartners
- Funktionsweise der Sprachübertragung

1.3 Marktentwicklung

Geschätzte globale Ausgaben für Sprachtechnologie:



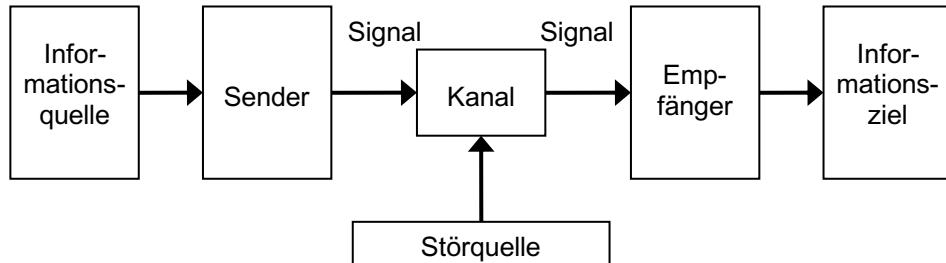
Geschätzte Ausgaben für Sprachtechnologie nach Anwendern (Datamonitor, 2010).



Geschätzte Ausgaben für Sprachtechnologie nach Applikationen (Datamonitor, 2010).

1.4 Kommunikationsmodelle

Kommunikationsmodell aus nachrichtentechnischer Sicht (Shannon und Weaver, 1949):



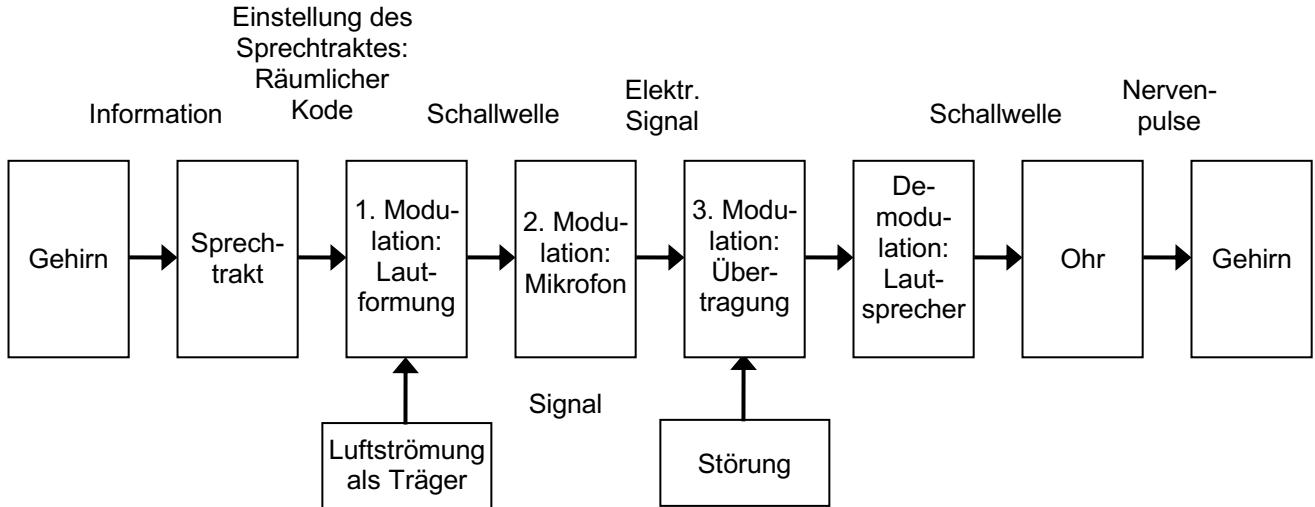
Kommunikationsmodell nach Shannon und Weaver (1949).

Dieses Modell ist rein technisch orientiert; es berücksichtigt weder die Bedeutung (den Inhalt, die Semantik) der Information (der Mitteilung oder Botschaft) noch den kommunizierenden Menschen, der meist fast gleichzeitig Sender und Empfänger (Kodierer und Dekodierer) von Mitteilungen ist, der bestimmte Gedanken und Vorstellungen hat und kommunikative Zwecke verfolgt, die er gegenüber dem Empfänger nur mit Hilfe einer gemeinsamen Sprache bzw. eines gleichen oder ähnlichen Kodes (Zeichenvorrats) realisieren kann. (Lewandowski, 1994, 556)

Schritte der Informationsübertragung, nach Dudley (1940):

- 1) Im Gehirn bereitgestellte Information wird umgesetzt in einen räumlichen Kode (Stellung von Rachen-Mund-Nasen-Trakt, Zunge, Kiefer, Gaumensegel, Wangen, Zähne, Lippen, ...) → enthält bereits charakteristische Informationen, die durch „Abtastung“ wieder gewonnen werden könnten
- 2) Nachrichtentechnisches Hilfsmittel: Modulation eines geeigneten Trägers (hier: Luftschwingungen) durch den Sprechtrakt → Anpassung an den Übertragungskanal und den Empfänger, Reichweitenerhöhung

Daraus lässt sich ein genaueres Modell der Kommunikationsstrecke ableiten, vgl. die folgende Abbildung (aus Heute, 1990).



Erweitertes Kommunikationsmodell, vgl. Heute (1990).

(Tele-) Kommunikation: Austausch von Informationen (Mitteilungen oder Botschaften) über eine bestimmte Entfernung

Dabei: Überbrückung der Entfernung durch

- Übertragung des akustischen Signals: Raumeinfluss, Störgeräusche, etc.
- Umsetzung akustisches Signal → elektrisches Signal: Mikrophon (z.B. Sprechkapsel beim Telefon)
- Übertragung des elektrischen Signals:
 - Telefonnetz, z.B. analoges (PSTN) oder digitales (ISDN) leitungsgebundenes Netz, mobiles Netz (GSM), Voice-over-IP, etc.
 - dabei wichtig: störsichere und effiziente Übertragung durch Kodierung
- Umsetzung elektrisches Signal → akustisches Signal: Lautsprecher (z.B. Hörkapsel)
- Übertragung des akustischen Signals: Raumeinfluss, Störgeräusche, etc.

Dann (oder währenddessen): Reaktion des Gesprächspartners, Formulierung einer Antwort → Rückübertragung der Antwort, s.o.

Einflussfaktoren auf die Interaktion:

- (Sprech-) Verhalten des menschlichen Kommunikationspartners
- (Sprech-) Verhalten des maschinellen Interaktionspartners
- Eigenschaften des Übertragungskanals
- Kommunikationssituation, Zweck der Kommunikation, Motivation, Erfahrung, etc.

Dabei können verschiedene *Störungen* auftreten (vgl. auch Vary et al., 1998, 7):

- Störungen des Übertragungskanals, z.B. Hintergrundgeräusche, Leitungsrauschen, Echos, Paketverluste, Verzögerungen, etc.
- Störungen der Sprachproduktion: Beim Menschen z.B. durch anatomische (Gaumenspalte) oder funktionale (z.B. als Folge eines Schlaganfalls)

Sprechstörung; bei der Maschine z.B. durch eine nicht verständliche Synthesisierung von Sprachsignalen

- Störungen der Sprachrezeption: Beim Menschen z.B. durch Schwerhörigkeit, der Maschine z.B. durch unzureichende Spracherkennung
- Kein gemeinsames Zeichensystem: Beim Menschen z.B. Fremdsprachen, Dialekte, Sozialekte, etc., bei der Maschine z.B. unzureichendes Vokabular oder unzureichende Grammatik des Spracherkenners

1.5 Literatur

- Blauert, J. (1994). Kommunikationsakustik II: Audiokommunikation und virtuelle Realität. Skriptum zur Vorlesung am Institut für Kommunikationsakustik, Ruhr-Universität, Bochum.
- Bußmann, H. (1990). Lexikon der Sprachwissenschaft. Verlag Körner, Stuttgart.
- Datamonitor (2006). Profiting from Evolving Speech Applications (Review Report). Reference Code DMTC1634, www.datamonitor.com.
- Dudley, H. (1940). The Carrier Nature of Speech. Bell Systems Technical Journal 19, 494-515.
- Gartner (2006). Hype Cycle for Enterprise Speech Technologies.
- Heute, U. (1990). Sprachverarbeitung. Skriptum zur Vorlesung der Arbeitsgruppe Digitale Signalverarbeitung, Ruhr-Universität, Bochum.
- Jekosch, U. (2000). Sprache hören und beurteilen: Ein Ansatz zur Grundlegung der Sprachqualitätsbeurteilung. Habilitationsschrift (unveröffentlicht), Fachbereich 3, Universität/GH Essen. Englische Version: Jekosch, U. (2005). Voice and Speech Quality Perception. Assessment and Evaluation. Springer, Berlin.
- Lewandowski, Th. (1994). Linguistisches Wörterbuch. 6. Auflage, Quelle & Meyer, Heidelberg.
- Möller, S. (2005). Quality of Telephone-Based Spoken Dialogue Systems. Springer, New York NY.
- Nöth, W. (2000). Handbuch der Semiotik. 2. Auflage, Verlag J.B. Metzler, Stuttgart.
- Shannon, C.E., Weaver, W. (1949). The Mathematical Theory of Communication. University of Illinois Press, Champaign IL, 1999.
- Vary, P., Heute, U., Hess, W. (1998). Digitale Sprachsignalverarbeitung. B.G. Teubner, Stuttgart.

2. Sprachsignaldarstellung und -eigenschaften

In diesem Kapitel beschäftigen wir uns mit Sprachsignalen. Bei der Erzeugung liegen diese zunächst in akustischer Form, d.h. als Schallwellen, die vom Sprecher ausgehen, vor. Die akustischen Signale können mittels elektroakustischer Wandler in elektrische Signale (d.h. Verläufe von Strom bzw. Spannung über der Zeit) um- und wieder zurückgewandelt werden. Die Funktionsprinzipien dieser Wandler werden z.B. in der Vorlesung „Kommunikationsakustik“ behandelt.

Die bei der Wandlung am Mikrofon entstehenden Signale sind zeitlich und bezüglich ihres Wertebereiches (Amplitude) kontinuierlich. Möchte man solche Signale mittels Digitalrechner verarbeiten, so muss man sie zunächst in eine Zeit- und Amplituden-diskrete Darstellung überführen; man muss sie also digitalisieren. Methoden zur Darstellung digitaler Signale werden ebenfalls in diesem Kapitel behandelt.

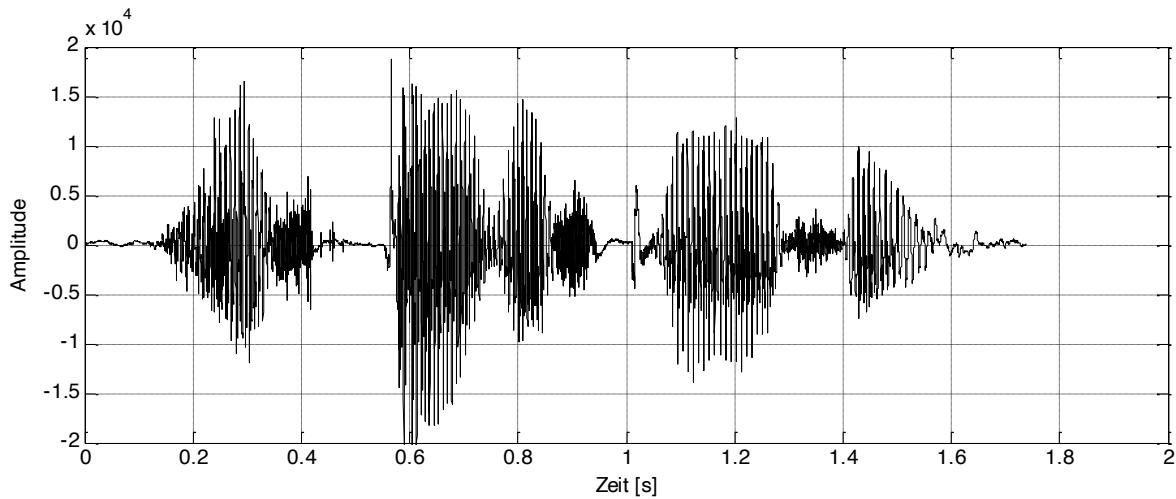
Wir beschäftigen uns also im Weiteren mit elektrischen – analog oder digital vorliegenden – Sprachsignalen. Diese können zunächst bezüglich ihres Verlaufes über der Zeit, d.h. im Zeitbereich beschrieben werden. Daneben bietet sich aber auch eine Zerlegung in Sinus- bzw. Kosinus-förmige Komponenten verschiedener Frequenz, das sogenannte Spektrum, an. Signaldarstellungen im Zeit- und Frequenzbereich sind prinzipiell identisch, allerdings bietet sich je nach Anwendungsfall die eine oder andere Art der Darstellung an. Dies gilt insbesondere dann, wenn Signale über irgendwelche Systeme übertragen werden sollen. Einige spezielle Systeme werden ebenfalls in diesem Kapitel behandelt.

Die hier behandelten Methoden sind Werkzeuge zur Beschreibung von Sprachsignalen, und einigen Studierenden wahrscheinlich aus anderen Vorlesungen bekannt. Spezielle Eigenschaften von Sprache, die sich aus deren Erzeugung beim Menschen ergeben, werden in Kapitel 3 behandelt. Darauf basieren denn verschiedene speziellere Verfahren zu deren Analyse, die sich an die Themen dieses Kapitels anschließen, jedoch erst nach Erläuterung der Spracherzeugung, also in Kapitel 4 behandelt werden können.

Dieses Kapitel beruht in weiten Teilen auf den Betrachtungen von Blauert (1994) und Heute (1990), vgl. auch Vary et al. (1998).

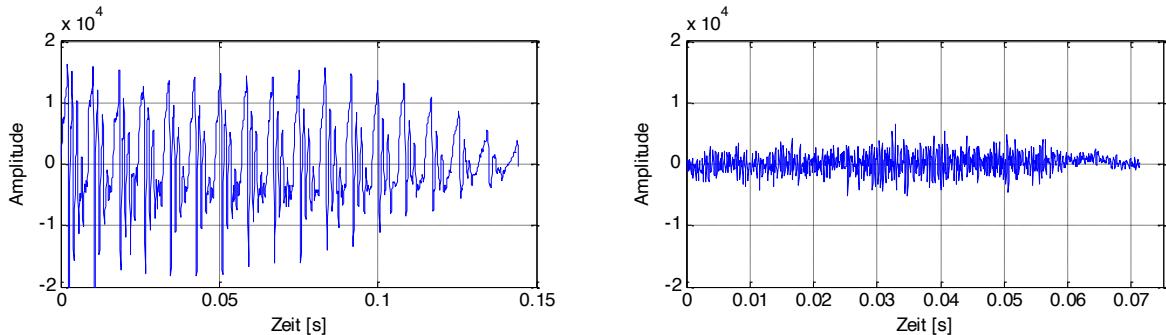
2.1 Darstellung kontinuierlicher Signale im Zeitbereich

Betrachtet man den Verlauf eines Mikrofon-Ausgangssignals über der Zeit, so zeigt sich beispielsweise folgendes Bild:



Sprachsignal zur Äußerung „Sichtbare Sprache“.

Bei genauer Betrachtung erkennt man Bereiche *quasi-periodischen* Verhaltens sowie *nicht-periodische* (z.B. rauschförmige) Abschnitte. Stimmhafte Laute (insbes. Vokale) sind kurzzeit-periodisch oder quasi-periodisch; stimmlose Laute (insbes. Zischlaute) sind nicht-periodisch. Darüber hinaus gibt es aber auch Mischlaute, z.B. stimmhafte Zischlaute.



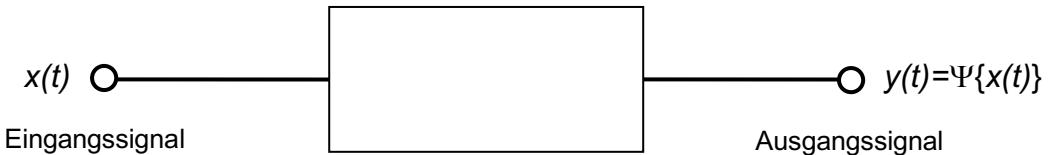
Sprachsignalabschnitte.

Links: Quasi-periodischer Abschnitt; rechts: rauschförmiger Abschnitt.

Die betrachteten Sprachsignalabschnitte lassen sich mittels einer geeigneten Bewertungsfunktion aus dem Gesamtsignal herausschneiden; man spricht dann von einer „gefensterten“ Version des Signals, und bezeichnet die Bewertungsfunktion als „Fensterfunktion“. Die einfachste Fensterfunktion ist z.B. eine Rechteckfunktion, welche ein Rechteck einer bestimmten Länge (z.B. 20 ms) des Signals einblendet.

2.2 LTI-Systeme

Das Sprachsignal $x(t)$ soll nun durch ein System übertragen werden. Mathematisch entspricht dies der Transformation oder Abbildung, bei der die Ausgangsfunktion $y(t)$ und die Eingangsfunktion $x(t)$ eindeutig über die Abbildung Ψ zugeordnet sind:



Symbolische Darstellung eines Systems.

In der Kommunikationstechnik haben wir es häufig mit Systemen zu tun, die sich durch Differentialgleichungen 1. Ordnung mit konstanten Koeffizienten beschreiben lassen. Solche Systeme verhalten sich linear und zeitinvariant, d.h. es gelten der Überlagerungssatz und der Verschiebungssatz; man bezeichnet ein solches System als *lineares, zeitinvariantes System (LTI-System, linear time-invariant system)*.

Überlagerungssatz:

Jede Linearkombination $x_i(t)$ von Eingangssignalen führt zu einer Linearkombination $y_i(t)$ der Ausgangssignale, die sich bei jedem der Eingangssignale einzeln ergeben:

$$\Psi \left\{ \sum_i a_i \cdot x_i(t) \right\} = \sum_i a_i \Psi \{x_i(t)\} = \sum_i a_i y_i(t) \quad (2.1)$$

Verschiebungssatz:

Für beliebige reelle Verschiebungszeiten T gilt, dass die Form des Ausgangssignals unabhängig vom Bezugszeitpunkt des Eingangssignals ist:

$$\Psi \{x(t-T)\} = y(t-T) \quad (2.2)$$

Bei Linearität und Zeitinvarianz kann man das Ausgangssignal des Systems dadurch ermitteln, dass man das Eingangssignal zunächst in Teilsignale zerlegt, zu diesen nacheinander die Ausgangssignale ermittelt, und diese anschließend additiv überlagert.

2.3 Darstellung kontinuierlicher Signale im Frequenzbereich, Spektrum

Eine für die Kommunikationstechnik besonders interessante Zerlegung von Signalen ist die Zerlegung in sinusförmige bzw. harmonisch-exponentielle Komponenten verschiedener Frequenzen:

$$c_\nu \cdot e^{j\omega_\nu t}, \quad c_\nu \in C$$

Diese Zerlegung wird als *Spektrum des Signals* bezeichnet; c_ν ist der Spektralwert zur Frequenz ω_ν . Die Zerlegung ist insbesondere deshalb günstig, da bei LTI-Systemen die Exponentialanteile in ihrer Art nicht verändert werden (sie sind ihre „Eigenfunktionen“); nur Amplitude und Phase werden u.U. verändert.

Je nachdem, ob nur die Beträge, die absoluten zeitlichen Zuordnungen oder die Betragsquadrate der Komponenten interessieren, spricht man von Amplituden- oder Betragsspektren, von Phasenspektren, oder von Leistungs(-dichte)-Spektren. Betrags- und Phaseninformationen finden sich im komplexen Spektrum vereint.

Wegen der Gültigkeit des Überlagerungssatzes für lineare Systeme spielt es nur eine untergeordnete Rolle, ob die Zerlegung in endlich viele Frequenzkomponenten gelingt, oder ob unendlich viele Anteile zu einer genauen Beschreibung des Zeitsignals benötigt werden. Am Ausgang des Systems ergibt sich stets die entsprechende – endliche oder unendliche – Überlagerung der nur linear, d.h. in ihren Amplituden und Phasen beeinflussten Exponentialkomponenten.

Fourier-Reihe:

Für eine periodische Zeitfunktion $x(t) = x(t+T)$ mit der Periodendauer T – reell oder komplexwertig, in einer Periode betragsintegrierbar, mit endlich vielen Sprungstellen und Extremwerten innerhalb einer Periode – kann man folgende Reihendarstellung angeben:

$$x(t) \approx g_n(t) = \sum_{\nu=-n}^n c_\nu e^{j\nu \frac{2\pi}{T} t} = \sum_{\nu=-n}^n c_\nu e^{j\omega_\nu t} \quad (2.3)$$

mit $\omega_\nu = \nu \frac{2\pi}{T}$. Das Ungefähr-Zeichen bedeutet, dass zwischen $x(t)$ und $g_n(t)$ im Allgemeinen eine Abweichung besteht. Der Wunsch, diese Abweichung in ihrem quadratischen Mittel möglichst klein werden zu lassen, d.h. die Forderung

$$\varepsilon_n = \int_{t_0}^{t_0+T} |x(t) - g_n(t)|^2 dt \stackrel{!}{=} \min \quad (2.4)$$

führt auf speziell zu wählende Reihenkoeffizienten c_ν . Diese Werte

$$c_\nu = \frac{1}{T} \int_{t_0}^{t_0+T} x(t) \cdot e^{-j\nu \frac{2\pi}{T} t} dt \quad (2.5)$$

heißen *Fourier-Koeffizienten* von $x(t)$.

Die Wahl des Fehlerkriteriums nach (2.4) ist sinnvoll, da es auf die einfache Bestimmung der Fourier-Koeffizienten nach (2.5) führt. Es ergibt sich daraus weiterhin die Konvergenz-Eigenschaft der Fourier-Reihendarstellung: Für beliebige, auch unstetige Funktionen $x(t)$ gilt

$$\lim_{n \rightarrow \infty} \{\varepsilon_n\} = 0 \quad (2.6)$$

also – mit Ausnahme von eventuellen Unstetigkeitsstellen –

$$\lim_{n \rightarrow \infty} \{g_n(t)\} = x(t) \quad (2.7)$$

Für endliche Werte von n gilt mit (2.3) und (2.7)

$$x(t) - g_n(t) = \sum_{\nu=n+1}^{\infty} \left[c_\nu \cdot e^{j\nu \frac{2\pi}{T} t} + c_{-\nu} \cdot e^{-j\nu \frac{2\pi}{T} t} \right] \quad (2.8)$$

Wegen der Orthogonalität der Summanden gilt gemäß (2.4)

$$\varepsilon_n = T \cdot \sum_{\nu=n+1}^{\infty} [|c_\nu|^2 + |c_{-\nu}|^2] \quad (2.9)$$

Im Grenzfall $n \rightarrow \infty$ folgt wegen (2.6), dass die Werte c_v mit wachsendem Index v immer weniger zum Fehler und damit auch zu $g_n(t)$ beitragen. Das Spektrum nimmt zu hohen Frequenzen hin dem Betrage nach immer mehr ab:

$$\lim_{|v| \rightarrow \infty} |c_v| = 0 \quad (2.10)$$

Ein solches Spektrum, das nur bei diskreten Frequenzen $v \cdot \frac{2\pi}{T}$ existiert, heißt *diskretes Spektrum* oder *Linienspektrum*.

Fourier-Transformation:

Für eine weitgehend allgemeine (also nicht unbedingt periodische) Zeitfunktion $x(t)$ – reell- oder komplexwertig, betragsintegrierbar und von beschränkter Variation – kann man eine Integral-Darstellung wie folgt angeben:

$$x(t) \approx g_\Omega(t) = \frac{1}{2\pi} \cdot \int_{-\Omega}^{+\Omega} X(j\omega) \cdot e^{j\omega t} d\omega \quad (2.11)$$

Wir verwenden hierbei anstelle der Frequenz f (in Hz) die Kreisfrequenz $\omega = 2\pi \cdot f$ (in s^{-1}); eine Darstellung in f ist aber gleichfalls möglich.

Eine ähnliche Argumentation wie bei der Fourier-Reihe führt wiederum dazu, dass (mit Ausnahme von Unstetigkeitsstellen) gilt:

$$\lim_{\Omega \rightarrow \infty} g_\Omega(t) = \frac{1}{2\pi} \lim_{\Omega \rightarrow \infty} \int_{-\Omega}^{+\Omega} X(j\omega) \cdot e^{j\omega t} d\omega = x(t) \quad (2.12)$$

sofern die Frequenzgangsfunktion $X(j\omega)$ wie folgt gewählt wird:

$$X(j\omega) = \int_{-\infty}^{+\infty} x(t) \cdot e^{-j\omega t} dt \quad (2.13)$$

Die Funktion $X(j\omega)$ beschreibt die i.a. kontinuierlich über der Frequenz verteilten Exponentialanteile. Man spricht von einem *kontinuierlichen Spektrum*, der Spektraldichte oder der *Fourier-Transformierten* zu $x(t)$:

$$X(j\omega) = F\{x(t)\} \quad (2.14)$$

Dieser Zusammenhang wird auch wie folgt symbolisiert:

$$X(j\omega) \bullet\circ x(t) \quad (2.15)$$

Die Funktion $X(j\omega)$ enthält wiederum Betrags- und Phaseninformationen. Sie kann als Verallgemeinerung der komplexen Amplituden c_v aus Gleichung (2.5) für unendlich große Periodendauern T , und damit verschwindend kleinen Abständen $\frac{2\pi}{T}$ der Spektralanteile gedeutet werden.

Die Darstellung des Spektrums wird häufig in einen Betragsanteil und einen Phasenanteil getrennt. Um Spektren über einen größeren Wertebereich einheitlich darzustellen empfiehlt sich darüber hinaus eine Logarithmierung, üblicherweise mit der Basis 10:

$$\log_{10} X(j\omega) = \log_{10} |X(j\omega)| + j \arg\{X(j\omega)\}$$

Die Betragsinformation des Spektrums wird häufig in der (pseudo-) Einheit Dezi-Bel (dB) angegeben; man bezeichnet sie auch als Signalpegel:

$$20 \cdot \log_{10}|X(j\omega)| \quad [dB]$$

Für die Fourier-Transformation gelten vier wesentliche Sätze, die hier (ohne Herleitung) nur kurz angegeben werden sollen:

Symmetrie-Satz:

Wenn $X(j\omega)$ die Fourier-Transformierte von $x(t)$ ist, dann gilt

$$X(jt) \circlearrowright 2\pi \cdot x(-\omega) \quad (2.16)$$

Mit diesem Satz lassen sich Fourier-Transformierte durch Ausnutzung von Hin- und Rücktransformation teilweise recht einfach bestimmen. Der Satz besagt auch, dass sich die Fourier-Transformation und die Fourier-Rücktransformation *nicht grundsätzlich* (bis auf ein Vorzeichen und den Vorfaktor 2π) unterscheiden. Dies lässt auch Folgerungen auf die Form von Signalen und zugehörigen Spektren zu:

- Ein *periodisches Signal* besitzt (wie oben gezeigt) ein *diskretes Spektrum* (Linienspektrum).
- Ein *diskretes* (z.B. digitalisiertes) *Signal* besitzt ein sich *periodisch wiederholendes Spektrum*.

Verschiebungssatz:

Wenn $X(j\omega)$ die Fourier-Transformierte von $x(t)$ ist, dann gilt

$$\begin{aligned} x(t) \cdot e^{j\omega_0 t} &\circlearrowright X(j(\omega - \omega_0)) \\ x(t - t_0) &\circlearrowright X(j\omega) \cdot e^{-j\omega t_0} \end{aligned} \quad (2.17)$$

Faltungs- und Multiplikationssatz (zum Begriff der Faltung vgl. Abschnitt 2.4):

Wenn $X_1(j\omega)$ die Fourier-Transformierte von $x_1(t)$ und $X_2(j\omega)$ die Fourier-Transformierte von $x_2(t)$ ist, dann gilt

$$\begin{aligned} x_1(t) * x_2(t) &\circlearrowright X_1(j\omega) \cdot X_2(j\omega) \\ x_1(t) \cdot x_2(t) &\circlearrowright \frac{1}{2\pi j} X_1(j\omega) * X_2(j\omega) \end{aligned} \quad (2.18)$$

Satz von Parseval:

Wenn $X(j\omega)$ die Fourier-Transformierte von $x(t)$ ist, dann gilt

$$\int_{-\infty}^{+\infty} |x(t)|^2 dt = \frac{1}{2\pi} \int_{-\infty}^{+\infty} |X(j\omega)|^2 d\omega \quad (2.19)$$

d.h. die Energie im Zeitbereich ist gleich der Energie im Frequenzbereich.

2.4 Impulsantwort und Übertragungsfunktion

Betrachten wir nun als Teilsignal einer Signalzerlegung das spezielle Signal $\delta(t)$, den sog. *Dirac-Impuls*. Dieses Signal ist über eine Distribution (diese ordnet einer Funktion einen bestimmten Wert über einen Integralausdruck zu) wie folgt definiert:

$$x(\tau) = \int_{-\infty}^{+\infty} x(t) \cdot \delta(t - \tau) dt \quad (2.20)$$

wobei

$$\int_{-\infty}^{+\infty} \delta(t) dt = 1 \quad (2.21)$$

Durch den Dirac-Impuls an der Stelle $t = \tau$ wird also der Wert der Funktion $x(t)$ an der Stelle $t = \tau$ „ausgeblendet“. Diese „Ausblendeigenschaft“ kann man sich dadurch veranschaulichen, dass man sich den Dirac-Impuls als *Rechteck mit verschwindender Breite und konstanter Fläche 1* vorstellt.

Überlagerungsintegral:

Aus (2.20) folgt durch Vertauschen von t und τ und durch die Symmetrieeigenschaft der Dirac-Funktion $\delta(t) = \delta(-t)$ und damit $\delta(\tau-t) = \delta(t-\tau)$ das sog. *Überlagerungsintegral*:

$$x(t) = \int_{-\infty}^{+\infty} x(\tau) \cdot \delta(t - \tau) d\tau \quad (2.22)$$

Dieses Integral beschreibt die Zeitfunktion $x(t)$ als unendliche Überlagerung von Dirac-Impulsen, die jeweils mit dem Funktionswert $x(\tau)$ bewertet sind.

Impulsantwort und Faltungsintegral:

Wenn ein LTI-System mit einem Dirac-Impuls angeregt wird, ist das entstehende Ausgangssignal $h(t)$ *charakteristisch für das System*. Man bezeichnet $h(t)$ deshalb als Stoßantwort oder *Impulsantwort* des Systems.

Für ein allgemeines Eingangssignal $x(t)$ ergibt sich das Ausgangssignal $y(t)$ nach (2.22) zu

$$\begin{aligned} y(t) &= \int_{-\infty}^{+\infty} x(\tau) \cdot h(t - \tau) d\tau \\ &= x(t) * h(t) \end{aligned} \quad (2.23)$$

Man bezeichnet dieses Integral als das *Faltungsintegral* und den Operator $*$ als Faltungsoperator. Die Faltung eines Signals mit $\delta(t)$ ergibt das Signal selber (Überlagerungsintegral).

Anschaulich gesprochen integriert das Faltungsintegral alle vergangenen, jeweils mit $x(\tau)$ bewerteten Impulsantworten des Systems auf, an die sich das System zum Zeitpunkt t noch „erinnert“. Der „erinnerte“ Anteil ergibt sich jeweils durch Bewertung mit der zeitlich verschobenen und im Zeitverlauf umgekehrten Impulsantwort (deshalb der Begriff „Faltung“). Die Faltungsoperation ist kommutativ, d.h. $x(t)*h(t) = h(t)*x(t)$.

Übertragungsfunktion:

Wir betrachten nun ein harmonisches Eingangssignal

$$x(t) = X(j\omega_0) \cdot e^{j\omega_0 t} \quad (2.24)$$

Mit Hilfe des Faltungsintegrals ergibt sich für das Ausgangssignal eines LTI-Systems, welches mit $x(t)$ angeregt wird

$$\begin{aligned} y(t) &= h(t) * x(t) = X(j\omega_0) \cdot \int_{-\infty}^{+\infty} h(\tau) \cdot e^{j\omega_0(t-\tau)} d\tau \\ &= X(j\omega_0) \cdot e^{j\omega_0 t} \cdot \int_{-\infty}^{+\infty} h(\tau) \cdot e^{-j\omega_0 \tau} d\tau \end{aligned} \quad (2.25)$$

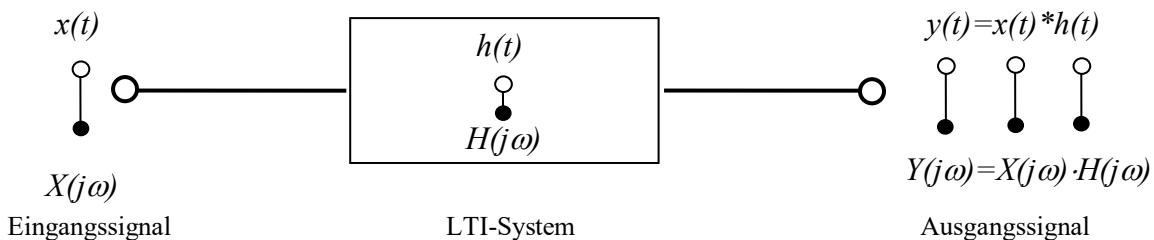
oder

$$y(t) = x(t) \cdot H(j\omega_0) \quad (2.26)$$

mit

$$H(j\omega) = |H(j\omega)| \cdot e^{j\phi(\omega)} = \int_{-\infty}^{+\infty} h(t) \cdot e^{-j\omega t} dt = F\{h(t)\} \quad (2.27)$$

Für das betrachtete LTI-System reduziert sich also die Faltungsoperation auf eine Multiplikation mit dem komplexen Faktor $H(j\omega_0)$. Man bezeichnet $H(j\omega)$ als die *komplexe Übertragungsfunktion* des Systems. Der *Zusammenhang zwischen der Impulsantwort $h(t)$ und der Übertragungsfunktion $H(j\omega)$ ist die Fourier-Transformation*. Die folgende Abbildung veranschaulicht die Situation:



Zusammenhänge zwischen Eingangs- und Ausgangssignal eines LTI-Systems und ihre Fourier-Transformierten.

2.5 Statistische Beschreibung von Sprachsignalen

Bislang sind wir von einer deterministischen Beschreibung von Sprachsignalen ausgegangen. Wie wir an den Darstellungen in Kapitel 2.1 erkennen können, sind Sprachsignale jedoch nicht wirklich deterministisch – ihre Entstehung ist für eine deterministische Beschreibung zu komplex, und daher sind zufällige Aspekte zu berücksichtigen. Daher ist neben der deterministischen auch eine statistische Betrachtungsweise angebracht.

Wir führen hier zunächst den Begriff der Wahrscheinlichkeit ein (vgl. Heute, 1990). Dabei betrachten wir eine Menge $\{x_v\}$ aller möglichen, sich gegenseitig ausschließenden Elementarereignisse:

$$\{x_v\} = \{x_1, x_2, x_3, \dots, x_n\} \quad (2.28)$$

wobei $n \rightarrow \infty$ zugelassen ist. Darüber hinaus betrachten wir eine Zufallsvariable x , die zufällig irgendeinem, aber sicher genau einem der möglichen Elementarereignisse entspricht. Die Werte x_v repräsentieren dann die tatsächlichen Ereignisse, z.B. die abgelesenen Spannungswerte des Mikrophonsignals o.ä.

Für die Zufallsvariable x und die Elementarereignisse x_v gelten folgende Axiome:

1. Jedem Ereignis $x = x_v$ wird eine nicht-negative Zahl zugeordnet. Sie heißt *Wahrscheinlichkeit* von x_v :

$$W\{x_v\} \geq 0 \quad (2.29)$$

Eine solche Zuordnung ist auch für nicht-elementare Ereignisse möglich.

2. Die Wahrscheinlichkeit des *sicheren Ereignisses* E ist

$$W\{E\} = 1 \quad (2.30)$$

3. Für Elementarereignisse oder sich gegenseitig ausschließende Ereignisse $y = y_v \in y_1, y_2, y_3, \dots, y_n$ gilt:

$$W\{y_1 \vee y_2 \vee y_3 \dots \vee y_r\}_{r \leq m} = \sum_{i=1}^r W\{y_i\} \quad (2.31)$$

Aus diesen Axiomen folgen weitere Eigenschaften des so eingeführten Wahrscheinlichkeitsbegriffes. So sind die komplementären Ereignisse $x = x_v$ und $x = \bar{x}_v$ miteinander unvereinbar und ergänzen sich zum sicheren Ereignis E :

$$W\{x_v \vee \bar{x}_v\} = W\{x_v\} + W\{\bar{x}_v\} = W\{E\} = 1 \quad (2.32)$$

und daher

$$W\{\bar{x}_v\} = 1 - W\{x_v\} \quad (2.33)$$

Das Komplement zum sicheren Ereignis E ist ein unmögliches Ereignis:

$$W\{\bar{E}\} = 1 - W\{E\} = 0 \quad (2.34)$$

Aus dem ersten Axiom und den o.a. Gleichungen folgt für irgendwelche Ereignisse x_v :

$$W\{x_v\} \in [0; 1] \quad (2.35)$$

Für eine geordnete Ereignismenge $x_1 < x_2 < x_3 < x_4 < \dots < x_m$, $m \geq r$ lässt sich der Fall $x \leq x_r$ erfassen:

$$W\{x \leq x_r\} = W\{x_1 \vee x_2 \vee x_3 \dots \vee x_r\} = \sum_{i=1}^r W\{x_i\} \quad (2.36)$$

Wegen Gl. (2.35) muss für eine endliche Menge mit m Ereignissen

$$\lim_{r \rightarrow m} W\{x \leq x_r\} = 1 \quad (2.37)$$

gelten, und für eine unendliche große Menge

$$\lim_{X \rightarrow \infty} W\{x \leq X\} = 1 \quad (2.38)$$

Als Komplement gilt

$$\lim_{X \rightarrow -\infty} W\{x \leq X\} = 0 \quad (2.39)$$

Diese Wahrscheinlichkeit, dass die Zufallsgröße x unterhalb einer bestimmten Schranke bleibt, lässt sich auch für kontinuierliche Variablen angeben, die kontinuierliche Werte zwischen $+\infty$ und $-\infty$ einnehmen können. Man bezeichnet hierbei

$$W\{x(t) \leq X\} = P_x(X, t) \quad (2.40)$$

als die *Verteilungsfunktion*. Durch $P_x(X, t)$ wird nicht nur eine einzige Variable x beschrieben, sondern ein ganzes „Ensemble“ gleichartiger Variablen, die den gleichen Erzeugungsvorschriften gehorchen. Eine dieser Vorschriften ist die Verteilungsfunktion, und alle Vorschriften zusammen definieren einen *Zufallsprozess*, dem die Variable x als Repräsentant angehört.

Die Variable x hängt i. Allg. von der Beobachtungszeit t ab, es handelt sich also um eine Zufallsfunktion $x(t)$. Daneben können sich aber auch die Prozesseigenschaften mit der Zeit ändern; dies ist bei der Variablen $P_x(X, t)$ bereits durch den Index t angedeutet. Die Änderung der Prozesseigenschaften mit der Zeit lässt sich besonders bei Sprache leicht nachvollziehen: Durch die Änderungen der Stellung des Vokaltraktes ändern sich auch die Eigenschaften des sich daraus ausbreitenden Sprachsignals. Trotzdem kann man auch für Sprache „insgesamt“ mittlere Eigenschaften bestimmen, wenn man von Stationarität ausgeht. In diesem Fall entfällt die Abhängigkeit von t in Gl. (2.40).

Die Variable $P_x(X, t)$ ist der Definition nach eine Wahrscheinlichkeit. Daher muss sie wegen Gl. (2.36) monoton nicht-fallend (also steigend oder zumindest konstant bleibend) verlaufen. Nach Gl. (2.38) und (2.39) gilt weiterhin

$$\lim_{X \rightarrow -\infty} P_x(X, t) = 0 \quad (2.41)$$

sowie

$$\lim_{X \rightarrow \infty} P_x(X, t) = 1 \quad (2.42)$$

Sofern $P_x(X, t)$ zumindest stückweise stetig ist kann man diese Funktion differenzieren. Man bezeichnet den Differenzialquotienten

$$\frac{\partial P_x(X, t)}{\partial X} = p_x(X, t) \quad (2.43)$$

als *Verteilungsdichte-Funktion (VDF)*. Die VDF wird i. Allg. zur Kennzeichnung eines Prozesses benutzt. Man erhält aus ihr die Verteilungsfunktion $P_x(X, t)$ durch Integration

$$P_x(X, t) = \int_{-\infty}^X p_x(\xi, t) d\xi \quad (2.44)$$

Die Wahrscheinlichkeit, dass die Variable x zum Zeitpunkt t im Intervall $[X - \Delta X/2; X + \Delta X/2]$ um den Wert X herum liegt, lässt sich bestimmen als

$$\begin{aligned} W\left\{x \in \left[X - \frac{\Delta X}{2}; X + \frac{\Delta X}{2}\right], t\right\} &= \int_{-\infty}^{X + \frac{\Delta X}{2}} p_x(\xi, t) d\xi - \int_{-\infty}^{X - \frac{\Delta X}{2}} p_x(\xi, t) d\xi \\ &= P_x\left(X + \frac{\Delta X}{2}, t\right) - P_x\left(X - \frac{\Delta X}{2}, t\right) \\ &= \Delta P_x(X, t) \end{aligned} \quad (2.45)$$

Andererseits ist der Differentialquotient in Gl. (2.43) auch ungefähr gleich dem Differenzen-Quotienten:

$$p_x(X, t) \approx \frac{\Delta P_x(X, t)}{\Delta X} \quad (2.46)$$

Die Verteilungsdichtefunktion ist also eine inkrementelle, auf das Intervall ΔX bezogene Wahrscheinlichkeit. Wegen der Monotonie der Verteilungsfunktion gilt zusätzlich

$$p_x(X, t) \geq 0 \quad (2.47)$$

Bislang sind wir immer von einem unabhängigen Signal (z.B. dem Verlauf des Mikrophonsignals über der Zeit) ausgegangen. Betrachtet man zwei Signale im Zusammenhang, so kann man die Definitionen nach Gl. (2.40) und (2.43) auf zweidimensionale Funktionen verallgemeinern:

$$P_{x_1 x_2}(X_1, X_2, t_1, t_2) = W\{(x_1(t_1) \leq X_1) \wedge (x_2(t_2) \leq X_2)\} \quad (2.48)$$

bezeichnet die sog. Verbund-Verteilungsfunktion, und

$$\frac{\partial^2 P_{x_1 x_2}(X_1, X_2, t_1, t_2)}{\partial X_1 \partial X_2} = p_{x_1 x_2}(X_1, X_2, t_1, t_2) \quad (2.49)$$

die Verbund-VDF. Von der Verbund-VDF gelangt man zur eindimensionalen VDF mittels der Randverteilung, bei der der betrachtete Wert (z.B. X_1) festgehalten wird und alle Werte der zweiten Größe (z.B. X_2) aufaddiert werden:

$$p_{x_1}(X_1, t_1) = \int_{X_2=-\infty}^{\infty} p_{x_1 x_2}(X_1, X_2, t_1, t_2) dX_2 \quad (2.50)$$

Die Verbund-Verteilungsfunktion $P_{x_1 x_2}(X_1, X_2, t_1, t_2)$ und die Verbund-VDF $p_{x_1 x_2}(X_1, X_2, t_1, t_2)$ hängen i. Allg. immer von beiden Schranken X_1 und X_2 ab. Allerdings gibt es Sonderfälle, in denen sich diese Funktionen in zwei ein-dimensionale Ausdrücke separieren lassen:

$$P_{x_1 x_2}(X_1, X_2, t_1, t_2) = P_{x_1}(X_1, t_1) \cdot P_{x_2}(X_2, t_2) \quad (2.51)$$

$$p_{x_1 x_2}(X_1, X_2, t_1, t_2) = p_{x_1}(X_1, t_1) \cdot p_{x_2}(X_2, t_2) \quad (2.52)$$

Man bezeichnet die beiden Variablen $x_1(t_1)$ und $x_2(t_2)$ dann als statistisch unabhängig. Aussagen zur Wahrscheinlichkeit einer der Variablen lassen sich dann – bis auf einen konstanten Faktor – unabhängig von der Größe der anderen Variablen treffen.

Im Folgenden sollen einige Kenngrößen ein-dimensionaler und zwei-dimensionaler Zufallsgrößen behandelt werden. Wir definieren zunächst den *Erwartungswert* einer Größe y , die gemäß einer Abbildung $f(x)$ von einer Zufallsgröße x abhängt:

$$E\{f(x(t))\} = \int_{-\infty}^{\infty} f(x) \cdot p_x(X, t) dX \quad (2.53)$$

Diese Definition bedeutet anschaulich eine Mittelung über alle möglichen Werte von $f(x)$ gewichtet mit der „Auftritts-Häufigkeit“ der Werte $x(t)$ – ähnlich z.B. einem Notendurchschnitt.

In der Praxis interessieren insbesondere folgende Sonderfälle:

$$1. \quad y = f(x) = x: \quad E\{x(t)\} = \int_{-\infty}^{\infty} X \cdot p_x(X, t) dX = \mu_x(t) \quad (2.54)$$

Man bezeichnet $\mu_x(t)$ als den *linearen Mittelwert* von $x(t)$ (auch: 1. Moment bezüglich 0).

$$2. \quad y = f(x) = x^2: \quad E\{x^2(t)\} = \int_{-\infty}^{\infty} X^2 \cdot p_x(X, t) dX = \bar{x}^2(t) \quad (2.55)$$

Man bezeichnet $\bar{x}^2(t)$ als den *quadratischen Mittelwert* von x (auch: 2. Moment bezüglich 0). Wenn man $x(t)$ mit einer physikalischen Größe wie dem Strom oder der Spannung in Verbindung bringt kann der quadratische Mittelwert als normierte Leistung interpretiert werden.

$$3. \quad y = f(x) = (x - \mu_x)^2:$$

$$\begin{aligned} E\{(x(t) - \mu_x)^2\} &= E\{x^2(t)\} - 2 \cdot \mu_x \cdot E\{x(t)\} + \mu_x^2 \\ &= \bar{x}^2(t) - \mu_x^2(t) = \sigma_x^2(t) \end{aligned} \quad (2.56)$$

Man bezeichnet $\sigma_x^2(t)$ als die *Varianz* von x (auch: 2. zentrales Moment). Sie beinhaltet bei einer physikalischen Größe im Prinzip die Leistung nach Abzug des Mittelwertes oder „Gleichanteiles“, also die Leistung des um den Mittelwert schwankenden Anteiles von $x(t)$. Die *Streuung* $\sigma_x(t)$ kann als „Effektivwert“ dieses Anteiles angesehen werden.

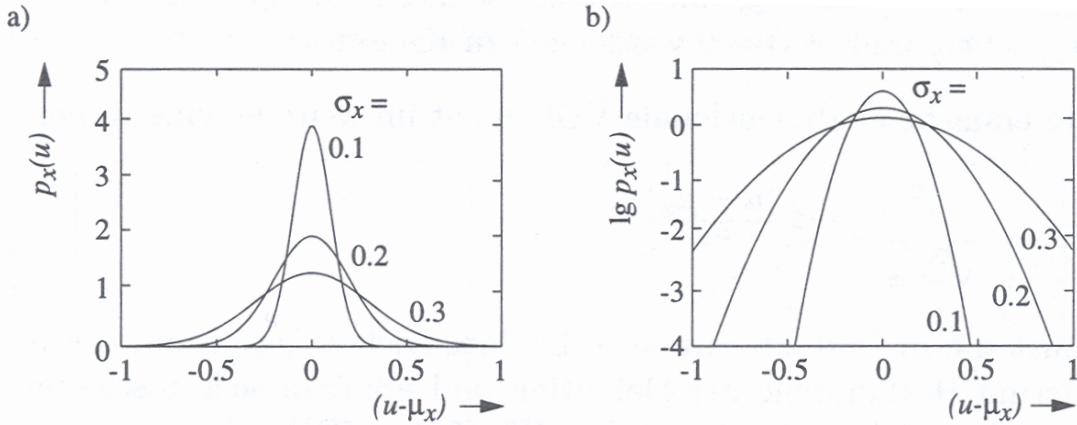
In der Praxis werden häufig die folgenden Standardverteilungen verwendet, die aber – wie wir noch sehen werden – nicht unbedingt realistisch für Sprachsignale sind:

1. Gauss- oder Normalverteilung: Diese wird beschrieben durch die VDF

$$p_x(X, t) = \frac{1}{\sqrt{2\pi}\sigma_x(t)} \cdot e^{-\frac{(X - \mu_x(t))^2}{2\sigma_x^2(t)}} \propto e^{-X^2} \quad (2.57)$$

Die Breite der dadurch beschriebenen „Glockenkurve“ ist beiderseits der durch μ_x bestimmten Symmetrielinie proportional zu σ_x , ihre Höhe bei $X = \mu_x$ umgekehrt proportional dazu.

Die Gauss-Verteilung entsteht als Grenzfall der Summation unendlich vieler unabhängiger Zufallsgrößen. Sie ist deshalb kaum real – geschweige denn so „normal“, wie es ihr Name suggeriert. Auch Sprachsignale beschreibt sie ausgesprochen schlecht! Dass sie trotzdem häufig verwendet wird liegt daran, dass Gauss-Verteilungen oft eine viel bessere und weitreichendere mathematische Behandlung statistischer Aufgaben erlaubt als andere Verteilungen; in vielen Fällen lassen sich nur mit der Gauss-Verteilung manche Rechnungen überhaupt ausführen.

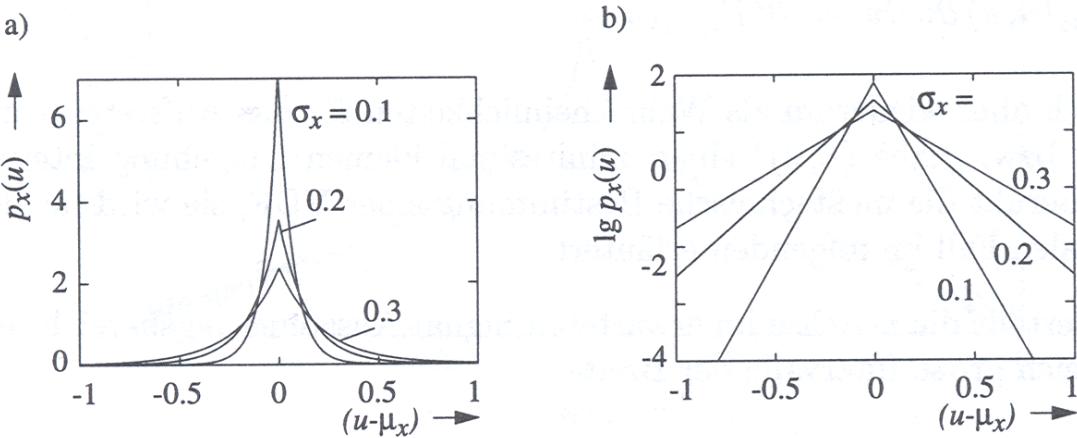


Gauss- oder Normal-VDF. a) Linearer, b) logarithmischer Maßstab
(aus Vary et al., 1998, 149).

2. Laplace-Verteilung: Diese wird beschrieben durch die VDF

$$p_x(X, t) = \frac{1}{\sqrt{2\sigma_x(t)}} \cdot e^{-\frac{|X-\mu_x(t)|}{\sigma_x(t)}} \propto e^{-|X|} \quad (2.58)$$

Für die Breite und Höhe der dadurch beschriebenen Verteilung gelten die gleichen Regeln wie für die Normalverteilung. Bei Darstellung im logarithmischen Maßstab fällt die Laplace-Verteilung beiderseits von μ_x linear ab.



Laplace- VDF. a) Linearer, b) logarithmischer Maßstab
(aus Vary et al., 1998, 151).

Wie wir noch sehen werden lässt sich Sprache durch eine Kombination solcher Verteilungen mäßig gut annähern. Wir kommen darauf in Kapitel 2.10 zurück.

Auch für zwei zusammenhängende Zufallsvariablen x_1 und x_2 lassen sich Erwartungswerte angeben. Für $y = f(x_1, x_2)$ gilt analog zu Gl. (2.53)

$$E\{f(x_1(t_1), x_2(t_2))\} = \int_{-\infty}^{\infty} f(X_1, X_2) \cdot p_{x_1 x_2}(X_1, X_2, t_1, t_2) dX_1 dX_2 \quad (2.59)$$

Bei zweidimensionalen VDFs interessieren meist die folgenden Sonderfälle:

1. $y = f(x_1, x_2) = x_1 \cdot x_2$:

$$E\{x_1(t) \cdot x_2(t)\} = \int_{X_1=-\infty}^{\infty} \int_{X_2=-\infty}^{\infty} X_1 X_2 \cdot p_{x_1 x_2}(X_1, X_2, t_1, t_2) dX_1 dX_2 = \varphi_{x_1 x_2}(t_1, t_2) \quad (2.60)$$

Man bezeichnet $\varphi_{x_1 x_2}(t_1, t_2)$ als die (*Kreuz-*) *Korrelationsfunktion (KKF)*. Die Kreuzkorrelation drückt die Ähnlichkeit zweier Signale bei Verschiebungen um $\lambda = t_2 - t_1$ aus.

2. $y = f((x_1(t) - \mu_{x_1})(x_2(t) - \mu_{x_2}))$:

$$\begin{aligned} E\{(x_1(t) - \mu_{x_1})(x_2(t) - \mu_{x_2})\} &= \psi_{x_1 x_2}(t_1, t_2) \\ &= \varphi_{x_1 x_2}(t_1, t_2) - \mu_{x_1}(t_1) \cdot \mu_{x_2}(t_2) \end{aligned} \quad (2.61)$$

Man bezeichnet $\psi_{x_1 x_2}(t_1, t_2)$ als die (*Kreuz-*) *Kovarianzfunktion*. Korrelations- und Kovarianzfunktion unterscheiden sich nur durch den Mittelwert-Term; d.h. im Fall von Mittelwert-Freiheit sind Korrelation und Kovarianz identisch.

Häufig interessieren Zufallswerte x_1 und x_2 , die ein und demselben Prozess – aber zu verschiedenen Zeiten t_1 und $t_2 = t_1 + \tau$ – entnommen wurden. In diesem Fall wird aus der (*Kreuz-*) Korrelationsfunktion eine *Auto-Korrelationsfunktion (AKF)*

$$\varphi_{xx}(\tau) = E\{x(t) \cdot x(t+\tau)\} = \int_{X_1=-\infty}^{\infty} \int_{X_2=-\infty}^{\infty} X_1 X_2 \cdot p_{xx}(X_1, X_2, \tau) dX_1 dX_2 \quad (2.62)$$

und aus der (*Kreuz-*) Kovarianzfunktion eine *Auto-Kovarianzfunktion*

$$\psi_{xx}(\tau) = \varphi_{xx}(\tau) - \mu_x^2 \quad (2.63)$$

Im Falle $\tau = 0$ ist die AKF

$$\varphi_{xx}(0) = \psi_{xx}(0) + \mu_x^2 = \bar{x}^2 = \sigma_x^2 + \mu_x^2 \quad (2.64)$$

Aus der Definition in Gl. (2.62) folgt, dass die AKF symmetrisch zum Punkt $\tau = 0$ ist. Es gilt also

$$\varphi_{xx}(-\tau) = \varphi_{xx}(\tau) \quad (2.65)$$

Dieser Punkt stellt auch das Maximum des Betrages der AKF dar:

$$|\varphi_{xx}(\tau)| \leq \varphi_{xx}(0) = \bar{x}^2 = \sigma_x^2 + \mu_x^2 \quad (2.66)$$

Für die Autokovarianz gilt Entsprechendes.

Wenn man sich die AKF als Ähnlichkeit des zwischen dem Signal $x(t)$ und dem um τ verschobenen Signal $x(t+\tau)$ vorstellt, wird die Bedeutung der Gl. (2.66) sofort klar: Die Ähnlichkeit muss bei der Verschiebung 0 maximal sein, da die Signale dann ja identisch sind. Das Gleiche gilt für periodische Signale bei Verschiebungen, die ein Vielfaches der Periodenlänge sind.

Bei instationären Prozessen ist zu beachten, dass die Korrelationen und Kovarianzen und auch die Mittelwerte vom Zeitpunkt t , nicht nur von der Verschiebung τ abhängen (oder alternativ von den beiden Zeitpunkten t_1 und t_2).

Man bezeichnet zwei Signale $x_1(t_1)$ und $x_2(t_2)$ als *unkorreliert*, wenn ihre Kreuzkovarianz zu 0 wird. Verschwindet die KKF ganz, so sind die Signale *orthogonal*. Beide Eigenschaften können punktuell für bestimmte Werte t_1 und t_2 (oder t und τ) auftreten, oder global für alle Beobachtungszeitpunkte. Für die Auto-Kovarianz eines mit sich selbst unkorrelierten Signals gilt also

$$\psi_{xx}(\tau) = \sigma_x^2 \cdot \delta(\tau) = \begin{cases} \sigma_x^2 & \text{für } \tau = 0 \\ 0 & \text{für } \tau \neq 0 \end{cases} \quad (2.67)$$

Die zugehörige AKF weist neben der impulsförmigen Komponente bei $\tau = 0$ noch eine additive Komponente μ_x^2 auf:

$$\varphi_{xx}(\tau) = \sigma_x^2 \cdot \delta(\tau) + \mu_x^2 \quad (2.68)$$

Die in Gl. (2.53) und (2.59) eingeführten Erwartungswerte beschreiben für einen statistischen Prozess das mittlere Verhalten der Variablen, also des sog. „Ensembles“ oder der „Schar“ der Zufallsgrößen. Neben diesen *Ensemble-Mittelwerten* lassen sich für eine Zufallsgröße oder einen Verbund von Zufallsgrößen aber auch *zeitliche Mittelwerte* angeben. Für den ein-dimensionalen Fall gilt für eine Funktion $f(x)$ einer Zufallsvariablen $x(t)$ als Zeitmittelwert

$$\langle f[x(t)] \rangle = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T f[x(t)] dt \quad (2.69)$$

Im zwei-dimensionalen Fall gilt mit der Funktion $f[x_1(t), x_2(t + \tau)]$

$$\langle f[x_1(t), x_2(t + \tau)] \rangle = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T f[x_1(t), x_2(t + \tau)] dt \quad (2.70)$$

Im ein-dimensionalen Fall ist das Ergebnis also unabhängig vom Zeitpunkt t (es wurde ja gerade über die Zeit gemittelt). Im zwei-dimensionalen Fall hängt das Ergebnis nur von der Zeit-Verschiebung τ , nicht jedoch vom absoluten Zeitpunkt t ab.

Bislang sind wir i. Allg. von einer Zeitabhängigkeit der Ensemble-Größen ausgegangen. Allerdings gibt es auch Prozesse, die immer nach gleichbleibenden Vorschriften ablaufen. In diesem Fall entfällt bei den ein-dimensionalen Größen die Zeitabhängigkeit, und bei den zwei-dimensionalen Größen spielt nur noch die Zeitdifferenz, nicht aber die absoluten Zeitpunkte eine Rolle. Man nennt solche Prozesse *stationär*.

Noch spezieller ist folgende Sonderfall: Es kann vorkommen, dass ein willkürlich gewählter Repräsentant eines Prozesses – wenn man ihn über der Zeit in seinem zufälligen Verlauf betrachtet – alle Kennzeichen des Prozesses insgesamt aufweist. Sein linearer Zeit-Mittelwert ist dann identisch mit dem linearen Erwartungswert, seine zeitliche Varianz ist dann identisch mit dem quadratischen Erwartungswert, etc. Solche Prozesse heißen *ergodisch*.

2.6 Energiedichte- und Leistungsdichespektrum

Die im vorherigen Kapitel eingeführten statistischen Größen beschreiben ein Signal im Zeitbereich – wie es auch die deterministischen Beschreibungen im Zeitbereich getan haben. Es liegt nahe, dass es auch hierzu eine äquivalente Beschreibung im Frequenzbereich (Spektralbereich) geben müsste. In der Tat existiert neben der Spektralanalyse durch direkte Anwendung der Fourier-Transformation auch eine spektrale Analyse für nicht deterministische Signale. Diese Analyseart liefert allerdings keine Information über das Phasenspektrum, sondern nur über das Betragsspektrum.

Man geht aus vom Betragsquadrat der Fourier-Transformierten $X(j\omega)$ des Signals $x(t)$

$$|X(j\omega)|^2 = X(j\omega) \cdot X^*(j\omega) \quad (2.71)$$

Die inverse Fourier-Transformierte dieses Ausdrucks ergibt das folgende Faltungsintegral:

$$x(-t) * x(t) = \int_{-\infty}^{+\infty} x(-\tau) \cdot x(t-\tau) d\tau = \int_{-\infty}^{+\infty} x(\tau) \cdot x(t+\tau) d\tau = \varphi_{xx}(t) \quad (2.72)$$

Durch inverse Fourier-Transformation des *Betragsquadrates* des Signalspektrums gelangt man also wiederum zur Autokorrelationsfunktion (AKF) des Signals $x(t)$.

Der obige Ausdruck konvergiert leider nur für Signale endlicher Energie, für die gilt:

$$\int_{-\infty}^{+\infty} x(t) \cdot x(t-0) dt < \infty \quad (2.73)$$

Das zugehörige Spektrum $|X(j\omega)|^2$ heißt *Energiedichespektrum*.

Bei Signalen unendlicher Energie, aber endlicher Leistung (Energie pro Zeiteinheit), bei denen also gilt

$$\lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^{+T} x(t) \cdot x(t-0) dt < \infty \quad (2.74)$$

kann man entsprechend eine Autokorrelationsfunktion durch Grenzwertbildung definieren:

$$\varphi_{xx}(t) = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^{+T} x(\tau) \cdot x(t+\tau) d\tau \quad (2.75)$$

Das zugehörige Spektrum heißt dann *Leistungsdichespektrum*.

Da die Autokorrelationsfunktion eine gerade Funktion ist, ist das zugehörige Energie- bzw. Leistungsdichespektrum rein reell (vgl. hierzu Vary et al., 1998, 139-145). Das ursprüngliche Signal lässt sich daraus leider nicht mehr eindeutig zurückgewinnen (im Gegensatz zum Fourier-Spektrum).

2.7 Darstellung diskreter Signale im Zeit- und Frequenzbereich

Zur Verarbeitung und zur Übertragung von Sprachsignalen (und von Signalen überhaupt) bedient man sich heutzutage fast ausschließlich einer digitalen Signaldarstellung, d.h. einer Signaldarstellung, die

- *zeitdiskret* und
- *wertdiskret*

ist. Eine solche Darstellung hat mehrere Vorteile:

- *Störsicherheit*: Störungen können – im Prinzip beliebig – klein gehalten werden, da die diskreten Signale bei längeren Übertragungswegen ohne Verlust wieder erzeugt werden können.
- *Universalität*: Signale unterschiedlicher Quellen (z.B. Ton, Bild, Daten) können auf demselben Weg übertragen werden, mit universellen Datennetzen, vgl. Voice-over-IP (VoIP).
- *Einfache Verarbeitung*: Digitale Signale gestatten eine einfache Erzeugung, Speicherung und Verarbeitung durch Digitalrechner.

Leider hat die digitale Signalverarbeitung auch einige Nachteile:

- Der begrenzte Wertevorrat bei der wertdiskreten Darstellung erfordert eine Quantisierung, mit der zwangsläufig ein *Quantisierungsfehler* einhergeht.
- Um ein Signal fester Bandbreite f_g zu übertragen muss es zunächst mit mindestens doppelter Frequenz abgetastet werden. Dadurch ergibt sich ein *höherer Bandbreitenbedarf* als bei der analogen Darstellung. Dieser Nachteil kann allerdings durch eine intelligente Kodierung ausgeglichen werden.

Dennoch überwiegen die Vorteile bei weitem und haben zum Durchbruch der digitalen Signalübertragung geführt.

Wenn ein Sprechschall auf ein Mikrofon trifft, so ergibt sich am Ausgang des Mikrofons zunächst ein analoges (zeit- und wertkontinuierliches) Signal. Das Spektrum dieses Signals wird im folgenden Unterkapitel noch weiter erläutert; hier sei aber schon vorweggenommen, dass dieses Spektrum begrenzt ist auf einen Frequenzbereich von 0 bis ca. 8 kHz, viele Sprachanteile auch auf maximal 4 kHz. Oberhalb von 8 kHz finden sich nur wenige Signalkomponenten, die nur eine sehr geringe Energie aufweisen. Sprache ist also ein Tiefpass-Signal, in dem nennenswerte Komponenten nur unterhalb einer sog. „Grenzfrequenz“ vorkommen.

Dieses Signal soll nun zunächst abgetastet werden, d.h. die kontinuierliche Signalform wird durch einen Impulszug ersetzt, der unmittelbar der Signalform folgt.

Nach dem *Abtasttheorem* kann ein bandbegrenztes Signal ohne Informationsverlust durch Abtastwerte in den Abständen

$$T = \frac{1}{f_A} \quad (2.76)$$

dargestellt werden, wenn die Abtastfrequenz mindest doppelt so hoch wie die höchste im Signal vorkommende Frequenz (Grenzfrequenz f_g) ist:

$$f_A \geq 2 \cdot f_g \quad (2.77)$$

Durch die Abtastung wird das Signal zeitdiskret. Aus dem Vorangegangenen ist bekannt, dass ein periodisches Signal ein Linienspektrum besitzt. Wegen der Umkehrbarkeit der Fourier-Transformation gilt umgekehrt auch, dass ein zeitdiskretes Signal ein periodisches Spektrum aufweist. Durch die Abtastung wird das ursprüngliche Signalspektrum also periodisch wiederholt, und zwar mit der Frequenz f_A . Im Bereich $f_A/2 \leq f \leq f_A$ wird das ursprüngliche Spektrum umgekehrt. Wenn das Signal zur Frequenz f_g noch nicht komplett vernachlässigbare Signalkomponenten enthält würden sich diese also mit dem ursprünglichen Spektrum überlagern (sog. *Aliasing*). Um dies zu vermeiden sollte man solche Komponenten durch eine vorherige Tiefpassfilterung mit der Grenzfrequenz f_g aus dem Signal entfernen; man bezeichnet diese Filterung als Anti-Alias-Filterung.

Das ursprüngliche Signal liegt jetzt in abgetasteter Form, d.h. zu diskreten Zeitpunkten $k \cdot T$ vor. Dadurch ist allerdings noch nichts gewonnen, da die Amplitude weiterhin kontinuierliche Werte annehmen kann und damit nicht gegen Störungen geschützt ist. Man geht deshalb auf eine diskrete Darstellung der Amplituden (zu den bereits diskreten Zeitpunkten) über, d.h. man *quantisiert die Amplituden* nach einer noch zu definierenden Vorschrift. Wie bereits ausgeführt entsteht dadurch ein *Quantisierungsfehler*, der minimal gehalten werden sollte. Zur Quantisierung und zur Größe dieses Quantisierungsfehlers vgl. Kapitel 6.

Ein abgetastetes Signal $x(k)$ besteht zunächst aus einer Folge von Zahlen, die mit k „durchnummeriert“ sind. Wenn man die Quantisierung zunächst außer Acht lässt, können diese Zahlen beliebige reelle – und verallgemeinert auch komplexe – Werte annehmen. Entstanden sind sie aus der Abtastung des kontinuierlichen Signals $x(t)$, wobei wir hier annehmen wollen, dass diese Abtastung zu äquidistanten Zeitpunkten $t = k \cdot T$ stattfindet. D.h.

$$x(k) = x(k \cdot T)$$

mit der Abtastfrequenz

$$f_A = \frac{1}{T}$$

Auch ein diskretes Signal $x(k)$ kann in sinusförmige Anteile zerlegt werden. Man verwendet hierfür wiederum die Fourier-Transformation, die für diskrete Signale wie folgt definiert ist:

$$x(k) \circ \bullet X(e^{j\Omega}) = \sum_{k=-\infty}^{\infty} x(k) \cdot e^{-jk\Omega} = F\{x(k)\} \quad (2.78)$$

Ihre Umkehrung

$$X(e^{j\Omega}) \bullet \circ x(k) = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(e^{j\Omega}) \cdot e^{jk\Omega} \quad (2.79)$$

stellt die Folge $x(k)$ als Überlagerung sinus- und kosinusförmiger Anteile

$$e^{jk\Omega} = \cos(k\Omega) + j \cdot \sin(k\Omega) \quad (2.80)$$

mit den komplexen Amplituden $X(e^{j\Omega})$ dar. Die Fourier-Transformation nach Gl. (2.78) gibt dagegen an, wie diese Amplituden als Funktion der normierten Frequenz Ω zu bestimmen sind. $X(e^{j\Omega})$ ist 2π -periodisch; mit

$$\Omega = 2\pi \cdot \frac{f}{f_A} = 2\pi \cdot f \cdot T \quad (2.81)$$

entspricht das der schon oben angesprochenen Periodizität des Spektrums abgetasteter Signale mit der Frequenz f_A .

Eine Erweiterung zur Fourier-Transformation diskreter Signale stellt die z-Transformation dar. Sie erlaubt die Spektraldarstellung von vielen Signalen, deren Fouriertransformierte nicht existiert. Auf diese Transformation soll allerdings der Übersichtlichkeit halber nicht weiter eingegangen werden.

Sofern das diskrete Signal $x(k)$ eine endliche Länge M hat oder – hier gleichwertig – periodisch mit der Periodenlänge M ist, kann man die sogenannte *Diskrete Fourier-Transformierte* (DFT) definieren. Sie enthält nur M diskrete Komponenten:

$$x(k) \circ \bullet X_\mu = \sum_{k=0}^{M-1} x(k) \cdot e^{-j \frac{2\pi}{M} \mu k} = \sum_{k=0}^{M-1} x(k) \cdot w_M^{\mu k} = DFT\{x(k)\} \quad (2.82)$$

wobei $w_M = e^{-j \frac{2\pi}{M}}$. Ihre Umkehrung lautet:

$$X_\mu \bullet \circ x(k) = \frac{1}{M} \sum_{\mu=0}^{M-1} X_\mu \cdot e^{j \frac{2\pi}{M} \mu k} = \frac{1}{M} \sum_{\mu=0}^{M-1} X_\mu \cdot w_M^{-\mu k} \quad (2.83)$$

Der Vergleich zwischen der Fourier-Transformation eines (unendlich langen) diskreten Signals nach G. (2.78) und der Diskreten Fourier-Transformation einer endlich langen Folge nach Gl. (2.82) zeigt, dass

$$X_\mu = X\left(e^{j\mu \frac{2\pi}{M}}\right) \quad (2.84)$$

D.h. die Diskrete Fourier-Transformation bestimmt Abtastwerte eines kontinuierlichen Frequenzspektrums einer endlich langen Folge, welches die Fourier-Transformation zur Verfügung stellt, in den Frequenzpunkten

$$\Omega_\mu = \mu \cdot \frac{2\pi}{M} \quad (2.85)$$

bzw.

$$f_\mu = \mu \cdot \frac{f_A}{M}, \quad \mu = \{0, 1, \dots, M-1\} \quad (2.86)$$

Zur Berechnung der Diskreten Fourier-Transformation (und ihrer Inversen) sind schnelle Algorithmen bekannt, die bei Digitalrechnern und speziellen Prozessoren erheblich Rechenzeit einsparen können. Man bezeichnet diese Realisierungen auch als *Fast Fourier Transform (FFT)*. Algorithmen hierzu sind z.B. bei Vary et al. (1998, 76-81) beschrieben. Nicht zuletzt deshalb ist die DFT bei der Analyse diskreter Signale sehr beliebt.

Durch die Einführung der Fourier-Transformation nach Gl. (2.78) und (2.79) für unendlich lange Zahlenfolgen sowie der Diskreten Fourier-Transformation (DFT) nach Gl. (2.82) und (2.83) für endlich lange oder periodisch sich wiederholende Zahlenfolgen lässt sich die Beschreibung im Zeit- und Frequenzbereich auch auf diskrete Signale anwenden. Dabei sind

insbesondere Signale am Eingang und Ausgang linearer, zeitinvarianter Systeme (LTI-Systeme) von Interesse. Ähnlich wie bei kontinuierlichen Signalen gilt hier:

$$\begin{aligned} y(k) &= x(k) * h_0(k) = \sum_{\kappa=-\infty}^{\infty} x(\kappa) \cdot h_0(k-\kappa) \\ &= h_0(k) * x(k) = \sum_{\kappa=-\infty}^{\infty} x(k-\kappa) \cdot h_0(\kappa) \end{aligned} \quad (2.87)$$

und im Frequenzbereich

$$Y(e^{j\Omega}) = H(e^{j\Omega}) \cdot X(e^{j\Omega}) \quad (2.88)$$

wobei

$$H(e^{j\Omega}) = F\{h_0(k)\} \quad (2.89)$$

die Fourier-Transformierte der *Impulsantwort* $h_0(k)$, d.h. den *Frequenzgang* des LTI-Systems darstellt.

Ein Vergleich der Formeln (2.87) und (2.89) mit den entsprechenden Formeln (2.23) und (2.27) für kontinuierliche Signale zeigt die Gleichwertigkeit der beiden Transformationen.

Anders verhält es sich leider bei der Diskreten Fourier-Transformation (DFT). Wegen der impliziten Periodizität der DFT besagt der Faltungssatz für

$$Y_\mu = H_\mu \cdot X_\mu = DFT\{h_0(k)\} \cdot DFT\{x(k)\} \quad (2.90)$$

dass

$$\begin{aligned} y(k) &= IDFT\{Y_\mu\} = x(k) \otimes h_0(k) = h_0(k) \otimes x(k) \\ &= \sum_{\kappa'=0}^{M-1} x(\kappa) \cdot h_0([k-\kappa]_{\text{mod } M}) \end{aligned} \quad (2.91)$$

aus einer *zyklischen Faltung* (Kennzeichen \otimes) entsteht. D.h., dass die endlich lange Zahlenfolge $x(k)$ mit der M -periodisch wiederholten Impulsantwort $h_0(k)$ gefaltet wird. Die dabei entstehende Folge $y(k)$ stimmt daher i. Allg. nicht überein mit dem Ergebnis einer „normalen“ (linearen) Faltung nach Gl. (2.87).

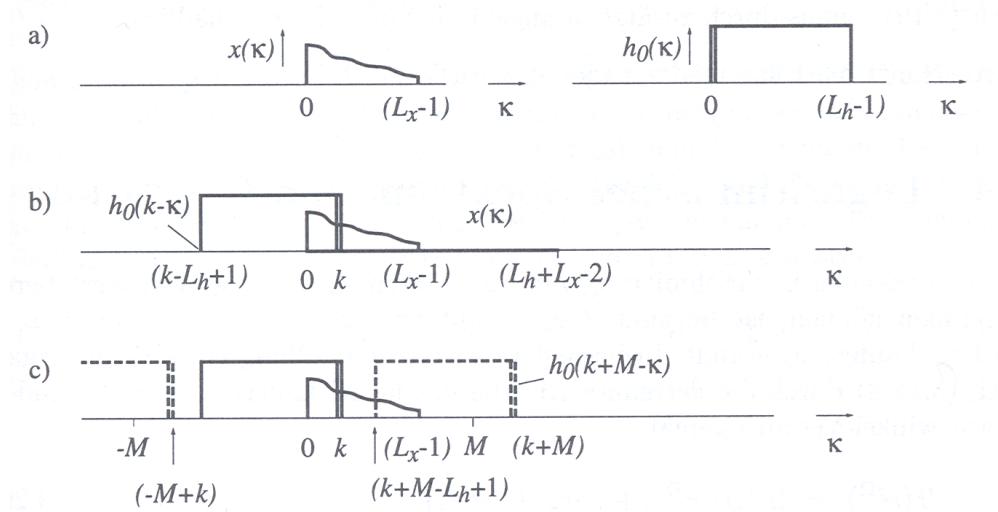
Falls $x(k)$ und $h_0(k)$ ungleiche Längen L_x und L_h aufweisen muss die DFT-Länge M mindestens

$$M_{\min} = \max\{L_x, L_h\} \quad (2.92)$$

sein. Kürzere Folgen sind entsprechend mit Nullwerten zu verlängern; man bezeichnet dies als *zero padding*. Das Ergebnis der Faltung hat nach Gl. (2.91) immer die Länge M .

Die Unterschiede zwischen linearer und zyklischer Faltung sind in untenstehender Abbildung gezeigt. Teilbild a) stellt die an der Faltung beteiligten Folgen $x(k)$ und $h_0(k)$ zum Zeitpunkt k dar. Teilbild b) zeigt, dass sich als Ergebnis der linearen Faltung eine Folge der Länge $L_y = L_x + L_h - 1$ ergibt (entsprechend dem fett eingezeichneten Teil der κ -Achse). Die zyklische Faltung nach Teilbild c) ergibt jedoch definitionsgemäß immer eine Folge der Länge M . Für die gezeichnete Situation entstehen $L_y - M$ Werte, die nicht denen der linearen Faltung entsprechen. Für Werte von k , die zwischen $L_x + L_h - 1 - M$ und $M - 1$ liegen, sind die Ergebniswerte der linearen Faltung gleich denen der zyklischen Faltung. Durch Wahl von $M_{\text{lin}} = L_y = L_x + L_h - 1$ liefert die zyklische Faltung die gleichen Werte wie die lineare Faltung. M

kann darüber hinaus durch *zero padding* vergrößert werden, ohne dass das einen Einfluss auf das Ergebnis der Faltung hätte.



Lineare und zyklische Faltung zweier endlich langer Signale $x(k)$ und $h_0(k)$.

- a) beteiligte Signale als kontinuierliche Einhüllende; b) lineare Faltung;
- c) zyklische Faltung (aus Vary et al., 1998, 67).

Die in Kapitel 2.5 angegebenen statistischen Kennwerte für kontinuierliche Signale lassen sich natürlich auch auf diskrete Zufalls-Signale erweitern. So sind Verteilungsfunktion

$$P_x(u, k) = W\{x(k) \leq u\} \quad (2.93)$$

und Verteilungsdichtefunktion (VDF)

$$\frac{\partial P_x(u, k)}{\partial u} = p_x(u, k) \quad (2.94)$$

im ein-dimensionalen Fall sowie die Verbund-Verteilungsfunktion

$$P_{x_1 x_2}(u_1, u_2, k_1, k_2) = W\{(x_1(k_1) \leq u_1) \wedge (x_2(k_2) \leq u_2)\} \quad (2.95)$$

und die Verbund-VDF

$$\frac{\partial^2 P_{x_1 x_2}(u_1, u_2, k_1, k_2)}{\partial u_1 \partial u_2} = p_{x_1 x_2}(u_1, u_2, k_1, k_2) \quad (2.96)$$

im zwei-dimensionalen Fall völlig analog definiert. Für die Erwartungswerte gilt

$$E\{f(x(k))\} = \int_{-\infty}^{\infty} f(x) \cdot p_x(u, k) du \quad (2.97)$$

und

$$E\{f(x_1(k_1), x_2(k_2))\} = \int_{-\infty}^{\infty} f(u_1, u_2) \cdot p_{x_1 x_2}(u_1, u_2, k_1, k_2) du_1 du_2 \quad (2.98)$$

bzw. mit Einführung der Verschiebung $\lambda = k_2 - k_1$ auch

$$E\{f(x_1(k), x_2(k + \lambda))\} = \int_{-\infty}^{\infty} f(u_1, u_2) \cdot p_{x_1 x_2}(u_1, u_2, \lambda) du_1 du_2 \quad (2.99)$$

Die dazu gehörigen Spezialfälle der Kreuz-Korrelationsfolge

$$\varphi_{x_1 x_2}(\lambda) = E\{x_1(k) \cdot x_2(k + \lambda)\} = \int_{u_1=-\infty}^{\infty} \int_{u_2=-\infty}^{\infty} u_1 u_2 \cdot p_{x_1 x_2}(u_1, u_2, \lambda) du_1 du_2 \quad (2.100)$$

und der Kreuz-Kovarianzfolge

$$\psi_{x_1 x_2}(\lambda) = \varphi_{x_1 x_2}(\lambda) - \mu_{x_1} \cdot \mu_{x_2} \quad (2.101)$$

sowie die Auto-Korrelationsfolge

$$\varphi_{xx}(\lambda) = E\{x(k) \cdot x(k + \lambda)\} = \int_{u_1=-\infty}^{\infty} \int_{u_2=-\infty}^{\infty} u_1 u_2 \cdot p_{xx}(u_1, u_2, \lambda) du_1 du_2 \quad (2.102)$$

und die Auto-Kovarianzfolge

$$\psi_{xx}(\lambda) = \varphi_{xx}(\lambda) - \mu_x^2 \quad (2.103)$$

folgen denen der kontinuierlichen Signale. Das Spektrum der AKF ist entsprechend durch

$$\Phi_{xx}(e^{j\Omega}) = F\{\varphi_{xx}(\lambda)\} \quad (2.104)$$

definiert.

Wir betrachten nun einen Prozess $x(k) = x_i(k)$ zu speziellen Zeitpunkten $k_i = i$ mit $i = 1, 2, \dots, n$ und einem willkürlich gewählten Zeitnullpunkt. Für diese Beobachtungszeitpunkte k_i lässt sich nun eine $(n \times n)$ -Matrix mit allen Korrelationen der Zufallsvariablen zu den entsprechenden Zeitpunkten angeben:

$$R = [\varphi_{xx}(i, j)]_{n \times n} = [\varphi_{xx}(j, i)]_{n \times n} \quad (2.105)$$

Diese Matrix enthält alle Auto-Korrelationen des Prozesses zu den betrachteten Zeitpunkten. Interessant sind die Symmetrieeigenschaften dieser Matrix. Die Werte auf der Hauptdiagonalen sind gemäß Gl. (2.62) und (2.64)

$$\varphi_{xx}(i, i) = \bar{x}^2(i) = \sigma_x^2(i) + \mu_x^2(i) \quad (2.106)$$

die zeitlich variablen Kenngrößen des Prozesses. Wenn man *Stationarität* annimmt, gilt statt Gl. (2.105)

$$\varphi_{xx}(i, j) = \varphi_{xx}((i + \nu), (j + \nu)) \stackrel{\text{def}}{=} \varphi_{xx}(i - j) = \varphi_{xx}(j - i) \quad (2.107)$$

und anstelle von Gl. (2.106)

$$\varphi_{xx}(i, i) = \varphi_{xx}(0) = \bar{x}^2 = \sigma_x^2 + \mu_x^2 \quad \forall i \quad (2.108)$$

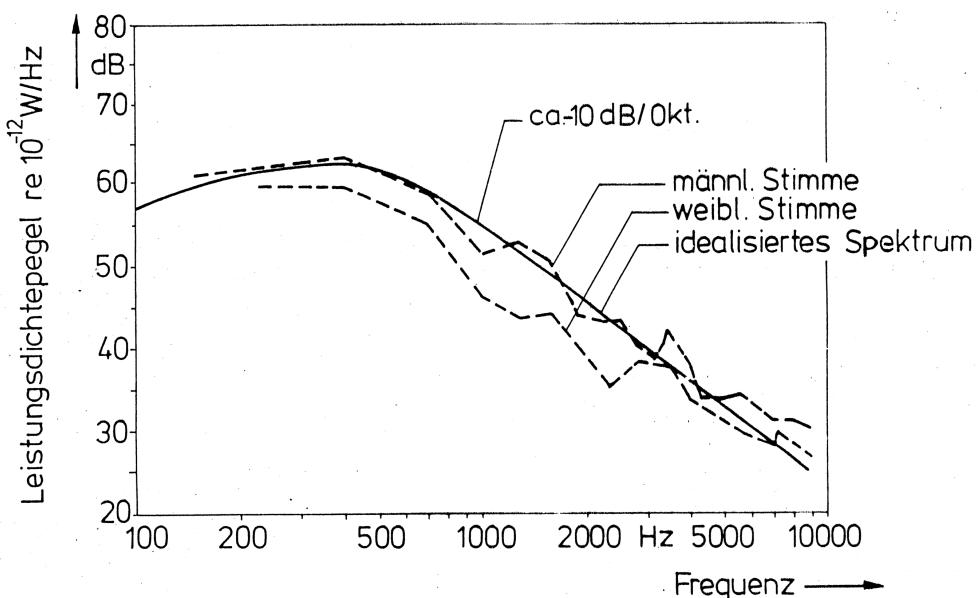
Die Auto-Korrelationsmatrix R wird dann zu einer symmetrischen Toeplitz-Matrix: In der Hauptdiagonalen und in den Diagonalen parallel dazu stehen jeweils identische Werte. Die Matrix R hat dann nur n unterschiedliche Werte. Solche symmetrischen, reellwerten AKF-Matrizen in Toeplitz-Struktur besitzen einige sehr interessante und vorteilhafte Eigenschaften: Sie sind stets invertierbar (regulär), sind positiv definit (erfüllen also die Eigenschaft

$v^T R v > 0 \forall v \neq 0$), und besitzen n positive Eigenwerte mit den dazu gehörigen n linear unabhängigen Eigenvektoren. Wir werden diese Matrix noch benötigen, wenn wir den Einfluss des menschlichen Vokaltraktes auf die Spracherzeugung analysieren.

2.8 Langzeit- und Kurzzeit-Signaleigenschaften

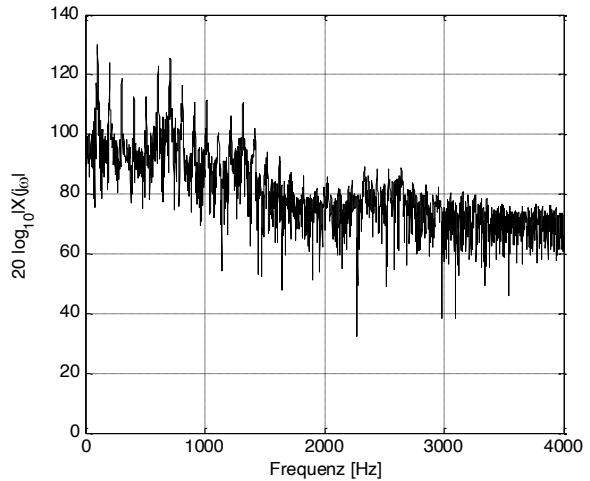
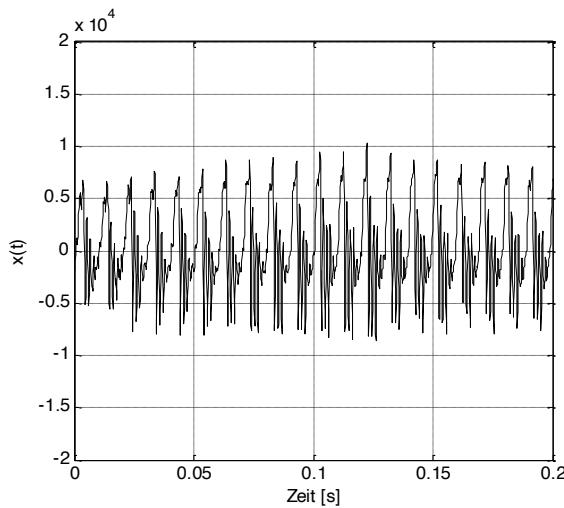
Die bislang erläuterten Größen sollen nun für ein typisches Sprachsignal betrachtet werden.

Ermittelt man das Leistungsdichtespektrum für Sprache durch Mittelung über einen langen Zeitraum (sog. *Langzeit-Leistungsdichtespektrum*), so stellt man fest, dass Sprache praktisch ein Bandpasssignal ist, d.h. es treten nur Frequenzkomponenten in einem bestimmten Frequenzband auf. Aus der Bandbegrenzung folgt auch, dass Sprache korreliert sein muss (ein völlig unkorreliertes Signal hat ein konstantes, ein sog. „weißes“ Spektrum). Diese Eigenschaft kann man sich bei der effizienten Kodierung von Sprache zunutze machen, vgl. Kapitel 6.

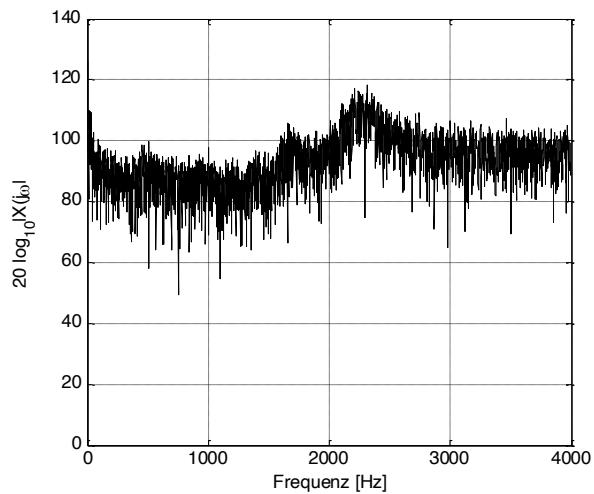
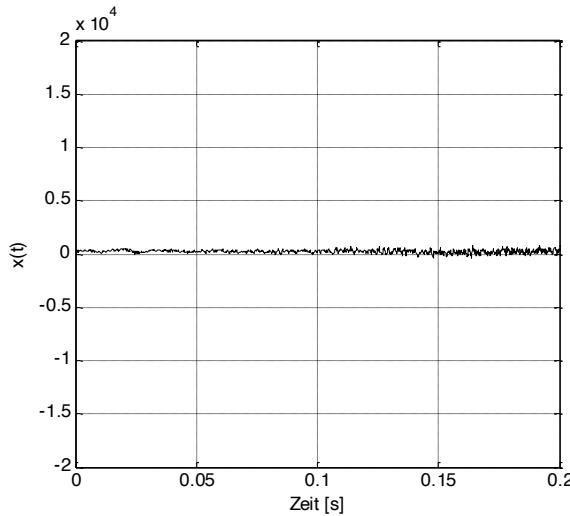


Langzeit-Leistungsdichtespektrum von Sprache (schematisiert, gemessen bei normaler Sprachlautstärke 30 cm vor dem Mund; aus Blauert, 1994).

Nun ist ein Sprachsignal aber – wie bereits gezeigt – nicht stationär, sondern ändert seine Form (in Abständen von ca. 20-30 ms) permanent. In einem solchen fortlaufenden Sprachsignal lassen sich quasi-periodische und nicht-periodische Anteile erkennen, die oftmals ohne deutliche Abgrenzung aufeinander folgen. Wenn man diese Abschnitte mittels einer geeigneten Bewertungsfunktion herausschneidet („fenstert“) und Fourier-transformiert, so ergeben sich *Kurzzeit-Leistungsdichtespektren*, die die betrachteten Signalabschnitte genauer charakterisieren. Die folgenden Abbildungen zeigen solche Kurzzeit-Spektren.



Zeitsignal und Kurzzeit-Leistungsdichtespektrum eines quasiperiodischen Sprachsignalabschnitts (stimmhafter Laut).



Zeitsignal und Kurzzeit-Leistungsdichtespektrum eines nichtperiodischen Sprachsignalabschnitts (stimmloser Laut).

Bei quasiperiodischen Signalabschnitten (bei stimmhaften Lauten) zeigt das Spektrum deutliche harmonische Komponenten, d.h. Komponenten bei Vielfachen einer Grundfrequenz. Die niedrigste dieser Komponenten liegt bei der *Sprachgrundfrequenz* f_0 , die bei Männern etwa 125 Hz beträgt, bei Frauen aber deutlich höher liegt (etwa 250 Hz). Sie bestimmt auch die wahrgenommene Tonhöhe der Sprachlaute.

Die Einhüllende des Betragsspektrums lässt bei quasi-periodischen wie auch bei nicht-periodischen Signalabschnitten deutlich einen oder mehrere „Gipfel“ erkennen. Diese Gipfel kennzeichnen Spektralbereiche relativ hoher Energiedichte; man nennt sie *Formanten* und nummeriert sie mit steigendem Abstand vom Ursprung durch (F_1, F_2, F_3 , etc.). Die Formanten sind zum einen charakteristisch für die Sprachsignalabschnitte und damit auch für die

zugehörigen Sprachlaute; sie können zum anderen auch charakteristisch für den Sprecher oder die Sprecherin sein (vor allem höhere Formanten).

Daneben fällt beim Zeitsignal auf, dass die quasi-periodischen Signalabschnitte *wesentlich energiereicher* sind als z.B. rauschförmige Abschnitte.

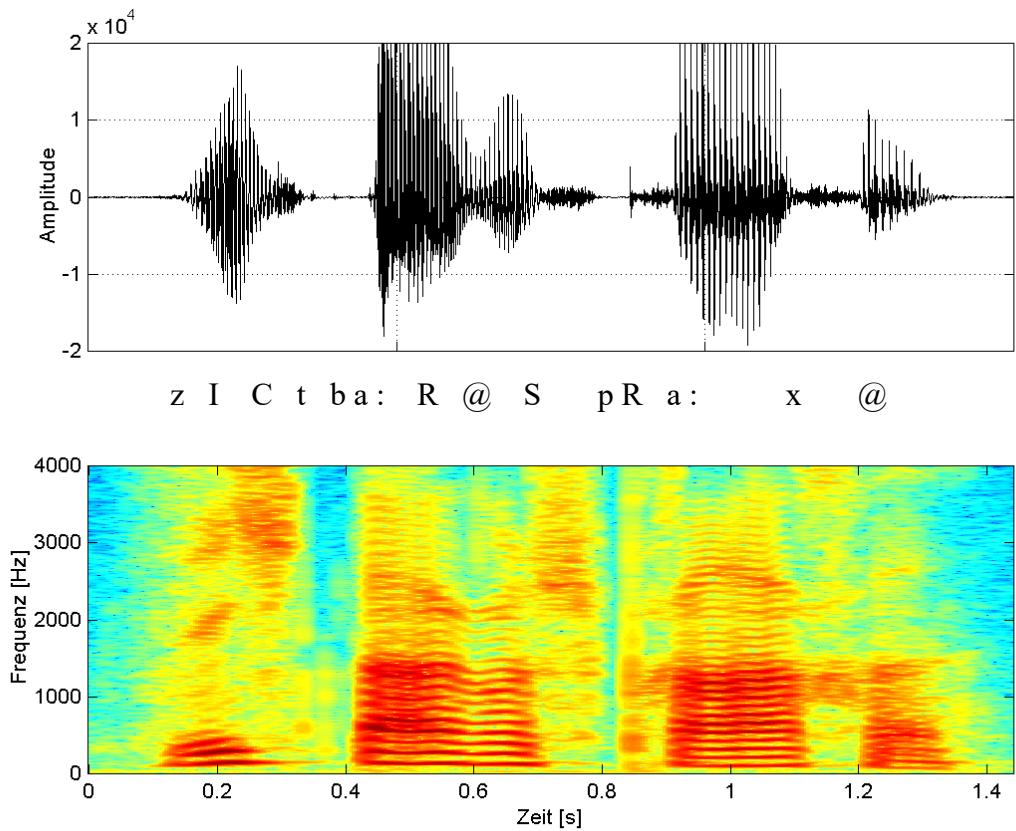
Es sei angemerkt, dass das Langzeit-Leistungsdichtespektrum keine Informationen über einzelne Sprachlaute liefert, da die entsprechende Information durch Mittelung über mehrere Laute „verschmiert“. Umgekehrt liefert ein Kurzzeit-Leistungsdichtespektrum keine Information über den einzelnen Sprachlaut hinaus.

2.9 Spektrogramm

Um sowohl einzelne Abschnitte wie auch den Verlauf des Sprachsignals über der Zeit spektral darzustellen verwendet man häufig eine gleitende Fourier-Transformation. Das heißt, dass das Sprachsignal sukzessive gefenstert wird, zu dem durch das Fenster eingebundenen Sprachsignalabschnitt jeweils ein Kurzzeit-Leistungsdichtespektrum berechnet wird, und diese Kurzzeit-Leistungsdichtespektren sukzessive über der Zeit dargestellt werden. Hierbei müssen praktisch 3 Dimensionen grafisch erfasst werden:

- Die Frequenz f oder die Kreisfrequenz ω
- Die Zeit t
- Die spektrale Leistungsdichte zum Zeitpunkt t bei der Frequenz f bzw. ω

Da im normalen Diagramm nur zwei Dimensionen abbildbar sind wird die dritte Dimension durch eine Farbkodierung oder durch Graustufen dargestellt. Eine solche quasidreidimensionale Darstellung von Signalen bezeichnet man als *Spektrogramm* (auch Visible-Speech-Diagramm). Ein Beispiel ist in der folgenden Abbildung gegeben.



Zeitsignal und Spektrogramm des Abschnitts „sichtbare Sprache“.

Es sei angemerkt, dass die spektrale Analysierschärfe reziprok zur zeitlichen ist, die durch die Länge des Analysefensters bestimmt wird. Es lassen sich also nicht gleichzeitig eine hohe zeitliche und spektrale Auflösung erzielen; für jeden Anwendungsfall muss ein individueller Kompromiss gesucht werden.

2.10 Amplitudenverteilung

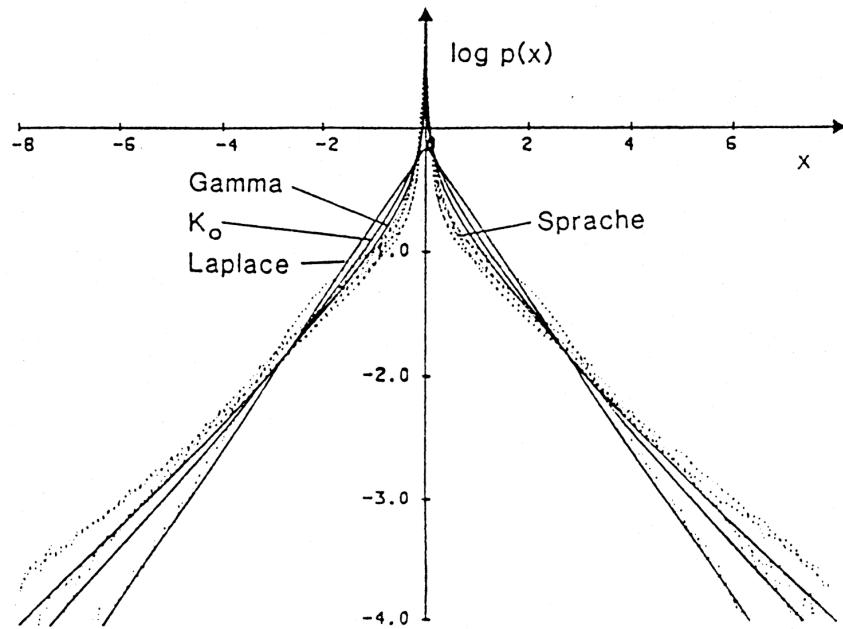
Neben den spektralen Eigenschaften lassen sich auch die Amplituden der Sprachsignale statistisch untersuchen. Man kommt dann auf die Verteilungsdichte der Amplituden, d.h. die relative Häufigkeit, mit der eine Amplitude im Sprachsignal auftritt. Bei normaler Sprache lässt sich diese Verteilungsdichtefunktion (VDF) nur schlecht durch eine Normalverteilung beschreiben. Eine Laplace-Verteilung modelliert zwar den Abfall zu höheren Amplituden (linear in logarithmischen Maßstab) besser, nicht jedoch die Spitze bei Null.

Man kann deshalb auf eine Kombination von Gauss- und Laplace-Verteilung übergehen, wie es z.B. Dunn und White (1940) vorgeschlagen haben:

$$p_x(X) = \frac{0.6}{\sqrt{2}\sigma_1} \cdot e^{-\frac{\sqrt{2}|X|}{\sigma_1}} + \frac{0.4}{\sqrt{2\pi}\sigma_2} \cdot e^{-\frac{X^2}{2\sigma_2^2}} \quad (2.109)$$

Allerdings stimmt auch der damit erzielbare Verlauf nicht wirklich gut mit gemessenen Verteilungen überein. Neuere Ansätze modellieren daher die Spitze bei $X = 0$ durch

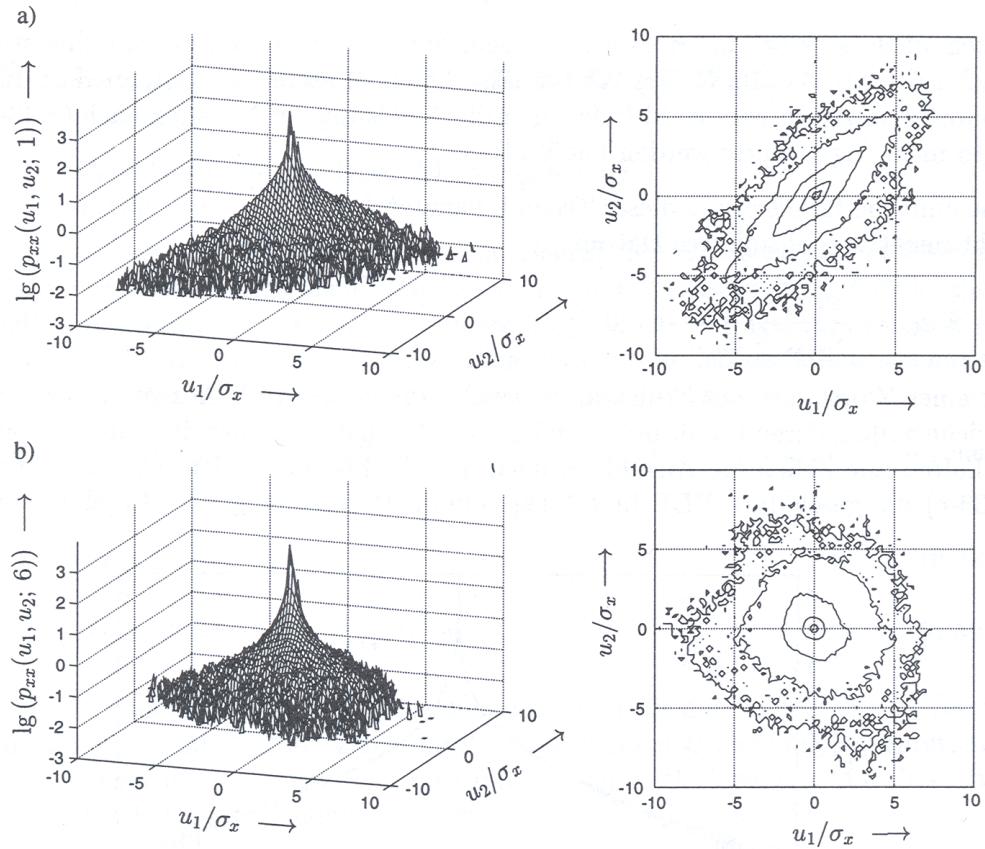
Singularitäten vom Typ $|\ln x|$ (K_0 -Verteilung) bzw. $1/\sqrt{x}$ (Gamma-Verteilung). Solche Ansätze sind in untenstehender Abbildung gezeigt.



Gemessene Amplitudenverteilungen und VDF-Modelle (aus Heute, 1990).

Die Spitze der Verteilungsdichtefunktion zeigt, dass kleine Amplituden um Null bei Sprache weitaus am häufigsten vorkommen. Dazu kommt, dass bei einer normalen Konversation (z.B. am Telefon) jeder Gesprächsteilnehmer im Mittel 60% der Zeit schweigt. Diese Eigenschaften von Sprache lassen sich zur effizienten Übertragung ausnutzen, bspw. durch genauere Quantisierung der Werte nahe Null und ungenauere Quantisierung größerer Amplituden (vgl. Kapitel 6), oder auch durch Zeitmultiplex-Verfahren, bei denen die Pausenabschnitte nicht übertragen und die so gewonnenen Zeitfenster für die Übertragung anderer Informationen genutzt werden.

Um die Korrelation, die ein Sprachsignal kennzeichnet, sichtbar zu machen, kann man zweidimensionale VDFen als Histogramme aufzeichnen. Hierbei werden die verbundenen Wahrscheinlichkeiten für $x_1(k) = x(k)$ und $x_2(k) = x(k+\lambda)$ gemessen. Die untenstehende Abbildung zeigt solche Verteilungen für $\lambda = 1$ und $\lambda = 6$. Bei $\lambda = 1$ sieht man, dass die „Höhenlinien des Histogramms“ Ellipsen sind; ihre Halbachsenverhältnisse hängen von σ_{x_1} und σ_{x_2} ab, ihre Halbachsenlage von der normierten Kreuz-Korrelationsfunktion. Bei $\lambda = 6$ sind die Ellipsen zu Kreisen geworden; die beiden Variablen sind dekorreliert.



Bivariate Histogramme für Sprachsignale. A) $\lambda = 1$ (hohe Korrelation);
b) $\lambda = 6$ (Dekorrelation). Aus Vary et al. (1998, 154).

2.11 Literatur

- Blauert, J. (1994). Kommunikationsakustik II: Audiokommunikation und virtuelle Realität. Skriptum zur Vorlesung am Institut für Kommunikationsakustik, Ruhr-Universität, Bochum.
- Dunn, H.K., White, S.D. (1940). Statistical Measurements on Conversational Speech. J. Acoust. Soc. Am. 11, 278-288.
- Heute, U. (1990). Sprachverarbeitung. Skriptum zur Vorlesung der Arbeitsgruppe Digitale Signalverarbeitung, Ruhr-Universität, Bochum.
- Vary, P., Heute, U., Hess, W. (1998). Digitale Sprachsignalverarbeitung. B.G. Teubner, Stuttgart.

3. Grundlagen der menschlichen Spracherzeugung

Nach einer ersten Betrachtung von Sprachsignalen im vorangegangenen Kapitel sollen nun die Mechanismen der menschlichen Spracherzeugung erläutert werden, die zu diesen Sprachsignaleigenschaften führen.

Sprache ist das Produkt einer willkürlichen, nach Regeln ablaufenden Bewegung des Sprechapparates (Blauert, 1994). Diese Regeln müssen erlernt werden, und zwar ab dem Beginn des frühen Kindesalters. Der motorische Ablauf der Sprachproduktion wird über Rückkoppelwege gesteuert und geregelt, sowohl

- auf akustischem Wege (durch Eigen- und Fremdstimulation über das eigene Gehör) als auch
- kinästhetisch (über die Sprech- und Atmungsmuskulatur).

Bei Störungen dieser Rückkoppelwege kann es zu Einschränkungen des Spracherwerbes kommen. Beispiele hierfür sind

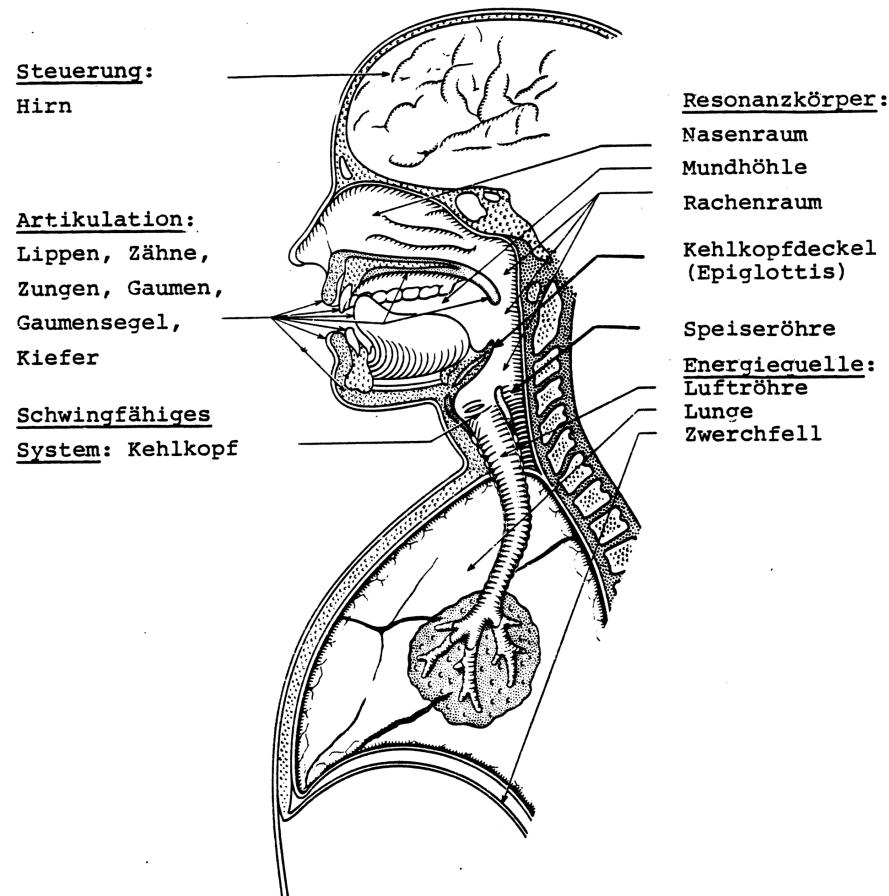
- angeborene oder erworbene Schwerhörigkeit
- Lokalanästhesie (z.B. nach zahnärztlicher Behandlung)
- anderweitige Störung der auditiven Rückkopplung, z.B. bei Anwesenheit störender Echos

Zum ungestörten Spracherwerb ist also die auditive Rückkoppelung entscheidend; sollte sie entfallen oder eingeschränkt sein (bspw. bei angeborener Schwerhörigkeit) ist auch die Sprachentwicklung gefährdet. Es gibt Anhaltspunkte dafür, dass dies bereits im Säuglingsalter (mit dem Schreien) beginnen kann.

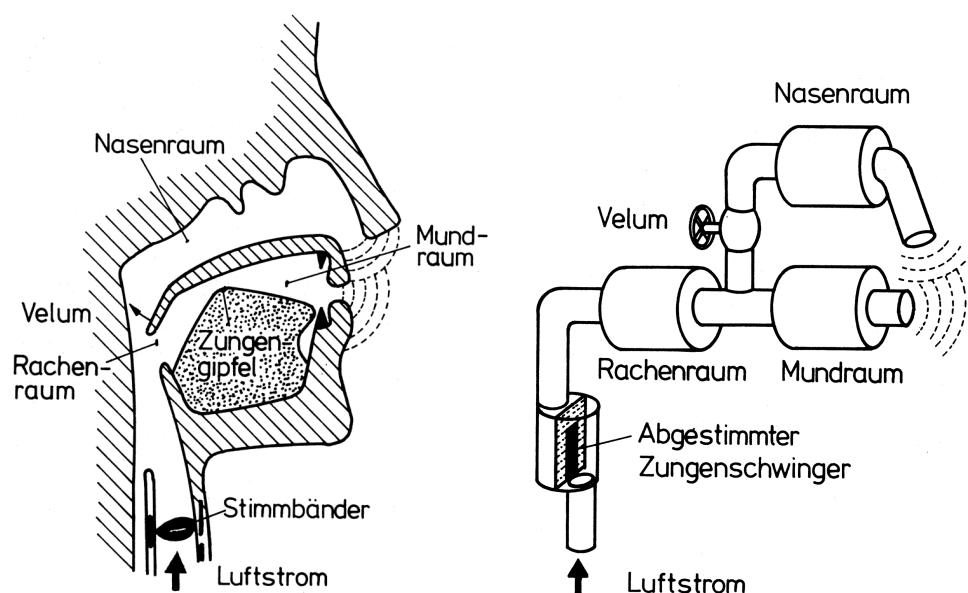
3.1 Anatomie des menschlichen Sprechapparates

Die nachstehende Abbildung gibt einen Überblick über die Anatomie des menschlichen Sprechtraktes.

Zur Erzeugung von Schallwellen muss zunächst ein Luftstrom generiert werden. Dies geschieht in den Lungen, durch Kontraktion von Brustkorb und Oberbauch mittels der Atmungsmuskulatur. Diese bilden die „Energiequelle“ des Systems. Der Luftstrom wird über die Luftröhre durch den Kehlkopf mit den Stimmbändern – d.h. durch ein schwingungsfähiges System – geleitet. Er wird dann im nachfolgenden Vokaltrakt (Mund-, Rachen- und Nasenraum, bilden einen Resonanzkörper) durch die Stellung der Artikulationsorgane moduliert und kann durch Mund und Nase entweichen. Die Situation ist vergleichbar mit der eines Blasinstrumentes, wie unten stehende Abbildung zeigt. Die Lunge entspricht dabei dem Winderzeuger, der Kehlkopf dem Mundstück, und der Vokaltrakt dem Ansatzrohr.



Anatomie des menschlichen Sprechtraktes (aus Heute, 1990, nach Flanagan, 1972).



Vergleich zwischen dem menschlichen Sprechtrakt und einem Blasinstrument
(aus Blauert, 1994).

Im Folgenden sollen die zwei wichtigsten Prozesse der Sprachproduktion näher betrachtet werden: Die Erzeugung des Anregungssignals (Luftstrom) und die Lautformung im Vokaltrakt.

3.2 Anregung

Energiequelle: Lunge (als Speicher)

Brustkorb, Zwerchfell, Muskulatur (also „Pumpe“)

→ erzeugt einen Druck von etwa 400 (leise Sprache) bis 2000 N/m² (laute Sprache)

Dieser Luftstrom wird nun durch den Kehlkopf geführt. Dabei können die Stimmbänder (Stimmlippen) entweder geöffnet sein, sodass die Luft weitgehend ungehindert entweichen kann, oder sie sind zunächst geschlossen und werden durch den Luftdruck aufgepresst. Hierdurch können zwei unterschiedliche Arten von „Anregungssignalen“ entstehen:

1) Periodische Anregung

Dieses Signal wird wie folgt erzeugt:

- Die Stimmbänder sind zunächst gespannt und verschließen die Luftröhre
- Durch die Kontraktion der Atmungsmuskulatur erhöht sich der Luftdruck unterhalb der Stimmbänder
- Bei einem bestimmten Druck springt die Glottis (Stimmritze) sprunghaft auf, und es kommt zu einem plötzlichen Entweichen der Luft
- Dadurch entsteht ein rascher Druckabfall; die Bernoulli-Kraft schließt die Glottis wieder

Durch Wiederholung dieser Schritte kommt es zu einem periodischen Öffnen und Schließen der Glottis mit der Periodenlänge $T_0 = 1/f_0$. f_0 ist die schon beobachtete *Grundfrequenz* des Sprachsignals; man bezeichnet sie (ungenau) auch als „Pitch-Frequenz“.

Die Grundfrequenz f_0 hängt ab von

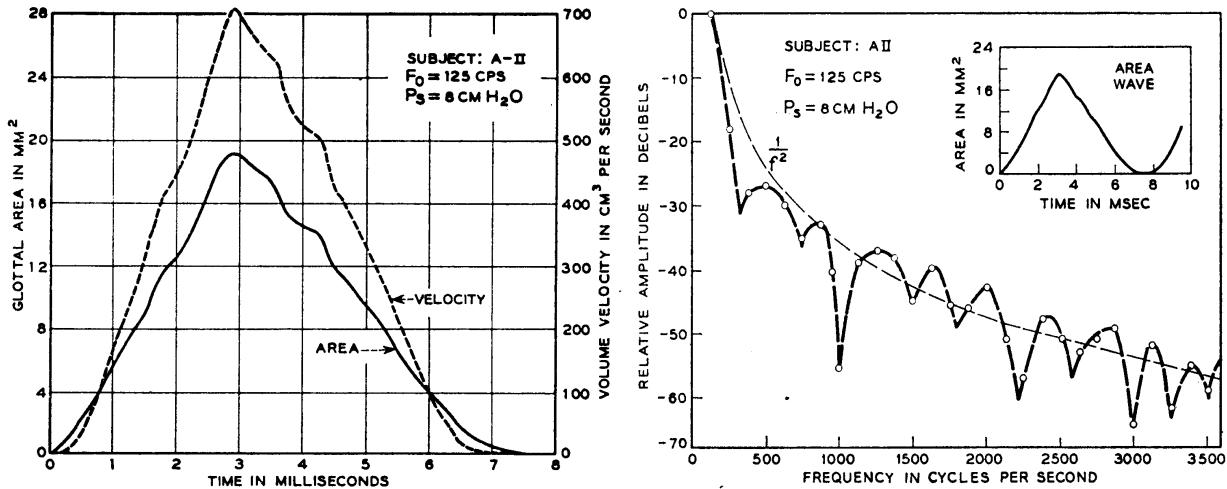
- der Stimmbandlänge und -masse (Männer ca. 20-24 mm, Frauen ca. 15-18 mm; Wachstum des Kehlkopfes während der Pubertät)
- der Spannung der Stellknorpel
- dem Druck unterhalb der Stimmbänder (höhere Grundfrequenz beim Schreien)

Typische Werte von f_0 sind 100-125 Hz bei Männern, 200-250 Hz bei Frauen, und noch höhere Werte bei Kindern (bei Säuglingen bis über 500 Hz).

Das dabei entstehende Signal entspricht im Zeitbereich *idealerweise* einer Impulsfolge. Transformiert in den Spektralbereich ergibt sich daraus ein weißes Spektrum (konstant über alle Frequenzen, aus der Fourier-Transformierten des Dirac-Impulses), das zudem noch diskret ist (Spektrum eines periodischen Signals), also ein *weißes Linienspektrum*. Ein solch breitbandiges Spektrum ist gut zum „Aufprägen“ von vielen Informationen (durch den Vokaltrakt) geeignet.

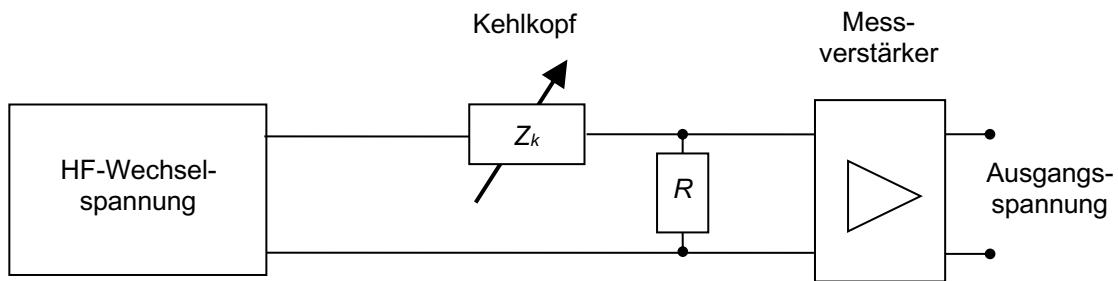
In der *Realität* gibt es aber keine Dirac-Impulse, sondern etwa dreieckige Impulse (Tastverhältnis 0,3 bis 0,7), bedingt durch die Nichtlinearität bei der Glottisbewegung und dem Druckverlauf. Das Spektrum dieser Dreiecks-Impulse fällt oberhalb etwa 500 Hz

proportional zu $1/\omega^2$ (im logarithmischen Maßstab entspricht das 12 dB pro Frequenzverdoppelung) ab, ist allerdings weiterhin sehr oberwellenhaltig.



Flächen- und Schnelleverlauf des Glottissignals über der Zeit (links) und zugehöriges Spektrum (rechts), aus Flanagan (1972, 49).

Das periodische Öffnen und Schließen der Glottis lässt sich z.B. mit einer Hochgeschwindigkeitskamera oder einer stroboskopischen Aufnahme beobachten. Indirekt lässt es sich mit dem sog. Laryngographen messen, siehe untenstehende Abbildung. Dieser misst die Kehlkopf-Impedanz mit Hilfe einer hochfrequenten Wechselspannung über zwei außen am Hals befestigte Elektroden; wenn die Glottis geschlossen ist, ist die Impedanz minimal, also wird die Spannung am Messwiderstand R maximal.



Messprinzip eines Laryngographen.

Periodische Anregung tritt auf bei Vokalen, stimmhaften Konsonanten und Nasalen, vgl. Abschnitt 3.4.

2) Aperiodische Anregung

Hier sind zwei unterschiedliche Anregungen zu unterscheiden:

- 2a) Ist die Glottis bereits geöffnet kann die Luft zwar entweichen, dies führt aber zu Turbulenzen unterhalb oder oberhalb der Glottis; es entsteht ein *rauschförmiges*

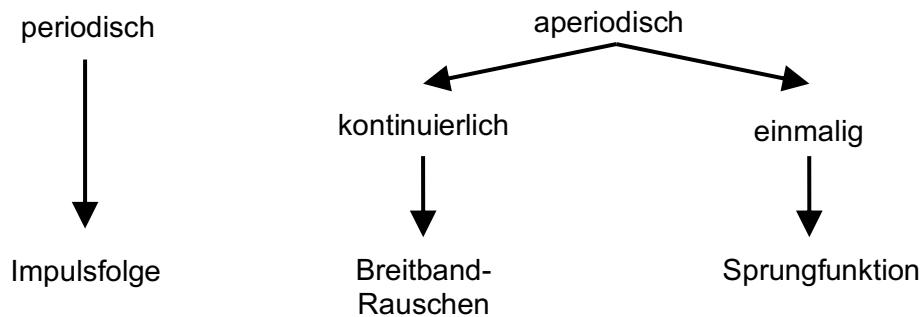
Anregungssignal, welches wiederum recht breitbandig (idealerweise „weiß“) ist, auf das sich also wiederum viele Informationen aufprägen lassen.

- 2b) Die Glottis ist zwar geöffnet, aber an einer anderen Stelle im Vokaltrakt ist der Luftauslass verschlossen; der Verschluss wird durch den Luftdruck plötzlich aufgepresst, was zu einer *Sprungfunktion* des Zeitsignals führt. Das Spektrum dieser Sprungfunktion fällt annähernd mit $1/\omega$ ab.

Neben diesen beiden idealen (stimmhaften und stimmlosen) Anregungen gibt es aber auch noch Mischformen, bspw. bei stimmhaften Konsonanten (z.B. /m/, /n/). Die stimmhafte Anregung kann auch fast komplett durch eine stimmlose ersetzt werden, ohne dass die lautliche Information verloren geht; dies geschieht beim Flüstern. Allerdings ist dadurch der Energiegehalt und damit die Reichweite stark eingeschränkt.

Rauschanregung tritt auf bei stimmlosen und stimmhaften Zischlauten, Sprunganregung bei sog. Plosiven oder Explosivlauten (z.B. /b/, /p/, /d/, /g/; stimmhaft oder stimmlos).

Die bislang behandelten Anregungssignale lassen sich wie folgt klassifizieren:



Klassifikation von Anregungssignalen.

Es sei angemerkt, dass die Anregungssignale – trotz zu hohen Frequenzen abfallenden Betragsspektrums – i.a. recht breitbandig sind. Dadurch lassen sich im nachgeschalteten Vokaltrakt viele Informationen auf das Anregungssignal aufprägen und somit übertragen. Dieser Vorgang wird als Lautformung bezeichnet.

3.3 Lautformung

Die bislang vorgestellten Anregungssignale enthalten zwar einige Variationsmöglichkeiten; diese reichen aber bei weitem nicht aus, um alle Sprachlaute darstellen zu können. Daher müssen weitere Informationen auf das Anregungssignal aufgeprägt werden; dies geschieht bei der sog. Lautformung.

Die Lautformung geschieht im *Vokaltrakt*, bestehend aus dem Mund-, Rachen- und Nasenraum, dessen Volumen und Abmessungen durch die Artikulationsorgane (Lippen, Unterkiefer, Zungengipfel, Gaumen, etc.) gesteuert werden. Diese bilden einen röhrenförmigen Raum mit veränderbarer Querschnittsfläche (Länge ca. 17 cm, Querschnitt ca. 0-20 cm²), der am unteren Ende mit der Glottis, am oberen Ende mit dem Mund

abgeschlossen ist. An dieses Rohr kann durch das Gaumensegel (Velum) der Nasenraum angekoppelt werden (Velumöffnung ca. 5 cm^2 , Nasenraum ca. 60 cm^3).

Der Vokaltrakt wirkt als Resonator, d.h. bestimmte Frequenzen des (breitbandigen) Anregungssignals werden verstärkt, andere werden abgeschwächt. Deshalb kann man den Vokaltrakt als lineares Filter mit der Übertragungsfunktion $H(j\omega)$ beschreiben. Die verstärkten Bereiche entsprechen den Formanten (Gipfeln des Spektrums), die abgeschwächten den Antiformanten (Täler des Spektrums).

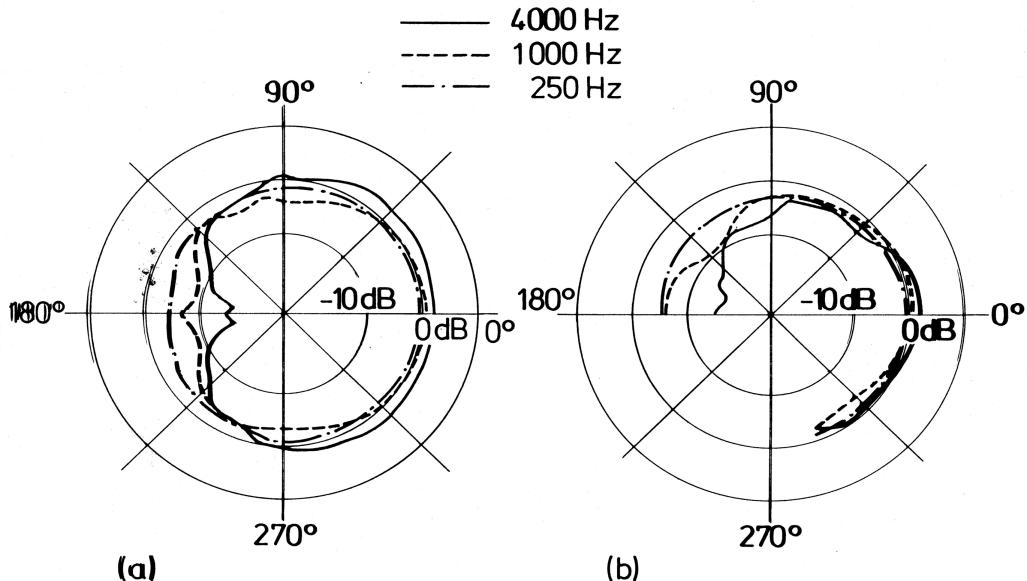
Für die Übertragungsfunktion des Filters lässt sich vereinfacht folgende Allpolstruktur angeben:

$$H(j\omega) = \frac{1}{1 - \sum_{k=1}^n b_k e^{-j\omega k\tau_0}} \quad (3.1)$$

Die Formanten sind die Pole dieses Filters. Das Filter lässt sich als Kaskade oder Parallelschaltung einfacher Teilfilter realisieren.

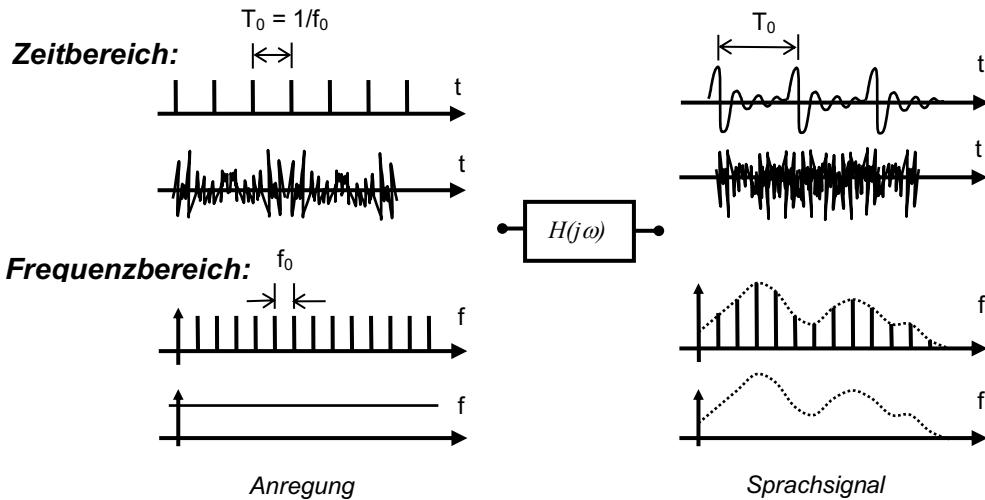
Es ist zu beachten, dass das Filter zeitvariant ist, denn die Stellung der Artikulationsorgane ändert sich permanent (Stationaritätsdauer ca. 20-30 ms).

Die Schallabstrahlung erfolgt über Mund- und Nasenöffnung. Die Richtcharakteristik ist bei tiefen Frequenzen nahezu kugelförmig, bei höheren Frequenzen immer stärker nach vorn gerichtet (durch Abschattung am Kopf), vgl. untenstehende Abbildung.



Richtcharakteristik der Abstrahlung menschlicher Sprache (schematisiert).
Links: Horizontalebene; rechts: Vertikalebene; vgl. Blauert (1994) und Flanagan (1972).

Die Einflüsse von Anregung und Lautformung finden sich direkt im Sprachsignal wieder. Die nachstehende Abbildung zeigt vereinfacht die Zeitsignale und Spektren bei quasi-periodischer bzw. bei rauschförmiger Anregung und Filterung mit dem Vokaltrakt-Filter.



Zeitsignale (obere Hälften) und zugehörige Betragsspektren (untere Hälften)
bei quasi-periodischer und rauschförmiger Anregung.
Links: Anregungssignal; rechts: Sprachsignal.

Im Sprachsignal bei periodischer Anregung finden sich die Formanten als Anteile einer gedämpften Schwingung wieder, die dem periodischen Signal überlagert ist. Beteiligt an diesen Schwingungen sind in erster Linie der erste und der zweite Formant; F_3 ist schon stark gedämpft.

3.4 Sprachlaute

Trotz des kontinuierlichen Charakters des Sprachsignals (und des Spektrums, vgl. das Spektrogramm) lassen sich in gesprochener Sprache einzelne Elemente auditiv unterscheiden, d.h. fortlaufende Sprache kann segmentiert werden. Diese Segmentierung gelingt mit technischen Mitteln allerdings bislang nur unzureichend. Im Folgenden sollen die Elemente von Sprache kurz vorgestellt und den akustischen bzw. artikulatorischen Korrelaten zugeordnet werden.

In der Phonetik werden Sprachlaute klassischerweise zunächst in Vokale (Selbstlaute) und Konsonanten (Mitlaute) unterschieden:

Vokale und Umlaute:

Diese werden durch rein stimmhafte Anregung des Vokaltraktes gebildet. Sie unterscheiden sich bzgl. der Lage des Zungengipfels (vorn, Mitte, hinten) und des Grades der Mundöffnung (offen, halboffen, geschlossen), wie untenstehende Abbildung zeigt. Diese beiden (orthogonalen) Unterscheidungsmerkmale korrelieren auch mit der Lage der ersten beiden Formanten.

Lage des Zungengipfels:		vorn	Mitte	hinten
Öffnungsgrad				
fast geschlossen	i nie I bitte	y Tüte Y Hütte	u Uhr U Mutter	
halboffen	e See ɛ Bett	ø Öse œ Götter ə Gabe	o Not ɔ Otto	
offen	a Mann			a: Tat

Außerdem die Diphthonge aɪ , aʊ , ɔʏ

Unterscheidung der Vokale nach Lage des Zungengipfels und Öffnungsgrad des Mundes (nach Blauert, 1994).

Konsonanten:

Diese werden unterschieden z.B. nach der Anregungsart (stimmhaft, stimmlos), der Lage der Anregungsstelle (Lippen, Zähne, Gaumen, etc.), oder danach, ob der Nasaltrakt angekoppelt ist oder nicht. Zur systematischen Klassifikation lassen sich Systeme von distinktiven, d.h. unterscheidenden Merkmalen aufstellen. Die folgende Abbildung gibt ein solches System zur Unterscheidung deutscher Konsonanten an.

Artikulationsstelle	Lippenlaute (labial)		Zahnlaute (dental)		Zahndammlaute (alveolar)		Gaumenlaute (palatal)		Gaumensegellte. (velar)		Stimmhaftlaute (glottal)	Stimmlosglottal
Anregungsart	sth	stl	sth	stl	sth	stl	sth	stl	sth	stl	sth	stl
Verschlusslaute (Plosivae)	b	p			d	t			g	k		
Enge-laute (Fricativae)	zentral		z ... singen	s ... Haß		f ... rasch	j		ç ... ich	x ... ach		
lateral			v ... Wiege	f ... Affe			l					
Zitterlaute (r- Laute)									Rachen	R		
Öffnungslaute (h- Laut)												h
Nasenlaute (Nasale)	m		n				ŋ ...					
	sth... stimmhaft		stl... stimmlos									

Klassifikation der deutschen Konsonanten (nach Blauert, 1994).

Die akustischen Korrelate verschiedener Sprachlaute im Zeit- und Frequenzbereich wurden bereits in Abschnitt 2.8 diskutiert. Im Frequenzbereich sind vor allem die Gipfel der Einhüllenden der Spektren von Interesse. Diese Formanten entstehen durch die Resonanzen des Vokaltraktes.

Bei einem neutralen Sprechtrakt (d.h. bei konstantem Querschnitt über die gesamte Länge von 17 cm; ungefähr realisiert beim Schwa-Laut /ə/) ergeben sich Formanten bei

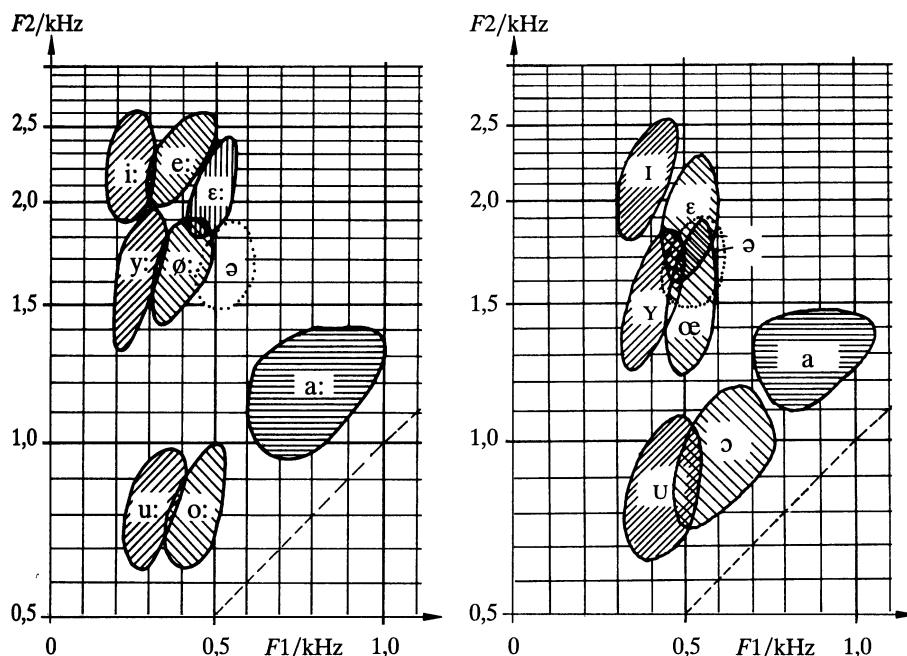
$$l = \frac{\lambda_{F_1}}{4} \rightarrow$$

$$f_{F_1} = \frac{c}{4l} \approx \frac{340 \text{ m/s}}{4 \cdot 0,17 \text{ m}} \approx 500 \text{ Hz} \quad (3.1)$$

$$f_{F_2} = 3 \cdot 500 \text{ Hz} = 1500 \text{ Hz}$$

$$f_{F_3} = 5 \cdot 500 \text{ Hz} = 2500 \text{ Hz}$$

Bei Variation des Sprechtraktes verschieben sich diese Formanten zu lautspezifischen Punkten, die durch Umgebungsläute zu Gebieten verbreitert werden. Die unteren Formanten sind charakteristisch für den zu artikulierenden Laut und lassen sich in sog. „Formantkarten“ darstellen. Hierbei werden F_1 gegen F_2 (und u.U. auch F_3) aufgetragen. Die folgende Abbildung zeigt solche Formantkarten. Die höheren Formanten sind relativ konstant und typisch für den Sprecher; sie können z.B. zur Sprecheridentifizierung eingesetzt werden.



Formantkarten der deutschen Vokale. Links: lange Vokale; rechts: kurze Vokale.
Aus Vary et al. (1998, 47).

Die Sprachlaute unterscheiden sich aber nicht nur bzgl. Ihres Zeitsignals und Spektrums (akustisch), sondern auch durch die Art ihrer Erzeugung (artikulatorisch):

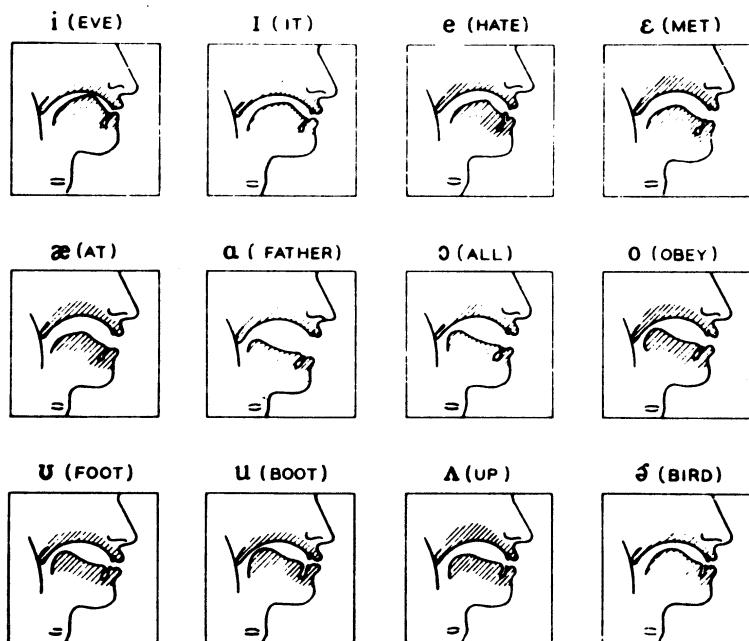
- Die Art der Anregung: stimmhaft, rauschförmig, sprunghaft
- Die beteiligten Elemente des Sprechtraktes
- Der Einstellung der Elemente des Sprechtraktes

Auf die Art der Anregung wurde bereits eingegangen. Nach dem Ort der Erregung können bspw. Zischlaute (Lage der Turbulenz) oder Explosivlaute (Lage des Verschlusses) eingeteilt werden:

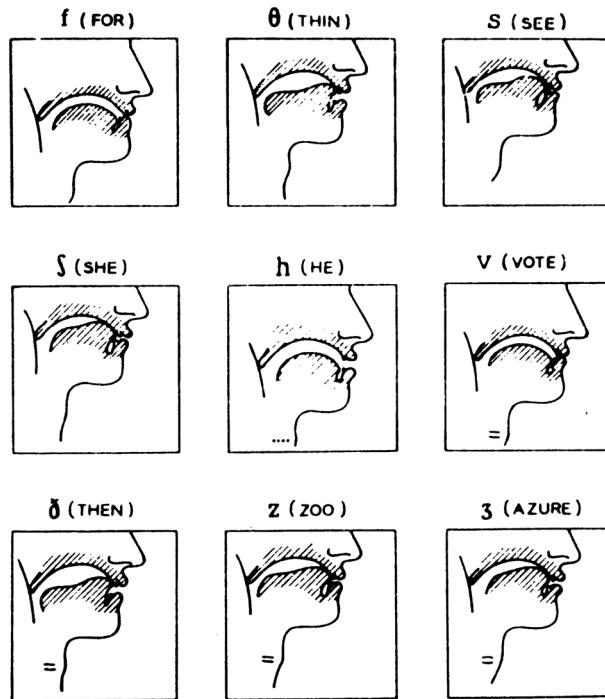
- Zischlaute: Lage der Turbulenz stimmhaft stimmlos
 Lippen / Zähne /v/ (Wasser) /f/ (Fass)
 Zungenspitze/Vordergaumen /z/ (Sonne) /s/ (Fass)
 Glottis /h/ (halt)
 Zunge/Mittelgaumen /ð/ (engl. the) /θ/ (engl. with)
 - Explosivlaute: Lage des Verschlusses stimmhaft stimmlos
 Lippen /b/ (Blei) /p/ (Papier)
 Zungenspitze/Vordergaumen /d/ (du) /t/ (Tag)
 Gaumen/Zungenberg /g/ (Gras) /k/ (kurz)

Bei Lauten mit unterschiedlicher Lage der Erregung spielt nur der Teil des Vokaltraktes, der oberhalb der Erregungsstelle liegt, eine wesentliche Rolle für die Filterung des Anregungssignals. Deshalb sind bei Vokalen, die eine tiefer liegende Erregungsstelle haben, die Formanten ausgeprägter als bspw. bei Zischlauten.

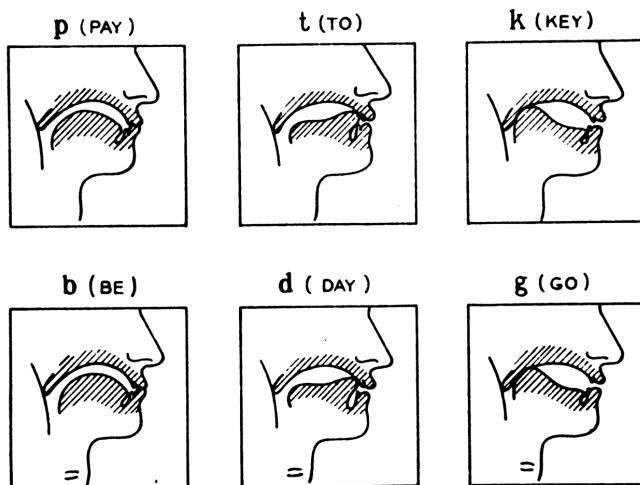
Die folgenden Abbildungen zeigen die Einstellung des Vokaltraktes bei der Artikulation verschiedener Vokale, Frikative und Plosive.



Profile des Sprachtraktes bei der Artikulation unterschiedlicher Vokale (aus Flanagan, 1972, 18).



Profile des Sprachtraktes bei der Artikulation unterschiedlicher Frikative (aus Flanagan, 1972, 20).



Profile des Sprachtraktes bei der Artikulation unterschiedlicher Plosive (aus Flanagan, 1972, 20).

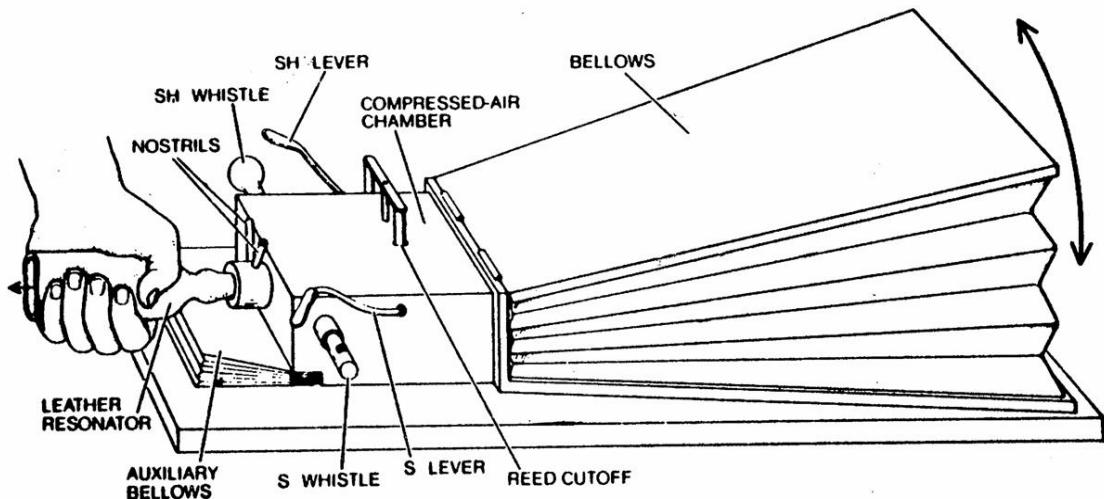
Solche Bilder wurden in der Vergangenheit z.B. aus Röntgenfilmen abgeleitet. Heutzutage gibt es verbesserte Methoden zur messtechnischen Erfassung der Artikulation, bspw. mittels kleiner Spulen, die auf der Zunge und anderen Artikulationsorganen befestigt werden, und deren Position dann im Magnetfeld verfolgt wird.

3.5 Modelle der Spracherzeugung

Es liegt nahe, aus den dargestellten Eigenschaften der Spracherzeugung beim Menschen Modelle zu entwickeln, mit denen sich Sprache synthetisieren, d.h. künstlich herstellen lässt.

Obwohl diese Modelle nur noch selten direkt zur Sprachsynthese eingesetzt werden (vgl. Kapitel 7.2), finden sich viele Grundgedanken dennoch in Verfahren zur Generierung, Übertragung, Kodierung und Aufbereitung von Sprachsignalen enthalten. Daher sollen diese Modelle im Folgenden kurz vorgestellt werden.

Aus den physikalischen Gegebenheiten bei der Spracherzeugung wurden schon sehr früh Modelle zur maschinellen Generierung einzelner Sprachlaute erdacht. Die bekannteste Apparatur wurde von Wolfgang von Kempelen Ende des 18 Jahrhunderts vorgestellt. Es existieren verschiedene Varianten dieser Maschine; eine klassische Variante ist in nachfolgender Abbildung illustriert.



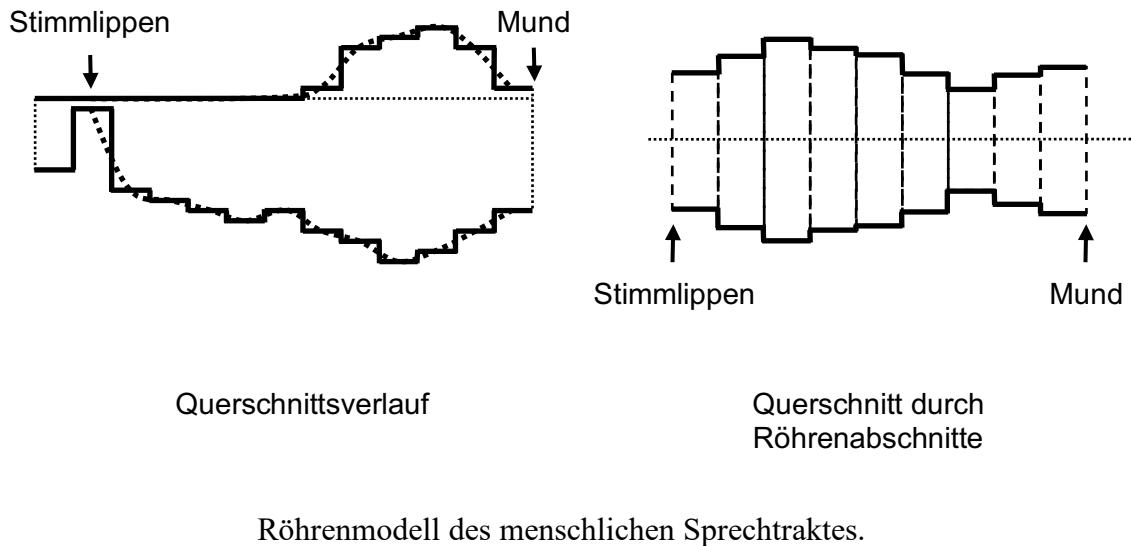
Beispiel für ein mechanisches Modell der Spracherzeugung nach von Kempelen, (1791), aus Blauert (1994).

Ausgehend von der Querschnittsdarstellung des menschlichen Sprechtraktes (vgl. Abschnitt 3.1) lässt sich der Vokaltrakt auch als *akustische Röhre mit veränderlichem, aber stückweise konstantem Querschnitt* darstellen.

Für ein Rohr konstanten Querschnitts ($A = A_0 = \text{konst.}$) wurden im vorangegangenen Abschnitt bereits die (theoretischen) Resonanzfrequenzen berechnet. Für ein Rohr veränderlichen Querschnitts hängt die Querschnittsfläche allgemein von der Orts-Koordinate x ab; die zeitliche Änderung wird hierbei zunächst nicht betrachtet. Für ein solches schallhartes Rohr lässt sich der Schalldruck p aus der Webster'schen Differentialgleichung berechnen:

$$\frac{\partial^2 p}{\partial x^2} + \frac{1}{A} \cdot \frac{dA}{dx} \cdot \frac{\partial p}{\partial x} = \frac{1}{c^2} \cdot \frac{\partial^2 p}{\partial t^2} \quad (3.2)$$

mit den Randbedingungen, dass am Mund der Druck gleich Null sein muss (Abstrahlung in ein schallweiches Medium) und an der Glottis die Schnelle gleich Null (Glottis ist idealerweise schallhart). Die Lösung dieser Gleichung ist im allgemeinen Fall recht aufwändig; sie lässt sich aber vereinfachen, wenn man die Querschnittsfläche als stückweise konstant annimmt; der i.A. beliebig geformte Sprachtrakt wird durch eine Aneinanderschaltung von Röhrenabschnitten konstanten Querschnitts simuliert. Die nachstehende Abbildung veranschaulicht diese Approximation.



Für diese Anordnung vereinfacht sich die Gleichung wie folgt:

$$\frac{\partial^2 p}{\partial x^2} = \frac{1}{c^2} \cdot \frac{\partial^2 p}{\partial t^2} \quad (3.3)$$

Zur Lösung dieser Differentialgleichung setzt man für jeden Abschnitt 1, 2, n eine in positive und eine in negative x-Richtung laufende Welle an:

$$p(x, t) = p_+ (t - \frac{x}{c}) + p_- (t + \frac{x}{c}) \quad (3.4)$$

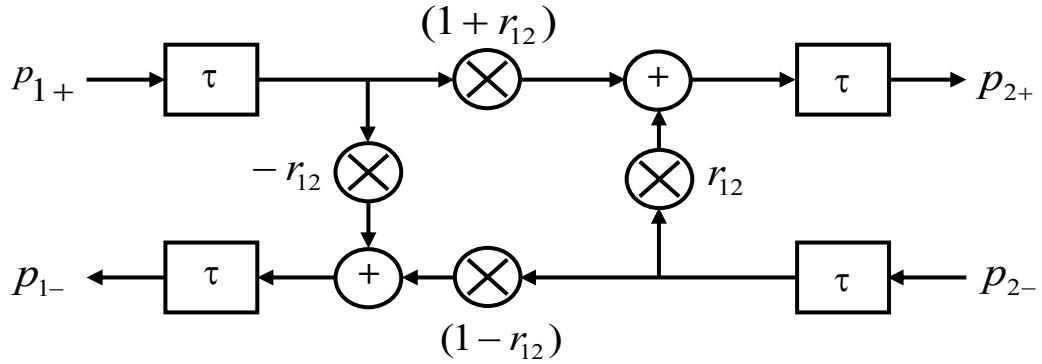
wobei an der Schnittstelle zwischen den Abschnitten 1 und 2 Kontinuität herrschen muss, d.h. $p_1 = p_2$. Mit Einführung des Reflexionsfaktors

$$r_{12} = \frac{A_1 - A_2}{A_1 + A_2} \quad (3.5)$$

lässt sich der Druck an den Rohrendpunkten wie folgt berechnen:

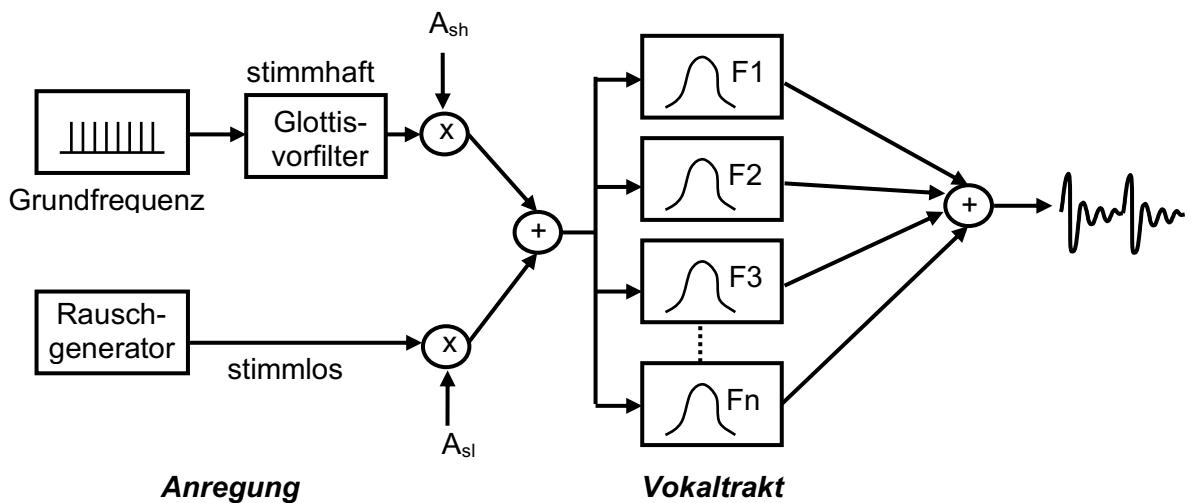
$$\begin{aligned} p_{2+} &= (1 + r_{12}) \cdot p_{1+} + r_{12} p_{2-} \\ p_{1-} &= (1 - r_{12}) \cdot p_{2-} - r_{12} p_{1+} \end{aligned} \quad (3.6)$$

Diese Struktur lässt sich gut als digitales Filter bestehend aus Verzögerungsgliedern τ und Multiplizierern \otimes darstellen (sog. Kelly-Lochbaum-Struktur), vgl. nachstehende Abbildung.



Realisierung des Röhrenmodells des menschlichen Sprechtraktes
als digitales Filter in Kelly-Lochbaum-Struktur.

Wenn man von der physikalischen Struktur des Vokaltraktes abweicht, lässt sich der gesamte Prozess der Spracherzeugung auch als vereinfachtes Modell beschreiben, bei dem ein Vokaltrakt-Filter von einem Anregungssignal angesteuert wird. Man bezeichnet diesen – in der Sprachtechnologie sehr weit verbreiteten – Ansatz als *Quelle-Filter-Modell der Spracherzeugung*. Das Modell ist in folgender Abbildung skizziert.



Quelle-Filter-Modell der menschlichen Spracherzeugung.

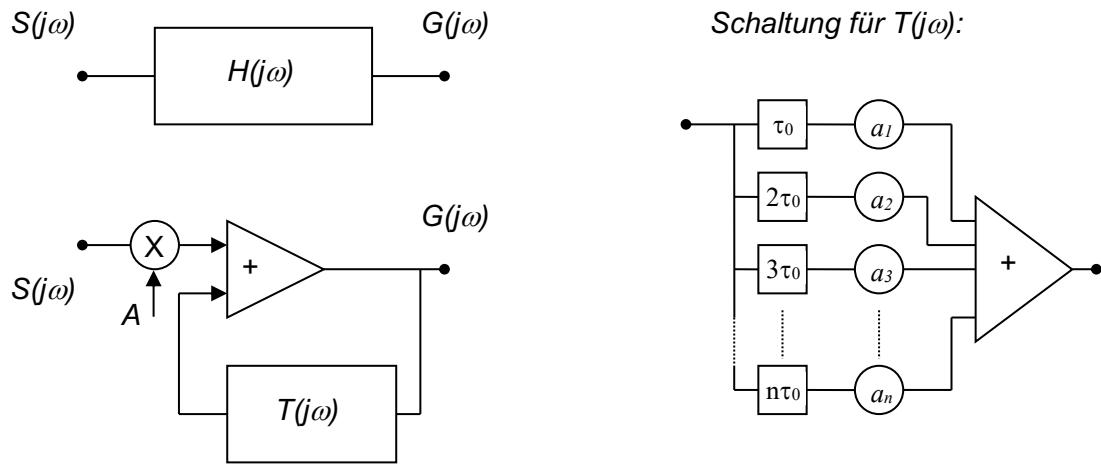
Das Anregungssignal des Quelle-Filter-Modells ergibt sich entweder als Impulsfolge der Frequenz f_0 (Grundfrequenz), die anschließend mit einem Glottis-Vorfilter gefiltert wird (Erzeugung des dreieckförmigen Verlaufs), oder durch einen Rauschgenerator. Diese Anregungssignale können nun in ihrer Amplitude verändert (Multiplizierer) und auch additiv überlagert werden (gemischt stimmhafte/stimmlose Anregung). Das so entstandene Anregungssignal wird nun auf ein Vokaltrakt-Filter gegeben.

Obwohl das Vokaltraktfilter eigentlich sowohl Pol- als auch Nullstellen aufweist, hat es sich für viele technische Anwendungen bewährt, es als Allpolfilter nachzubilden. Dies gelingt allerdings nur dem Betrage nach. In oben stehender Abbildung wurde das Vokaltraktfilter als eine Parallelschaltung einzelner Teilfilter (realisieren jeweils einen Formanten) implementiert.

Alternativ kann es auch als Reihenschaltung einzelner Teilträger (entspricht dann dem Röhrenmodell) realisiert werden:

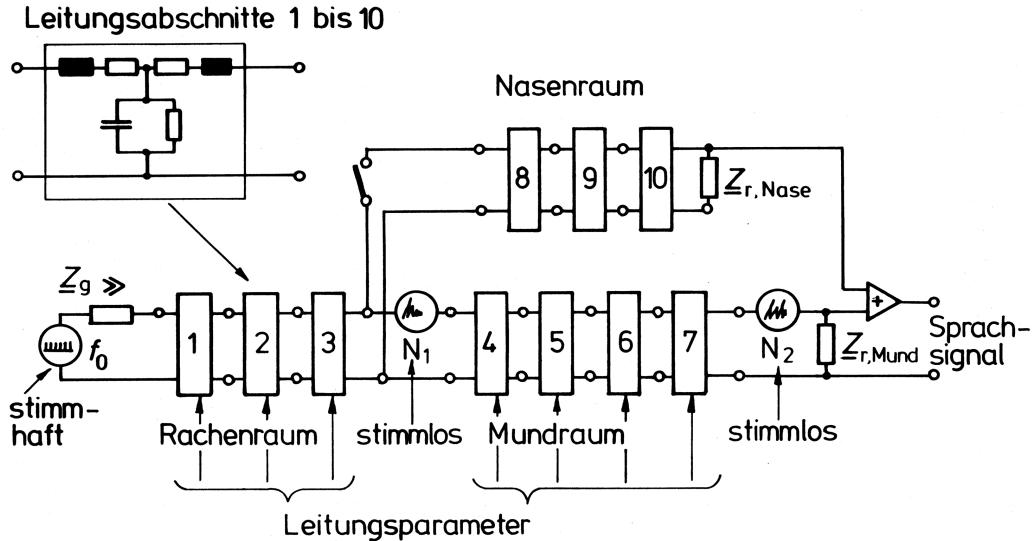
$$\begin{aligned}
 H(j\omega) &= \frac{A}{\prod_{k=1}^m (1 - b_k e^{-j\omega k \tau_0})} \\
 &= \frac{A}{1 - \sum_{k=1}^n a_k e^{-j\omega k \tau_0}} \\
 &= \frac{A}{1 - T(j\omega)}
 \end{aligned} \tag{3.7}$$

Das Filter $T(j\omega)$ wird als Transversalfilter bezeichnet; es lässt sich in folgender Struktur realisieren:



Realisierungsmöglichkeiten des Vokaltraktfilters (nach Blauert, 1994).

Dieses Modell beinhaltet noch etliche Vereinfachungen; bspw. ist keine Sprunganregung integriert, die Erregung findet immer am unteren Ende des Vokaltraktes statt, und die Ankopplung der Signalquelle und des Mundes werden nicht berücksichtigt. Das unten gezeigte – etwas komplexere – Modell berücksichtigt diese Aspekte. Es ist aus Leitungsabschnitten konstanter Länge (aber variablen Querschnitts) aufgebaut, die z.B. als elektrische Schaltungen wie im Inlet gezeigt realisiert werden können. Die Anregung findet am unteren Ende (Glottis), zwischen Mund- und Rachenraum oder am Mund statt; zusätzlich kann der Nasaltrakt über einen Schalter an- und abgekoppelt werden.



Leitungsmodell der menschlichen Spracherzeugung (nach Blauert, 1994).

Bei der Erzeugung fortlaufender Sprache müssen die Steuerparameter (Signalquellen, Parameter oder Bauelemente der Teilfilter) alle ca. 2,5 bis 20 ms nachgeführt werden, damit sich der Eindruck eines kontinuierlichen Signals ergibt. Innerhalb dieser Intervalle können die betrachteten Elemente als konstant angesehen werden.

3.6 Literatur

- Blauert, J. (1994). Kommunikationsakustik II: Audiokommunikation und virtuelle Realität. Skriptum zur Vorlesung am Institut für Kommunikationsakustik, Ruhr-Universität, Bochum.
- Flanagan, J.L. (1972). Speech Analysis, Synthesis and Perception. Springer Verlag, Berlin.
- Heute, U. (1990). Sprachverarbeitung. Skriptum zur Vorlesung der Arbeitsgruppe Digitale Signalverarbeitung, Ruhr-Universität, Bochum.
- Vary, P., Heute, U., Hess, W. (1998). Digitale Sprachsignalverarbeitung. B.G. Teubner, Stuttgart.

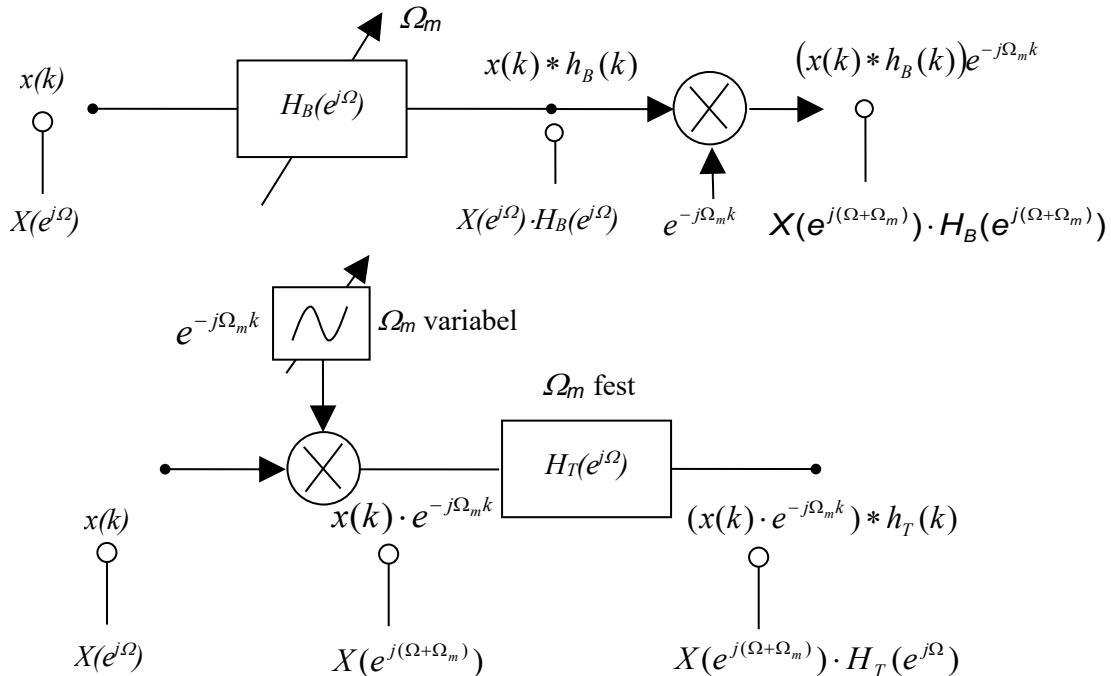
4. Sprachsignalanalyse

In diesem Kapitel sollen kurz einige ausgewählte Verfahren zur Analyse von Sprachsignalen vorgestellt werden, die im Verlaufe der Vorlesung noch weiter verwendet werden. Hierzu zählt zunächst die Spektralanalyse von Sprachsignalen, die ja schon in Kapitel 2 und 3 vorgestellt wurde. Darüber hinaus gibt es Verfahren, die sich explizit auf das im vorangegangenen Abschnitt vorgestellte Quelle-Filter-Modell der Spracherzeugung beziehen. Hierzu zählen das sog. Cepstrum sowie die lineare Prädiktion.

4.1 Spektralanalyse

Das Prinzip der Spektralanalyse mittels Fourier-Transformation wurde bereits in Kapitel 2 behandelt. Allerdings ergeben sich noch zwei Alternativen zur direkten Transformationsberechnung, die zuvor kurz betrachtet werden sollen.

Zur Berechnung der spektralen Anteile bei allen Frequenzen f kann das interessierende Sprachsignal z.B. gespeichert und durch einen durchstimmbaren Bandpass, d.h. ein Filter, dass nur Frequenzen in einem bestimmten engen Bereich (Frequenzband $\Delta\Omega$) um die Mittenfrequenz Ω_m durchlässt, geschickt werden. Die nachfolgende Abbildung zeigt ein solches Analysesystem. Man erhält dann am Ausgang des Bandpasses eine Schwingung mit der Bandpass-Mittenfrequenz und einer (komplexen) Amplitude, die proportional dem Spektralwert des zu untersuchenden Signals – allerdings modifiziert durch die normalerweise nicht-ideale Übertragungsfunktion des Bandpasses – bei der Frequenz $\Omega = \Omega_m$ ist. Um nicht die komplette Schwingung, sondern nur ihre (komplexe) Amplitude $X(e^{j\Omega_m})$ zu bestimmen, kann am Ausgang des Bandpasses ein Demodulator nachgeschaltet werden.



Spektralanalyse mittels durchstimmbarem Bandpassfilter (oben) und Suchtonanalyse (unten).

Zur Vermeidung eines durchstimmbaren Filters kann das Eingangssignal auch mittels des Modulators auf den Durchlassbereich $\Delta\Omega$ des Filters verschoben werden; hierbei wird der Modulationssatz der Fourier-Transformation ausgenutzt. Man bezeichnet das Prinzip als Suchtonanalyse. Das Ergebnis ist dasselbe, sofern

$$H_T(e^{j\Omega}) = H_B(e^{j(\Omega+\Omega_m)}) \quad (4.1)$$

bzw.

$$H_T(e^{j(\Omega-\Omega_m)}) = H_B(e^{j\Omega}) \quad (4.2)$$

d.h. wenn ihre Frequenzgänge durch Verschiebung auseinander hervorgehen. Für die Impulsantworten der Filter heißt dies

$$h_T(k) \cdot e^{j\Omega_m k} = h_B(k) \quad (4.3)$$

Dies lässt sich leicht aus dem Verschiebungssatz der Fouriertransformation ableiten.

Am Ausgang der Anordnung ergibt sich dann jeweils

$$\begin{aligned} y(k) &= (x(k) * h_B(k)) \cdot e^{-j\Omega_m k} = (x(k) \cdot e^{-j\Omega_m k}) * h_T(k) \\ &= \sum_{\kappa=-\infty}^{\infty} (x(\kappa) \cdot h_T(k-\kappa)) \cdot e^{-j\Omega_m \kappa} \end{aligned} \quad (4.4)$$

Man bezeichnet das Ausgangssignal $y(k)$ als das *Kurzzeitspektrum* des Signals $x(k)$. Diese Größe ist offensichtlich eine Fouriertransformierte – allerdings nicht die des Signals $x(k)$, sondern die des mit der Folge $h_T(k-\kappa)$ gewichteten Signals – an der Stelle $\Omega = \Omega_m$. Da die Tiefpass- und Bandpass-Filter kausal und stabil sein sollten, nimmt $h_T(k)$ mit steigendem k ab. Durch die Multiplikation $x(k) \cdot h_T(k-\kappa)$ wird also ein Stück der „Signalvergangenheit“ aus dem Signal ausgeblendet; weiter zurückliegende Signalwerte werden schwächer gewichtet und gehen schwächer in das Ausgangssignal ein.

In bestimmten Fällen ist $h_T(k)$ sogar endlich lang: Dann blendet $h_T(k-\kappa)$ ein Stück des Signals ein, welches für die momentane Spektralwertbestimmung verwendet wird, also „sichtbar“ ist. Man spricht dann von einer Signal-Fensterung. Das Fenster $h_T(k-\kappa)$ gleitet mit wachsendem k über das gesamte Signal $x(k)$; es führt eine *gleitende Kurzzeit-Spektralanalyse* aus.

Um das Speichern und sequentielle Berechnen des Spektrums zu vermeiden kann man auch eine sog. Filterbank mit parallelen Bandpässen unterschiedlicher Mittenfrequenzen (Durchlassfrequenzen) verwenden. Man bezeichnet diese als eine *Bandpassfilterbank*. Allerdings wird dabei der Aufwand zur Realisierung der Filter größer, da die Filteranzahl steigt.

Zwei Arten von Bandpassfilterbänken sind üblich:

- Filter konstante *absoluter* Bandbreite: $\Delta\Omega = \text{konst.}$
- Filter konstanter *relativer* Bandbreite: $\Delta\Omega / \Omega = \text{konst.}$

Je nachdem, welche Filter man verwendet, sind die Messergebnisse unterschiedlich zu interpretieren. Insbesondere die Filter konstanter relativer Bandbreite sind bei der Sprachsignalanalyse von Interesse, da sie in etwa der Frequenzanalyse im menschlichen Ohr (vgl. nächstes Kapitel) entsprechen.

Bei der Analyse zeitlich variabler Signale (wie bei Sprachsignalen) ist eine vorhergehende Fensterung des Signals wichtig, wenn man Informationen über einzelne Signalabschnitte (und zugehörige Sprachlaute) erhalten möchte. Die Multiplikation mit der Fensterfunktion im Zeitbereich (Ausblenden des interessierenden Signalabschnitts) führt allerdings zur Faltung mit der Fourier-Transformierten der Fensterfunktion im Frequenzbereich; dadurch wird der spektrale Anteil des zu analysierenden Signals verändert. Diese sog. Fenstereffekte sind bei der Interpretation der erhaltenen Spektrallinien zu berücksichtigen.

Zur Bestimmung des gesamten Spektrums – nicht nur einzelner Spektralkomponenten – eines diskreten Signals bietet sich natürlich die direkte Auswertung der Gl. (2.78) bzw. (2.82) an. Aus Gründen der Realisierbarkeit muss dabei die Anzahl der Frequenzkomponenten begrenzt sein, wie es in Gl. (2.82) gegeben ist. Die Berechnung aller Werte X_μ für alle $\mu \in \{0, 1, 2, \dots, M-1\}$ liefert M Spektralwerte zu den Komponenten $\Omega_\mu = \mu \frac{2\pi}{M}$. Diese Komponenten beschreiben $x(k)$ exakt, wenn es sich um ein Signal endlicher Länge oder um ein periodisches Signal der Periodenlänge M handelt. Sind die verwendeten Abtastpunkte des Signals $x(k)$ allerdings nur ein Ausschnitt eines ursprünglich längeren (oder gar unendlich langen) Signals, so erhält man nur eine Näherung.

Wir betrachten ein solches längeres Signal und lassen es durch eine Bandpassfilterbank laufen. Am Ausgang eines einzelnen Bandpasses mit dem Index μ ergibt sich nach Gl. (4.4) dann das Signal

$$y_\mu(k) = \sum_{\kappa=-\infty}^{\infty} (x_\mu(\kappa) \cdot h_T(k-\kappa)) \quad (4.5)$$

wobei

$$x_\mu(k) = x(k) \cdot e^{-j\mu\Omega_0 k} \quad (4.6)$$

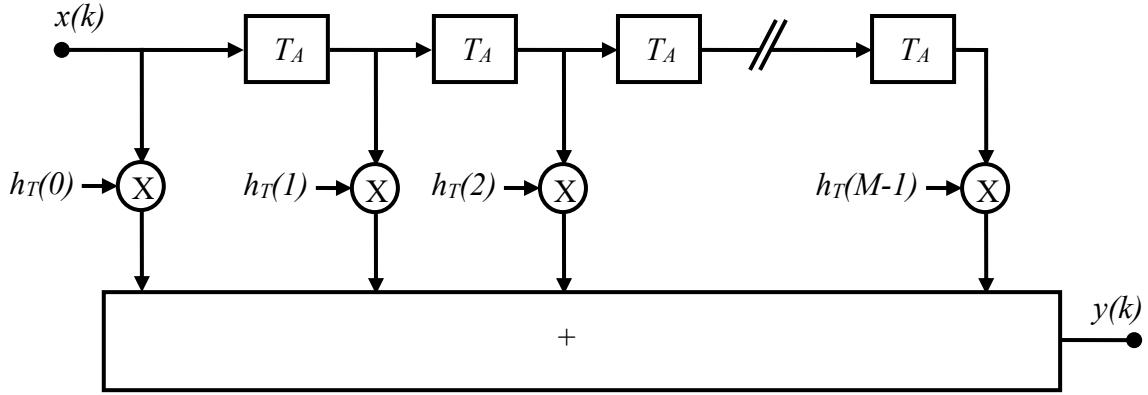
die de-modulierten Signale des unteren Teils der vorangegangenen Abbildung darstellen. Wählt man als Spezialisierung nun für $h_T(k)$ ein nichtrekursives Filter vom Grad $M-1$, d.h. ein Filter mit

$$h_T(k) \begin{cases} \neq 0 & \text{für } k \in \{0, 1, \dots, M-1\} \\ = 0 & \text{für } k \notin \{0, 1, \dots, M-1\} \end{cases} \quad (4.7)$$

so ergibt sich für das Ausgangssignal

$$y_\mu(k) = \sum_{\kappa=k-(M-1)}^k (x_\mu(\kappa) \cdot h_T(k-\kappa)) \quad (4.8)$$

d.h. ein Kurzzeitspektrum in einem endlich langen Fenster. Eine Realisierung hierfür ist in folgender Abbildung gezeigt.



Realisierung einer DFT als Filterbank.

Wir betrachten das Ausgangssignal nun zu festen Zeitpunkten $k = n \cdot M - 1$, d.h. alle M Takte. Eine solche Betrachtung lässt eine Abtastratenreduktion um den Faktor M zu, und ist erlaubt, sofern wir Filter verwenden, deren Bandbreite $\Delta\Omega$ kleiner als $2\pi/M$ ist. Wir nehmen ferner an, dass

$$h_T(k) \begin{cases} = 1 & \text{für } k \in \{0, 1, \dots, M-1\} \\ = 0 & \text{für } k \notin \{0, 1, \dots, M-1\} \end{cases} \quad (4.9)$$

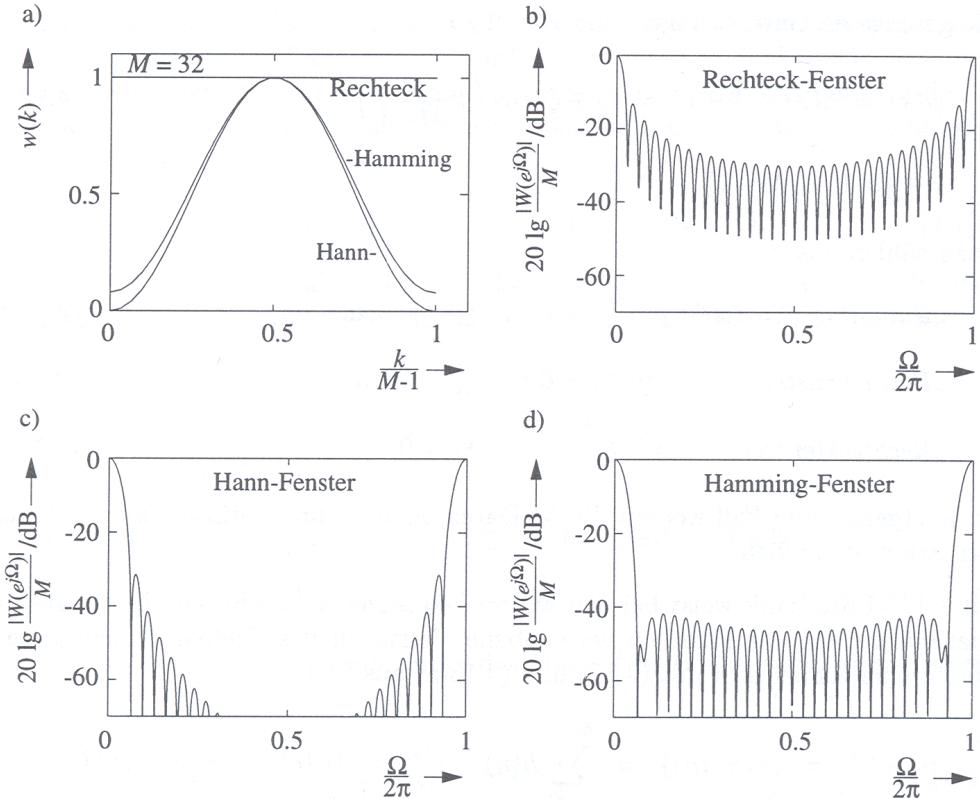
ein Recheckfenster darstellt. In diesem Falle ergibt sich das Ausgangssignal $y_\mu(k)$ zu

$$\begin{aligned}
 y_\mu(k) &= \sum_{\kappa=0}^{M-1} (x_\mu(\kappa) \cdot 1) \\
 &= \sum_{\kappa=0}^{M-1} x(\kappa) \cdot e^{-j\mu\kappa \frac{2\pi}{M}} \\
 &= DFT\{x(k)\}
 \end{aligned} \tag{4.10}$$

Wir können die DFT also auffassen als eine spezielle Filterbank mit M äquidistanten Kanälen, einer Unterabtastung um den Faktor M , sowie mit speziellen Filtern, die als Impulsantwort eine Rechteckfunktion aufweisen. Der Frequenzgang des Tiefpassfilters lässt sich durch Berechnung der DFT des Rechtecksignales bestimmen; er ergibt sich zu

$$H_T(e^{j\Omega}) = e^{-j\frac{M-1}{2}\Omega} \cdot \frac{\sin\left(\frac{M}{2}\Omega\right)}{\sin\left(\frac{\Omega}{2}\right)} \quad (4.11)$$

Der Verlauf dieses Frequenzganges ist in folgender Abbildung gezeigt.



Impulsantworten (oben links) und Übertragungsfunktionen (unten links und rechts) des Rechteck-, Hann- und Hamming-Fensters. Nach Vary et al. (1998, 84).

Offenbar ist das bei der DFT verwendete Filter nicht optimal: Es weist nur eine geringe Sperrdämpfung auf, insbesondere bei weit abseits der Frequenz 0 gelegenen Frequenzen. Die Durchlassbreite ist mit $2 \cdot \Omega_0$ ebenfalls recht breit. Man kann das Sperrverhalten verbessern, indem man die Rechteckfunktion ersetzt durch eine allgemeine Gewichtung, etwa vom Typ

$$h_T(k) = \alpha + \beta \cdot \cos\left(k \frac{2\pi}{M}\right) \quad (4.12)$$

Üblicherweise verwendet man folgende Fensterfunktionen:

- *Hamming-Fenster:* $\alpha = 0.54, \beta = -0.46$
- *Hann-Fenster:* $\alpha = 0.5, \beta = -0.5$
- *Rechteck-Fenster:* $\alpha = 1, \beta = 0$

Die zugehörigen Zeitfunktionen und Frequenzgänge sind ebenfalls in o.a. Abbildung gezeigt. Man sieht, dass sich die Sperrdämpfung beim Hamming- und Hann-Fenster gegenüber dem Rechteck verbessert hat, dass aber die Durchlassbreite im Gegenzug verdoppelt wurde. Um bei gleicher Durchlassbreite die Sperrdämpfung zu verbessern kann man die Fensterlänge bei gleicher Kanalzahl vergrößen; man kommt dann zu sogenannten Polyphasen-Filterbänken. Details hierzu finden sich z.B. bei Vary et al. (1998).

Möchte man für ein Signal $x(k)$ alle M Werte X_μ der Diskreten Fourier-Transformation nach Gl. (2.82) berechnen, so erfordert das bei allgemeinen (komplexen) Signalen M^2 komplexe Multiplikationen und ebenso viele Additionen. Bei typischen Werten $M \approx 1000$ und Abtastraten im Bereich 8...48 kHz ergibt das etwa $8 \dots 48 \cdot 10^6$ Operationen (Multiplikationen kombiniert mit Additionen). Um diesen Rechenaufwand zu verringern gibt es deshalb schon

seit längerem Verfahren, die i. Allg. unter dem Begriff *Fast Fourier Transform (FFT)* zusammengefasst werden.

Ein inzwischen schon „klassischer“ Weg nutzt die Periodizität der Koeffizienten $w_M^{\mu k}$ aus Gl. (2.82) und (2.83) aus. Weiter setzen wir voraus, dass M eine grade Zahl ist. In diesem Falle kann man die Koeffizienten X_μ in zwei Teilen – einem mit geraden und einem mit ungeraden k -Werten – berechnen:

$$X_\mu = \sum_{k=0}^{M-1} x(k) \cdot w_M^{\mu k} = \sum_{k=0}^{M/2-1} x(k) \cdot w_M^{2\mu k} + \sum_{k=0}^{M/2-1} x(2k+1) \cdot w_M^{2\mu k} w_M^\mu \quad (4.13)$$

Wegen $w_M = e^{-j\frac{2\pi}{M}}$ gilt $w_M^{2\mu k} = w_M^{\mu k}$, und mit

$$\begin{aligned} x_1(k) &= x(2k) \quad , k \in \left\{0, 1, \dots, \frac{M}{2}-1\right\} \\ x_2(k) &= x(2k+1) \quad , k \in \left\{0, 1, \dots, \frac{M}{2}-1\right\} \end{aligned} \quad (4.14)$$

lässt sich Gl. (4.5) umschreiben in

$$X_\mu = \sum_{k=0}^{M/2-1} x_1(k) \cdot w_{\frac{M}{2}}^{\mu k} + w_M^\mu \cdot \sum_{k=0}^{M/2-1} x_2(k) \cdot w_{\frac{M}{2}}^{\mu k} \quad (4.15)$$

Anstelle einer DFT mit der Länge M lassen sich also 2 DFTs mit der Länge $\frac{M}{2}$ berechnen.

Dabei ergibt sich ein Aufwand von $2 \cdot \left(\frac{M}{2}\right)^2 = \frac{M^2}{2}$ komplexen Multiplikationen und

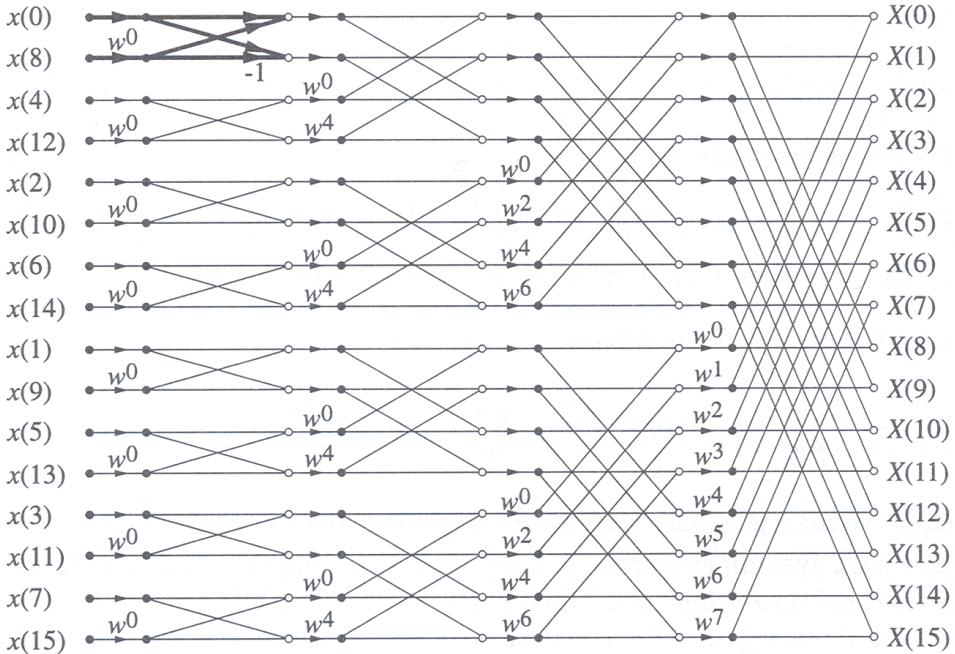
Additionen, plus M komplexen Multiplikationen mit dem Faktor w_M^μ , und die M Additionen der Teilergebnisse. Bezogen auf den ursprünglichen Aufwand von M^2 komplexen Multiplikationen und Additionen ergibt sich eine Aufwandsreduktion um

$$\frac{2 \cdot \left(\frac{M}{2}\right)^2 + M}{M^2} = \frac{1}{2} + \frac{1}{M} \quad (4.16)$$

Bei großen M entspricht dies fast einer Aufwandsreduktion um die Hälfte.

Im Falle, dass $\frac{M}{2}$ ebenfalls gerade ist, lässt sich die Prozedur weiter fortsetzen, d.h. die beiden Teilsummen lassen sich wiederum in zwei DFTs der Länge $\frac{M}{4}$ zerlegen. Das

Verfahren lässt sich am weitesten treiben, wenn $M = 2^n$ ist, d.h. wenn die DFT-Länge eine Zweierpotenz ist. Dieses Verfahren der fortgesetzten Halbierung der DFT-Berechnung im Zeitbereich heißt Radix-2/Decimation-in-Time-Algorithmus. Der Signalflußgraph für diese Berechnung ist in folgender Abbildung dargestellt. Das darin wiederkehrende (fett hervorgehobene) Element nennt man *Butterfly*.



FFT-Berechnung mittels Radix-2/Decimation-in-Time-Algorithmus. Nicht explizit bezeichnete Zweige tragen den Gewichtungsfaktor 1 (nach Vary et al., 1998, 78).

Dieser Algorithmus benötigt zur Berechnung einer DFT der Länge M

$$\frac{M}{2} \cdot m = \frac{M}{2} \cdot \log_2 M \quad (4.17)$$

Multiplikationen. Insbesondere bei großen Werten M führt dies zu drastischen Rechenzeiteinsparungen. Für $M = 16$ erreicht man eine Verringerung um den Faktor $\frac{m}{2M} = \frac{1}{8}$; bei $M = 1024$ bleiben statt über 10^6 nur noch 5120 Operationen, d.h. etwa 0.5% der Operationen. Hierbei ist bereits berücksichtigt, dass man die Symmetrie der $w_M^{\mu k}$ ausnutzen kann, um vor dem Grundelement, dem Butterfly, nur eine Multiplikation rechnen zu müssen. Die Gesamtzahl komplexer Additionen ist doppelt so groß.

Neben dem Radix-2/Decimation-in-Time-Verfahren gibt es andere effiziente Algorithmen, die zum Teil ähnlich Strukturen erzeugen. Details dazu findet man z.B. bei Oppenheim und Schafer (1999). Da in der Anwendung häufig nur reellwertige Signale vorkommen (z.B. bei Sprachsignalen) kann man die Nullen der Imaginärteile vorteilhaft nutzen, indem man $x_1(k)$ und $x_2(k)$ in einem Signalblock rechnet. Man bildet dazu künstlich ein komplexwertiges Signal $x_0(k) = x_1(k) + j \cdot x_2(k)$.

4.2 Cepstrum

Ausgangspunkt dieser Signalanalysetechnik ist zunächst das Quelle-Filter-Modell der Spracherzeugung:

$$G(j\omega) = S(j\omega) \cdot H(j\omega) \quad (4.18)$$

wobei $G(j\omega)$ die Ausgangsfunktion des Vokaltraktfilters ist, $S(j\omega)$ die Anregungsfunktion, und $H(j\omega)$ die Übertragungsfunktion des Vokaltraktfilters, vgl. Kapitel 3.5.

Logarithmierung des Betrags des Ausgangssignals ergibt:

$$\ln|G(j\omega)| = \ln|S(j\omega) \cdot H(j\omega)| = \ln|S(j\omega)| + \ln|H(j\omega)| \quad (4.19)$$

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

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

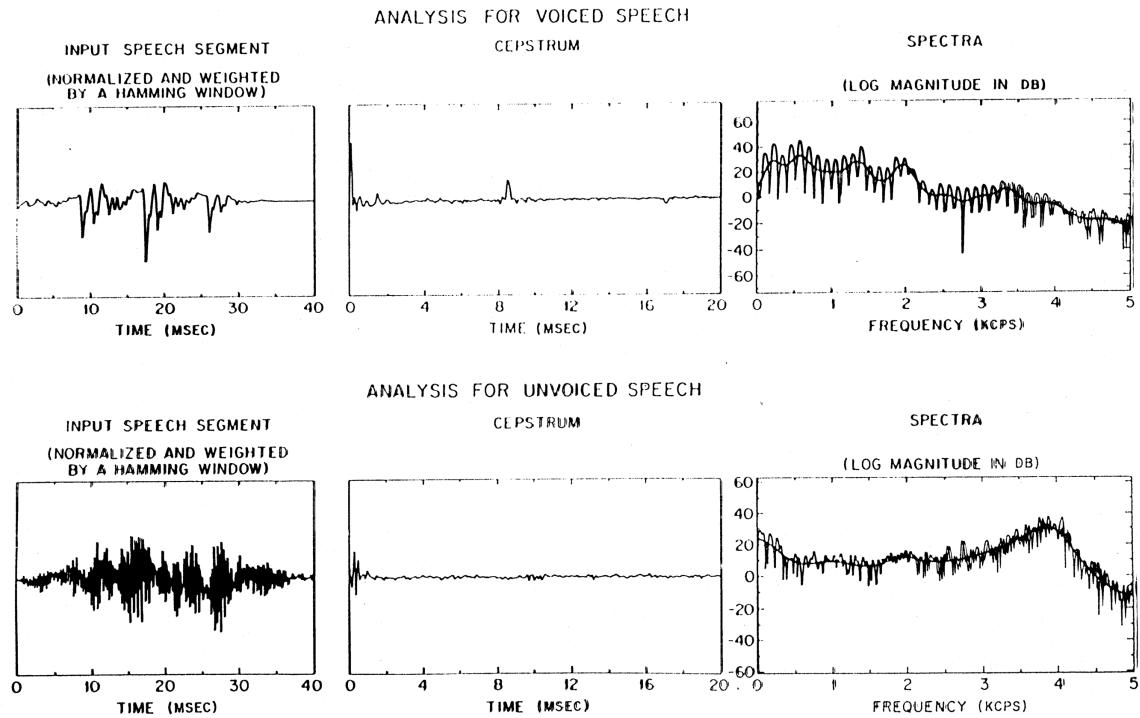
Die Variable x entspricht ihrer Dimension nach einer Zeit; man bezeichnet sie als *quefrency* (umgedreht aus frequency). Gleichermaßen stammt der Begriff *Cepstrum* aus einer Verballhornung des Spektrums; man bezeichnet Filterungen im Quefrency-Bereich als *liftering* (statt filtering), und entsprechend einen Tiefpass als short-pass lifter bzw. einen Hochpass als long-pass lifter. Die Filterung im x -Bereich wird auch als homomorphe Filterung bezeichnet, denn durch die nichtlineare Transformation werden die Signale so dargestellt, dass eine verallgemeinerte lineare Filterung möglich wird.

Neben diesen amüsanten Aspekten hat das Cepstrum aber eine für die Sprachsignalanalyse sehr bedeutende Eigenschaft: 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. Die nachfolgende Abbildung verdeutlicht diese Eigenschaft am Beispiel je eines stimmhaften und eines stimmlosen Signalabschnitts.

Das Cepstrum kann natürlich auch in der digitalen Form angegeben werden. Man verwendet dann normalerweise die DFT (bzw. IDFT) und schreibt:

$$\begin{aligned} C(n) &= IDFT\{\ln|G_\mu|\} \\ &= IDFT\{\ln|S_\mu| + \ln|H_\mu|\} \\ &= C_1(n) + C_2(n) \end{aligned} \quad (4.21)$$

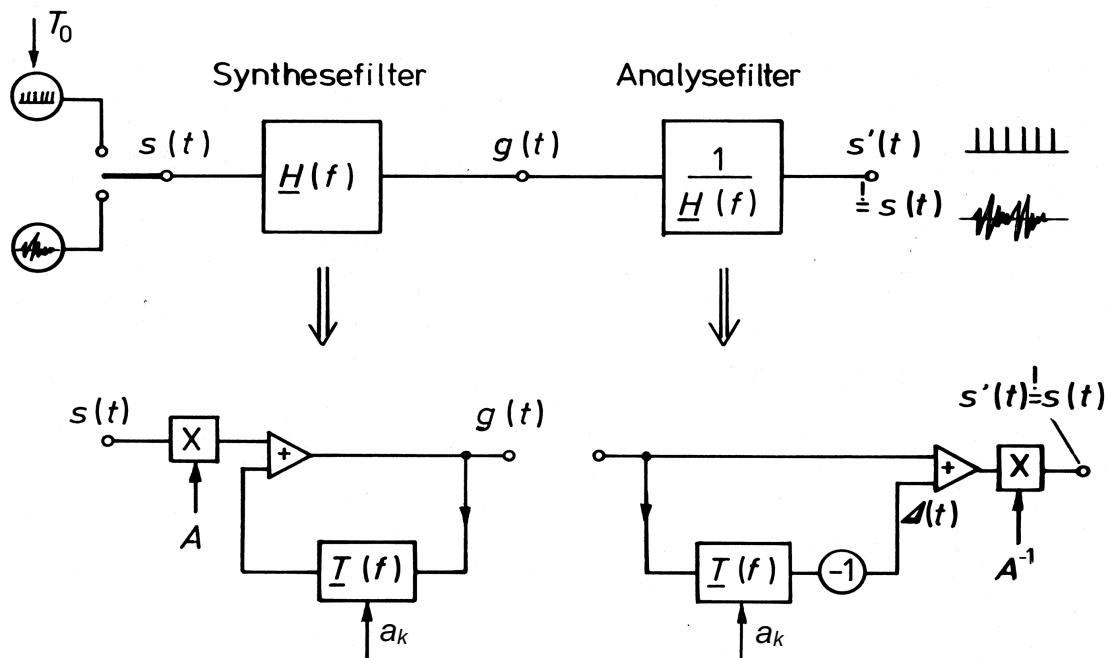
Das entstehende Cepstrum ist (aufgrund der Betragsbildung vor dem Logarithmus) reell.



Cepstrum eines stimmhaften und eines stimmlosen Signalabschnitts (nach Schäfer und Rabiner, 1970, zitiert nach Flanagan, 1972, 175).

4.3 Lineare Prädiktion

Auch die dritte hier behandelte Analysemethode geht vom Quelle-Filter-Modell der Spracherzeugung aus, wobei man den Vokaltrakt zunächst als zeitlich konstant annimmt. 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). Bei Annahme einer Allpolstruktur für das Vokaltraktfilter $H(j\omega)$ erhält das dazu inverse Filter eine besonders einfache Struktur, wie der folgenden Abbildung zu entnehmen ist.



Prinzip der Sprachsignalanalyse durch lineare Prädiktion, nach Blauert (1994).

Die Analyseaufgabe besteht nun darin, die Koeffizienten a_k 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.

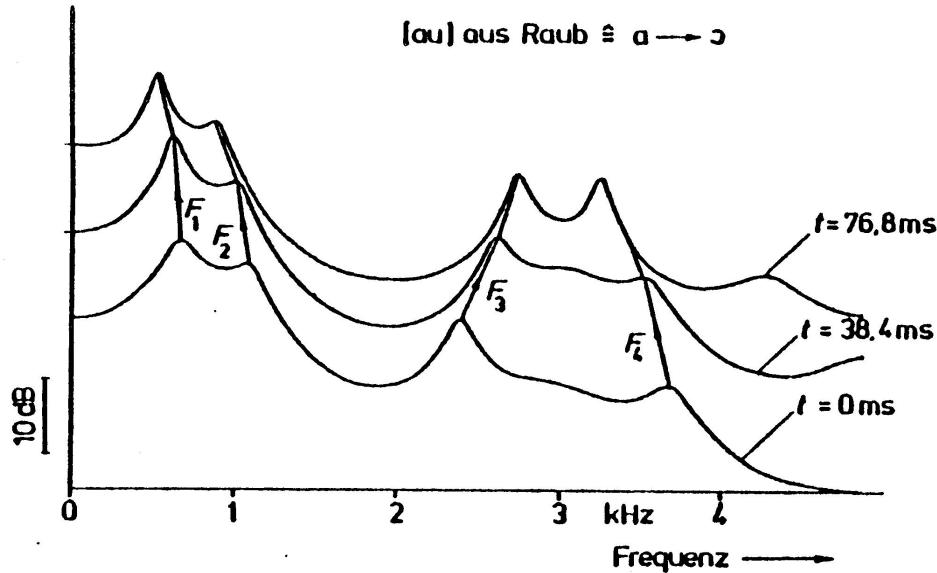
Das Filter $T(j\omega)$ kann – wie in Abschnitt 3.5 dargestellt – als rein nichtrekursives Filter realisiert werden; es blendet aus dem Sprachsignal $g(t)$ die vergangenen Werte zu den Zeitpunkten $k \cdot \tau_0$ aus und berechnet daraus – mit geeigneten Gewichtungen a_k – ein Signal $g(t) - A \cdot s(t)$ zum aktuellen Zeitpunkt als Linearkombination. 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 a_k werden so bestimmt, dass die Differenz zwischen $s(t)$ und $s'(t)$ im Zeitbereich bzw. die Differenz zwischen $S(j\omega)$ und $S'(j\omega)$ im Frequenzbereich minimal wird. Hierzu stehen verschiedene Methoden zur Verfügung, die meist eine Minimierung des mittleren quadratischen Fehlers anstreben. Die Berechnung muss – da das Filter zeitlich variant ist – ständig wiederholt werden, um zu jedem Signalabschnitt ein passendes Filter zu erhalten. Zur Vermeidung einer kompletten Neuberechnung sind Methoden zur blockweisen oder sequentiellen Adaption der Filterkoeffizienten entwickelt worden, vgl. Vary et al. (1998, 174-187).

Durch die inverse Filterung mit dem „Analysefilter“ in oben stehender Abbildung wird das Spektrum des Sprachsignals seiner spektralen Einhüllenden beraubt; es wird somit „weiß“ gemacht, d.h. bei stimmlosen Abschnitten wird ein konstantes kontinuierliches Spektrum und bei stimmhaften Abschnitten ein konstantes Linienspektrum angestrebt. Durch die „weißmachende“ Eigenschaft der linearen Prädiktion wird auch ein großer Teil der Information aus dem Sprachsignal extrahiert; man kann die Prädiktorkoeffizienten daher gut zur effizienten

Kodierung und Übertragung von Sprache verwenden. Die damit erzielbare Bitratenreduktion ist in Kapitel 6 beschrieben.

Im Prädiktorfilter (Synthesefilter) ist die Struktur des Vokaltraktes, d.h. die Formanten, gespeichert, möglichst ohne Anteile des Eingangssignals zu berücksichtigen. Mittels der linearen Prädiktion lassen sich daher gut Formanten extrahieren und Formantverläufe über der Zeit beobachten. Ein Beispiel dafür ist in untenstehender Abbildung angegeben.



Beispiel für Formanterkennung und -verfolgung mittels linearer Prädiktion.
Prädiktorfilter 10. Ordnung (aus Blauert, 1994).

Bei einem Prädiktorgard n von typischerweise 8...10 steht zur Prädiktion ein Signalabschnitt von $8\dots10\cdot\tau_0 \approx 1 \dots 1,25 \text{ ms}$ zur Verfügung. Dieser Zeitbereich ist klein gegen den Bereich der Grundperiode von Sprache ($50 \dots 250 \text{ Hz}$ entsprechend $4 \dots 20 \text{ ms}$). Im Restsignal des Prädiktorfilters (Analysefilter) bleibt daher auch nach optimaler Bestimmung der Prädiktorkoeffizienten bei stimmhaften Signalabschnitten ein periodisches Anregungssignal zurück. Dies ist zwar interessant zur Signalanalyse, nicht jedoch bei der effizienten Kodierung von Sprache.

Man schaltet daher manchmal dem bislang behandelten *Kurzzeit-Prädiktor* noch einen *Langzeit-Prädiktor* nach. Dieser berechnet ein Differenzsignal $s''(t)$ aus $s'(t)$ wie folgt:

$$s''(t) = s'(t) - b \cdot [s'(t - T_0)] \quad (4.22)$$

Hierzu ist eine Schätzung der Länge der Grundperiode T_0 notwendig. Die beiden Parameter b und T_0 werden so gewählt, dass die Energie des Fehlersignals $s''(t)$ über ein begrenztes Zeitintervall minimal wird. Das entstehende Ausgangssignal ist dann allerdings nicht mehr ähnlich dem Anregungssignal; es fehlt hier vor allem die Grundfrequenz des Anregungssignals.

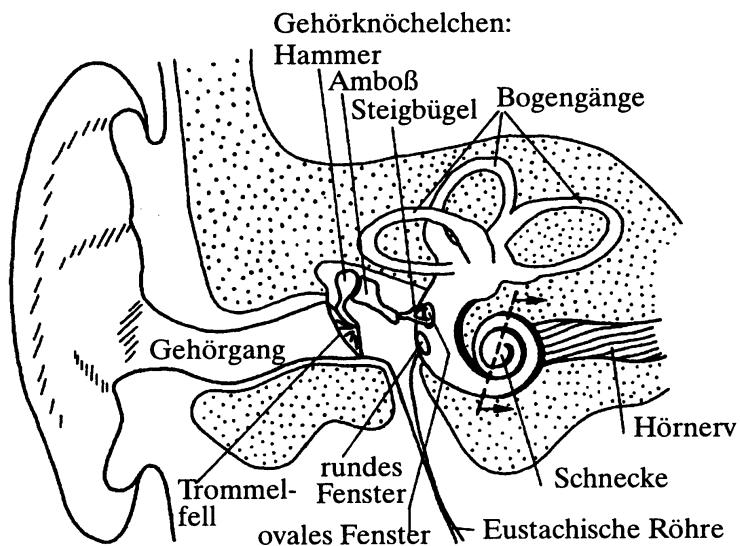
4.4 Literatur

- Blauert, J. (1994). Kommunikationsakustik II: Audiokommunikation und virtuelle Realität.
Skriptum zur Vorlesung am Institut für Kommunikationsakustik, Ruhr-Universität,
Bochum.
- Heute, U. (1990). Sprachverarbeitung. Skriptum zur Vorlesung der Arbeitsgruppe Digitale
Signalverarbeitung, Ruhr-Universität, Bochum.
- Oppenheim, A. V., Schafer, R. W. (1999). Discrete-time Signal Processing. Prentice Hall
International Editions.
- Vary, P., Heute, U., Hess, W. (1998). Digitale Sprachsignalverarbeitung. B.G. Teubner,
Stuttgart.

5. Grundlagen der auditiven Wahrnehmung

In diesem Kapitel sollen die Grundlagen der auditiven Wahrnehmung behandelt werden. Dazu gehört zunächst die Anatomie und Funktionsweise des menschlichen Gehörs, das sich in drei Teile (Außen-, Mittel- und Innenohr) gliedert. Darüber hinaus sollen zwei wichtige Aspekte der menschlichen Wahrnehmung behandelt werden, nämlich die Wahrnehmung der Tonhöhe und der Lautheit. Das Kapitel beruht zum großen Teil auf den Darstellungen von Blauert (1994) und Zwicker und Fastl (1999).

Die Anatomie des Ohres ist in untenstehender Abbildung skizziert. Man erkennt deutlich das Außenohr, welches die Ohrmuschel (Pinna) und den Gehörgang (ear canal) bis zum Trommelfell (eardrum) umfasst. Daran schließt sich das Mittelohr mit den drei Gehörknöchelchen Hammer (Malleus), Amboss (Incus) und Steigbügel (Stapes) an. Der Steigbügel drückt auf das ovale Fenster, welches die Grenze zum Innenohr bildet. Dieses besteht aus der Schnecke (Cochlea), in der sich die Nervenzellen befinden. Der Hörnerv stellt die Verbindung zum Gehirn her.



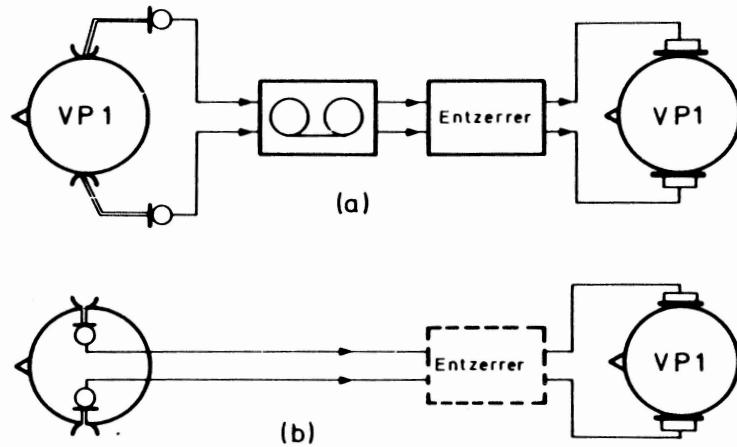
Schematische Darstellung von Außen-, Mittel- und Innenohr (nach Zwicker, 1982, 22).

5.1 Außenohr

Zum Außenohr gehören der Kopf (soweit akustisch wirksam), die Ohrmuschel sowie der Gehörgang bis zum Trommelfell. Das Außenohr ist vor allem wichtig wegen seiner *Richtwirkung*; es ermöglicht das räumliche Hören. Dies ist umso erstaunlicher, da die genaue Form der Ohrmuschel von Mensch zu Mensch stark variiert. Dazu kodiert das Außenohr die Information über die räumliche Position der Schallquelle in Frequenz- und Zeitinformationen um.

Wenn das Schallsignal am Eingang beider Gehörgänge aufgezeichnet und entzerrt (zu einem anderen Zeitpunkt, in einem anderen Raum) wiedergegeben wird, so kann beim Hörer dasselbe Hörereignis – inklusive seinem Ort, seiner Ausdehnung und seiner Eigenschaften – entstehen. Voraussetzung hierfür wäre allerdings, dass der Kopf fixiert ist, dass das visuelle System keinen Einfluss hat (z.B. durch Verdunklung), und dass die Vorgeschichte gleich ist (vgl. Kapitel 1). Man kann dieses Prinzip zur originalgetreuen Übertragung und Wiedergabe

von Audio-Signalen verwenden. Allerdings ist eine direkte Messung im individuellen Außenohr meist nicht möglich; man verwendet dann einen Kunstkopf mit einem vereinfachten Außenohrmodell.



Beispiele kopfbezogener Übertragungssysteme (nach Blauert, 1998, 50). Oben: Elektroakustische Übertragungskette mit Sondenmikrophon und entzerrter Kopfhörerwiedergabe; unten: Kopfbezogenes Übertragungssystem mit Kunstkopf.

Die Ohrmuschel besteht aus Knorpelgerüst, das straff mit Haut überzogen ist. Sie hat eine charakteristische, individuell ausgeprägte Reliefform. Der Gehörgang ist ein komplett mit Haut ausgekleideter, leicht gekrümmter Kanal mit einer mittleren Länge von ca. 25 mm und einem Durchmesser von ca. 7-8 mm. Das Trommelfell ist rund oder leicht elliptisch geformt und befindet sich in einem Winkel von ca. 40-50° zum Gehörgang. Es ist ca. 0,1 mm dick, besteht aus einer häutigen Membran, die durch Bindegewebe verstärkt ist, und hat eine effektive Fläche von ca. 50 mm². Auf der Rückseite ist das Trommelfell durch den Hammer (1. Gehörknöchelchen) belastet, außerdem schwingt es auf einem Luftpolster, welches durch die Paukenhöhle und angrenzende Höhlen gebildet wird (Luftausgleich durch die Eustachische Röhre).

Die Ohrmuschel stellt zusammen mit dem Gehörgang ein *Resonatorsystem* dar, d.h. es verstärkt einzelne Frequenzen und schwächt andere ab. Dabei hängt die Erregbarkeit der einzelnen Resonanzen des Systems von der Richtung und Entfernung der Schallquelle ab; das System kann also als Filter beschrieben werden, dessen Übertragungsfunktion von der Position der Schallquelle relativ zum Empfänger (Ohr) abhängt. Die Hauptresonanz liegt bei ca. 3 kHz ($\lambda/4$ -Resonanz des Gehörgangs), eine weitere Resonanz (die der Haupthöhle der Ohrmuschel) bei ca. 5 kHz. Das Trommelfell zeigt darüber hinaus eine Resonanz bei ca. 1 kHz. Diese Resonanzen liegen im Bereich größter Hörempfindlichkeit. Der Kopf übt ebenfalls einen Einfluss auf die Übertragung aus, da er den Schall beugt (Näherungsrechnung: Beugung an einer Kugel).

Die Übertragungsfunktion des Außenohres lässt sich messtechnisch erfassen. Hierzu misst man im Freifeld (angenähert z.B. im reflexionsarmen Raum) den Schall, den ein Im-Ohr-Mikrophon erfasst, in Abhängigkeit von der Schallquelle (Schallsignal und Position). Man bekommt somit einen Satz von Übertragungsfunktionen, eine für jede Schallquellenposition relativ zum Kopf. Man bezeichnet diese als *Außenohr-Übertragungsfunktionen* oder *head-related transfer functions, HRTFs*.

Die Außenohr-Übertragungsfunktionen werden zu beiden Ohren bestimmt. Der Mensch verwendet beide Ohrsignale, um aus den Unterschieden bzgl. des Signalpegels (interaurale Pegeldifferenzen) bzw. der Eintreffzeitpunkte am Ohr (interaurale Zeitdifferenzen) weitere Rückschlüsse auf die Position und Entfernung der Schallquelle zu ziehen. Details hierzu sind bei Blauert (1998) beschrieben. 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. Dies lässt sich z.B. bei Hörbehinderten beobachten, die nur auf einem Ohr mit einem Hörgerät versorgt sind; das Fehlen des zweiten Ohrsignals führt u.a. dazu, dass Gespräche in einer Multi-Sprecher-Umgebung nicht mehr gut verfolgt werden können (fehlender Cocktail-Party-Effekt).

5.2 Mittelohr

Das Mittelohr stellt einen Hebelapparat dar, der den Druck auf das Trommelfell auf das ovale Fenster der Schnecke überträgt. Er besteht aus den *drei miteinander verbundenen Gehörknöchelchen*, die die Kette zwischen Trommelfell und ovalem Fenster bilden. Die Gehörknöchelchen sind in der Paukenhöhle beweglich aufgehängt, und zwar so, dass durch Kopfbewegungen eine möglichst geringe Bewegung der Gehörknöchelchen angeregt wird; dazu trägt auch die geringe Masse der Knöchelchen bei.

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 (Faktor ca. 1,3), zum anderen über das Verhältnis der Flächen von Trommelfell und ovalem Fenster (Zahlenbeispiel $0,7 \text{ cm}^2 / 0,05 \text{ cm}^2 \approx 14$). Daraus gibt sich eine theoretische Gesamtverstärkung von ca. 16-18 (gemessene Werte nach Békésy: 15). Die Verstärkung ist jedoch auch frequenzabhängig, mit angenähertem Tiefpassverhalten und einer ersten Resonanz bei ca. 1,2 kHz.

Zusätzlich zur Impedanztransformation hat das Mittelohr auch noch eine (allerdings minimale) Schutzfunktion: Der Hebelapparat kann bei extremen Schalldrücken ausweichen und damit das Innenohr vor zu hohen Anregungen schützen. Dieser sog. Stapedius-Reflex wird z.B. zur Diagnostik verwendet.

5.3 Innenohr und Nervensystem

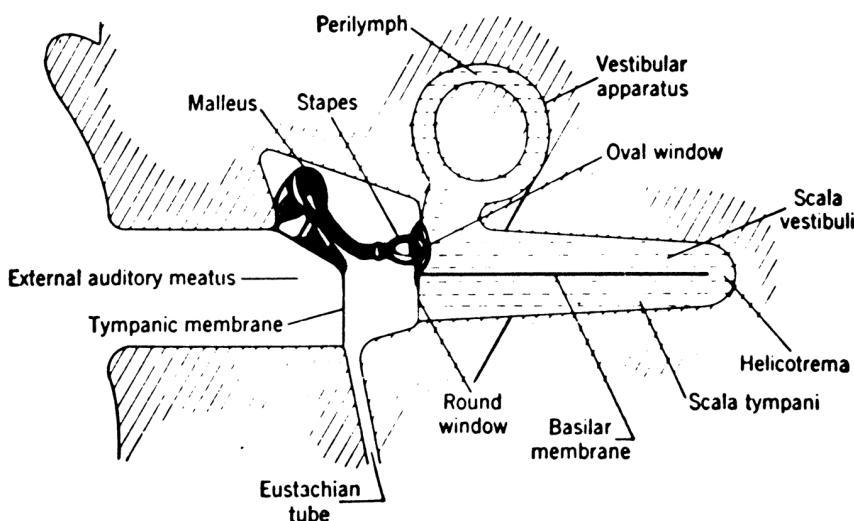
Da das menschliche Gehör in der Lage ist, aus zusammengesetzten Geräuschen Teiltöne herauszuhören, muss das Gehör eine *Spektralanalyse nach Art der Fourier-Transformation* durchführen. Dies bezeichnet man als *Ohm'sches Gesetz der Psychoakustik*. Außen- und Mittelohr scheiden als Ort für die Durchführung dieser Spektralanalyse aus, da sie beide breitbandig arbeiten (d.h. ein großer Frequenzbereich wird übertragen) und keine Verzweigungen oder Reiztransportorgane aufweisen. Die Fourier-Analyse muss also im Innenohr bzw. in den nachgeschalteten (nicht-mechanischen) Teilen des Nervensystems stattfinden.

Der Hauptbestandteil des Innenohres ist die sog. *Schnecke (Cochlea)*, d.h. ein System aus drei langen Kammern, von denen zwei am äußersten Ende miteinander verbunden sind, und die schneckenförmig aufgewickelt sind (2 ½ Windungen). Die Schnecke ist in das Felsenbein, einen sehr harten Knochen, eingebettet. In der Schnecke findet der messtechnische Teil des Hörvorganges statt; allerdings lässt sich ein großer Teil der menschlichen Hörfähigkeit nicht

allein damit erklären. Es müssen daher noch weitere – hochgradig nichtlineare – Vorgänge im Hirn eine Rolle spielen, die noch nicht vollständig erforscht sind. Daneben befinden sich über der Schnecke noch halbkreisförmige Kanäle (sog. Bogengänge) in nahezu exakter 90°-Anordnung, die für die Gleichgewichts-Empfindung wichtig sind.

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.

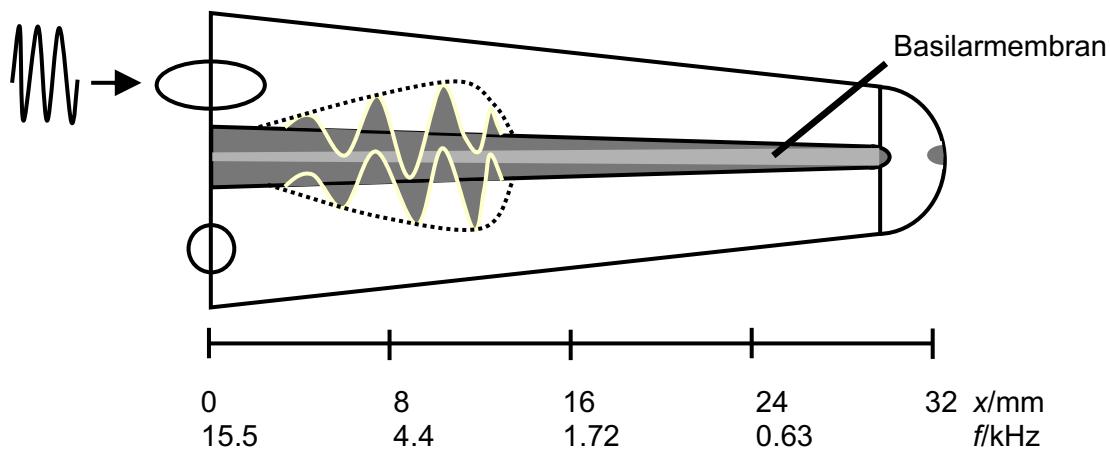
Die unten stehende Abbildung zeigt eine vereinfachte Darstellung des Ohres mit einer „abgewickelten“ Schnecke. Wie aus der Zeichnung ersichtlich besteht der Hauptkanal der Schnecke aus der Skala Vestibuli (wird durch das ovale Fenster ausgelenkt), die über das Helikotrema (kleine Öffnung am äußersten Ende der abgewickelten Schnecke) mit der Scala Tympani verbunden ist. Beide Kanäle werden durch die Basilarmembran getrennt. In abgewickelter Form ist die Schnecke etwa 35 mm lang und zeigt einen nach innen von ca. 4 mm² auf 1 mm² abnehmenden Querschnitt. Am anderen Ende der Scala Tympani befindet sich das runde Fenster, was einen Druckausgleich gegenüber dem ovalen Fenster zulässt.



Vereinfachte Darstellung des Ohres mit einer abgewickelten Schnecke
(nach Blauert, 1994).

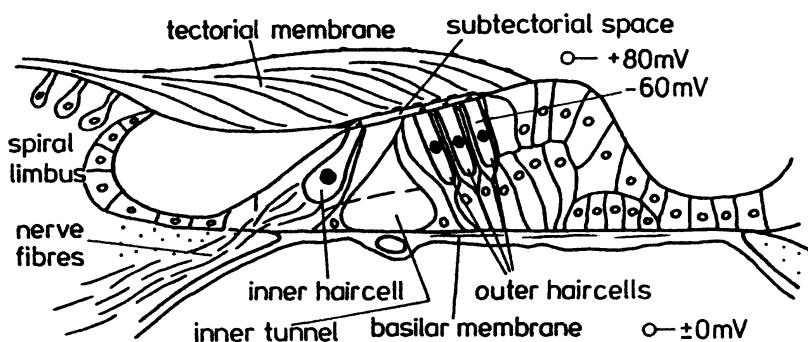
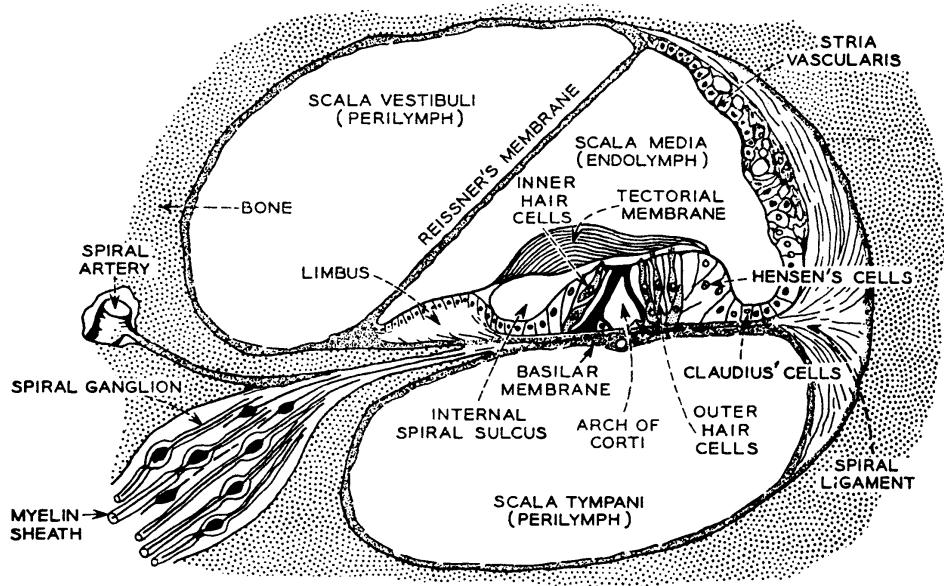
Entgegen der Verjüngung der Schnecke zum Ende hin wird die Basilarmembran zum Helikotrema hin breiter (von ca. 0,05...0,1 auf 0,5 mm), schwerer und nachgiebiger. Auf der Basilarmembran findet ein großer Teil der Frequenzauflösung des Gehörs statt. Dazu werden die Frequenzen des Anregungssignals (Druck 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. Früher ging man davon aus, dass die einzelnen Abschnitte der Basilarmembran wie schwingende Saiten in Resonanz geraten könnten, und dass dadurch die ortsspezifische Anregung stattfinden würde. Allerdings ist die Breitenvariation der Basilarmembran zu gering und die Besetzung der Basilarmembran mit Hörzellen zu hoch, als dass damit die Frequenzauflösung des Ohres erklärt werden könnte. Diese Hypothese wurde also inzwischen verworfen.

Die unten stehende Abbildung zeigt eine schematische Darstellung der abgewickelten Basilarmembran. Durch die Auslenkung des ovalen Fensters wird eine Welle in der Schnecke angeregt, die sich auf der Basilarmembran ausbreitet. Dabei handelt es sich um eine *Wanderwelle*, die ihr Maximum bei einem für jede Frequenz typischen Ort auf der Basilarmembran erreicht. Hierdurch wird die Frequenz-Orts-Transformation realisiert. Hierbei zeigt sich, dass das Maximum der Einhüllenden etwa logarithmisch über dem Ort x verteilt ist. Außerdem steigt die Phase über dem Ort an.



Frequenz-Orts-Transformation auf der Basilarmembran.

Die eigentliche Erregung der Nerven findet im sog. *Corti'schen Organ* statt, welches sich auf der Basilarmembran befindet. Das Corti'sche Organ besteht aus den eigentlichen Sinneszellen, den sog. *Haarzellen*, sowie weiteren Hilfszellen. Es gibt zwei Arten von Haarzellen: Die äußeren Haarzellen, die in drei Reihen etwa in der Mitte des Corti'schen Organs auf der Basilarmembran befestigt sind, sowie den inneren Haarzellen, die sich auf der inneren Seite des Corti'schen Organs befinden. Die Haarzellen werden von der Deckmembran (Tectorialmembran) abgedeckt, welche einen inneren Bereich (unterhalb der Scala Media) von der Basilarmembran trennt. Unten stehende Abbildung zeigt einen Querschnitt durch den Cochlea-Kanal sowie die Anordnung der Haarzellen auf der Basilarmembran.



Schematische Darstellung des Innenohres. Oben: Querschnitt durch den Cochlea-Kanal (nach Davis, 1957, aus Flanagan, 1972, 92); unten: Anordnung der Haarzellen auf der Basilarmembran (nach Zwicker und Fastl, 1999, 26).

Obwohl die genaue Funktion beider Arten von Haarzellen noch nicht vollständig bekannt ist geht man davon aus, dass die inneren Haarzellen wahrscheinlich auf eine Scherung an der Deckmembran reagieren, wenn sich die Basilarmembran bewegt. An den Haarzellen liegt ein Potentialunterschied von 140 mV an, zu dem sich bei Beschallung weitere Mikrophonpotentiale addieren. Am oberen Ende der Haarzellen befinden sich kleine Haare (die Stereozilien), deren Enden in der Tectorialmembran verankert sind. Durch die *Scherbewegung* wird der Membranwiderstand gesteuert. Dadurch fließt ein Strom, der den Transport von Transmittersubstanz zur Synapse bedingt. Durch die höhere Konzentration an Transmittersubstanz wird die Wahrscheinlichkeit erhöht, dass *Nervenimpulse* (spikes) abgegeben werden. Diese werden in den affarenten (zum Gehirn führenden) Nervenfasern weitergeleitet. Zu den äußeren Haarzellen führen efferente (vom Gehirn kommende) Nervenfasern, die auf eine Rückkoppelung schließen lassen (sog. Cochlearer Verstärker). Durch die Rückkoppelung kann die Abstimmeigenschaft des Gehörs – d.h. seine Frequenzselektivität – verbessert werden.

Nervenimpulse sind kurze elektrische Impulse fast konstanter Amplitude. Die „Feuerrate“ dieser Impulse bestimmt die Stärke der Nervenerregung. Im Corti'schen Organ findet also eine *Analog-Digital-Wandlung* statt, bei der die analog in der Schallwelle vorliegende Information über die Basilmembran-Bewegung (die ihrerseits die Frequenz kodiert) in Spikefolgen der Nervenfasern umkodiert wird. Die dabei entstehenden Potentiale können in den Nervenfasern mit Hilfe von Mikroelektroden gemessen werden.

Der Hörnerv führt von der Cochlea durch den inneren Gehörgang (inner tunnel) zum Hirnstamm, wo er in den Nucleus Cochlearis mündet. Von diesem Hirnnervenkern führen eine Reihe von Verbindungen zu anderen Kerngebieten im Hirnstamm (auch Verbindungen zwischen den beiden Ohren), sowie weiter aufsteigende Bahnen bis zum primären *auditorischen Kortex*, der Hörrinde im Großhirn. Die auditorische Kortex liegt in unmittelbarer Nähe des Sprachzentrums und der Körperfühlsphäre.

Die *Frequenz-Orts-Darstellung* (tonotopische Repräsentation) wurde auch in höheren Stadien der auditorischen Verarbeitung nachgewiesen; die Frequenz-Selektivität bleibt also in weiteren Stadien der auditorischen Verarbeitung erhalten. Darüber hinaus wird Information aus der Zeitfunktion der Spike-Folgen gezogen (Periodizität, Spike-Rate); sie geht allerdings bei höheren Frequenzen allmählich verloren. In höheren Stadien wurden auch Neuronen nachgewiesen, die auf spezielle Reizmuster reagieren, z.B. auf bestimmte interaurale Zeitdifferenzen oder auf bestimmte Spektren.

5.4 Frequenzauflösung und Tonhöhenwahrnehmung

Wie oben dargestellt findet im Innenohr also eine spektrale Zerlegung der eintreffenden Schalle statt. Dabei kann das Gehör Frequenzen von ca. 50 bis 16000 Hz wahrnehmen; der genaue Bereich ist abhängig vom Alter, vom evtl. vorhandenen Störgeräusch, von dauerhaften Schädigungen etc.

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. *Hörfächen*. Untenstehende Abbildung zeigt eine solche Hörfäche. Die unterste Linie ist die *Hörschwelle*, d.h. der Schalldruckpegel, bei dem ein entsprechender Schall gerade noch wahrgenommen wird. Die oberen Linien sind sog. *Isophonen*, d.h. Linien gleichen Lautstärkepegels (oder gleicher Pegellautstärke). Letzterer ist wie folgt definiert:

Lautstärkepegel: Pegel des als gleich laut empfundenen 1 kHz-Tones (Einheit phon). Der Bezugswert ist der kleinste hörbare Schalldruck bei 1 kHz, $p_0 = 2 \cdot 10^{-5}$ Pa.

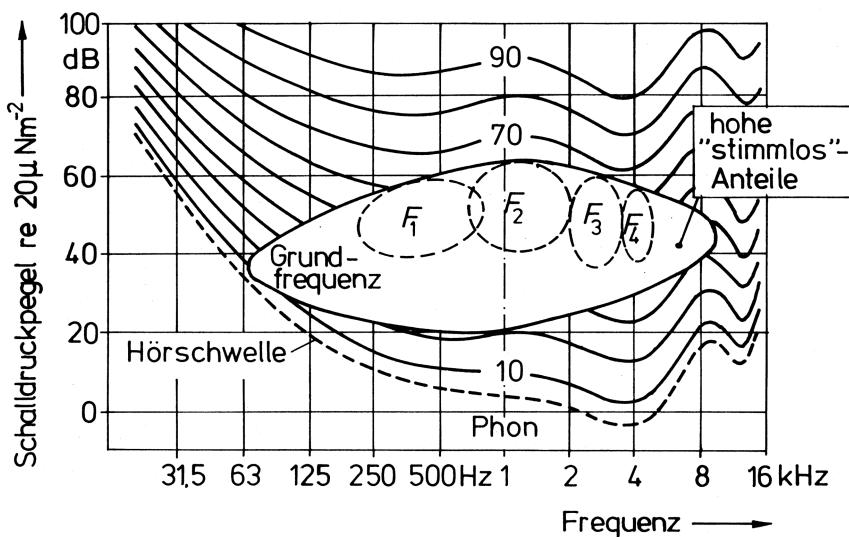
$$L_N = 20 \cdot \log_{10} \frac{p_{1\text{kHz}}}{p_0} \quad (5.1)$$

Die Isophonen werden also durch Paarvergleich mit einem (regelbaren oder vorgegebenen) 1-kHz-Ton bestimmt. Man erkennt, dass ein Ton konstanten Schalldruckpegels unterschiedlich laut wahrgenommen wird, wenn die Frequenz variiert. Nach oben hin wird die Hörfäche von der Schmerzgrenze begrenzt, die etwa 120-130 phon entspricht.

Die in der Abbildung eingezeichnete Hörschwelle entspricht einer mittleren Schwelle bei Versuchspersonen im Alter von etwa 20 bis 25 Jahre. Mit zunehmendem Alter tritt ein

Hörverlust vor allem bei hohen Frequenzen ein, z.B. bei 10 kHz: 0 dB bei 20 Jahren, -15 dB bei 40 Jahren, -25 dB bei 60 Jahren (Blauert, 1994).

In die Abbildung ist auch der typische Bereich, den Sprachsignale ausmachen, eingezeichnet. Man sieht, dass das menschliche Gehör gerade im Bereich von Sprache besonders empfindlich ist; Sprache und Gehör sind also offenbar speziell aneinander angepasst. Sprache nutzt aber nur einen Teil der Hörläche aus; andere Bereiche werden z.B. durch Umweltgeräusche verwendet. Bei der Telefonübertragung wird nur ein Teil des Sprachbereiches übertragen (etwa 300-3400 Hz, vgl. Kapitel 6). Insbesondere fallen die Grundfrequenz sowie die oberen Formanten weg. Trotzdem wird Telefonsprache i.a. als verständlich angesehen (auch wenn einzelne Laute nicht immer zu unterscheiden sind).



Hörläche des Menschen. Untere Linie: Hörschwelle; obere Linien: Isophonen; mittlerer Bereich: Bereich der Sprache. Aus Blauert (1994).

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 Tonhöhenkala)

Dabei entsprechen gleiche Intervalle auf der Skala Verdoppelungen der Frequenz (Oktavschrifte, das heißt gleiche musikalische Intervalle). Es handelt sich also um eine logarithmische Frequenzmaß-Skala:

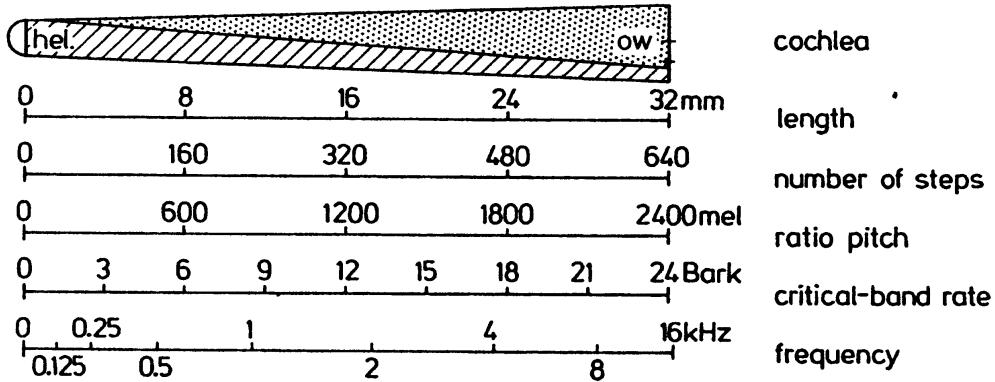
$$m \text{ [okt]} = \text{ld} \frac{f}{f_0} \quad \text{mit} \quad f_0 = 131 \text{ Hz} \quad (5.2)$$

In der Technik benutzt man meist gerundete Frequenzwerte 32,5/63/125/250/500 Hz etc.

2. *Hörereignisskala* der Tonhöhe (melodische Tonhöhenkala)

Hierbei wird die Frequenzauflösung der Basilarmembran, d.h. eine natürliche Tonhöhenkala abgebildet. Die Zuordnung ergibt sich in etwa wie in der folgenden Abbildung gezeigt. Hier ist insbesondere die *Mel-Skala* von Interesse; man bezeichnet die darauf gemessene Tonhöhe als Verhältnistonhöhe (ratio pitch) oder auch Tonheit. Ein Vergleich mit der linearen Frequenzskala zeigt, dass der Zusammenhang zwischen der Verhältnistonhöhe und der Frequenz im oberen Bereich logarithmisch ist, im

unteren jedoch nicht. Jedoch besteht ein linearer Zusammenhang zwischen Position auf der Basilarmembran x , der Anzahl gerade noch wahrnehmbarer Tonschritte (gemessen mittels Frequenz-Modulation), der Verhältnistonhöhe (gemessen in mel) und der Tonheit (critical band rate, gemessen in Bark). Letztere wird im folgenden Kapitel noch erläutert. Insbesondere gilt 1 Bark = 100 mel.



Skalen der Tonhöhenwahrnehmung in Bezug zur Basilarmembranausdehnung
(aus Zwicker und Fastl, 1999, 162).

Die Skala „number of steps“ in vorstehender Abbildung zeigt, dass das menschliche Gehör ungefähr 600 Tonhöhen unterscheiden kann. Dies geschieht, wie unten noch gezeigt wird, innerhalb nur weniger Gruppen (sog. Frequenzgruppen) auf der Basilarmembran. Das zeigt, dass neben der Frequenz-Orts-Umsetzung auf der Basilarmembran eine nichtlineare Weiterverarbeitung stattfinden muss, um so viele Tonhöhen unterscheiden zu können.

Die bisherigen Betrachtungen galten für reine (Sinus-) Töne. Bei zusammengesetzten Klängen kann das menschliche Ohr die Tonheit bestimmen

- aus der 1. harmonischen Schwingung
- falls diese fehlt – aber die 3. bis 5. Harmonische vorhanden sind – durch Rückschluss von Obertönen auf den Grundton (sog. Residuum-Pitch)
- durch die Auswertung der zeitlichen Wiederholung (sog. Repetition-Pitch).

5.5 Lautheitswahrnehmung

Die oben angegebene Definition des Lautstärkepegels weist den Nachteil auf, dass sie nur aus Punkten *gleicher* Wahrnehmung besteht; sie sagt also nichts darüber aus, wie stark sich zwei Töne unterschiedlichen Lautstärkepegels in ihrer wahrgenommenen Lautheit unterscheiden.

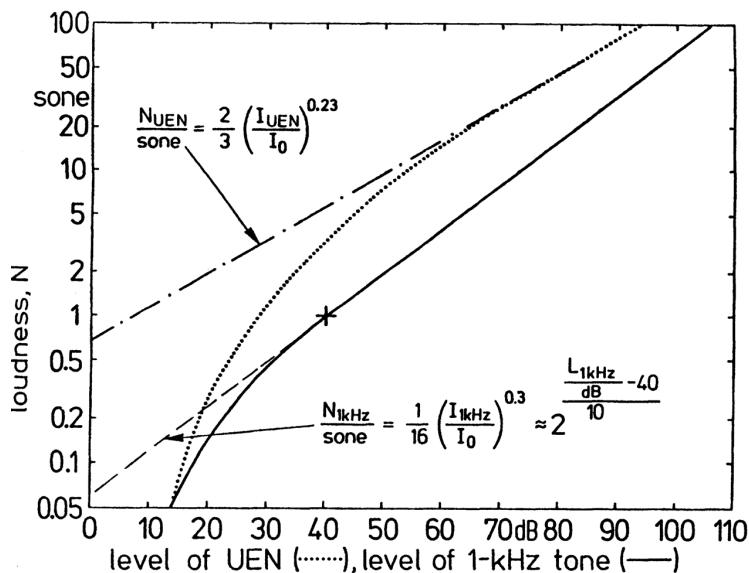
Befragt man Versuchspersonen, wie stark sich zwei Töne unterschiedlichen Schalldruckpegels in ihrer Lautheit unterscheiden (bspw. halb oder doppelt so laut), so kommt man auf die *Lautheit N* mit der Einheit sone. Diese hängt mit dem Schalldruckpegel p bzw. mit der Schallintensität I (Einheit W/m^2) über ein exponentielles Gesetz zusammen:

$$N[\text{sone}] = \text{const.} \cdot \left(\frac{I}{I_0} \right)^k = \text{const.} \cdot \left(\frac{p}{p_0} \right)^{2k} \quad (5.3)$$

Wie das phon ist auch die Einheit sone eine Pseudo-Einheit, da die Quotientenbildung mathematisch die Einheit 1 ergibt. Die Bezugspegel sind hier zu $I_0 = 10^{-12} \text{ W/m}^2$ bzw. $p_0 =$

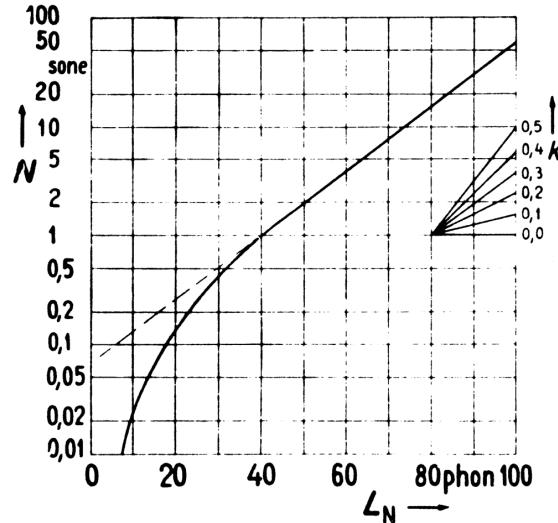
$2 \cdot 10^{-5}$ Pa gewählt. Ein ähnliches Gesetz gilt für fast alle Skalen der Stärke des Wahrgenommenen. Früher vermutete man hier einen logarithmischen Zusammenhang (vgl. Weber-Fechnersches Gesetz).

Der Zusammenhang zwischen dem Pegel eines 1 kHz-Tones und der Lautheit ist in untenstehender Abbildung skizziert. Man erkennt, dass das Exponentialgesetz erst ab einer bestimmten Mindest-Lautheit von 1 sone gilt. Darüber ergibt sich ein Exponent von ca. 0,3. Darunter fällt die Lautheit stärker ab, als es durch den exponentiellen Zusammenhang beschrieben ist. In gleicher Abbildung erkennt man, dass der Exponent offenbar von der Art des Signals abhängt; für ein gleichmäßig anregendes Rauschen ergibt sich ein Exponent von nur etwa 0,23.



Lautheit eines 1-kHz-Tones und eines gleichmäßig anregenden Rauschens (UEN) als Funktion des Schalldruckpegels (aus Zwicker und Fastl, 1999, 207).

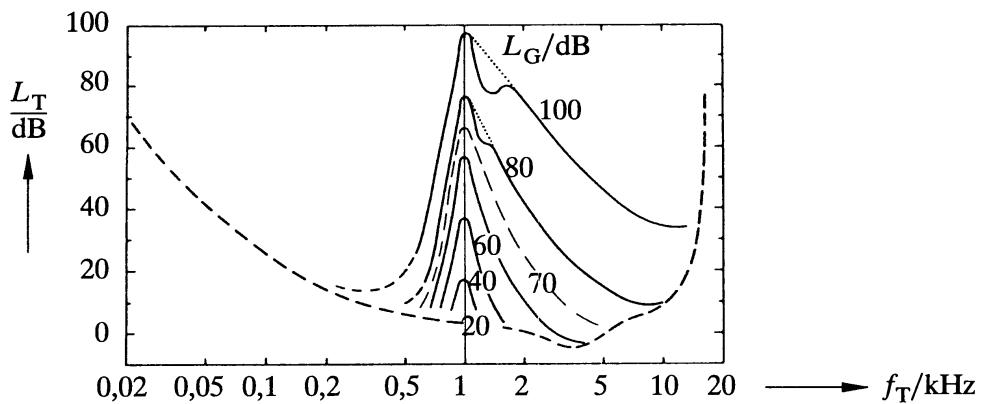
Da der *Lautstärkepegel* in phon dem Schalldruckpegel eines 1-kHz-Tones entspricht, können die Isophonen (Konturen gleichen Lautstärkepegels) auch in der Einheit sone, d.h. als *Lautheit* skaliert werden. Oberhalb von 40 phon ergibt sich dabei ein exponentieller Zusammenhang; logarithmisch ausgedrückt heißt dies, dass ein Lautstärkepegelanstieg von 10 phon einer Verdoppelung der Lautheit in sone entspricht. Dieser Zusammenhang ist in folgender Abbildung dargestellt.



Zusammenhang zwischen Lautstärkepegel L_N und Lautheit N (aus Blauert, 1994).

Bislang haben wir uns mit der Wahrnehmung von Schallen in Ruhe beschäftigt. Ein Schall kann aber durch einen anderen Schall (Maskierer) verdeckt und dadurch unhörbar werden. Man bezeichnet diesen Effekt als *Maskierung* und den Pegel des verdeckten Schalls, bei dem er in Anwesenheit des verdeckenden Schalls gerade wieder hörbar wird, als *Mithörschwelle*.

Zur Veranschaulichung der *spektralen Verdeckung* betrachten wir folgendes Experiment: Durch ein Bandpassfilter wird schmalbandiges Rauschen unterschiedlichen Pegels erzeugt. Um die Mittenfrequenz des Schmalbandrauschen wird nun ein Testton variabler Frequenz und variablen Pegels angeboten, der „abtastet“, ob zusätzlich zum Schmalbandrauschen etwas gehört werden kann. Die dabei entstehenden Kurven der Mithörschwelle sind in folgender Abbildung gezeichnet.

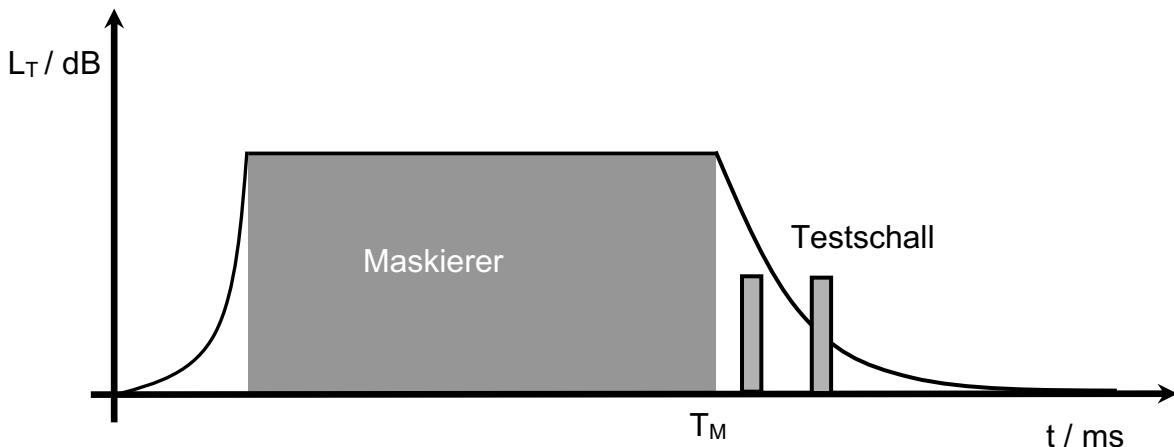


Maskierung im Frequenzbereich. Mithörschwellen bei Schmalbandrauschen der Mittenfrequenz 1 kHz und Pegel L_G als Maskierer und Sinuston der Frequenz f_T und Pegel L_T als Testton (aus Zwicker, 1982, 41, nach Vary et al., 1998, 37).

Man erkennt zum einen die Frequenz-Selektivität des Gehörs. Zum anderen ist die Hörschwelle bzgl. des Testtons gegenüber der Ruhehörschwelle auch bei Frequenzen angehoben, bei denen der Maskierer überhaupt keine spektralen Anteile aufweist. Die Maskierung ist unsymmetrisch: An der unteren Flanke sinkt die Mithörschwelle stark ab, an

der oberen Flanke ist der Abfall flacher. Die Maskierung ist darüber hinaus auch pegelabhängig: Für lautere Maskierer wird die obere Flanke stärker ausgeprägt. Die Doppelgipfel bei hohen Pegeln des Maskierers entstehen durch Nichtlinearitäten im Gehör.

Neben dieser spektralen gibt es auch eine *zeitliche Verdeckung*: Ein Testton wird unhörbar, wenn er sich in unmittelbarer zeitlicher Nähe zu einem Maskierer befindet. Dabei unterscheidet man zwischen einer kurzen Periode der Vorverdeckung und der Nachverdeckung. Eine Erklärung hierfür ist, dass die Hörempfindung nicht sofort mit Einsetzen des Reizes einsetzt, sondern eine bestimmte Zeit braucht, um sich aufzubauen. Dadurch kann die Hörempfindung des früheren Testtons durch die nachfolgende – und größere – des Maskierers verdeckt werden. Umgekehrt klingt die Hörempfindung nach dem Abschalten des Maskierers nicht sofort ab, sodass nachfolgende Testtöne überdeckt werden. Die hierbei auftretenden Zeiten liegen in der Größenordnung von 200 ms. Die folgende Abbildung veranschaulicht das Phänomen der Nachverdeckung.



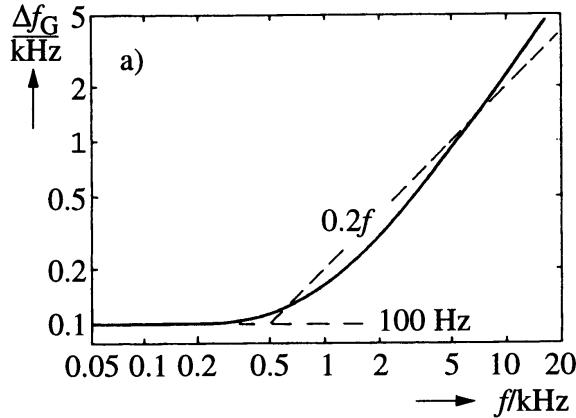
Vor- und Nachverdeckung bei breitbandigem Rauschen als Maskierer und kurzen Tonimpulsen als Testschalle.

Bei aus verschiedenen Frequenzkomponenten *zusammengesetzten Schallen* stellt sich die Frage, wie das Gehör die unterschiedlichen Komponenten zu einer Gesamt-Lautheit zusammensetzt. Hierzu sei zunächst folgendes Experiment betrachtet.

Es soll die Lautheit eines Hörereignisses bestimmt werden, wenn als Schall eine Reihe nahe benachbarter Sinustöne dargeboten wird. Zunächst wird ein Ton bei 1 kHz so eingestellt, dass man ihn gerade hört (Pegel 0 dB). Nun werden anstelle des einen zwei Töne gleichen Pegels (z.B. bei 1000 und 990 Hz) so eingestellt, dass man sie ebenfalls gerade noch hört. Man erhält dabei für die Pegel der beiden Teiltöne -3 dB. Dieses Verfahren wird mit mehr Teiltönen fortgesetzt; man erhält zunächst eine Absenkung um jeweils 3 dB bei Verdoppelung der Anzahl der Teiltöne. Dies funktioniert jedoch nur bis zu einer bestimmten „kritischen“ Bandbreite, die die Teiltöne überstreichen; wenn diese kritische Bandbreite überschritten wird nimmt der benötigte Pegel der Teiltöne nicht weiter ab.

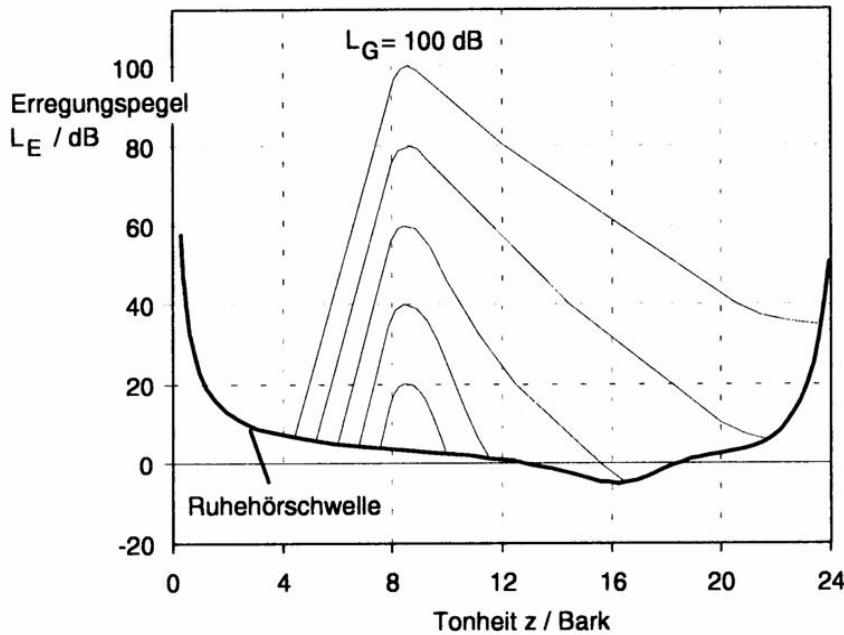
Offenbar scheint das Gehör bei der Lautheitsbildung über bestimmte Bereiche des Frequenzkontinuums zu integrieren. Man bezeichnet diese Bereiche als *Frequenzgruppen* (*critical bands*); dieser Begriff wurde von Zwicker eingeführt, und eine genauere Betrachtung findet sich bei Zwicker und Fastl (1999). Reicht 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. Der Zusammenhang

zwischen Frequenz f (in Hz) und Breite der Frequenzgruppe Δf_G ist in nachfolgender Abbildung dargestellt. Für niedrige Frequenzen ist die Breite der Frequenzgruppen zunächst konstant (100 Hz) und steigt ab etwa 500 Hz proportional (20% der Bandmittenfrequenz) an. Unabhängig davon entspricht eine Frequenzgruppe aber einem konstanten Längsbereich auf der Basilarmembran, vgl. Kapitel 5.4.



Breite der Frequenzgruppen über der Frequenz
(nach Zwicker, 1982, 51, aus Vary et al., 1998, 34).

Zur Lautheitsbildung zusammengesetzter Schalle misst das Gehör also offenbar die Leistung der spektralen Komponenten mit Filtern in Frequenzgruppenbreite; die in diesen Frequenzgruppen anfallenden Leistungen werden addiert, wobei allerdings die spektrale Maskierung berücksichtigt wird. Dies kann man mathematisch durch Einführung der sog. Erregung E beschrieben, die im folgenden Bild (als Pegel $L_E = 10 \cdot \log_{10}(E/E_0)$) dargestellt ist.



Zusammenhang zwischen Erregungspegel und Tonheit z für ein Schmalbandrauschen von 1 Bark Breite und 1 kHz Mittenfrequenz.

Der Erregungspegel ist wiederum mit der Lautheit N über die spezifische Lautheit N' verknüpft:

$$N' \sim \left(\frac{E}{E_0} \right)^{0,23} \quad (5.4)$$

und

$$N = \int_0^{24 \text{ Bark}} N' dz \quad (5.5)$$

Somit lässt sich die Gesamtlautheit durch Integration der spezifischen Lautheiten errechnen.

Das beschriebene Verfahren kann auch zur Messung der Lautheit von (technischen) Geräuschen eingesetzt werden. Wegen des damit verbundenen Aufwands misst man allerdings bislang meist auf einfachere Weise mittels sog. *bewerteter Schalldruckpegel*. Hierbei werden Bewertungskurven definiert, mit denen das Spektrum eines Schalls direkt (im Spektralbereich) gewichtet wird und daraus ein sog. bewerteter Pegel berechnet wird. Üblich sind Bewertungskurven, die invers zu bestimmten Isophonen sind und somit die Empfindlichkeit des menschlichen Gehörs in einem bestimmten Pegelbereich berücksichtigen. Am häufigsten wird die sog. A-Bewertung verwendet, die im Prinzip einer vereinfachten und umgedrehten Isophone für den Bereich 30-60 phon entspricht; der damit bewertete Pegel wird dann mit dB(A) abgekürzt. Diese Bewertungsverfahren sind z.B. in Schallpegelmessern bereits integriert.

Bei unseren bisherigen Betrachtungen haben wir uns auf das Betragsspektrum von Sprache beschränkt. Dies deshalb, da das menschliche Ohr bei der Sprachwahrnehmung gegenüber Phasenverzerrungen relativ unempfindlich ist. Allerdings kann die Phase wichtig werden, wenn mehrere Spektralanteile in eine Frequenzgruppe fallen (bei Sprache z.B. bei Zischlauten). Hier beeinflusst die Phase die Klangfarbe des Gehörten.

5.6 Literatur

- Blauert, J. (1994). Kommunikationsakustik II: Audiokommunikation und virtuelle Realität. Skriptum zur Vorlesung am Institut für Kommunikationsakustik, Ruhr-Universität, Bochum.
- Flanagan, J.L. (1972). Speech Analysis, Synthesis and Perception. Springer Verlag, Berlin.
- Heute, U. (1990). Sprachverarbeitung. Skriptum zur Vorlesung der Arbeitsgruppe Digitale Signalverarbeitung, Ruhr-Universität, Bochum.
- Vary, P., Heute, U., Hess, W. (1998). Digitale Sprachsignalverarbeitung. B.G. Teubner, Stuttgart.
- Zwicker, E. (1982). Psychoakustik. Springer, Berlin.
- Zwicker, E., Fastl, H. (1999). Psychoacoustics: Facts and Models. Springer, Berlin.

6. Sprachsignalübertragung und -kodierung

Im folgenden Kapitel soll die Übertragung von Sprachsignalen behandelt werden. Diese Übertragung geht meist nicht ohne eine Veränderung des ursprünglichen Signals vonstatten, welche zu einer Beeinträchtigung der Qualität führen kann. Eine möglichst hohe Qualität ist aber nur *eine* Anforderung an die Übertragung: Daneben sollte die Übertragung auch bei möglichst niedriger Bitrate erfolgen, um die Kanalkapazität zu maximieren, die verwendeten Algorithmen sollten eine möglichst geringe Komplexität aufweisen (um die erforderliche Rechenleistung zu minimieren), und das Signal sollte möglichst unverzögert beim Empfänger ankommen, um die Konversation nicht zu beeinträchtigen.

Da das Gebiet der Übertragung von Sprach- und Audiosignalen äußerst umfangreich ist, soll hier nur ein – allerdings grundlegend wichtiger – Teilaспект behandelt werden, nämlich die *Kodierung* von Sprachsignalen, und etwas allgemeiner von Audio-Signalen. Hierbei beschränken wir uns wiederum auf die Kodierung der Signalform, d.h. auf die sog. *Quellenkodierung*. Die Quellenkodierung ist vor allem deshalb interessant, da dort Prinzipien der menschlichen Spracherzeugung, der Sprachsignalanalyse und der Sprachwahrnehmung verwendet werden, die schon in Kapitel 2-5 beschrieben wurden. Neben der Quellenkodierung findet eine *Kanalkodierung* statt, die den zu übertragenden Datenstrom an den Übertragungskanal anpasst und ihn z.B. robust gegenüber Übertragungsfehlern macht.

Wie Kapitel 1 zu entnehmen ist, erfährt das Sprachsignal noch eine Reihe weiterer Beeinträchtigungen; bspw. können Leitungsrauschen (auf analogen Leitungen) oder Hintergrundgeräusche (an der akustischen Schnittstelle) eingekoppelt werden, oder es können Laufzeiten und Echos auftreten. Die über die Kodierung hinausgehenden Elemente der Übertragungsstrecke werden hier nicht weiter behandelt; Details für den Bereich der leitungsvermittelten oder paketvermittelten Telefonie finden sich z.B. bei Möller (2000) oder Raake (2006).

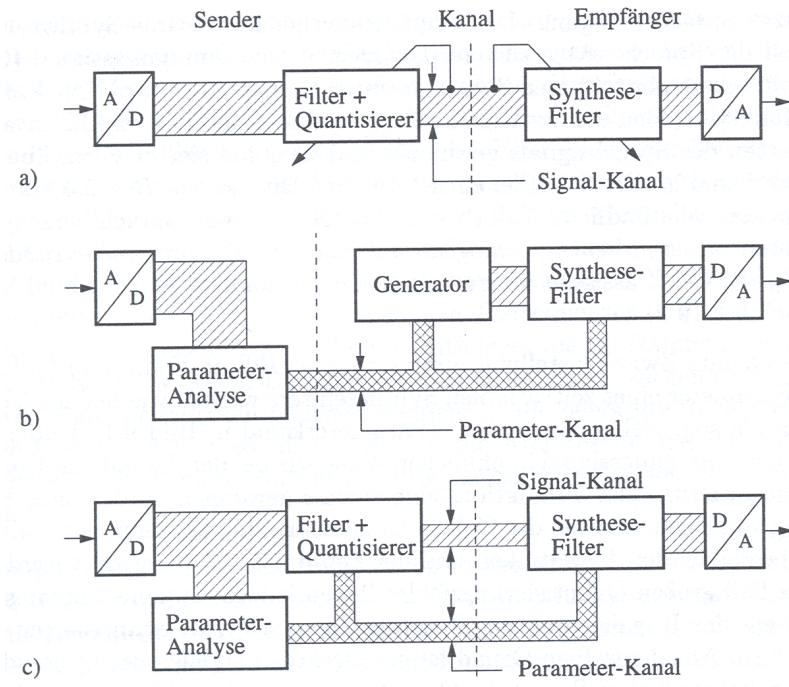
Die folgenden Betrachtungen sind teilweise aus Heute (1990) und Vary et al. (1998) entnommen.

6.1 Klassen von Sprach- und Audiosignalkodierern

Zur Übertragung von Sprach- und Audiosignalen könnte man natürlich zunächst auf die Idee kommen, die kontinuierlichen oder die zeitdiskreten Signale unverändert zu übertragen. Dies ist in der Tat möglich; um jedoch die erforderliche Störsicherheit zu erzielen und die Störungen im Verhältnis zum Signal möglichst gering zu halten, gleichzeitig aber die Bitrate gering zu halten ist es vorteilhafter, die Signale zunächst „geschickt“ zu komprimieren. Dabei muss „geschickt“ nicht unbedingt heißen, dass das Signal unverändert übertragen wird: Da sich Sprach- und Audiosignale letztendlich meist an den wahrnehmenden Menschen richten ist es vielmehr wichtig, dass die entstehenden Störungen nicht oder kaum wahrnehmbar sind.

Zur effektiven Kodierung sollten daher – neben dem Wissen über die menschliche Spracherzeugung und die daraus resultierenden Sprachsignaleigenschaften – auch Wissen über die menschliche Wahrnehmung wie bspw. Verdeckungseffekte berücksichtigt werden. Audiokodierer nutzen vor allem die Wahrnehmungseffekte aus, da Wissen über das zu kodierende Quellsignal fehlt. Bei Sprache lässt sich – durch Einbeziehen dieses Wissens – die Bitrate stark verringern, und somit die Effizienz der Kodierung steigern.

Trotz der Vielzahl an Algorithmen zur Sprachkodierung lassen sich die verwendeten Prinzipien grob in nur drei Klassen einteilen. Diese Prinzipien sind in der folgenden Abbildung dargestellt.



Prinzipien der Sprachsignalkodierung, aus Vary et al. (1998, 235).

a) Signalformkodierung; b) Parametrische Kodierung; c) Hybrid-Kodierung.

Bei der *Signalformkodierung* wird versucht, die Informationen bereits sendeseitig zu reduzieren. Dies kann z.B. durch eine geschickte Quantisierung (wie im folgenden Kapitel 6.2 gezeigt) geschehen. Besser versucht man jedoch, das Signal prädiktiv zu filtern oder nach einer Kurzzeit-Spektralanalyse zu normieren, sodass sich die Dynamik des Signals schon quellseitig reduziert und entsprechend weniger Informationen übertragen werden müssen. Diese Filterung und Normierung wird meist adaptiv ausgeführt, d.h. sie passt sich an die sich ändernden Signaleigenschaften an.

Auf der Empfängerseite wird das Signal aus den reduzierten Informationen wieder bestmöglich rekonstruiert. Dies kann z.B. durch eine inverse Filterung oder Normierung geschehen, wobei die entsprechenden Filter- oder Normierungsparameter aus dem übertragenen Signalstrom rückwärtsadaptiv gewonnen werden können; diese Informationen brauchen also meist nicht separat zum Empfänger (über den sog. Parameter-Kanal) übertragen werden. Entsprechende Verfahren werden in Kapitel 6.3 vorgestellt.

Im Gegensatz dazu wird bei der *parametrischen Kodierung* nicht die Sprachsignalform kodiert übertragen, sondern es werden Parameter übertragen, aus denen sich das Sprachsignal – oder ein ähnlich klingendes Signal – auf der Empfangsseite künstlich synthetisieren lässt. Hierbei kann z.B. das in Kapitel 3 ausgeführte Quelle-Filter-Modell der Sprachsignalerzeugung verwendet werden. Im Gegensatz zur Signalformkodierung bleibt der Signal-Kanal leer, stattdessen wird nur der Parameter-Kanal verwendet. Hiermit lassen sich sehr niedrige Bitraten – bei allerdings stark eingeschränkter Qualität – erzielen. Beispiele hierfür sind in Kapitel 6.4 angegeben.

Als Kompromiss zwischen diesen beiden Verfahren bietet sich schließlich eine Mischform – die sog. *Hybrid-Kodierung* an. Diese versucht zunächst, die Signalinformationen sendeseitig (z.B. unter Ausnutzung von prädiktiven Filtern) zu reduzieren, überträgt dann aber doch eine grobe Form des verbleibenden sog. Rest-Signals über den Signal-Kanal. Mit Hilfe dieses Prinzips, das heutzutage überaus häufig anzutreffen ist, lässt sich die Bitrate bei recht hoher Sprachqualität so weit drücken, wie es für Anwendungen im Mobilfunk oder über Internetkanäle notwendig ist. Wichtige Prinzipien und Beispiele hybrider Kodierer werden in Kapitel 6.5 behandelt.

Neben dieser Dreiteilung findet man in der Literatur häufig noch die Unterscheidung zwischen Kodierverfahren im *Zeit- und im Frequenzbereich*. Bei der Sprachsignalübertragung wendete man bislang hauptsächlich Zeitbereichsverfahren an. Durch den Übergang zu breitbandiger (50-7000 Hz und mehr) Übertragungstechnik werden jedoch auch Frequenzbereichsverfahren interessant, und teilweise auch mit Zeitbereichsverfahren kombiniert. Details hierzu finden sich in Kapitel 6.6.

Die Vielzahl der Kodierprinzipien lässt die Auswahl eines passenden Kodierers für einen bestimmten Anwendungsfall schwierig erscheinen. Jedoch gibt es Kriterien, die – je nach Anwendungsfall gewichtet – helfen, den jeweils optimalen Kodierer zu finden. Solche Kriterien sind in Kapitel 6.7 umrissen.

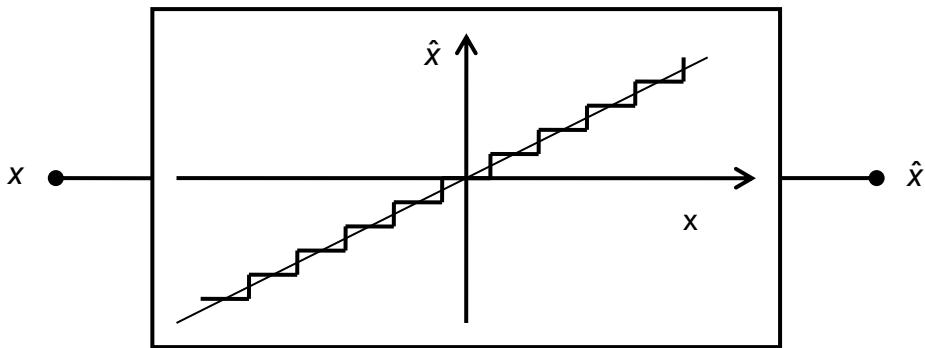
6.2 Quantisierung

Bislang haben wir uns ausschließlich mit analogen und mit zeitdiskreten Signalen beschäftigt. Ein wichtiger Grund für den Übergang zu diskreter Darstellung war aber die davon erhoffte höhere Störsicherheit. Nach der Abtastung liegt das ursprüngliche Signal zu diskreten Zeitpunkten $k \cdot T_A$ vor. Dadurch ist allerdings noch nichts gewonnen, da die Amplitude weiterhin kontinuierliche Werte annehmen kann und damit nicht gegen Störungen geschützt ist. Man geht deshalb auf eine diskrete Darstellung der Amplituden (zu den bereits diskreten Zeitpunkten) über, d.h. man *quantisiert die Amplituden* nach einer noch zu definierenden Vorschrift. Wie bereits ausgeführt entsteht dadurch ein *Quantisierungsfehler*, der minimal gehalten werden sollte.

6.2.1 Lineare Quantisierung, Pulse-Code-Modulation (PCM)

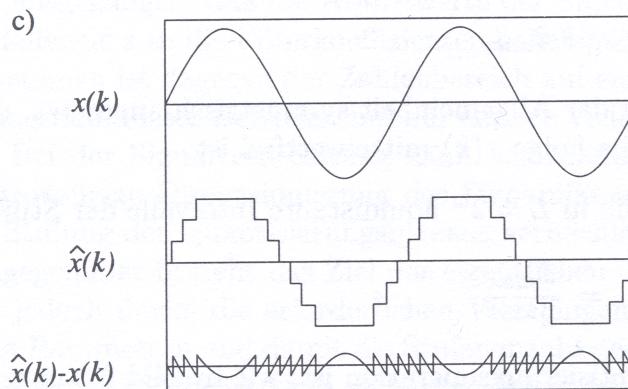
Wir gehen davon aus, dass wir Amplitudenwerte $x(k)$ in einem Aussteuerungsbereich $x_{min} \leq x \leq x_{max}$ durch (quantisierte) Zahlenwerte $-\hat{x}_{\max} \leq \hat{x} \leq \hat{x}_{\max}$ darstellen möchten; d.h. wir gehen davon aus, dass $x(k)$ mittelwertfrei ist.

In der einfachsten Form werden die Amplitudenwerte in äquidistanten Intervallen zusammengefasst. Diese Intervalle lassen sich dann z.B. binär darstellen. Bei Datenworten der Wortlänge w führt dies auf einen Wertevorrat von 2^w Möglichkeiten. Man bezeichnet dieses Verfahren mit *linearer Quantisierung* auch als *Pulse-Code-Modulation, PCM*. Die folgende Abbildung veranschaulicht die dabei verwendete Quantisierungskennlinie.



Kennlinie der linearen Quantisierung.

Durch die Quantisierung entsteht ein Fehlersignal $e(k) = \hat{x}(k) - x(k)$. Für ein einfaches sinusförmiges Signal sind diese Signale in folgender Abbildung dargestellt.



Beispiel eines sinusförmigen Signals bei linearer Quantisierung (Vary et al., 1998, 240).

Will man bei linearer Quantisierung n digitale Signale der Bandbreite f_g übertragen, so ergibt sich die benötigte Bandbreite des Kanals zu

$$f_K \geq w \cdot n \cdot f_g \quad (6.1)$$

Die Bandbreite ist also w mal größer als diejenige, die zur Übertragung des analogen Signals benötigt würde. Der Vorteil ist allerdings die größere Störsicherheit durch die digitale Darstellung: Bei der Übertragung binärer Signale – z.B. 1 und 0 – lassen sich Störungen bis zu $1/2$ durch die Wahl einer Schwelle komplett eliminieren.

Durch die Quantisierung entstehen Quantisierungsstufen Q , die sich aus dem maximalen Aussteuerungsbereich D (den Bereich, den die analogen Signalamplituden abdecken dürfen) und der Anzahl der Stufen 2^w , um diesen Bereich darzustellen, ergeben:

$$Q = \frac{\text{Aussteuerungsbereich}}{\text{Stufenzahl}} = \frac{D}{2^w} = D \cdot 2^{-w} \quad (6.2)$$

Der maximale Fehler, den man bei der Quantisierung eines Amplitudenwertes macht, ist $Q/2$. Der tatsächliche Fehler lässt sich nur berechnen, wenn man das Signal $x(k)$ in einer

geschlossenen Form kennt; das ist bei Sprachsignalen nicht der Fall, weshalb man eine statistische Betrachtung anstellen muss. Deshalb kann man nur annehmen, dass man die statistischen Eigenschaften (Verteilungsfunktion, VDF, etc.) des Signales $x(k)$ kennt.

Aus diesen Eigenschaften lassen sich dann die statistischen Eigenschaften des Fehlersignales $e(k)$ berechnen. Details hierzu sind z.B. bei Vary et al. (1998, 240-246) beschrieben. Hier interessiert insbesondere die Leistung des Fehlersignals $e(k)$, die sich (statistisch gesehen) als Erwartungswert berechnen lässt. Geht man z.B. von einer Gleichverteilung des Signales $x(k)$ aus, so ergibt sich die Leistung des Fehlersignals, das sog. *Quantisierungsrauschen*, zu

$$N = E\{e^2(k)\} = \frac{Q^2}{12} = \frac{D^2 \cdot 2^{-2w}}{12} \quad (6.3)$$

Dieser Wert gilt genau genommen nur für den Fall der Gleichverteilung von $x(k)$, die (wie bereits gezeigt) nicht vorliegt. Er gilt jedoch näherungsweise auch für den Fall, dass das Signal (und somit auch der Quantisierungsfehler) innerhalb der Quantisierungsintervalle jeweils gleichverteilt ist; diese Näherung ist bei genügend feiner Quantisierung bzw. genügend groß gewählter Wortlänge normalerweise recht gut erfüllt.

Als einen einfachen Indikator für die Störung, die man sich durch die lineare Quantisierung einhandelt, verwendet man üblicherweise das Verhältnis von Signalleistung S und Rauschleistung N , den sog. *Störabstand* oder *Signal-to-Noise Ratio*, abgekürzt *SNR*:

$$SNR [dB] = 10 \cdot \log_{10} \frac{S}{N} \quad (6.4)$$

Durch Einsetzen von N erhält man

$$SNR [dB] = 10 \cdot \log_{10} \frac{12 \cdot S}{D^2 \cdot 2^{-2w}} = 4,77 + 10 \cdot \log_{10} \frac{4 \cdot S}{D^2} + 6 \cdot w \quad (6.5)$$

Die genauen Werte sind hier nicht von Belang; es ist aber wichtig festzuhalten, dass

- der Störabstand mit 6 dB pro 1 bit Wortlänge steigt bzw. sinkt
- der Störabstand abhängig vom Aussteuerungsbereich D ist

Letztere Eigenschaft ist nicht optimal und kann durch eine nichtlineare Quantisierung vermieden werden.

Es sei angemerkt, dass wir davon ausgegangen sind, dass der Quantisierer nicht übersteuert wird, d.h. dass sich alle Eingangswerte im Bereich $x_{min} \leq x \leq x_{max}$ bewegen. Dies ist bei Sprache nicht immer erfüllt, wie die nicht verschwindende Amplitudenverteilung für größere Amplituden zeigt (vgl. Kapitel 2.10). Daher wird ein mit realen Sprachsignalen konfrontierter Quantisierer immer zu einem geringen Prozentsatz übersteuert werden. Bei Übersteuerung fällt der Störabstand dann mit steigender Signalleistung rapide ab. Man muss daher durch entsprechende Dimensionierung des Aussteuerungsbereiches darauf achten, dass dieser Fall nur sehr selten (z.B. im Promille-Bereich) auftritt.

Zahlenbeispiel:

Sprache enthält oberhalb von 4 kHz nur wenige Informationen, die für die Verständlichkeit von Bedeutung sind. Deshalb beschränkt man sich bei der normalen Telefonübertragung (ISDN, GSM, VoIP) normalerweise auf den Bereich 0-4 kHz, wobei eine weitere (technische) Beschränkung auf den Übertragungsbereich 300-3400 Hz vorgenommen wird. Nach dem

Abtasttheorem muss die Abtastfrequenz zur Übertragung eines 4 kHz breiten Signals mindestens 8 kHz betragen.

Verlangt man nun für einen Sinuston halber Vollaussteuerung ($\hat{x} = \hat{x}_{\max}/2$) einen Störabstand von 60 dB (das entspricht einem Verhältnis S/N von 10^6), so ergibt sich die benötigte Mindest-Wortlänge der linearen PCM zu etwa 11 bit. Zusammen mit der Abtastfrequenz ergibt das eine Bitfolgefrequenz (Bitrate) von $f_B = 88$ kbit/s. Allerdings ist hierbei der Störabstand von der Signalaussteuerung abhängig; bei geringerer Aussteuerung sinkt auch der Störabstand proportional zum Signalpegel. Bspw. ergibt sich bei einem Effektivwert des Signals von 1/10 des mittleren Effektivwertes (was bei einem natürlichen Sprecher durchaus real ist) nur noch ein SNR von ca. 40 dB.

6.2.2 Nichtlineare Quantisierung

Bislang wurde ein Quantisierer verwendet, dessen Intervalle den Aussteuerungsbereich D gleichmäßig abdecken. Die Quantisierungs-Kennlinie, mit der die Quantisierungsstufen-Nummer über der Eingangsamplitude dargestellt wird, ist also linear. Wie im vorangegangenen Abschnitt gezeigt wurde ist der damit erzielbare Störabstand aber von der Signalaussteuerung abhängig. Dies ist ungünstig, da sowohl laute wie auch leise Signalabschnitte gleich gut übertragen werden sollten.

Wir gehen aus von der Forderung

$$\frac{S}{N} = \frac{12 \cdot S}{Q^2} = \text{konstant} \quad (6.6)$$

Dies führt darauf, dass die Quantisierungsstufe Q proportional zur Signalleistung \sqrt{S} oder proportional zur Signalamplitude x sein muss.

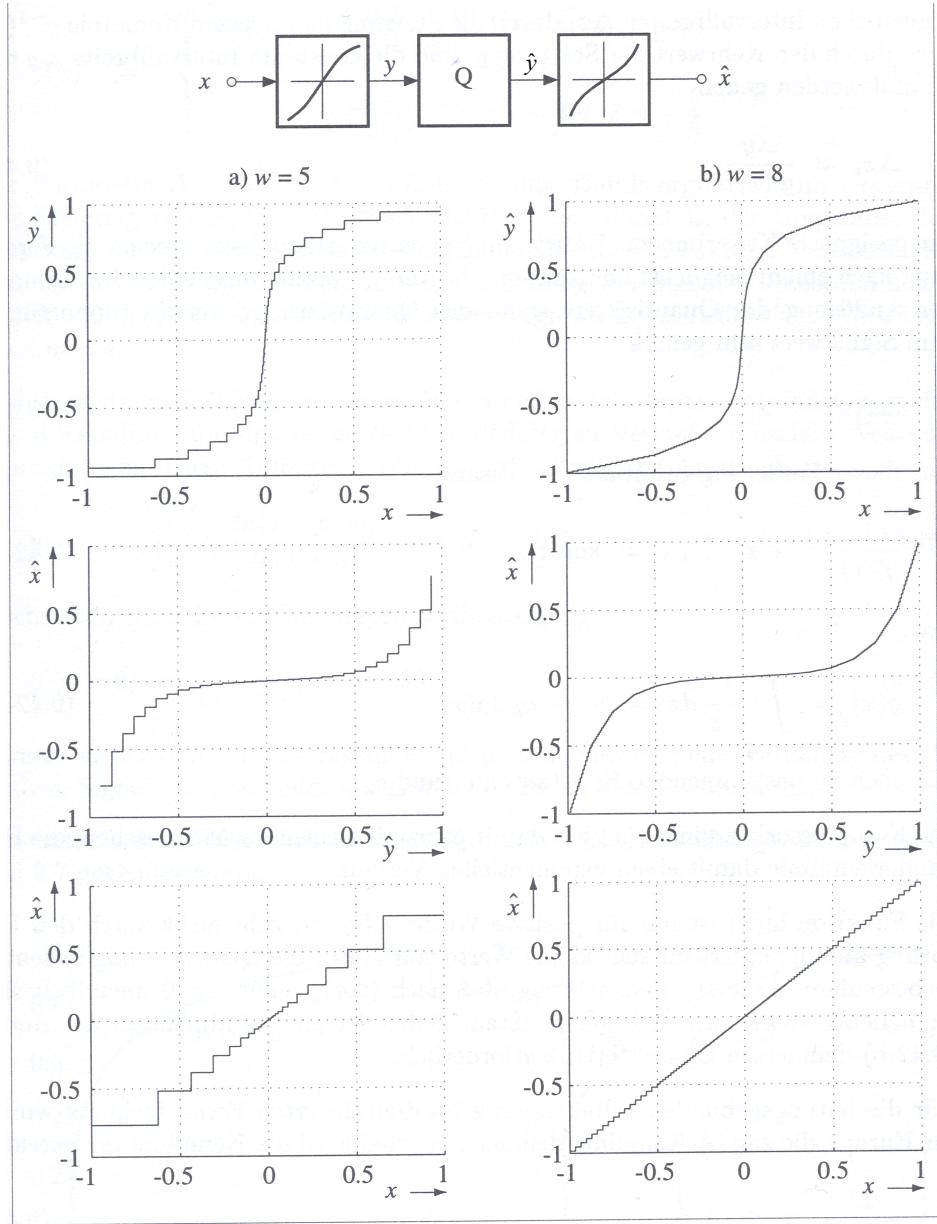
Daraus lässt sich eine Quantisierungskennlinie ableiten, die der Stufenummer y mit konstantem Stufenabstand $\Delta y = 1$ jeweils einen Wertebereich Δx wie folgt zuweist:

$$y'(x) = \frac{\Delta y}{\Delta x} = \frac{1}{Q(x)} \quad (6.7)$$

Dies führt mit $Q(x) \sim x$ und Integration auf

$$y(x) = c_0 + c_1 \ln(x) \quad (6.8)$$

d.h. eine *logarithmische Quantisierungskennlinie*. Eine solche Kennlinie lässt sich leider nur schwer in einem Analog-Digital-Wandler realisieren. Man führt deshalb – funktional gleichwertig – zunächst eine logarithmische Verzerrung (Kompression) des Signals durch, quantisiert dann linear, und führt nach erfolgter Dekodierung eine exponentielle Entzerrung (Expandierung) durch. Dieses Verfahren bezeichnet man auch als *logarithmische Kompaundierung*. Die folgende Abbildung zeigt beispielhaft den Zusammenhang zwischen den Eingangs- und Ausgangssignalen nach der eingangsseitigen Kompression, der anschließenden Quantisierung und der abschließenden Expansion.



Quantisierung mit Kompondierung für unterschiedliche Wortlängen w
(aus Vary et al., 1998, 249).

Leider ist die unmittelbare Realisierung nach Gleichung (6.8) nicht möglich, da sich für $x \rightarrow 0$ $y \rightarrow -\infty$ ergeben würde, d.h. es würden sich unendlich viele Intervalle in der Nähe der Null ergeben. Man kann dies umgehen, indem man

- entweder den logarithmischen Verlauf so verändert, dass der Ursprung $x = 0$ nicht erreicht wird; man kann dies z.B. mittels einer Ursprungs-Verschiebung erreichen:

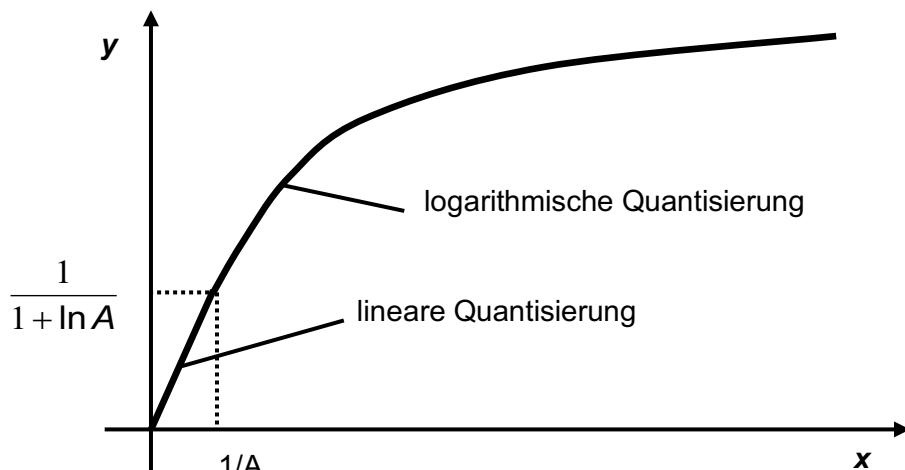
$$y_\mu(x) = \left. \frac{\ln(1 + \mu x)}{\ln(1 + \mu)} \right|_{\mu \approx 255} \quad (6.9)$$

Dies führt auf das sog. μ -law, ein Verfahren, das v.a. in den USA verwendet wird; oder

- den streng logarithmischen Verlauf nur bis zu einem kleinsten Wert $|x| \leq 1/A$ realisiert, und darunter auf eine lineare Kennlinie übergeht:

$$y_A(x) = \begin{cases} \text{sign}(x) \cdot \frac{1 + \ln(A|x|)}{1 + \ln(A)} & \text{für } \frac{1}{A} \leq |x| \leq +1 \\ \frac{A \cdot x}{1 + \ln(A)} & \text{für } -\frac{1}{A} \leq x \leq +\frac{1}{A} \end{cases} \quad (6.10)$$

Der Parameter A wird üblicherweise zu 87.56 gewählt. Die entstehende Kennlinie ist in nachstehender Abbildung gezeigt. Sie weist im Ursprung eine Steigung von 2^4 auf; dadurch werden im Ursprung die wirksamen Quantisierungsintervalle um den Faktor 2^4 verkleinert, was einer Erhöhung des SNR in diesem Bereich um $20 \cdot \log_{10}(2^4) \approx 24$ dB entspricht, bzw. einer Erhöhung der effektiven Wortlänge um 4. Eine solche Kennlinie wird in Europa näherungsweise durch die sog. 13-Segment-Kennlinie realisiert, die von der International Telecommunication Union (ITU-T) in der Empfehlung G.711 definiert wird.



Linear-logarithmische Quantisierungskennlinie.

Wenn man wiederum einen gleichverteilten Quantisierungsfehler annimmt, so lässt sich zeigen, dass der Störabstand in etwa durch folgende obere Schranke angenähert werden kann:

$$SNR [dB] \leq 10 \cdot \log_{10} \frac{3 \cdot 2^{2w}}{(1 + \ln A)^2} = (6w - 10,2) \quad (6.11)$$

Damit ergibt sich wiederum eine Abhängigkeit mit 6 dB/bit, dieses mal aber *unabhängig* von der Signalaussteuerung, solange $|x| \geq 1/A|_{A \approx 100}$.

Zahlenbeispiel:

Für einen Störabstand von 40 dB (vgl. vorangegangene Zahlenbeispiel) ergibt sich eine minimale Wortlänge von ca. 8 bit. Mit 8 kHz Abtastrate ergibt sich daraus eine Bitrate von $f_B = 64 \text{ kbit/s}$. Dies entspricht dem üblichen *ISDN-Standard*.

6.2.3 Optimalquantisierung

In Abschnitt 2.10 wurde bereits erläutert, dass die Amplituden eines Sprachsignals keineswegs gleichverteilt sind, sondern vor allem Werte um Null herum annehmen. Die kleineren Signalwerte kommen also viel häufiger vor und sollten deshalb genauer als große quantisiert werden, um den Quantisierungsfehler klein zu halten. Man könnte also versuchen, die Intervallgrenzen und die Repräsentanten eines jeden Intervalls so zu wählen, dass sich für eine gegebene Amplitudenverteilung ein möglichst hohes SNR ergibt. Dies führt auf die sog. *Optimalquantisierung*, deren genaue Realisierung von der angenommenen Verteilung der Sprachsignalamplituden (z.B. Laplace-, Gauss-Verteilung, etc.) abhängt.

Wir bestimmen zunächst die Leistung des Quantisierungsrauschen zu

$$\begin{aligned} N = E\{e^2(x)\} &= \int_{-\infty}^{+\infty} (\hat{x}_i - u)^2 \cdot p_x(u) du \\ &= \sum_{i=1}^{2^w} \int_{x_{i-1}}^{x_i} (\hat{x}_i - u)^2 \cdot p_x(u) du \end{aligned} \quad (6.12)$$

Die Bedingungen zur optimalen Bestimmung der Intervallgrenzen x_i und der Repräsentanten \hat{x}_i ergeben sich aus den partiellen Ableitungen. Für die Intervallgrenzen und $k = 1, 2, 3, \dots, 2^w-1$ gilt:

$$\frac{\partial N}{\partial x_k} = (\hat{x}_k - x_k)^2 p_x(x_k) - (\hat{x}_{k+1} - x_k)^2 p_x(x_k) = 0 \quad (6.13)$$

somit

$$x_k = \frac{\hat{x}_k + \hat{x}_{k+1}}{2} \quad (6.14)$$

Für die Repräsentaten \hat{x}_i und $k = 1, 2, 3, \dots, 2^w$ gilt:

$$\frac{\partial N}{\partial \hat{x}_k} = 2 \int_{x_{k-1}}^{x_k} (\hat{x}_k - u) p_x(u) du = 0 \quad (6.15)$$

somit

$$\hat{x}_k = \frac{\int_{x_{k-1}}^{x_k} u p_x(u) du}{\int_{x_{k-1}}^{x_k} p_x(u) du} \quad (6.16)$$

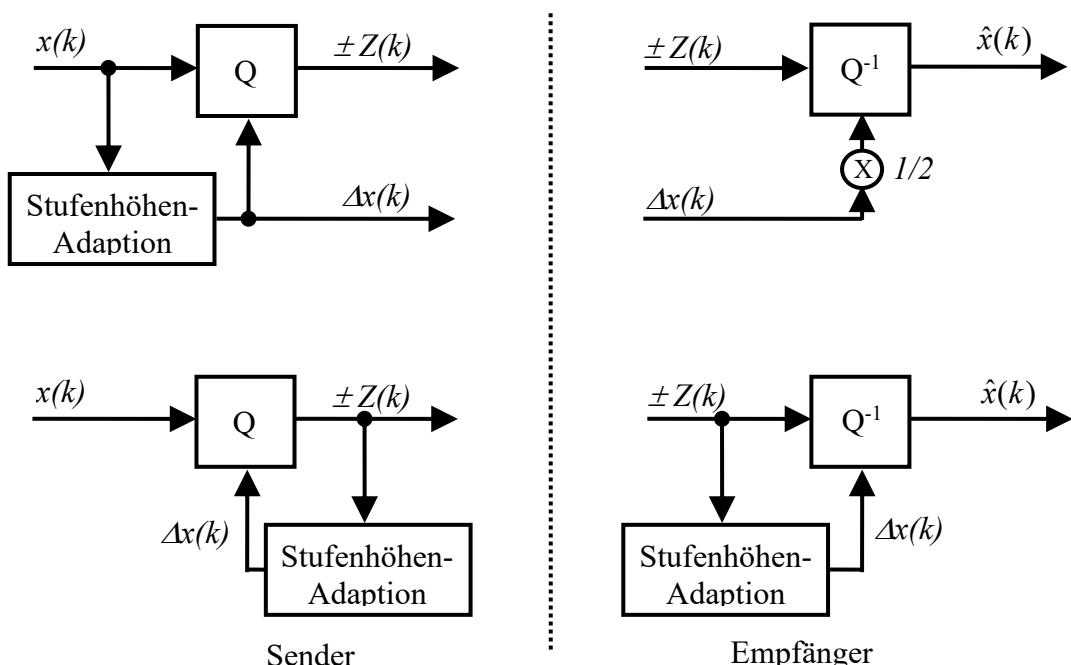
Die optimalen Repräsentanten eines jeden Quantisierungsintervalls entsprechen also den „Schwerpunkten“ dieser Intervalle; die optimalen Intervallgrenzen liegen genau in der Mitte zwischen diesen „Schwerpunkten“, ausgenommen die beiden Intervallgrenzen x_0 und x_L , die durch die Ränder der VDF vorgegeben sind.

Die Gleichungen (6.14) und (6.16) lassen sich numerisch für beliebige VDFen lösen. Die genaue Form der sich daraus ergebenden Kennlinien hängt von der gewählten VDF ab. Üblicherweise erzielen die Kennlinien zwar im optimalen Aussteuerungspunkt eine Verbesserung von ca. 3 dB gegenüber der log. Kompondierung, sind im Gegenzug allerdings

stärker aussteuerungsabhängig. Auch aus Gründen der Realisierbarkeit ist die Optimalquantisierung deshalb bislang nicht wirklich konkurrenzfähig.

6.2.4 Adaptive Quantisierung

Der Störabstand und seine Abhängigkeit von der Signalaussteuerung lässt sich ebenfalls verbessern, wenn man die Stufenhöhe des Quantisierers – bei Beibehaltung des Prinzips der gleichmäßigen, linearen Quantisierungskennlinie – an die momentane Aussteuerung anpasst. D.h. man wählt eine Stufenhöhe, die die momentane Signalaussteuerung (oder -leistung) widerspiegelt. Dies ist eine Alternative zur Optimalquantisierung, bei der die Kennlinie global für alle Signalabschnitte optimiert wurde.



Adaptive Quantisierung mit Vorwärtsadaption (oben) oder Rückwärtsadaption (unten),
vgl. Vary et al. (1998, 257).

Grundsätzlich sind zwei unterschiedliche Prinzipien vorstellbar, welche in o.a. Abbildung dargestellt sind. Im ersten Fall – der sogenannten Vorwärtsadaption oder *Adaptive Quantization Forward* (AQF) – wird die Stufenhöhe $\Delta x(k)$ blockweise für einen Signalabschnitt der Länge N berechnet und dann für diesen Abschnitt festgehalten. Der Wert der Stufenhöhe muss dann mit übertragen werden, da ansonsten am Empfänger das Signal nicht wieder korrekt generiert werden kann. Da dazu zusätzliche Informationen übertragen werden müssen wählt man die Blocklänge recht lang, z.B. alle $N = 128$ Werte. Im zweiten Fall – der sogenannten Rückwärtsadaption oder *Adaptive Quantization Backwards* (AQB) – entfällt die Notwendigkeit, die Stufenhöhe zu übertragen, da diese Information aus dem übertragenen Signal $Z(k)$ gewonnen werden kann, sofern die Übertragung fehlerfrei funktioniert.

Der quantisierte Wert ergibt sich in beiden Fällen zu

$$\hat{x}(k) = \text{sign}(x(k)) \cdot Z(k) \frac{\Delta x(k)}{2} \quad (6.17)$$

Die Stufenhöhe wird bei beiden Verfahren proportional zur geschätzten momentanen Varianz von $x(k)$ bzw. von $\hat{x}(k)$ eingestellt:

$$\Delta x(k) = c \cdot \hat{\sigma}_x(k) \quad , c = \text{const.} \quad (6.18)$$

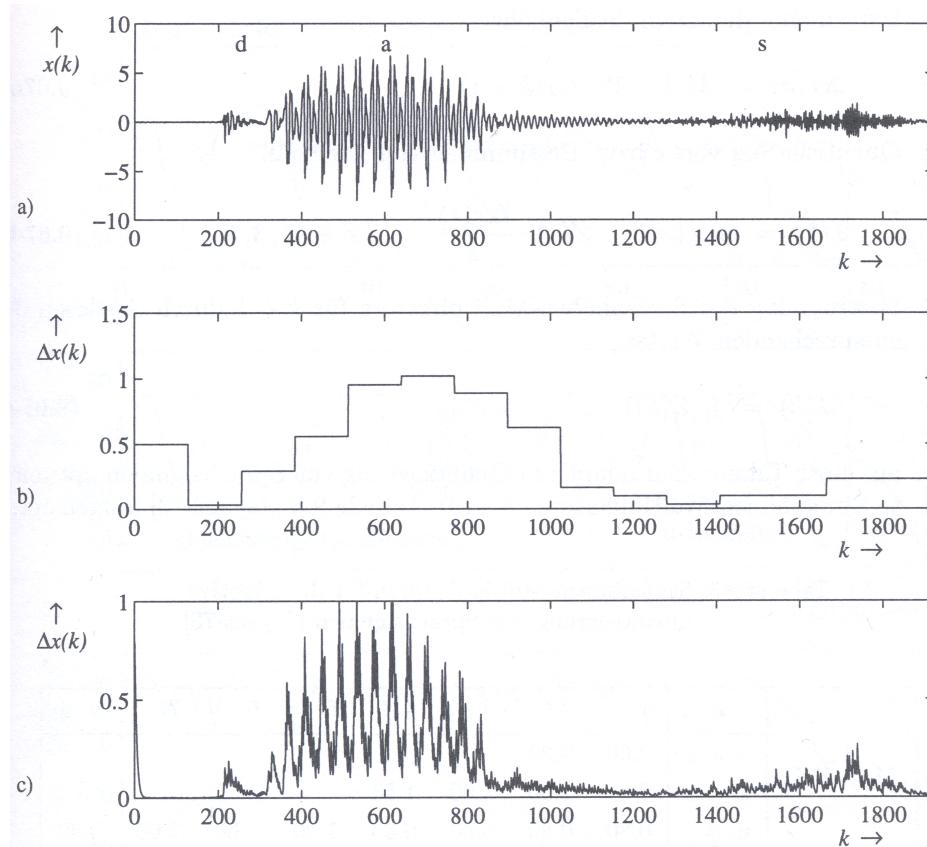
Beim AQF-Verfahren wird die Varianz blockweise aus N Signalwerten geschätzt:

$$\hat{\sigma}_x(k) = \sqrt{\frac{1}{N} \sum_{i=0}^{N-1} x^2(k_0 + i)} \quad \forall \quad k = k_0, k_0 + 1, \dots, k_0 + N - 1 \quad (6.19)$$

Beim AQB-Verfahren wird die Schätzung der Signalleistung iterativ anhand des letzten, bereits verfügbaren Wertes $\hat{x}(k-1)$ und der letzten geschätzten Signalleistung durchgeführt:

$$\hat{\sigma}_x^2(k) = \alpha \cdot \hat{\sigma}_x^2(k-1) + (1-\alpha) \cdot \hat{x}^2(k-1), \quad 0 < \alpha < 1 \quad (6.20)$$

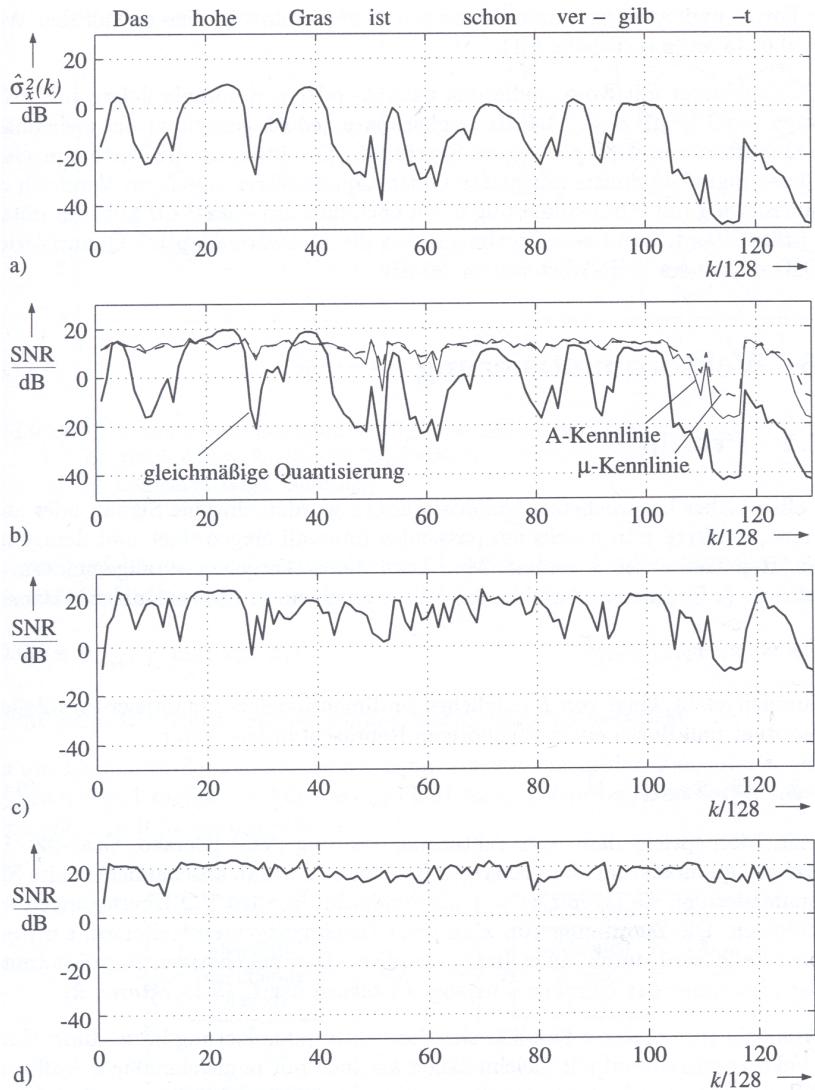
Durch die iterative Berechnung kann die Stufenhöhe schneller an die Aussteuerung des Signales angepasst werden. Unterstehende Abbildung veranschaulicht diese Eigenschaft.



Beispiel zur Adaption der Stufenhöhe. Oben: Zeitsignal „das“; mitte: Stufenhöhe bei Vorwärtsadaption; unten: Stufenhöhe bei Rückwärtsadaption. Aus Vary et al. (1998, 259).

Mit Hilfe der Vorwärtsadaption lässt sich ungefähr derselbe Störabstand wie bei der Optimalquantisierung erzielen. Bei Rückwärtsadaption lässt er sich nochmals um ca. 3 dB steigern. Die nachfolgende Tabelle und nachfolgende Abbildung, die Vary et al. (1998, 260-261) entnommen sind, zeigen beispielhaft die erzielbaren Störabstände bei einer Quantisierung mit der Wortlänge $w = 4$.

Quantisierung	SNR
linear	7,44
logarithmisch, A-Kennlinie	13,71
logarithmisch, μ -Kennlinie	13,43
Optimalquantisierer	16,53
adaptive Quantisierung (AQF)	16,73
adaptive Quantisierung (AQB)	19,88



Erzielbare Störabstände mit $w = 4$. a) Kurzzeitleistung des Signals; b) Blockweise berechnetes SNR bei linearer oder logarithmischer Quantisierung; c) Blockweise berechnetes SNR bei AQF; d) Blockweise berechnetes SNR bei AQB.

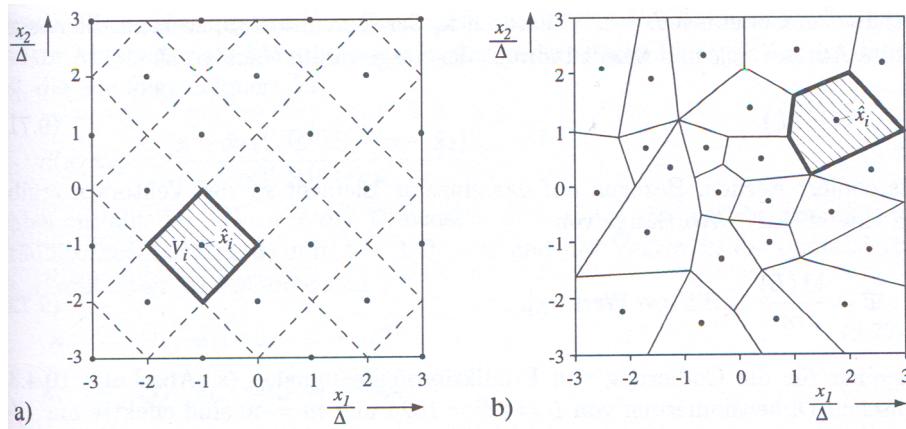
Aus Vary et al. (1998, 261).

Die logarithmische Quantisierung erzielt gegenüber der gleichmäßigen, linearen Quantisierung bereits einen Gewinn von ca. 6 dB. Dieser lässt sich durch Optimalquantisierung oder AQF nochmals um 3 steigern. Den höchsten Störabstand erzielt man mit AQB, das gegenüber der linearen Quantisierung in diesem Beispiel einen Gewinn von über 12 dB verzeichnet.

6.2.5 Vektorquantisierung

Bei den bisherigen Verfahren zur Quantisierung sind wir davon ausgegangen, dass wir einen Signalwert oder Parameterwert x durch einen Repräsentanten \hat{x} ersetzen wollen. Diese Idee kann man auch verallgemeinern: Es werden m aufeinanderfolgende (Signal- oder Parameter-) Werte zu einem Vektor $\mathbf{x} = (x_1, x_2, \dots, x_m)^T$ zusammengefasst und durch einen Repräsentantenvektor $\hat{\mathbf{x}} = (\hat{x}_1, \hat{x}_2, \dots, \hat{x}_m)^T$ ersetzt. Dieses Verfahren bezeichnet man als *Vektorquantisierung*. Die bisher betrachtete Quantisierung einzelner Werte kann als ein Spezialfall der Vektorquantisierung, als sogenannte Skalarquantisierung, betrachtet werden.

Wie im skalaren Fall kann die Quantisierung gleichmäßig oder ungleichmäßig, oder auch mit „optimal“ ausgewählten Repräsentanten, ausgeführt werden. An die Stelle eines ein-dimensionalen Repräsentanten tritt nun ein Repräsentantenvektor $\hat{\mathbf{x}}_i$; die Sammlung aller L Repräsentantenvektoren bezeichnet man als *Codebuch* der L Codevektoren $\hat{\mathbf{x}}_i$. An die Stelle des ein-dimensionalen Quantisierungsintervalls Δx tritt eine m -dimensionale sog. Voronoi-Zelle \mathbf{V}_i . Ein Beispiel für den 2-dimensionalen Fall ist in untenstehender Abbildung skizziert.



Vektorquantisierung mit $L = 25$ Codevektoren der Dimension $m = 2$. a) gleichmäßige Auflösung; b) ungleichmäßige Auflösung. Aus Vary et al. (1998, 263).

Die Aufgabe der Vektorquantisierung besteht nun darin, bei einem gegebenen Codebuch jeden Eingangsvektor \mathbf{x} durch einen möglichst ähnlichen Repräsentantenvektor $\hat{\mathbf{x}}_{i,opt}$ zu ersetzen. Die Auswahl erfolgt auf Basis eines Distanz- oder Ähnlichkeitsmaßes:

$$d(\mathbf{x}, \mathbf{x}_{i,opt}) = \min_i d(\mathbf{x}, \hat{\mathbf{x}}_i) \quad (6.21)$$

Sofern das Codebuch beim Empfänger bekannt ist muss nicht der gesamte Repräsentant $\hat{\mathbf{x}}_{i,opt}$, sondern nur seine „Adresse“ i_{opt} übertragen werden. Hierdurch lässt sich die zur Beschreibung notwendige Bitrate erheblich drücken. Enthält ein Codebuch bspw. $L = 2^w$ Repräsentanten $\hat{\mathbf{x}}_i$ der Dimension m , so kann die optimale Adresse i_{opt} durch $w = \text{ld}(L)$ bits kodiert werden. Bezogen auf ein einzelnes Element x_i des Vektors \mathbf{x} ergibt sich somit eine effektive Wortlänge

$$\bar{w} = \frac{\text{ld}(L)}{m} \text{ bit.} \quad (6.22)$$

Für eine typische Dimensionierung mit $L = 2^{10} = 1024$ und $m = 40$ ergibt sich somit effektiv nur eine Wortlänge von $\frac{1}{4}$ bit pro Wert x_i . Dies ist eine erhebliche Einsparung gegenüber der herkömmlichen Quantisierung; allerdings geht das auch auf Kosten des Störabstandes, wie wir noch sehen werden.

Die Wahl eines passenden Distanzmaßes richtet sich nach der Art der zu kodierenden Werte. Für die Kodierung von Sprachsignalabschnitten verwendet man häufig den mittleren quadratischen Fehler

$$\begin{aligned} d(\mathbf{x}, \hat{\mathbf{x}}_i) &= \frac{1}{m} (\mathbf{x} - \hat{\mathbf{x}}_i)^T (\mathbf{x} - \hat{\mathbf{x}}_i) \\ &= \frac{1}{m} \sum_{\mu=1}^m (x_\mu - \hat{x}_{i\mu})^2 \quad i = 1, 2, \dots, L \end{aligned} \quad (6.23)$$

Die Minimierung dieses Fehlers entspricht der Auswahl des zu \mathbf{x} nächsten Repräsentanten $\hat{\mathbf{x}}_i$ im m -dimensionalen Vektorraum. Alternativ kann man den quadratischen Fehler auch wie folgt wichten:

$$d(\mathbf{x}, \hat{\mathbf{x}}_i) = \frac{1}{m} (\mathbf{x} - \hat{\mathbf{x}}_i)^T \mathbf{W} (\mathbf{x} - \hat{\mathbf{x}}_i) \quad (6.24)$$

wobei \mathbf{W} eine Matrix der Dimension $m \times m$ darstellt. Durch geeignete Wahl von \mathbf{W} kann z.B. der Fehler so gewichtet werden, dass er sich auditiv am wenigsten auf die Qualität der so kodierten Sprache auswirkt.

Die Vektorquantisierung wird häufig nicht direkt zur Quantisierung von Sprachsignalabschnitten, sondern zur Quantisierung von Koeffizientensätzen – etwa derer der linearen Prädiktion – verwendet. Hierzu verwendet man andere Distanzen, wie z.B. die sog. Itakura-Saito-Distanz:

$$d(\mathbf{x}, \hat{\mathbf{x}}_i) = \frac{(\mathbf{x} - \hat{\mathbf{x}}_i)^T \mathbf{R}^{(n+1)} (\mathbf{x} - \hat{\mathbf{x}}_i)}{\mathbf{x}^T \mathbf{R}^{(n+1)} \mathbf{x}} \quad (6.25)$$

Dabei enthält der Vektor \mathbf{x} die zu quantisierenden Prädiktorkoeffizienten a_i , und der Vektor $\hat{\mathbf{x}}_i$ die quantisierten Repräsentanten dieser Koeffizienten. $\mathbf{R}^{(n+1)}$ ist die quadratische $(n+1) \times (n+1)$ AKF-Matrix des Sprachsignalsegmentes, für das die Prädiktorkoeffizienten berechnet wurden.

Es ist zu beachten, dass die Vektorquantisierung im Allgemeinen sehr rechenintensiv ist, insbesondere dann, wenn der Eingangsvektor \mathbf{x} mit allen L Codebuchvektoren verglichen werden muss (vollständige Suche). Wie nach kurzer Rechnung (vgl. Vary et al., 1998, 265) gezeigt werden kann benötigt man dann $(3m-1) \cdot 2^{m\bar{w}}$ Operationen. Für $L = 1024$ Vektoren und 8 kHz Abtastfrequenz sind das bspw. $24,6 \cdot 10^6$ Operationen pro Sekunde. Da es sich bei der Vektorquantisierung häufig nur um eine Teilaufgabe der Signalkodierung handelt ist dieser Aufwand recht erheblich. Man versucht deshalb, durch schnelle Suchalgorithmen und eine geschickte Codebuchauswahl diesen Aufwand zu verringern.

Ein Beispiel hierfür ist der sog. Lattice-Quantisierer. Dies ist ein Vektorquantisierer, dessen Codevektoren durch regelmäßige Gitterpunkte in einem m -dimensionalen Raum gegeben sind. Die eingangs in diesem Kapitel gegebene Abbildung (linkes Bild) zeigt einen solchen Lattice-Quantisierer für die Dimensionalität 2. Die Lage der Codevektoren dieses

Quantisierers lässt sich analytisch angeben, sodass das Codebuch selbst nicht gespeichert werden muss. Darüber hinaus erfüllen die Codevektoren noch die Bedingungen, dass sämtliche Vektorkomponenten ganzzahlige Vielfache einer kleinsten Einheit Δ sind, und dass die Komponentensummen gerade sind. Aufgrund dieser Bedingungen lassen sich die Distanzberechnungen auf einfache Rundungsoperationen zurückführen.

Den Vorteil der einfachen Beschreibung und Suche erkauft man sich beim Lattice-Quantisierer durch eine nicht optimale Auswahl der Codevektoren. In der Tat ist ein Lattice-Quantisierer nur bei gleichverteilten Vektoren \mathbf{x} optimal. Optimale Vektor-Codebücher kann man durch Verallgemeinerung der schon in Kapitel 6.2.3, Gleichung (6.12), aufgestellten Bedingung erzielen:

$$E\{d(\mathbf{x}, \hat{\mathbf{x}}_i)\} = \sum_{i=1}^L \int_{V_i} d(\mathbf{u}, \hat{\mathbf{x}}_i) \cdot p_x(\mathbf{u}) d\mathbf{u} \quad (6.26)$$

Hierbei ist $p_x(\mathbf{u})$ die m -dimensionale Verbundverteilungsdichte der Vektorfolge $\mathbf{x}(k)$; die Integration erfolgt über die i -te m -dimensionale Voronoi-Zelle. Optimale Repräsentanten erhält man wiederum durch analog zu Gl. (6.13) und (6.15) aufzustellende Bedingungen; sie stellen die „Schwerpunkte“ der Voronoi-Zellen dar. Da sich ein formelmäßiger Zusammenhang zwischen der Verbundverteilungsdichte und den Codevektoren meist nicht angeben lässt verwendet man iterative Suchalgorithmen für die optimalen Codevektoren; hier ist z.B. der Linde-Buzo-Gray-Algorithmus zu nennen (vgl. Vary et al., 1998, 267).

Die sich dabei ergebenden Codebücher hängen natürlich von der Verteilungsdichte ab. So passen sich die Größen der Voronoi-Zellen der Verteilungsdichte an, und eventuell im Signalblock vorhandene Korrelationen werden zur Verbesserung des Störabstandes genutzt. So lassen sich mit einer effektiven Wortlänge von 4 für unkorrelierte Signale Störabstände von 20,8 dB, für korrelierte Signale bis zu 24 dB erzielen. Allerdings ist diese Wortlänge schon sehr hoch; bei typischen Wortlängen von effektiv 0,25 bit ergeben sich nur Störabstände zwischen 1 und 6 dB. Diese Werte machen deutlich, weshalb der Vektorquantisierer meist nicht zur unmittelbaren Signalquantisierung eingesetzt wird. Er ist aber sehr gut zur Quantisierung von Prädiktionsfehlersignalen, von Kurzzeitspektren, oder von Koeffizientensätzen, wie bspw. der LPC-Koeffizienten, geeignet.

Bei Einsatz des Vektorquantisierers zur Signalquantisierung ist damit zu rechnen, dass Signalblöcke ähnlicher Form, aber unterschiedlicher Verstärkung, auftreten. Deshalb macht es Sinn, im Codebuch die optimalen Repräsentanten zu normieren und zusätzlich zur Adresse auch einen Skalierungsfaktor zu übertragen. Noch besser ist es allerdings, wenn jeder Codebuchvektor $\hat{\mathbf{x}}_i$ mit einem jeweils optimierten, aus \mathbf{x} und $\hat{\mathbf{x}}_i$ abgeleiteten Verstärkungsfaktor an den jeweiligen Eingangsvektor \mathbf{x} angepasst wird. Man bezeichnet dieses Verfahren als Gain-Shape-Vektorquantisierung. Es kommt in vielen handelsüblichen Kodieralgorithmen zum Einsatz.

6.3 Differentielle PCM

Bei der Kodierung von Sprachsignalen haben wir bislang nur die Quantisierung betrachtet und den dabei entstehenden Fehler abgeschätzt. Ziel der Quantisierung war es, eine störsichere und einheitliche Signaldarstellung zu erhalten, allerdings auf Kosten einer höheren Bandbreite im Vergleich zum analogen Signal.

Die Bandbreite lässt sich nun auf zwei prinzipielle Arten verringern:

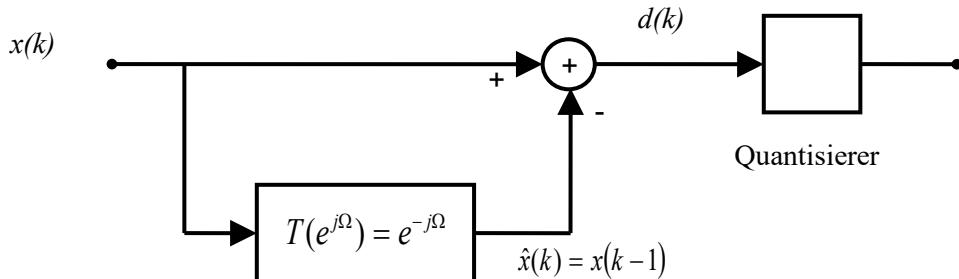
- *Redundanzreduktion*: Alles, was keine Information enthält, wird bei der Übertragung weggelassen.
- *Irrelevanzreduktion*: Alle Informationen, die für den gegebenen Anwendungsfall irrelevant sind, werden weggelassen. Z.B. könnte das Ziel sein, eine verständliche Sprache zu übertragen, ohne dass die Identität des Sprechers interessiert; in diesem Falle könnte sprechertypische Informationen weggelassen werden.

Die in diesem Abschnitt vorgestellten Verfahren versuchen vor allem, die *Redundanz im Sprachsignal zu reduzieren*. Dass eine solche vorhanden ist lässt sich leicht zeigen, wenn man die Autokorrelation eines Sprachsignalabschnitts berechnet: Einzelne Abtastwerte sind teilweise sehr stark korreliert, d.h. der folgende Abtastwert hängt stark vom vorangegangenen Wert ab. Dies zeigt sich auch in der Bandbegrenzung des Sprachsignals: Starke Sprünge von einem Abtastwert zum nächsten würden hohe Frequenzen erzeugen, die im Sprachsignal nicht enthalten sind.

Wegen der Korrelation zwischen einzelnen Abtastwerten ist es günstig, nicht den Abtastwert selbst, sondern nur die *Differenz zum vorangehenden Abtastwert* zu übertragen. Durch die Differenzbildung entsteht ein Signal, welches den Quantisierer weniger stark aussteuert, und welches deshalb mit einer geringeren Wortlänge (Bitrate) übertragen werden kann. Im einfachsten Fall führt dies auf die sog. *Differenz-PCM*. Dabei ergibt sich das zu quantisierende und übertragende Signal zu

$$d(k) = x(k) - x(k-1) \quad (6.27)$$

Die Differenz-PCM ist in nachfolgender Abbildung dargestellt.



Schematische Darstellung der Differenz-PCM.

Man kann die Differenz-Kodierung verbessern, indem man eine *gewichtete Differenz* zwischen zwei Abtastwerten überträgt:

$$d(k) = x(k) - a \cdot x(k-1) \quad (6.28)$$

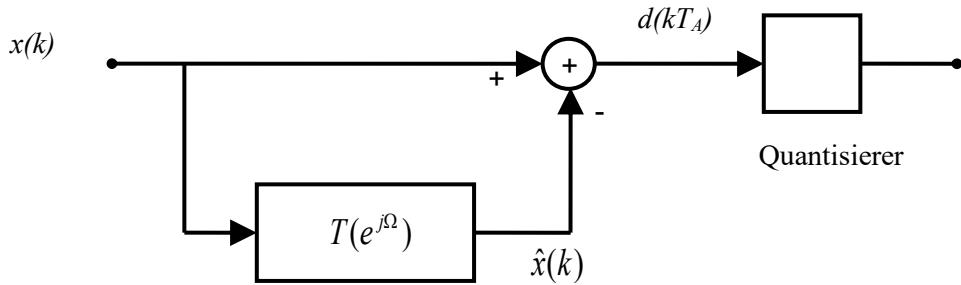
Damit lässt sich in etwa 1 bit Wortlänge einsparen, was bei 8 kHz Abtastrate schon eine Verringerung der Bitrate um 8 kbit/s entspricht.

Der Gewichtungsfaktor a kann natürlich auch adaptiv eingestellt werden, entweder blockadaptiv oder sequentiell. Bei der blockadaptiven Lösung – die in der Literatur meist als *Adaptive Predictive Coding* (APC) bezeichnet wird – wählt man als im Sinne des minimalen mittleren quadratischen Fehlers optimalen Koeffizienten

$$a_{opt} = \frac{\varphi_{xx}(1)}{\varphi_{xx}(0)} \quad (6.29)$$

d.h. das Verhältnis der Autokorrelationen bei einer Verschiebung von einem Abtastwert zur Autokorrelation von Null.

Diese gewichtete Differenz lässt sich weiter verallgemeinern: Statt des gewichteten – direkt vorangehenden – Abtastwertes lässt sich eine gewichtete Summe einer Reihe von vorangehenden Abtastwerten abziehen. Mit anderen Worten: Man versucht, den augenblicklichen Abtastwert als eine Linearkombination vorangegangener Abtastwerte darzustellen. Dies ist das Verfahren der *linearen Prädiktion*, die schon in Abschnitt 4.3 vorgestellt wurde. Es ist schematisch in folgender Abbildung gezeigt.



Schematische Darstellung der linearen Prädiktion.

Das Schätzsignal ergibt sich hierbei als Linearkombination aus n vorangegangenen Abtastwerten:

$$\hat{x}(k) = \sum_{i=1}^n a_i \cdot x(k-i) \quad (6.30)$$

Die Übertragungsfunktion des Prädiktorfilters ergibt sich hierbei zu

$$T(e^{j\Omega}) = \sum_{i=1}^n a_i e^{-ji\Omega} \quad (6.31)$$

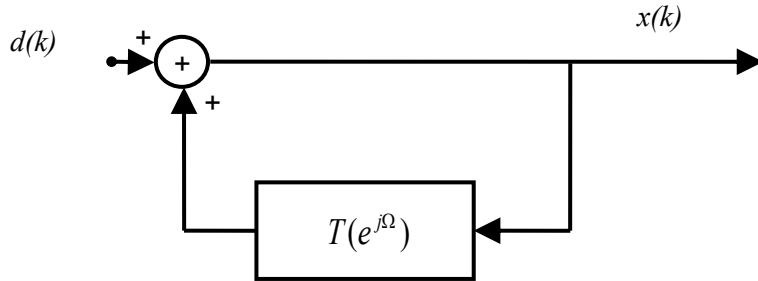
und die Übertragungsfunktion der Gesamtschaltung zu

$$H(e^{j\Omega}) = 1 - T(e^{j\Omega}) \quad (6.32)$$

Aus dem übertragenen Differenzsignal $d(k)$ lässt sich durch Invertierung der Übertragungsfunktion wieder das Sprachsignal $x(k)$ gewinnen:

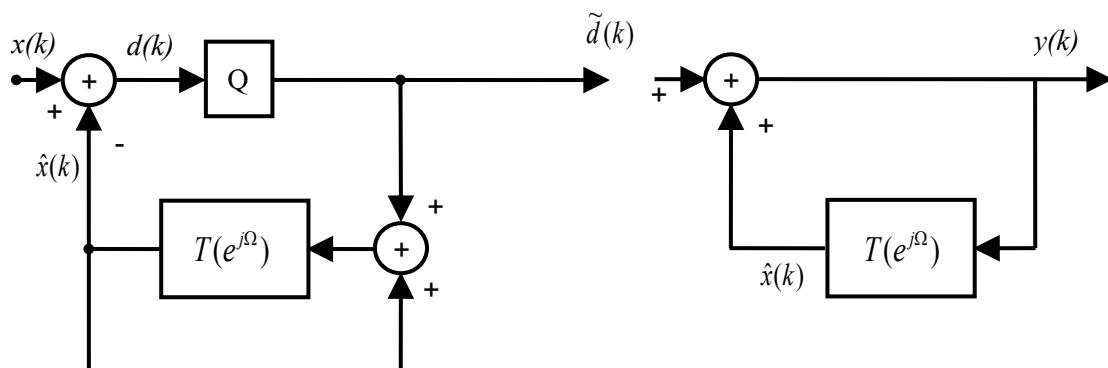
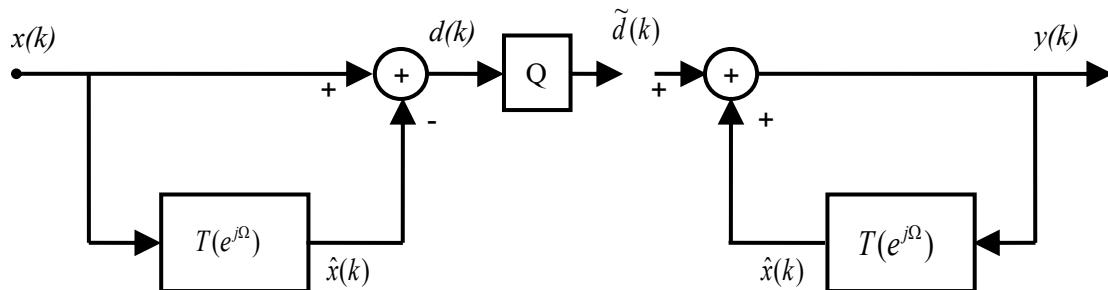
$$\frac{1}{H(e^{j\Omega})} = \frac{1}{1 - T(e^{j\Omega})} \quad (6.33)$$

Diese Funktion lässt sich wie unten dargestellt realisieren.



Empfangsfilter der linearen Prädiktion.

Die Koeffizienten a_i sollten normalerweise adaptiv aus dem Signal berechnet werden, d.h. sie hängen von der Zeit k ab ($a_i(k)$). Zur Rekonstruktion auf der Empfangsseite muss man sie deshalb zum Empfänger übertragen. Dadurch wird ein Teil der Einsparung, die die Prädiktion erbringen sollte, wieder zunichte gemacht. Diesen Nachteil umgeht man bei einer alternativen Form der Prädiktion, bei der das Differenzsignal „rückwärts“ mit Hilfe von Informationen, die beiden – Sender und Empfänger – vorliegen, berechnet wird. Beide Prinzipien sind in untenstehender Abbildung skizziert.



Lineare Prädiktionskodierung. Oben: Vorwärtsprädiktion; unten: Rückwärtsprädiktion; links: Sender; rechts: Empfänger.

Sofern das Restsignal nicht quantisiert wird stimmen Vorwärtsprädiktion (auch *open-loop prediction*) und Rückwärtsprädiktion (*closed-loop prediction*) überein. Wird jedoch das Restsignal quantisiert, so unterscheiden sich die sendeseitig geschätzten Signale $\hat{x}(k)$

voneinander. Der Quantisierungsfehler wirkt sich dadurch in unterschiedlicher Form auf das Ausgangssignal aus, wie folgende Betrachtungen zeigen werden.

Wir gehen zunächst von einem linearen Quantisierer mit Stufenhöhe Q aus, bei dem das Quantisierungsrauschen gleichverteilt ist. Bei der Vorwärtsprädiktion setzt sich das quantisierte Restsignal aus den beiden Komponenten

$$\tilde{d}(k) = d(k) + \Delta(k) \quad (6.33)$$

zusammen. Dieses Signal wird nun empfangsseitig durch das – lineare – Empfangsfilter geschickt, und es entsteht $y(k)$. Da das Sende- und Empfangsfilter zueinander invers sind entsteht dabei das Originalsignal $x(k)$ sowie eine mit dem Empfangsfilter gefilterte Version des Quantisierungsrauschens $\Delta(k)$. Das ansonsten weiße Quantisierungsrauschen wird also mit der Übertragungsfunktion des Empfangsfilters gefiltert. Das Empfangsfilter modelliert jedoch – sofern die Prädiktorkoeffizienten genügend gut bestimmt wurden – das Vokaltraktfilter. Das Quantisierungsrauschen wird also quasi durch ein angenähertes Vokaltraktfilter gewichtet. Dies wirkt sich auditiv ausgesprochen positiv aus. Man kann allerdings zeigen, dass sich der sog. Prädiktionsgewinn

$$G_p = \frac{\sigma_x^2}{\sigma_d^2} \quad (6.34)$$

bei dieser Anordnung nicht in einer Verbesserung des SNR niederschlägt. Durch die Prädiktion wird also der Störabstand nicht verbessert, allerdings wird das Quantisierungsrauschen durch die empfangsseitige Filterung auditiv maskiert. Dies ist eine direkte Anwendung der spektralen Maskierung aus Kapitel 5.

Bei der Rückwärtsprädiktion verhält es sich genau umgekehrt. Dort gilt nach oben stehender Abbildung

$$\begin{aligned} y(k) &= \hat{x}(k) + \tilde{d}(k) \\ &= \tilde{x}(k) \\ &= d(k) + \Delta(k) + \hat{x}(k) \\ &= x(k) + \Delta(k) \end{aligned} \quad (6.35)$$

Im Gegensatz zur Vorwärtsprädiktion macht sich das Quantisierungsrauschen also im Ausgangssignal als weisses Rauschen bemerkbar. Allerdings bewirkt der Prädiktionsgewinn eine Steigerung des SNR um den Faktor $10 \cdot \lg(G_p)$; das entspricht einer Wortlängenverkürzung um

$$\Delta w = \frac{1}{2} \lg G_p \quad (6.36)$$

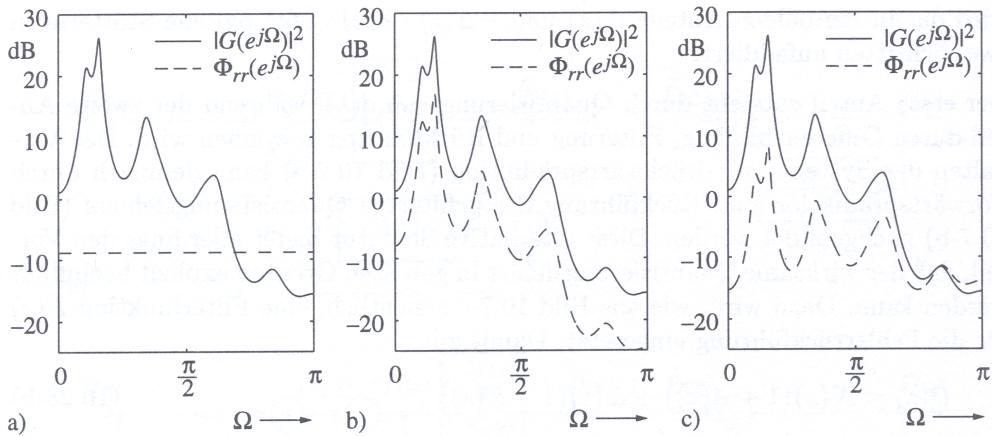
wenn man den Prädiktionsgewinn direkt zur Verringerung der Bitrate einsetzen und somit das SNR konstant halten möchte.

Die Unterschiede zwischen Vorwärts- und Rückwärtsprädiktion lassen sich an dem unten dargestellten Beispiel des Vokals „o“ zeigen, welches Vary et al. (1998, 286) entnommen ist. Im linken Bild (Rückwärtsprädiktion) liegt weisses Quantisierungsrauschen vor, welches in den energiearmen Signalabschnitten das Nutzsignal übersteigt, in den energiereichen aber deutlich unter dem Nutzsignal liegt. Umgekehrt ist es bei der Vorwärtsprädiktion im mittleren

Bild. Zwischen diesen beiden Extrema lässt sich das Rauschen durch ein dem Vorwärtsprädiktor nachgeschaltetes Filter spektral optimal beeinflussen, sodass

- der Störabstand ähnlich der Rückwärtsprädiktion verbessert wird, gleichzeitig jedoch
- der verbleibende Fehler spektral wie bei der Vorwärtsprädiktion an das Signalspektrum angenähert – und somit „unhörbar“ gemacht wird.

Man bezeichnet diese gezielte Beeinflussung des Quantisierungsrauschen als *Noise Shaping*.



Spektrale Formung des Quantisierungsrauschen. a) Rückwärtsprädiktion; b) Vorwärtsprädiktion; c) Noise Shaping. Dargestellt sind der Frequenzgang des Synthesefilters $|G(e^{j\omega})|$ sowie das Leistungsdichtespektrum des wirksamen Quantisierungsfehlers $\Phi_{rr}(e^{j\omega})$. Aus Vary et al. (1998, 286).

Eine weitere Reduktion der Bitrate lässt sich erzielen, wenn man die Prädiktion und die Quantisierung *adaptiv* vornimmt. Das heißt, dass die *Quantisierungsstufen an die jeweilige Signalamplitude angepasst* werden und die *Prädiktorkoeffizienten a_i für den jeweiligen Signalabschnitt bestimmt* werden. Man verwendet hier meist einen rückwärtsadaptiven Quantisierer (AQB) sowie einen Rückwärtsprädiktor, bei dem keine weiteren Informationen zum Empfänger übertragen werden müssen.

Dieses Verfahren bezeichnet man üblicherweise als *Adaptive Differenz-Pulse-Code-Modulation*, abgekürzt *ADPCM*. Es wird in schnurlosen Telefonen nach dem sog. DECT-Standard oder in sogenannten „Leitungsvervielfachern“ (z.B. für Seekabel) eingesetzt und ist von der International Telecommunication Union in der Empfehlung G.726 standardisiert. Allerdings werden hierbei besondere Prädiktoren (mit Pol- und Nullstellen) sowie besondere Adaptionsalgorithmen eingesetzt. Die damit erzielbare Sprachqualität kommt derjenigen einer logarithmischen PCM nach der A - oder μ -Kennlinie (ITU-T-Empfehlung G.711 mit 64 kbit/s) schon recht nahe, und das bei nur 32 kbit/s, d.h. einer effektiven Wortlänge von 4 bit (gegenüber 8 bit bei log. PCM). Gegenüber der logarithmischen PCM lassen sich also durch die lineare Prädiktion sowie durch adaptive Quantisierung und Prädiktion insgesamt fast die Hälfte an Informationen einsparen.

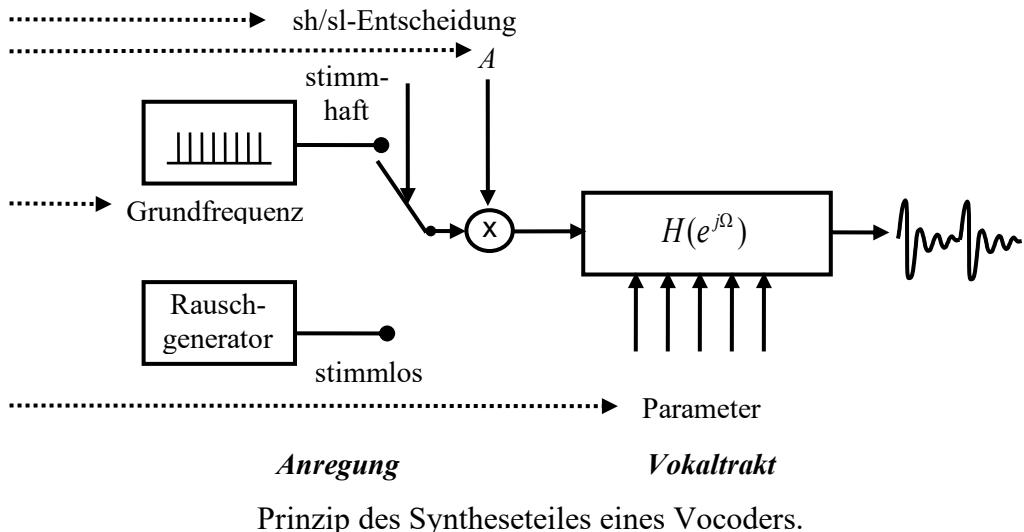
6.4 Parametrische Kodierung

Möchte man Sprache bei noch niedrigeren Bitraten kodieren, so lässt sich der Spracherzeugungsprozess modellhaft nachvollziehen. Man bezeichnet solche Verfahren als

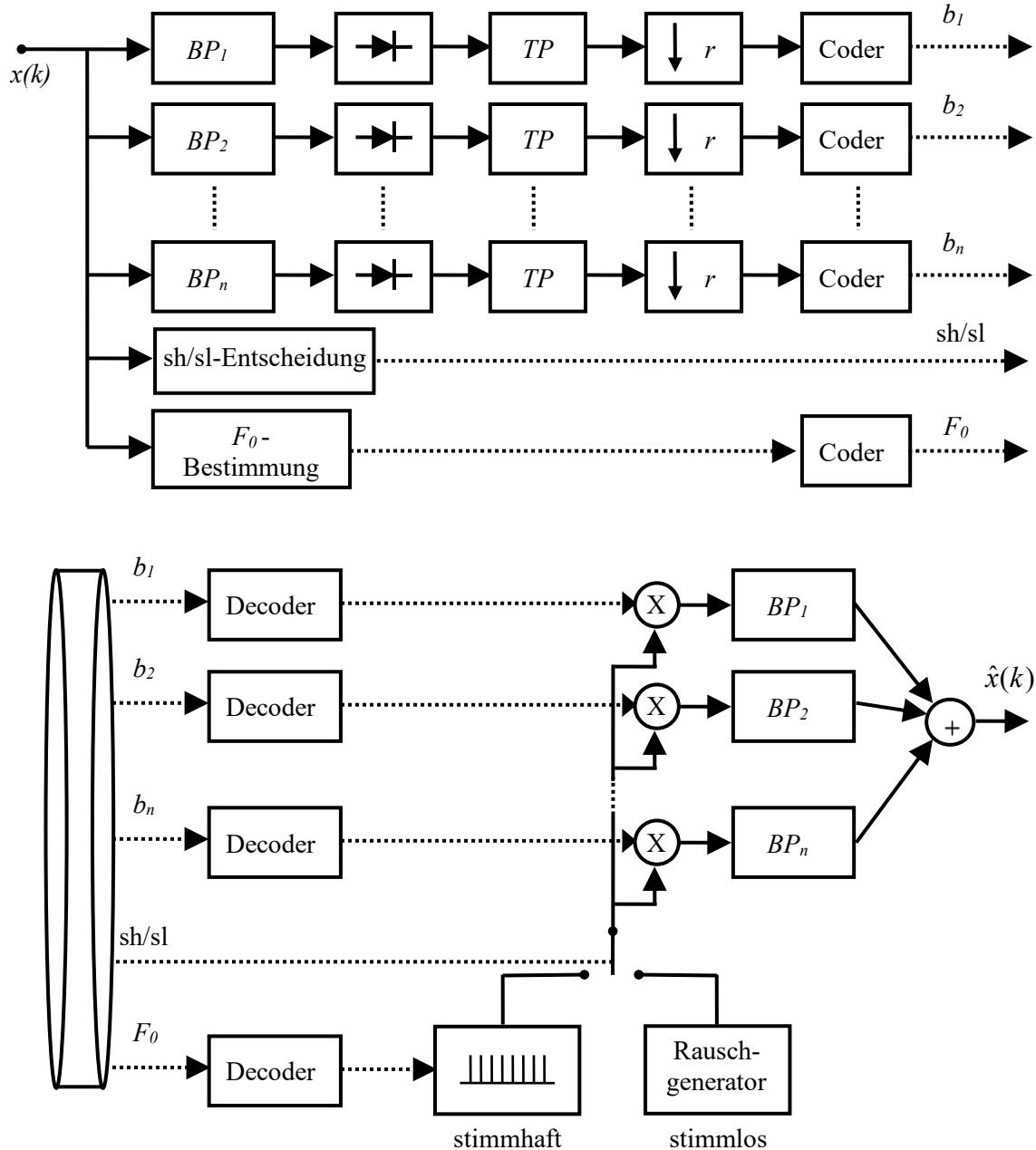
Analyse-Synthese-Systeme oder auch als *Vocoder (Voice Coder)*. Gemeinsam ist den Verfahren, dass das Sprachsignal bei Sender analysiert wird, dann nur die Parameter, die seine Entstehung beschreiben, übertragen werden, und daraus dann beim Empfänger ein neues Sprachsignal synthetisiert wird. Übertragen wird also nur eine parametrische Repräsentation des Sprachsignals, und nicht das Sprachsignal (oder ein äquivalentes Signal) selbst. Diese besteht im Prinzip aus

- seiner Amplitude A
- der Anregungsart „stimmhaft“ oder „stimmlos“
- bei stimmhafter Anregung zusätzlich aus der Grundfrequenz der Anregung

Der Syntheseteil eines solchen Vcoders ist seinem Prinzip nach unten dargestellt. Im Folgenden sollen exemplarisch drei Varianten der parametrischen Kodierung vorgestellt werden.



Eine erste direkte Art der Beschreibung des Vokaltraktfilters ist seine Einhüllende im Spektralbereich. Dieses Prinzip wird im sogenannten *Kanalvocoder*, der ältesten Vocoderform angewandt. Das Synthesefilter besteht dabei aus einer parallelen Anordnung von Bandpassen, die mit zeitlich variablen Faktoren skaliert und alle mit dem gleichen Anregungssignal angeregt werden. Die Skalierungsfaktoren werden sendeseitig durch Messung der Energie der Hüllkurve mit Hilfe der gleichen Bandpässe bestimmt. Da sich die Energie der Hüllkurve nur langsam – mit der Bewegung des Vokaltraktes – ändert, benötigt man nur eine geringe Bitrate zur Übertragung dieser Informationen. Dazu kommt dann noch die Information über das Anregungssignal, d.h. eine Stimmhaft-Stimmlos-Entscheidung, sowie in stimmhaften Abschnitten die Grundperiode. Das Prinzip ist in untenstehender Abbildung skizziert.



Prinzip des Kanalvocoders. Oben: Analyseteil; unten: Syntheseteil.

Eine zweite Art der Analyse-Synthese-Systeme wird noch expliziter bei der Behandlung der Sprachsynthese in Kapitel 7 vorgestellt: Der *Formant-Vocoder*. Er unterscheidet sich vom Kanalvocoder dadurch, dass das Sprachtraktfilter nicht direkt in seiner spektralen Einhüllenden dargestellt wird, sondern durch Formant-Mittenfrequenzen und zugehöriger Formant-Bandbreiten. Diese Formanten werden dann entweder als Parallelschaltung oder als Kaskade von Filtern realisiert. Die Schwierigkeit besteht hierbei darin, die Formantfrequenzen präzise und ohne Fehler zu bestimmen. Dies ist insbesondere dann problematisch, wenn Formanten eng beieinander liegen.

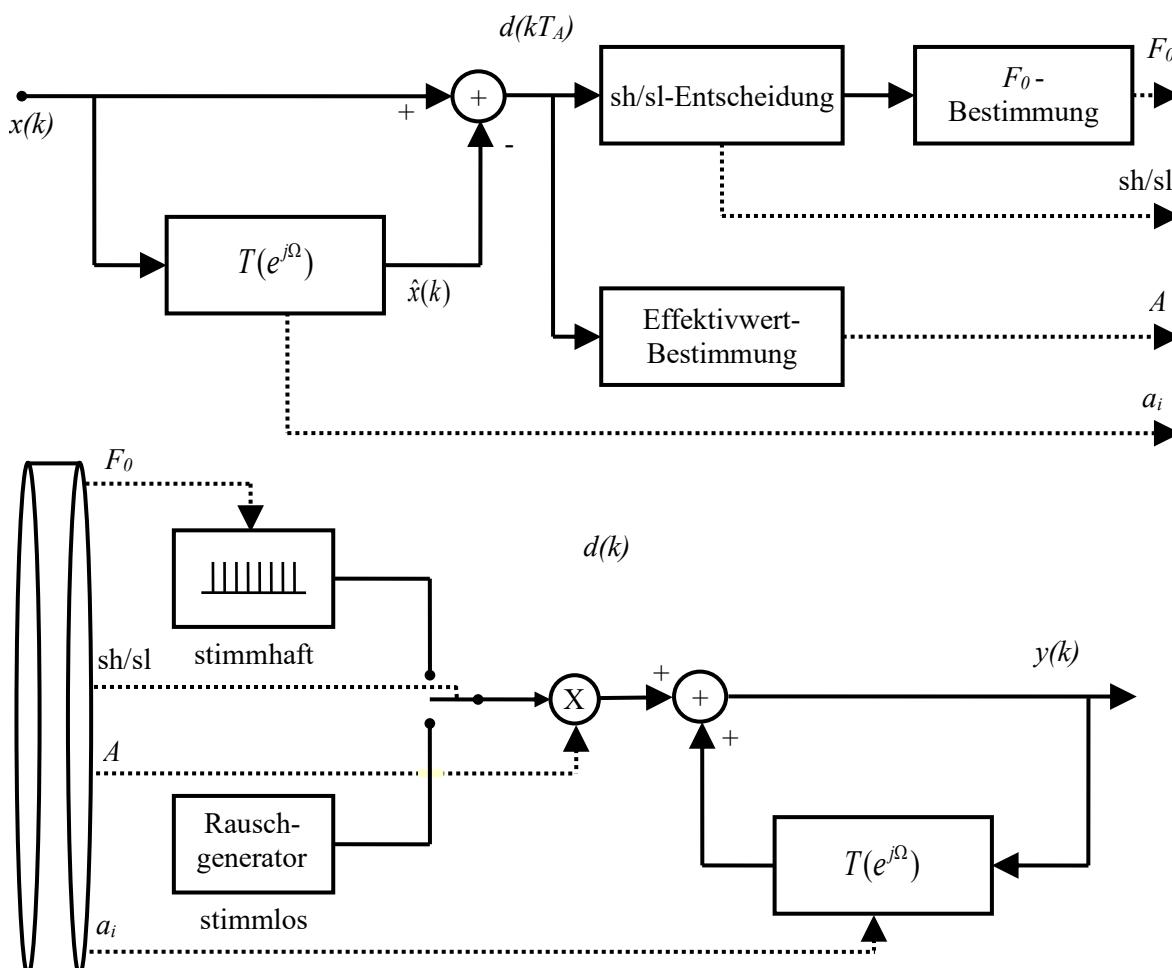
Mit dem Kanal- oder dem Formant-Vocoder lässt sich Sprache parametrisch bei sehr niedrigen Bitraten (typischerweise $f_B = 0,5 \dots 1,2 \text{ kbit/s}$) übertragen. Allerdings leidet hierbei

die Natürlichkeit sehr stark. Insbesondere die Sprecherinformation ist nur sehr eingeschränkt im synthetisierten Signal vorhanden.

Daneben kann man zur parametrischen Beschreibung natürlich auch das Prinzip der *linearen Prädiktion* verwenden, welches ja schon im vorangehenden Abschnitt zur Signalkodierung angewandt wurde. Man verwendet dabei meist ein einfaches Allpolmodell, welches einer unverzweigten Röhre ohne Nasaltrakt entspricht. Anstelle des Differenzsignals, welches bei einer idealen Prädiktion das Anregungssignal des Vokaltraktfilters nachempfinden sollte, wird dieses Signal nun künstlich generiert, und zwar

- durch seine Amplitude A
- durch die Anregungsart „stimmhaft“ oder „stimmlos“
- bei stimmhafter Anregung zusätzlich durch die Grundfrequenz der Anregung

Zusätzlich zu den Koeffizienten des Prädiktorfilters $H(e^{j\Omega})$ bzw. $T(e^{j\Omega})$, welche ja die Informationen des Vokaltraktes beinhalten, müssen also nur noch die Anregungsparameter zum Empfänger übertragen werden. Die folgende Abbildung zeigt die Struktur dieses *Prädiktionsvocoders*.



Prädiktionsvocoder. Oben: Analyseteil; unten: Syntheseteil.

Mit Hilfe des Prädiktionsvocoders lässt sich die Bitrate auf Werte von $f_B = 1,2 \dots 2,4$ kbit/s drücken. Allerdings ist dann das Sprachsignal auch nicht mehr dem ursprünglichen recht ähnlich, wodurch vor allem die Natürlichkeit und die Sprecher-Erkennbarkeit sehr stark

leiden. Trotzdem lassen sich mit solchen Verfahren noch verständliche Sprachsignale reproduzieren. Im Gegenzug ist anzumerken, dass der Aufwand, der Sender- und empfängerseitig betrieben werden muss, recht stark ansteigt. So muss bereits bei der adaptiven Analyse durch LPC für jeden Zeitrahmen ein Prädiktorfilter berechnet werden, hinzu kommt beim Vocoder die Bestimmung der Anregungsparameter.

Die mit einem Prädiktionsvocoder erzielbare Sprachqualität hängt stark von der Genauigkeit ab, mit der sich die spektrale Einhüllende durch das LPC-Filter im Empfänger nachbilden lässt. Diese wird insbesondere durch die Quantisierung der Prädiktorkoeffizienten beeinflusst. Zwar könnte man alle Prädiktorkoeffizienten möglichst genau übertragen. Bei einem Prädiktorgang von z.B. 10, einer Wortlänge $w = 16$ bit und einer Blocklänge von 20 ms würde das aber schon einer Bitrate von 160 bit/20 ms entsprechen, d.h. 8 kbit/s. Hinzu kommen die Anregungsinformationen.

Wenn man den Vokaltraktfilter direkt aus den Koeffizienten a_i bestimmen möchte so ist eine genaue Darstellung notwendig, da das entstehende Filter ansonsten leicht instabil wird. Untersuchungen von Kleijn und Paliwal zeigen, dass bei Verwendung von Optimalquantisierern für die Prädiktorkoeffizienten und einer Bitrate von 3 kbit/s für etwa 25% der Rahmen kein stabiles Synthesefilter entsteht. Diese Art der Realisierung wird deshalb in der Praxis meist nicht eingesetzt.

Stattdessen kann man das Vokaltraktfilter auch unter Bezugnahme auf das Röhrenmodell durch die Reflexionskoeffizienten r_{ij} beschreiben. Diese müssen im Intervall $-1 < r_{ij} < +1$ liegen, damit das entstehende Vokaltraktfilter stabil ist. Dabei müssen nicht alle Koeffizienten mit gleicher Genauigkeit dargestellt werden; häufig quantisiert man die ersten Reflexionskoeffizienten mit mehr bits als die letzten. Je größer der Betrag des Reflexionskoeffizienten ist, desto stärker ist sein Einfluss auf die spektrale Verzerrung, die bei ungenauer Darstellung durch die Quantisierung entsteht. Deshalb werden Reflexionskoeffizienten mit größerem Betrag genauer quantisiert. Man kann zu diesem Zweck z.B. Optimalquantisierer einsetzen, oder man transformiert die Werte der Reflexionskoeffizienten zunächst mit einer arcsin- oder artanh-Funktion und quantisiert dann gleichmäßig. Letzteres Verfahren wird z.B. auch beim GSM-Vollratencodec im Mobilfunk eingesetzt; allerdings wird das Anregungssignal dabei nicht vollständig synthetisiert, wie wir im kommenden Abschnitt noch sehen werden.

6.5 Sprachkodierung bei mittleren Bitraten, Hybrid-Kodierung

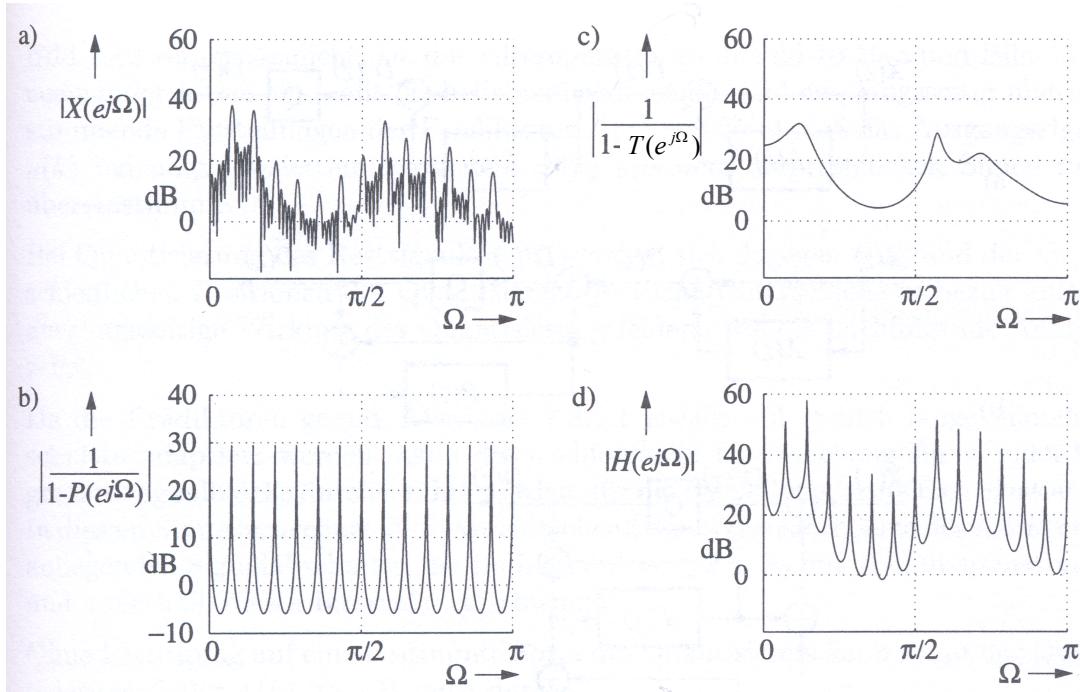
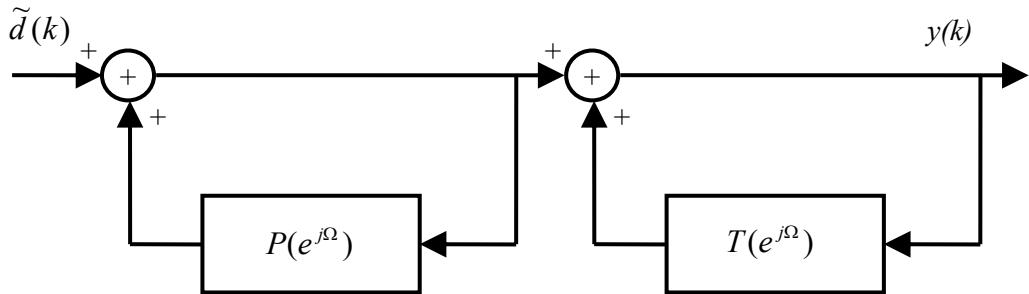
In den vorangegangenen Abschnitten wurde zwei unterschiedliche Verfahren zur Kodierung von Sprachsignalen vorgestellt: Zum einen die Signalformkodierer, die versuchen, ein dem ursprünglichen Signal möglichst ähnliches Signal zu erzeugen, zum anderen die parametrischen Kodierer, die ein Signal mit ähnlichen Eigenschaften (Anregung, Lautformung) generieren, welches aber in der Form dem ursprünglichen Signal nicht unbedingt ähnlich sein muss. Die Signalformkodierung erlaubt die Übertragung eines auf den Bereich 300-3400 Hz bandbegrenzten Signals mit 64...32 kbit/s, die parametrische Kodierung erzielt Bitraten von 0,5...2,4 kbit/s, allerdings bei deutlich verringelter Sprachqualität.

Zwischen diesen Verfahren – und den dazugehörigen Bitraten – existiert eine Lücke, die früher als sog. *coding gap* bezeichnet wurde. In der jüngeren Zeit wurde eine Reihe von

Verfahren entwickelt, die diese Lücke schließen. Kodierverfahren, die diese Lücke füllen, sind vor allem deshalb interessant, weil

- die Bitraten der Signalformkodierer zu hoch sind, um im Mobilfunk und anderen Anwendungen mit eingeschränkter Bandbreite einsetzbar zu sein
- die Qualität und vor allem die Natürlichkeit, die sich mit parametrischen Kodierverfahren erzielen lässt, für normale Anwendungen zu niedrig ist (parametrische Kodierer werden aber z.B. für militärische Anwendungen eingesetzt).

Gemeinsam ist diesen Verfahren, dass die Informationen über den Vokaltrakt (Synthesefilter) als Nebeninformationen übertragen werden, und dass das Vokaltraktfilter nicht von einem rein synthetischen, sondern von einem vereinfachten natürlichen Restsignal angeregt wird. Dieses wird durch geschickte Quantisierung so vereinfacht, dass sich eine recht niedrige Bitrate ergibt. Weiterhin benutzen die meisten der Hybrid-Kodierer neben der Kurzzeit-LPC-Analyse auch einen Langzeit-Prädiktor, welches die Periodizität des Anregungssignals bei stimmhaften Abschnitten modelliert. Unten stehende Abbildung zeigt die Struktur des Dekoders auf der Empfangsseite, sowie beispielhaft einen Signalabschnitt, welcher durch diese Struktur beschrieben wird.



Empfangsseite eines hybriden Kodierers. Oben: Struktur; unten: Signalbeispiel mit
 a) Betragsspektrum eines stimmhaften Abschnitts; b) Langzeitprädiktor;
 c) Kurzzeitprädiktor; d) Kaskade aus b) und c). Aus Vary et al. (1998, 303).

Die verschiedenen Kodierer unterscheiden sich dabei hauptsächlich in der Anordnung der Prädiktoren auf der Sendeseite sowie in der Quantisierung und sonstigen Aufbereitung des Restsignals, welches seriell oder vektoriell quantisiert werden kann.

Bei der „normalen“ (skalaren) Quantisierung des Restsignals können die sendeseitig notwendigen Prädiktoren (Kurzzeit-Prädiktor und Langzeitprädiktor) jeweils als Vorwärts- oder Rückwärts-Prädiktor ausgeführt werden. Wenn das Restsignal nicht quantisiert wird sind diese beiden Strukturen – wie in Kapitel 6.3 gezeigt wurde – äquivalent. Sofern das Restsignal jedoch quantisiert wird (was natürlich notwendig ist) so unterscheiden sich die Strukturen im Einfluss des Quantisierungsrauschens. Dabei können Vorwärts- und Rückwärtsprädiktion wiederum als Extrema eines allgemeinen Ansatzes mit *Noise Shaping* beschrieben werden. Details hierzu finden sich z.B. in Vary et al. (1998, 304-307).

Bei der vektoriellen Quantisierung des Restsignals wird im Prinzip das in Kapitel 6.2.5 vorgestellte Gain-Shape-Verfahren angewandt. Es kommt also ein Codebuch zum Einsatz, in dem normierte Repräsentantenvektoren für das Restsignal abgelegt sind. Eine Analyse solcher Restsignalvektoren zeigt, dass sie in guter Näherung sprecherunabhängig sind und einer Gauss-Verteilung (und nicht einer Gleichverteilung) folgen. Obwohl das Restsignal weitgehend unkorreliert lässt sich daher damit eine Reduktion der Bitrate erzielen. Typischerweise strebt man hier Bitraten von 0,5...1,5 bit pro Abtastwert an; allerdings lässt sich damit verständlicherweise nur ein geringer Störabstand des Restsignales erzielen.

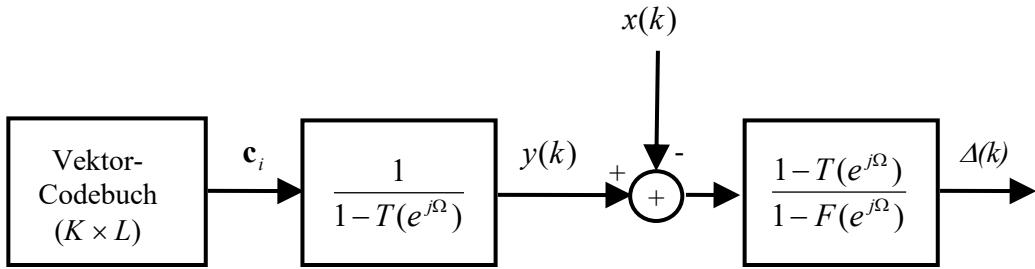
Zur Veranschaulichung des Prinzips der vektoriellen Quantisierung verzichten wir zunächst auf die Langzeitprädiktion und betrachten ein Codebuch mit K Codevektoren \mathbf{c}_i ($i = 1, 2, \dots, K$) der Länge L . Für jeden dieser Codevektoren wird versuchsweise ein Signalabschnitt $y(k)$ synthetisiert. Dazu muss der Codevektor mit der Übertragungsfunktion des Vokaltraktfilters

$$H(e^{j\Omega}) = \frac{1}{1 - T(e^{j\Omega})} \quad (6.37)$$

gefiltert werden. Zwischen dem Ausgangssignal des Vokaltraktfilters und dem Eingangssignal wird die Differenz $y(k) - x(k)$ gebildet. Dieses Differenzsignal wird nun wiederum spektral gefiltert, und zwar mit der Inversen der Übertragungsfunktion des Empfangsfilters $1 - T(e^{j\Omega})$, wie eine ausführliche Berechnung in Vary et al. (1998, 301-310) zeigt. Möchten man noch Noise-Shaping-Effekte berücksichtigen so kann man auch mit der Funktion

$$\frac{1 - T(e^{j\Omega})}{1 - F(e^{j\Omega})} \quad (6.38)$$

filtern, in der $F(e^{j\Omega})$ die Übertragungsfunktion des Noise-Shaping-Filters darstellt. Am Ausgang dieses Filters entsteht dann für jeden Codevektor die effektive (u.U. perzeptiv gewichtete) Abweichung zwischen einem Abschnitt des Eingangssignals $x(k)$ und einem durch \mathbf{c}_i generierten Abschnitt des Ausgangssignals $y(k)$. Durch Abtesten aller im Codebuch vorhandenen K Vektoren kann man nun den optimalen Repräsentantenvektor – im Sinne eines kleinsten mittleren quadratischen Fehlers – für das Restsignal herausfinden. Man bezeichnet dieses spezielle Verfahren der Kodierung mittels *Analyse-durch-Synthese* als *Code-Excited Linear Prediction*, oder *CELP*-Kodierung. Das Prinzip ist nochmals in der unten dargestellten Abbildung veranschaulicht.



Grundstruktur des CELP-Kodierers, vgl. Vary et al. (1998, 310).

Der Rechenaufwand bei diesem Verfahren ist allerdings beträchtlich. Um das Restsignal bspw. mit 0,5 bit pro Abtastwert zu kodieren (entsprechend 4 kbit/s) wird ein Codebuch mit

$$\frac{\lg(K)}{L} = 0,5 \quad (6.39)$$

benötigt. Für Abschnitte von 20 Abtastwerten (entsprechend 2,5 ms) sind das schon 1024 Vektoren; für jeden dieser Vektoren muss jeweils eine komplette Folge von Werten $y(k)$ und $\Delta(k)$ berechnet werden. Sobald die optimale Folge feststeht muss dann allerdings nur noch der Index des optimalen Codevektors übertragen werden.

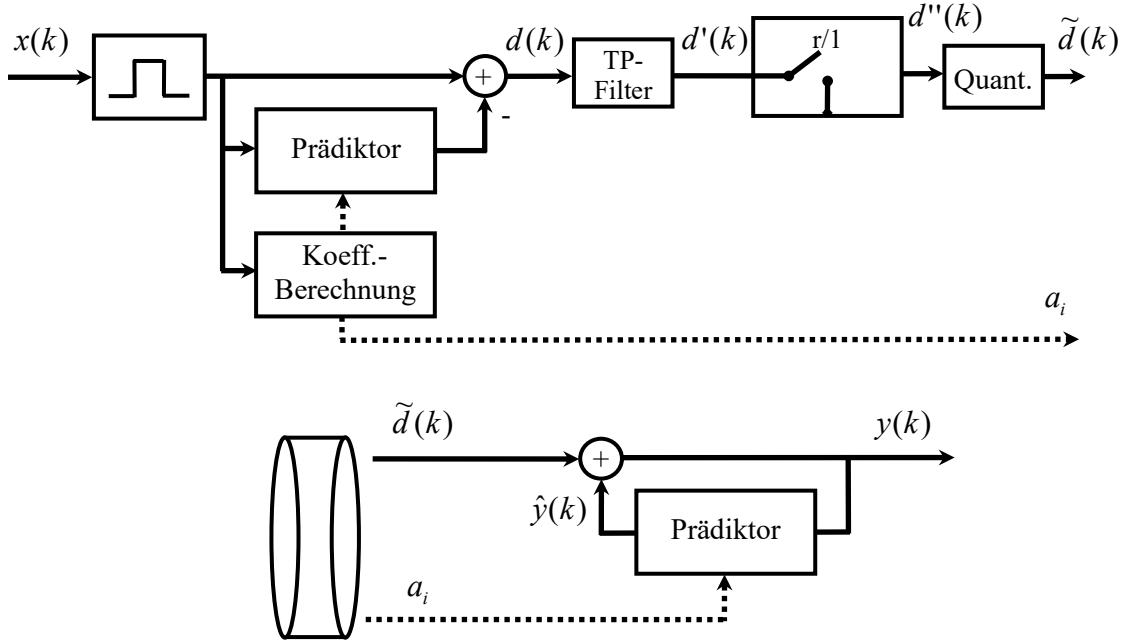
Sowohl die skalare als auch die Vektorquantisierung finden vielfache Anwendung in modernen Kodierern, wie sie im Mobilfunk oder bei der Übertragung über IP-basierte Netze benutzt werden. Im Folgenden sollen exemplarische einige Beispiele für Kodierer aus beiden Gruppen aufgezeigt werden. Dabei werden nicht immer alle Details erläutert, da dies meist zu weit in die Kodierungstheorie führen würde. Für ein allgemeines Verständnis reicht die hier angefügte Betrachtung jedoch aus.

6.5.1 RELP-Codierung

Das erste hier vorgestellte Konzept verwendet die skalare Quantisierung und schließt sich im Prinzip nahtlos an die adaptive Differenz-PCM mit Vorwärtsprädiktion und adaptiver Quantisierung (ADPCM) an. Allerdings wird nur eine Minimal-Version des Restsignals übertragen. Vergleicht man das Prinzip mit der rein parametrischen LPC-Kodierung so wird schnell klar, dass das Restsignal vor allem

- einen korrekten zeitlichen Energieverlauf (Verstärkungsfaktor A),
- eine richtige Periodizität in stimmhaften Signalabschnitten (Grundfrequenz F_0) sowie
- einen rauschhaften Charakter in stimmlosen Abschnitten

aufweisen sollte. Dies lässt sich auch mit einer deutlich bandbegrenzten Variante des Restsignals erzielen. Man muss also nicht das komplette Restsignal übertragen, sondern nur ein geschickt ausgewähltes Basisband. Man bezeichnet dieses Verfahren deshalb auch als *Basisband-RELP* mit RELP für *Residual Excited Linear Prediction*. In nachfolgender Abbildung ist das Grundprinzip unter Vernachlässigung des Langzeit-Prädiktors und evtl. für die Übertragung der einzelnen Signale und Parameter auf dem Kanal notwendige Multiplexer dargestellt.



Prinzipschema des Basisband-RELP-Kodierers, nach Vary et al. (1998, 312).
Oben: Analyseteil; unten: Syntheseteil.

Die Prädiktionskoeffizienten werden zunächst blockweise berechnet, wobei häufig Blocklängen von z.B. 160 Abtastwerten entsprechend 20 ms verwendet werden. Dadurch hat das Kodierverfahren eine Grundverzögerung von mindestens einer Blocklänge, realistischerweise jedoch das 1,25...2-fache der Blocklänge.

Wesentliches Merkmal des Basisband-RELP-Kodierers ist die Tiefpassfilterung und anschließende Taktreduktion um den Faktor r . Der Tiefpass besitzt eine Grenzfrequenz von

$$\Omega_g = \frac{\pi}{r} \quad (6.40)$$

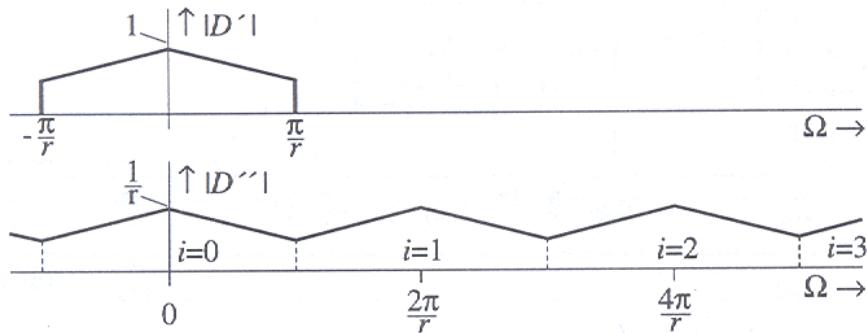
Damit kann man die Abtastrate am Ausgang dieses Zweiges um r reduzieren; bei einer Grenzfrequenz von ursprünglich 4 kHz im Telefonband wird also z.B. bei $r = 3$ effektiv nur noch ein Basisband bis zu einer Grenzfrequenz von $4/3$ kHz übertragen. Inklusive der nachfolgenden Quantisierung bewirkt dieser Zweig also eine zeitliche und wertmäßige Quantisierung.

Das nach der Abtastratenreduktion entstehende Restsignal lässt sich wie folgt beschreiben:

$$d''(k) = \begin{cases} d'(k) & \text{für } k = \lambda \cdot r \\ 0 & \text{für } k \neq \lambda \cdot r \end{cases} \quad \lambda \in \mathbb{N} \quad (6.41)$$

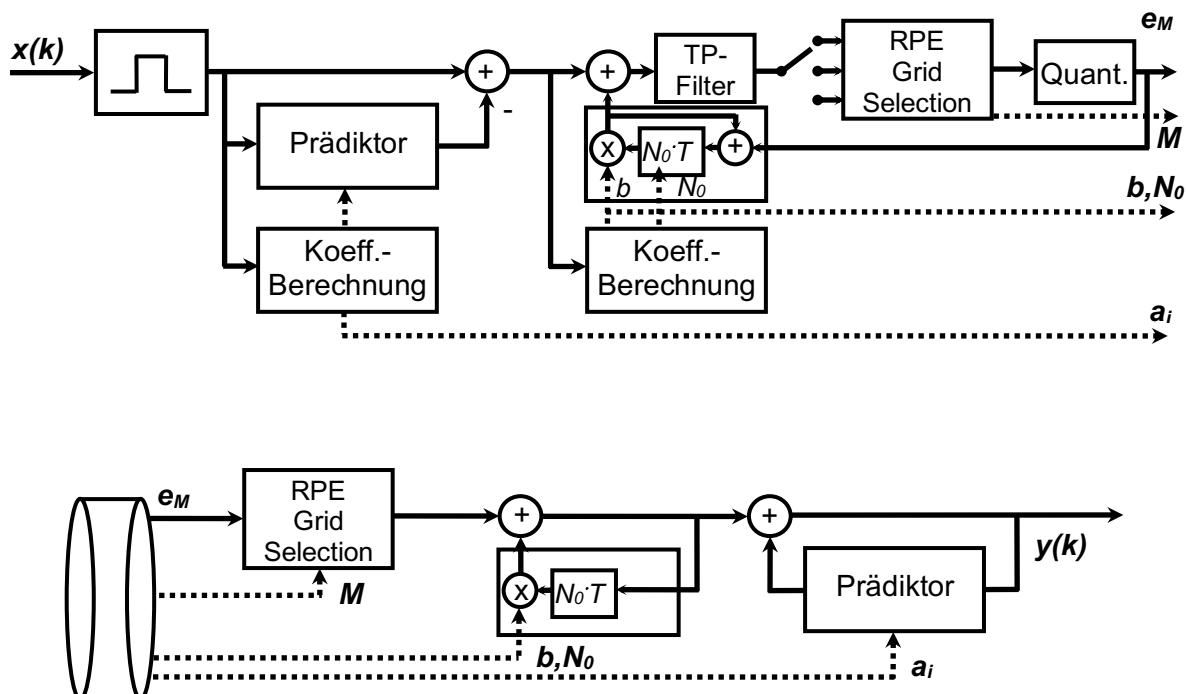
Die infolge der Unterabtastung entstehenden Nullwerte brauchen natürlich nicht mit übertragen zu werden und dienen so der Bitratenreduktion. Das Spektrum dieser Signale ist in unten stehender Abbildung skizziert; es ist breitbandig, ist aber im Basisband identisch mit demjenigen des Original-Restsignals. Insbesondere stimmt in diesem Band die Pitchstruktur, welche für die Natürlichkeit der Stimme recht wichtig ist. Allerdings ist dies nur bis zur o.a. Grenzfrequenz gewährleistet; an den Sprungstellen setzen die gepiegelten Spektralanteile die typische äquidistante Linienstruktur normalerweise nicht fort. Dies macht sich in einer unnatürlichen Stimme – insbesondere bei hohen Frauenstimmen – bemerkbar; bei

Männerstimmen befinden sich glücklicherweise noch genug Spektrallinien im Basisband. Auch zur Übertragung von Musiksignalen oder Modemsignalen ist ein solcher Kodierer deshalb schlecht geeignet.



Spektrum des Anregungssignales beim Basisband-RELP-Kodierer,
aus Vary et al. (1998, 313)

Dieses Problem lässt sich durch Einfügung eines Langzeitprädiktors reduzieren, wie es z.B. beim *Vollratenkodierer des GSM-Netzes* geschieht. Er entspricht in seiner Grundstruktur dem Basisband-RELP-Kodierer, allerdings mit zwei wichtigen Abweichungen. Seine Struktur ist in folgender Abbildung skizziert.



Prinzipschema des GSM-Vollratenkodierers, nach Vary et al. (1998, 546).
Oben: Analyseteil; unten: Syntheseteil.

Der Kodierer arbeitet wie folgt:

- Das Signal wird zunächst gefenstert.

- Für jeden Signalabschnitt (Fenster) werden die Koeffizienten des *Kurzzeit-Prädiktors* a_i nach dem Prinzip der Vorwärtsprädiktion bestimmt. Diese Koeffizienten werden – wie bei der parametrischen Kodierung – zum Empfänger übertragen. Am Ausgang des Kurzzeit-Prädiktors ergibt sich ein mit dem inversen geschätzten Vokaltraktfilter gefiltertes Signal, das in etwa dem Anregungssignal entspricht.
- An diese Kurzzeit-Prädiktion des Vokaltraktes schließt sich eine *Langzeit-Prädiktion* des Anregungssignals in Form einer Rückwärts-Prädiktion an. Diese bestimmt die Koeffizienten b (Amplitude) und N_0 (Anzahl der Abtastwerte in einer Grundperiode; $N_0 = f_A/f_0$). Beide Koeffizienten werden zum Empfänger übertragen. Am Ausgang des Langzeit-Prädiktors ergibt sich ein Anregungssignal, welches allerdings seiner Periodizität weitgehend beraubt ist.
- Das verbleibende Restsignal wird zunächst mit einem Tiefpass gefiltert, der die übertragene Bandbreite des Restsignals auf 1/3 der normalen Bandbreite (= 4/3 kHz) begrenzt. Die Filterung wird im Sinne einer Blockfilterung ausgeführt, bei der 40 Signalwerte um 10 Nullwerte ergänzt werden und die entstehenden 50 Werte dann blockweise durch ein einfaches nichtrekursives Filter der Länge 11 geschickt werden.
- Der Unterabtastung dient der Block „RPE Grid Selection“. Die Abkürzung RPE steht dabei für *Regular Pulse Excitation*, bei der das Restsignal durch eine Anregungsfolge in einem regelmäßigen Raster dargestellt wird. Dazu werden die 40 gefilterten Signalwerte der Folge $e_{TP}(k)$ in drei gegeneinander um einen Abtastwert verschobene Teilfolgen $e_i(k)$ zerlegt:

$$e_i(k) = \begin{cases} e_{TP}(k) & \text{für } k = i + \lambda \cdot 3 \\ 0 & \text{für } k \neq i + \lambda \cdot 3 \end{cases} \quad i = 0, 1, 2; \lambda = 0, 1, 2, \dots, 12. \quad (6.42)$$

Aus den 3 (bzw. wegen der Blocklänge $L = 40$ eigentlich 4) Teilfolgen $e_i(k)$ wird die im Sinne einer minimalen Energie optimale Teilfolge ausgewählt. Zusätzlich wird noch alle 5 ms die optimale Rasterposition M neu bestimmt. Durch diese quasi zufällige Variation des festen Rasters wird der tonal-metallische Effekt des Basisband-RELP weitgehend vermieden – allerdings auf Kosten einer höheren Rauigkeit bei hohen Stimmen.

- Die entstehende Restsignalfolge von 13 Werten wird anschließend nach dem AQF-Verfahren quantisiert und ebenfalls zum Empfänger übertragen. Dazu werden die 13 Werte zunächst auf ihr Maximum normiert und mit 3 bit gleichmäßig quantisiert. Das Blockmaximum wird mit 6 bit quantisiert, hinzu kommen noch die Rasterposition (2 bit), d.h. insgesamt $(3 \cdot 13 + 6 + 2)\text{bit} / 5 \text{ ms} = 9,4 \text{ kbit/s}$. Hierzu kommen noch alle 20 ms 36 bit für die LPC-Koeffizienten des Kurzzeit-Prädiktors (entspr. 1,8 kbit/s) sowie alle 5 ms die Parameter N_0 und b des Langzeitprädiktors (entspr. wiederum 1,8 kbit/s). Insgesamt kommt dieser Kodierer also auf eine Übertragungsrate von 13 kbit/s und liegt damit fast genau in der Mitte der „Coding Gap“.
- Im Empfänger wird zunächst das Restsignal wiederhergestellt. Dazu wird die (quantisiert übertragene) Restsignalfolge mittels eines inversen RPE-Grid-Selection-Algorithmus in das Restsignal überführt.
- Auf dieses Restsignal wird nun Anregungsinformation mithilfe eines inversen Langzeit-Prädiktors wieder aufgeprägt. Das so entstehende Anregungssignal trägt nun wieder die Information der Grundfrequenz.

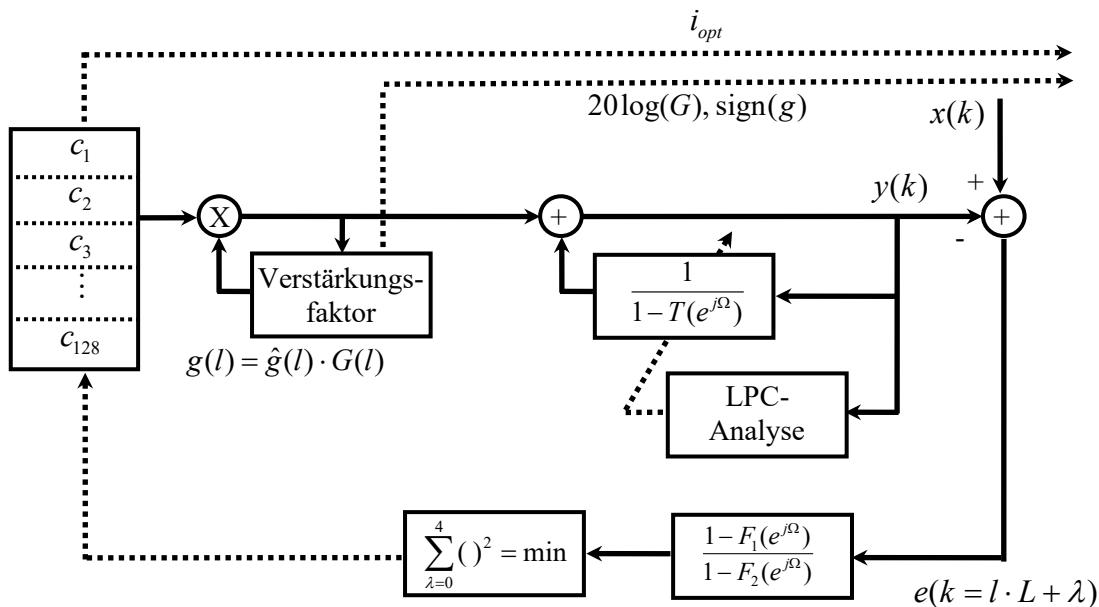
- Auf dieses Signal wird nun wiederum die Information des Vokaltraktes (inverser Kurzzeit-Prädiktor) aufgeprägt; es entsteht das übertragene Sprachsignal.

Dieses als RPE-LTP (Regular Pulse Excitation Long Term Prediction) bekannte Kodierungsprinzip wird (neben anderen) im GSM-Netz eingesetzt. Es zeigt, dass sich durch Kombination verschiedener Ideen die zur Übertragung eines 4 kHz breiten Sprachsignals erforderliche Bitrate von 64 kbit/s (logarithmische PCM, ISDN) auf etwa 1/5 (13 kbit/s) drücken lässt. Dies geht allerdings mit einer leichten Verschlechterung der Sprachqualität einher. Außerdem ist der Kodierer nicht für Modem- oder Musiksignale geeignet. Methoden zur Bestimmung der Sprachqualität derart kodierter Sprache werden in Kapitel 6.7 kurz umrissen, und in der Vorlesung „Usability Engineering“ ausführlich dargestellt.

6.5.2 CELP-Codierung

Das Prinzip der Vektor-orientierten CELP-Kodierung wurde bereits beschrieben. Das Problematische an diesem Prinzip ist insbesondere der hohe Rechenaufwand, da für jeden der K Codebuchvektoren \mathbf{c}_i der Dimension L jeweils L Werte des Signals $y_i(k)$ bestimmt werden müssen, aus denen dann der optimale Codebuchvektor ausgewählt wird. In der Literatur sind verschiedene Verfahren zur Aufwandsreduktion beschrieben, die z.B. bei Vary et al. (1998, 319-327) zusammengefasst sind. Dort wird auch gezeigt, dass sich ein Langzeit-Prädiktor als zweites – adaptives – Codebuch im CELP-Format darstellen lässt.

Hier soll im Folgenden nur ein Beispiel eines CELP-Kodierers vorgestellt werden, welcher z.B. bei „Leitungsvervielfachern“ eingesetzt wird. Der Kodierer ist von der ITU-T in ihrer Empfehlung G.728 standardisiert und in u.a. Abbildung skizziert.



Prinzipschema des Sendeteils eines LD-CELP-Kodierers nach ITU-T Rec. G.728
(aus Vary et al., 1998, 539).

Im Prinzip handelt es sich um einen CELP-Kodierer mit einem Codebuch mit 128 Einträgen. Allerdings wird die Blocklänge mit $L = 5$ Abtastwerten (entspricht 0,625 ms) sehr klein gewählt, um nur eine geringe Verzögerung zu erhalten. Dadurch kann die Gesamt-Signalverzögerung auf unter 2 ms gedrückt werden. Der Codebuchvektor wird zunächst mit einem adaptiven Verstärkungsfaktor $g(l)$ gewichtet. Dieser Verstärkungsfaktor wird

logarithmisch-differentiell quantisiert; d.h. er wird zunächst logarithmiert und in zwei Anteile zerlegt gemäß

$$20\lg|g(l)| = 20\lg|\hat{g}(l)| + 20\lg|G(l)| = \hat{v}(l) + \Delta v(l) \quad (6.43)$$

wobei dann $\Delta v(l) = v(l) - \hat{v}(l)$ mit 2 bit quantisiert wird, und $\hat{v}(l)$ mit Hilfe eines Prädiktors der Ordnung 10 aus vorangegangenen optimalen Codevektoren und Verstärkungsfaktoren berechnet wird. Der so gewichtete Codevektor wird durch einen rückwärts-adaptiven Kurzzeit-Prädiktor (Ordnung 50) geschickt; eine Langzeit-Prädiktion gibt es nicht. Für alle so behandelten Codebuchvektoren wird anschließend die Differenz zum aktuellen Signalrahmen gebildet, und der Fehler mit einem Filter der Ordnung 10 gewichtet. Aus der so gewichteten Differenz wird nun der Codebuchvektor mit der minimalen Fehlerenergie ausgewählt, und der entsprechende Index wird zum Empfänger übertragen. Der Empfänger ist im Prinzip im Sender enthalten, denn der mittlerer Signalfad berechnet ja – sofern er mit $\mathbf{c}_{i,opt}$ gefüttert wird – einen Schätzwert für den aktuellen Signalrahmen.

Insgesamt benötigt dieser Kodierer alle 5 Abtastwerte 7 bit zur Übertragung des optimalen Codebuchindex sowie (2+1) bit (2 bit für $\Delta v(l)$ und 1 bit für das Vorzeichen), d.h. 10 bit/0,625 ms = 16 kbit/s. Positiv ist insbesondere die sehr geringe Grundverzögerung unter 2 ms. Bezuglich der erzielbaren Sprachqualität ist er in etwa äquivalent zu einer ADPCM mit 32 kbit/s.

6.6 Sprachkodierung im Frequenzbereich

Bislang haben wir uns nur mit Kodierverfahren beschäftigt, die im Zeitbereich arbeiten. Daneben gibt es aber auch eine Klasse von Verfahren, die vor der Quantisierung und Übertragung eine Spektralanalyse durchführen. Nach der Übertragung wird das Signal mittels einer Spektralsynthese wieder zusammengesetzt. Dazwischen bietet es sich an, die spektral (d.h. in Frequenzbändern) vorliegenden Daten gezielt zu manipulieren, um somit Bits einzusparen. Dahinter steht zum einen der Gedanke, dass die Spektralanalyse im Prinzip der Verarbeitung im Innenohr entspricht, zum anderen, dass sich gezielt einzelne Bänder manipulieren lassen und somit psychoakustische Effekte (z.B. die Verdeckung) gezielt ausgenutzt werden können.

Dass sich überhaupt Bitrate einsparen lässt lag bei der linearen Prädiktion daran, dass Sprache korreliert ist. Ähnlich lässt sich mit einer Sprachkodierung im Frequenzbereich Bitrate sparen, da das Leistungsdichtespektrum nicht konstant ist, d.h. das Sprachsignal hat kein weisses Spektrum. Insbesondere sind Sprachanteile bei höheren Frequenzen, bei den Antiformanten sowie zwischen den Peaks bei Vielfachen der Grundfrequenz im Mittel weniger energiereich als z.B. bei den Formanten. Es wird dabei schnell klar, dass der Gewinn insbesondere dann deutlich ausfällt, wenn man die Kurzzeitspektren betrachtet.

Die Spektralanalyse mit Hilfe einer Filterbank, wie sie in Kapitel 4.1 beschrieben ist, stellt im Prinzip eine Analysefilterbank dar, wie sie Teil eines Frequenzbereichskodierers sein könnte. Voraussetzung dabei ist, dass man die Bandsignale unterabtastet; ansonsten würde sich die Bitrate ja vervielfachen, anstatt dass sie reduziert werden könnte. Man bezeichnet solche Verfahren allgemein als Frequenzbereichskodierung (*Frequency Band Coding*, FBC) und teilt sie in Ansätze der Teilbandcodierung (*Sub-Band Coding*, SBC) und der *Transformationskodierung* (TC) ein. Daneben gibt es noch weitere Verfahren, die ein Sprachsignal aus einer Reihe von Sinusgeneratoren mit zeitlich variablen Amplituden,

Frequenzen und Phasen durch Überlagerung erzeugen (sog. Sinusmodellierung oder Harmonische Kodierung). Diese sollen jedoch hier nicht betrachtet werden.

6.6.1 Transformationskodierung

Bei der Transformationskodierung (TC) wird ein Signalabschnitt von M Abtastwerten zunächst mittels einer Fensterung aus dem Datenstrom entnommen, als Datenvektor

$$\mathbf{x}_k = (x_k(0), x_k(1), \dots, x_k(M-1))^T \quad (6.44)$$

aufgefasst, und mittels einer noch zu spezifizierenden linearen Operation transformiert in einen Spektralvektor

$$\mathbf{X}(k) = \mathbf{A} \mathbf{x}_k = (X_k(0), X_k(1), \dots, X_k(M-1))^T \quad (6.45)$$

Nach der Quantisierung und Übertragung wird $\mathbf{X}(k)$ durch einen Vektor

$$\hat{\mathbf{X}}(k) = (\hat{X}_k(0), \hat{X}_k(1), \dots, \hat{X}_k(M-1))^T \quad (6.46)$$

mit im allgemeinen „verfälschten“ Spektralwerten ersetzt. Aus diesen lässt sich eine Schätzung des Signalblocks

$$\hat{\mathbf{x}}_k = \mathbf{A}^{-1} \hat{\mathbf{X}}(k) = (\hat{x}_k(0), \hat{x}_k(1), \dots, \hat{x}_k(M-1))^T \quad (6.47)$$

durch inverse Transformation gewinnen. Es ist zu beachten, dass sich das Signal dabei immer um die Blocklänge M verzögert.

Eine genauere Betrachtung des dabei gemachten Fehlers zeigt, dass die Energie des Fehlers im Zeitbereich gleich der mittleren Störleistung im Spektralbereich ist:

$$N_x = E \left\{ \frac{1}{M} \sum_{\kappa=0}^{M-1} |\hat{x}(k-\kappa) - x(k-\kappa)|^2 \right\} = E \left\{ \frac{1}{M} [\hat{\mathbf{X}}(k) - \mathbf{X}(k)]^H [\hat{\mathbf{X}}(k) - \mathbf{X}(k)] \right\} = N_T \quad (6.48)$$

Dies gilt für alle Transformationen, die unitär sind, d.h. bei denen $\mathbf{A}^{-1} = \mathbf{A}^H$, wobei das hochgestellte H eine Hermite'sche, also eine konjugiert-komplex und transponierte Matrix bezeichnet. Eine Überprüfung zeigt, dass die Diskrete Fourier-Transformation (DFT) bis auf einen konstanten Faktor M unitär ist; sie wäre unitär, wenn man den Vorfaktor $1/M$ auf Hin- und Rücktransformation gleichmäßig aufteilen würde. Daneben erweist sich auch eine Diskrete Cosinus-Transformation in einer modifizierten Form als unitär, ebenfalls eine Karhunen-Loëve-Transformation (KLT). Details hierzu findet man bei Vary et al. (1998, 342-343).

Durch die Transformationskodierung lässt sich Bitrate einsparen, da die Beträge der einzelnen Spektralwerte $X_\mu(k)$ im Allgemeinen unterschiedlich groß sind. Demnach können bei konstanter Quantisierungsstufe bei geringerer Aussteuerung führende Null-Bits weggelassen werden, und zwar bei jedem Spektralwert $X_\mu(k)$ unterschiedlich viele. D.h. es muss für jeden Spektralwert eine optimale Wortlänge w_μ gefunden werden.

Wenn man davon ausgeht, dass für die Frequenzbereichskodierung von M Spektralwerten genauso viele Bits zur Verfügung stehen wie für M Zeitwerten, wenn also

$$\sum_{\mu=0}^{M-1} w_\mu = M \cdot w_x = M \cdot w_T \quad (6.49)$$

mit w_T dem Mittelwert der Wortlängen w_μ ist, so kann man die optimale Verteilung der Bits auf die Spektralwerte durch Minimierung der Rauschleistung bzw. Maximierung des Störabstandes bestimmen. Es zeigt sich, dass

$$w_\mu = w_T + ld \frac{\sigma_\mu}{\sqrt[M]{\prod_{\lambda=0}^{M-1} \sigma_\lambda}} \quad (6.50)$$

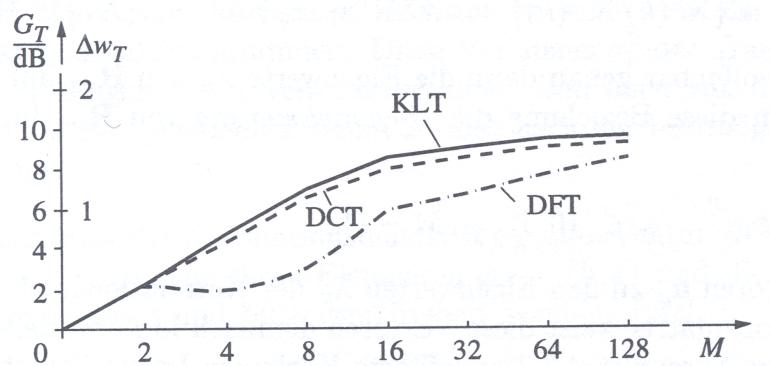
d.h. dass – ausgehend von der mittleren Wortlänge – einem Spektralwert $X_\mu(k)$ mehr oder weniger Bits zugeteilt werden sollten, je nachdem, ob sein Aussteuerungsbereich größer oder kleiner als das geometrische Mittel ist. Bspw. führt eine Verdoppelung von σ_μ zur Zuteilung einer weiteren Binärstelle. Das Quantisierungsrauschen, was sich dabei ergibt, ist für alle Frequenzbänder gleich, sofern man konstante Quantisierungsstufen wählt. Auch kann man zeigen, dass sich bei der Transformationskodierung ein weisses Quantisierungsrauschen ergibt, wie auch bei der direkten Signalquantisierung mit konstanter Stufenhöhe. Allerdings hängt die Gesamt-Rauschleistung von der Spektralform des Signals ab.

Betrachtet man das Verhältnis der Rauschleistungen zwischen direkter Signalquantisierung und Transformationskodierung, so ergibt sich der sogenannte Transformationsgewinn

$$\frac{N_Q}{N_T} = \frac{\frac{1}{M} \sum_{\mu=0}^{M-1} \sigma_\mu^2}{\sqrt[M]{\prod_{\mu=0}^{M-1} \sigma_\mu^2}} \cdot 2^{2(w_T - w_x)} \quad (6.51)$$

also das Verhältnis zwischen arithmetischem und geometrischem Mittelwert der spektralen Varianzen. Dieser Wert ist gleich 1, sofern $w_T = w_x$ und $\sigma_\mu^2 \equiv \sigma_x^2$, ansonsten ist er größer als 1. Der Störabstand wird also bei nicht-weissen Signalen (wie Sprache) besser. Alternativ kann man den Störabstand konstant halten und mit kleineren Wortlängen, also geringeren Bitraten, arbeiten.

Nach Gleichung (6.51) kann man den Transformationsgewinn optimieren, indem man das geometrische Mittel der Spektralvarianzen minimiert. Dies lässt sich durch eine geeignet gewählte Transformation unterstützen. Als „optimal“ erweist sich hier eine Karhunen-Loëve-Transformation, d.h. eine Transformation mit Koeffizienten, die vom Signal abhängen. Leider ist eine solche Transformation nicht praktikabel: Zu Ihrer Berechnung müsste man entweder „die allgemeine Korrelationsmatrix von Sprache“ kennen, oder aber eine Sprachen- oder Sprecher-spezifische Transformation benutzen, was in der Praxis nicht funktioniert. Nicht ganz optimal, aber immer noch recht gut funktioniert auch eine Diskrete Cosinus-Transformation. In unten stehender Abbildung sind die Transformationsgewinne dieser Transformationen aufgezeigt.



Transformationsgewinne verschiedener Transformationen in Abhängigkeit von der Blocklänge M , bestimmt mit Hilfe eines Modellsignals
(aus Vary et al., 1998, 350).

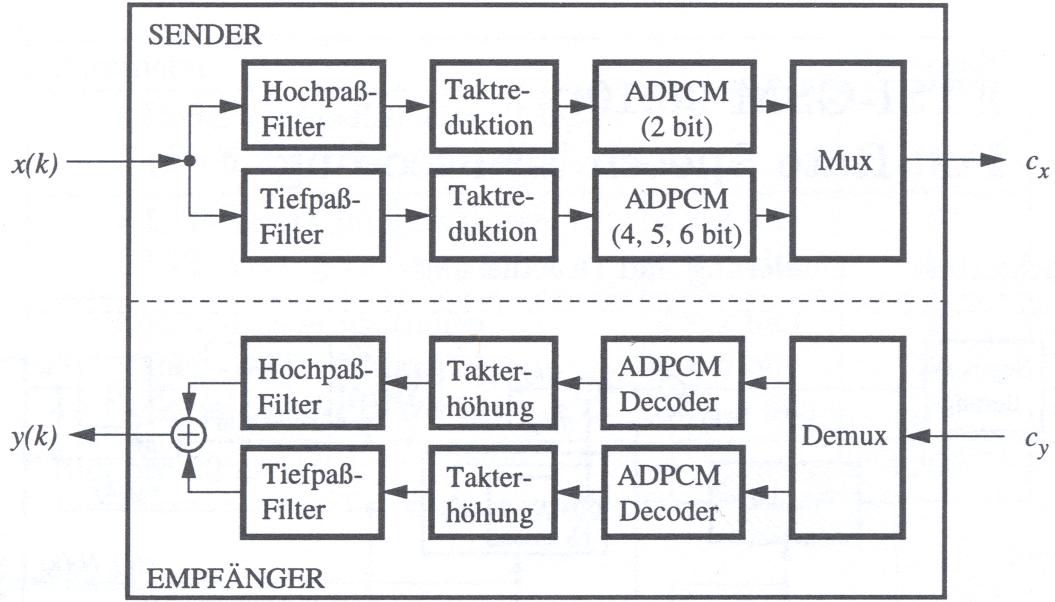
Wie auch bei der Signalformkodierung lassen sich größere Transformationsgewinne erst dann erzielen, wenn man die Wortlängen der Spektralwerte $X_\mu(k)$ adaptiv quantisiert, d.h. mit einer Wortlänge $w_\mu(k)$. Dieses Verfahren bezeichnet man als *Adaptive Transform Coding, ATC*. Allerdings muss dann die jeweilige Wortlänge ebenfalls mit übertragen werden. Dies kann im ungünstigsten Fall zu einer Erhöhung statt Erniedrigung der Bitrate führen. Das Problem kann man reduzieren, wenn man Gruppen von benachbarten Spektralkoeffizienten mit gleicher Wortlänge quantisiert; dies reduziert die Anzahl der zur Rekonstruktion benötigten Hilfswerte. Alternativ kann man die Einhüllende des Spektrums auch über LPC-Koeffizienten oder ein Cepstrum beschreiben und somit die Aussteuerung festlegen.

Obwohl man mit adaptiver Transformationskodierung eine Erhöhung des Störabstandes erreichen kann, werden solcherart kodierte und dekodierte Signale als dumpfer als z.B. vergleichbare RELP-Algorithmen im Bereich unter 16 kbit/s empfunden. Sie haben sich deshalb im Mobilfunkbereich nicht durchgesetzt. Bei höheren Bitraten zeigen sie aber einige Vorteile, die jedoch weniger für Sprache, als vielmehr für Audio-Signale (bei dann auch höherer Abtastfrequenz) interessant sind.

6.6.2 Teilbandkodierung

Teilbandkodierungen bestehen aus einer Analyse-Synthese-Filterbank und verwenden für die Quantisierung der verschiedenen Bänder unterschiedliche Verfahren bzw. Aussteuerungen. Sie folgen also dem Prinzip der im vorangegangenen Abschnitt beschriebene Transformationskodierung; unterschiedlich ist hier vor allem die viel geringere Anzahl der Bänder (typischerweise 4...8 bei SBC statt 128 bei der TC). Zudem sind die Bandbreiten nicht konstant, sondern wachsen mit steigender Frequenz an; damit lassen sich auf stark vereinfachte Weise die Frequenzgruppen im Innenohr nachbilden. Durch die unterschiedlich breiten Spektralbänder können auch unterschiedliche Taktreduktionen in den Teilbändern verwendet werden.

Wie die TC kann auch die SBC adaptiv ausgeführt werden. Durch die geringere Kanalzahl ist hier auch die Menge der „Hilfsinformationen“ geringer. Auch können die Teilbänder als Differenzen, z.B. in Form einer ADPCM oder APC, kodiert werden. Dies macht hier Sinn, da durch die geringe Kanalzahl und die größere Breite der Bänder auch innerhalb eines Bandes ein Prädiktionsgewinn zu erzielen ist.



Teilbandkodierer nach ITU-T Rec. G.722, aus Vary et al. (1998, 545).

Ein einfaches – aber derzeit sehr aktuelles – Beispiel für eine SBC stellt der von der ITU-T in Empfehlung G.722 standardisierte 2-Band-Kodierer dar. Er ist in unten stehender Abbildung skizziert. Der Kodierer wird zur Übertragung breitbandiger Sprache (50-7000 Hz) bei einer Abtastfrequenz von 16 kHz eingesetzt, kann jedoch auch andere Audiosignale vernünftig übertragen. Er arbeitet bei Bitraten von 48-64 kbit/s. Die Aufteilung der beiden Bänder erfolgt mit einem Quadrature-Mirror-Filter- (QMF-) Paar. Beide Teilbänder werden mittels ADPCM kodiert; das untere Teilband mit einer Wortlänge von 4...6 (entsprechend 32-48 kbit/s), das obere im einer Wortlänge von 2 (entsprechend 16 kbit/s). Die Signalverzögerung liegt bei nur 1,5 ms.

6.7 Kriterien zur Auswahl und Beurteilung von Kodierern

Bei den Beschreibungen der einzelnen Kodierverfahren hat sich bereits gezeigt, dass der Störabstand, also das Verhältnis von Nutzsignalleistung zur Leistung des Quantisierungsräuschens, als Kriterium für die Auswahl eines Kodierers nicht ausreicht. Insbesondere ist der Störabstand nicht unbedingt korreliert mit der wahrgenommenen Sprachqualität; dies ist vor allem bei komplexeren Kodierverfahren zu beobachten. Daher muss man zur validen Bestimmung der Sprachqualität zumeist auditive Tests durchführen. Alternativ wird aber derzeit auch stark an instrumentellen Verfahren geforscht, die einen Qualitätsindex (zumeist für die sogenannte „Gesamtqualität“, teilweise aber auch für einzelne perzeptive Dimensionen wie „Klang“ oder „Kontinuierlichkeit“) schätzen können. Es handelt sich dabei genau genommen um Vorhersagemodelle, die Beobachtungen aus auditiven Tests nachvollziehen zu versuchen; „besser“ als auditive Tests sind solche instrumentellen Verfahren daher nicht. Beispiele für auditive und instrumentelle Testverfahren werden in der Vorlesung „Usability Engineering“ behandelt.

Neben der Sprachqualität gibt es aber weitere Anforderungen an Sprachkodierer, die sie für einzelne Anwendungen geeignet oder ungeeignet machen. So ist für ein interaktives System (wie bspw. das Telefon oder VoIP) die Signalverzögerung von entscheidender Bedeutung. Obwohl von der ITU-T Einweg-Verzögerungen unter 150 ms als unkritisch erachtet werden,

kann doch bei interaktiven Szenarien auch unterhalb dieses Wertes schon eine Änderung von Konversationsverhalten beobachtet werden. Auch ist zu berücksichtigen, dass die Verzögerung des Sprachkodierers nur eine Komponente der Gesamtverzögerung darstellt. Wird ein kodiertes Signal bspw. in ein IP-Paket verpackt und asynchron auf den Übertragungsweg geschickt, so ist empfangsseitig ein Jitter-Puffer vorzusehen, der ebenfalls Verzögerungen mit sich bringt.

Neben der Sprachqualität und der Verzögerung sind natürlich die zur Verfügung stehende Kanalkapazität und die sich daraus ergebende Bitrate von zentraler Bedeutung. So bestimmt die Bitrate zunächst, welche Kodierverfahren überhaupt in Frage kommen. Allerdings kann die Bitrate – insbesondere bei paketvermittelten Netzen – auch variabel gehalten werden. So kann ein optimaler Kompromiss zwischen Sprachqualität und Bitrate – und damit auch Kanalzahl – adaptiv gewählt werden, z.B. in Funktion der Auslastung des Kanals. Die Komplexität des Algorithmus und die Signalverzögerung stellen in diesem Falle eher Grenzwerte dar, die für bestimmte Anwendungen nicht überschritten werden sollen.

6.8 Literatur

- Blauert, J. (1994). Kommunikationsakustik II: Audiokommunikation und virtuelle Realität. Skriptum zur Vorlesung am Institut für Kommunikationsakustik, Ruhr-Universität, Bochum.
- Heute, U. (1990). Sprachverarbeitung. Skriptum zur Vorlesung der Arbeitsgruppe Digitale Signalverarbeitung, Ruhr-Universität, Bochum.
- Möller, S. (2000). Assessment and Prediction of Speech Quality in Telecommunications. Kluwer Academic Publishers, Boston.
- Raake, A. (2006). Speech Quality of Voice over IP. Assessment and Prediction. John Wiley & Sons, Chichester.
- Vary, P., Heute, U., Hess, W. (1998). Quantisierung und Codierung. In: Digitale Sprachsignalverarbeitung, B.G. Teubner, Stuttgart, 233-270.
- Vary, P., Heute, U., Hess, W. (1998). Codierung im Zeitbereich. In: Digitale Sprachsignalverarbeitung, B.G. Teubner, Stuttgart, 271-336.
- Vary, P., Heute, U., Hess, W. (1998). Codierung im Frequenzbereich. In: Digitale Sprachsignalverarbeitung, B.G. Teubner, Stuttgart, 337-375.

7. Sprachtechnologische Systeme

In den vorangegangenen Kapiteln haben wir uns mit der Erzeugung und Wahrnehmung von Sprache durch den Menschen beschäftigt, sowie mit der Frage, wie Sprachsignale beschrieben und analysiert werden können. Wir hatten jedoch eingangs klargestellt, dass sich Sprache nicht nur als Kommunikationsmedium zwischen Menschen eignet, sondern auch zur Interaktion zwischen Mensch und Maschine eingesetzt werden kann. Dazu müssen Algorithmen entwickelt werden, mit deren Hilfe Sprachsignale erkannt und auf eine symbolische Ebene (z.B. Rechtschrift) umgesetzt werden können, und weitere Algorithmen, die aus einer symbolischen Darstellung sprachähnliche Signale generieren können. Diesen Algorithmen widmen sich die Abschnitte 7.1 (*Spracherkennung*) und 7.2 (*Sprachsynthese*).

Zur vollständigen Interaktion zwischen Mensch und Maschine bedarf es jedoch noch weiterer Module, die den Dialogverlauf zwischen Benutzer und System steuern. Diese bewirken z.B. eine Interpretation der erkannten Symbole (Sprachverständnis), die Ableitung einer angemessenen Reaktion des Systems (Dialogsteuerung), die Kommunikation mit externen Modulen (z.B. Datenbanken) sowie die Formulierung einer informativen Antwort für den Benutzer (Antwortgenerierung). Diese (und unter Umständen weitere) Module sind in einem *Sprachdialogsystem* vereint, welches in seinem prinzipiellen Aufbau in Abschnitt 7.3 besprochen werden soll.

7.1 Spracherkennung

Es ist das Ziel der Spracherkennung, Sprachsignale in eine Folge von linguistischen Komponenten zu dekodieren. Es findet also eine *Umsetzung von der Signal- auf die Symbolebene* statt. Die dazu verwendete Methode hängt von einer Vielzahl von Randbedingungen ab:

- Sprache:* Sprache, Standardsprache oder Dialekte, Sprechstil, etc.
Sprecher: Sprecherabhängig oder -unabhängig arbeitende Systeme, Sprecher bekannt oder unbekannt, Sprecher kooperativ oder gleichgültig
Zieleinheiten: Einzelwörter, verbundene Wörter, vorgelesener Fließtext, spontane Sprache, Schlüsselwörter in spontaner Sprache, etc.

Anzahl der

Zieleinheiten: 10 (Ziffern) bis > 500.000

Komplexität: Komplexität des Vokabulars (Perplexität)

Umgebung: Akustische Eigenschaften der Sprechumgebung (Störgeräusche, Echos) und evtl. beteiligter Übertragungskanäle (Telefon, Internet)

Auf Grundlage dieser Kriterien kann man unterscheiden zwischen sprecherabhängiger und sprecherunabhängiger Spracherkennung, Einzelworterkennung oder kontinuierlicher Spracherkennung, Schlüsselworterkennung, etc.

7.1.1 Problemstellung

Da Sprache beim Menschen nach bestimmten Regeln erzeugt wird (vgl. hierzu Kapitel 3) sollte es möglich sein, die Sprachlaute (d.h. Sprach-Hörereignisse) den verschiedenen Abschnitten des Sprachsignals zuzuordnen. Hier ergibt sich aber das Problem, dass im Sprachsignal die einzelnen Einheiten (Phone, aber auch Wörter) nicht in isolierter Form

vorliegen. Vielmehr wird die Aussprache einzelner Laute durch die benachbarten Laute stark beeinflusst; man bezeichnet diesen Effekt als *Koartikulation*. Zudem sind auch die Wörter (außer beim getrennten Vorlesen) i.a. nicht voneinander getrennt. Laute oder Silben werden beim schnellen Sprechen ausgelassen (reduziert), und längere Pausen finden sich eigentlich nur am Ende von Sätzen oder Äußerungen.

Zudem werden sich zwei Sprachsignale desselben Satzes, wenn er von zwei unterschiedlichen Sprechern vorgelesen wird, zum Teil deutlich voneinander unterscheiden. Auch wenn derselbe Satz zweimal vom selben Sprecher vorgelesen wird, wird das Sprachsignal nicht beide Male identisch sein.

Trotz dieser Variabilität ist der Mensch in der Lage, Sprache zu erkennen, selbst wenn sie z.B. durch einen Telefonkanal übertragen wurde (eingeschränkte Bandbreite des ankommenden Signals) oder mit einem starken Dialekt gefärbt ist. Dies deshalb, weil der Mensch Sprache nicht als scheinbar sinnlose Einheiten einfach „erkennt“, sondern sie gleichzeitig dem Sinn nach „interpretiert“.

Hierbei hilft uns unser Weltwissen. Vergleicht man die beiden Ausdrücke

1. die Kunst zu verstehen
2. dick uns Zufall stehen

so ist (bei Kenntnis der deutschen Sprache) sofort klar, dass nur die erste Äußerung gemeint sein kann, obwohl sich beide Äußerungen (vor allem beim schnellen Sprechen) in ihren Lauten weitgehend gleichen. 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* und die daraus folgenden Eigenschaften des Sprachsignals
- Wissen darüber, wie typische Sprachsignale einzelner Sprachlaute aussehen (sog. *akustisches Modell*)
- Wissen über das zu erkennende *Vokabular*, d.h. welche Folgen von Symbolen zulässig sind
- Wissen über die Abfolge von Wörtern in der zu erkennenden Sprache (*Sprachmodell* oder *Grammatik*)

Dieses Wissen wird in einem Spracherkennner zum Teil implizit berücksichtigt, zum Teil aber auch als explizite Datenbank gespeichert.

7.1.2 Aufbau eines Spracherkenners

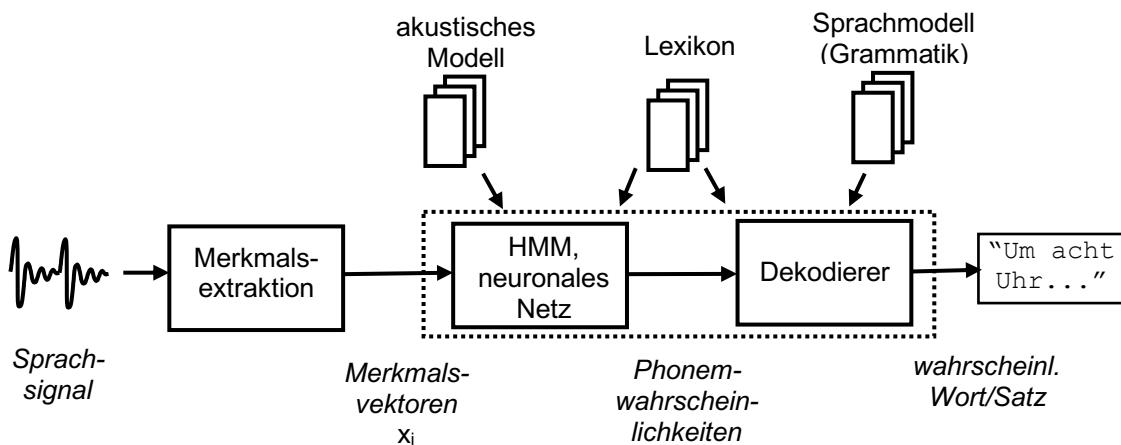
Um der Variabilität des Sprachsignals Rechnung zu tragen wird die maschinelle Spracherkennung (*automatic speech recognition, ASR*) im Allgemeinen als statistischer Prozess durchgeführt. Dabei wird eine beobachtete Folge von Lauten bezüglich ihrer Ähnlichkeit mit mehreren vortrainierten Lautfolgen verglichen, 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

Diese Bestandteile finden sich im nachfolgend skizzierten *Aufbau eines Spracherkenners* wieder.



Schematischer Aufbau eines Spracherkenners.

Aus dem Sprachsignal (Zeitsignal) werden zunächst Merkmale extrahiert. Dies geschieht kontinuierlich mittels einer gleitenden Fensterung des Signals; zu jedem Fenster wird also ein Vektor von Merkmalen berechnet. Verfahren hierzu werden in Abschnitt 7.1.3 vorgestellt. Bei einem phonembasierten Erkennung werden die extrahierten Merkmale zunächst bestimmten Phonemwahrscheinlichkeiten zugeordnet. Hierzu wird ein akustisches Modell benötigt, sowie ein Lexikon über mögliche Abfolgen von Phonemen in einem zu erkennenden Wort oder Satz.

Aus den Phonemwahrscheinlichkeiten werden anschließend mittels weiterer (supra-segmentaler) Informationen Wahrscheinlichkeiten für Phonemstrings und daraus Wahrscheinlichkeiten für Wörter und Sätze berechnet. Man bezeichnet letzteren Vorgang auch als Dekodierung. Zur Berechnung der Phonemwahrscheinlichkeiten werden üblicherweise entweder *Hidden-Markov-Modelle* (HMMs) oder sog. *neuronale Netze* eingesetzt; zur Dekodierung benutzt man meist HMMs. Beide Verfahren werden in Kapitel 7.1.4 erläutert. Die benötigten Informationen werden aus einem Lexikon (einer Auflistung von Phonemstrings und entsprechenden orthographischen Strings) sowie einem Sprachmodell bezogen. Sprachmodelle werden in Abschnitt 7.1.5 kurz diskutiert.

7.1.3 Merkmalsextraktion

Die Merkmale sind die eigentlichen Informationsträger für die Spracherkennung: Jegliche Information, die hier nicht in geeigneter Repräsentation enthalten ist, kann später nicht im Erkennungsprozess berücksichtigt werden. Es kommt also darauf an Merkmale zu finden, die

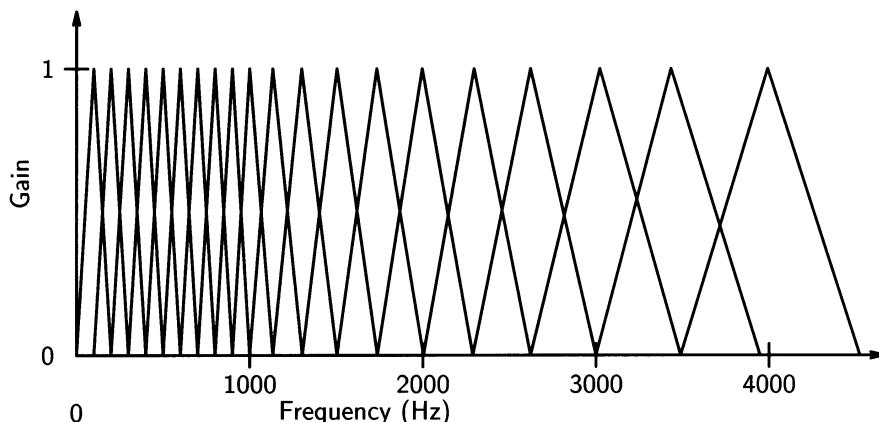
die zur Unterscheidung der Sprachlaute (und nicht etwa der Sprecher oder der Sprechumgebung) relevanten Informationen aus dem Sprachsignal extrahieren und in eine geeignete Repräsentation überführen.

Hierzu haben sich Verfahren bewährt, die explizit die Lautinformation im Vokaltrakt extrahieren, z.B. das in Kapitel 4 vorgestellte Cepstrum oder die lineare Prädiktion. Allerdings werden die Methoden meist nicht in ihrer ursprünglichen Form angewandt, sondern berücksichtigen noch Aspekte der menschlichen Wahrnehmung, wie sie in Kapitel 5 beschrieben wurden. Dem liegt die Annahme zugrunde, dass die menschliche Wahrnehmung relativ optimal auf die Erkennung von Sprache adaptiert ist; eine Nachahmung der relevanten Aspekte im Erkennungsalgorithmus sollte also die Erkennungsleistung erhöhen. Die wichtigsten Methoden sollen im Folgenden kurz umrissen werden.

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. Allerdings wird ein Cepstrum i.a. auf einer linearen Frequenz-Skala berechnet, die nicht der menschlichen Wahrnehmung entspricht.

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, vgl. untenstehende Abbildung. 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 ($\Delta\Delta$) zu den Koeffizienten der vorhergehenden Rahmen (Fenstern) berechnet wird. Die Dimension des Vektors steigt damit auf 26 oder 39 Einträge.



Gewichtungsfilter für die Berechnung Mel-skalierten cepstral Koeffizienten
(aus O'Shaughnessy, 2000, 215, nach Davis und Mermelstein, 1980).

Perceptual Linear Predictive (PLP) Coding:

Dieses Verfahren stellt eine Variante der LPC-Analyse dar, die an das menschliche Gehör angelehnt ist. Sie wurde von Hermansky (1990) entwickelt.

Das Sprachsignal wird zunächst mittels eines Hamming-Fensters gefenstert und spektral analysiert. Das Spektrum wird nun auf eine der Bark-Skala ähnlichen Skala transformiert und mit einer konstruierten Frequenzgruppen-Maskierungskurve gefaltet. Anschließend wird das Spektrum in 1-Bark-Schritten neu abgetastet, was eine weitere Glättung des Spektrums zur Folge hat. Als Ergebnis dieser Faltung erhält man ein Bark-skaliertes Spektrum, dessen Auflösung gegenüber dem linearen Spektrum deutlich reduziert ist. Durch eine Höhenanhebung wird die Frequenzabhängigkeit der Lautheit teilweise ausgeglichen. Anschließend wird durch Ziehen der Kubikwurzel die Intensitäts- in eine (angenäherte) Lautheits-Darstellung überführt. Diese Lautheits-Darstellung wird nun in den Zeitbereich rücktransformiert und daraus LPC-Koeffizienten berechnet.

Der Vorteil der PLP ist eine gegenüber der Standard-LPC erheblich reduzierte Modellordnung; Versuche haben gezeigt, dass eine Modellordnung von 5 (gegenüber 10-12 bei LPC) relativ optimal ist. Als positiver Nebeneffekt werden auch sprecherabhängige Merkmale – die den Erkennungsprozess stören könnten – unterdrückt.

Relative Spectral Analysis (RASTA):

Die PLP-Analyse ist zwar an das menschliche Gehör angepasst, ist allerdings nicht sonderlich robust gegenüber Störungen im Sprachsignal, wie sie z.B. durch unterschiedliche Mikrophontypen oder Hintergrundgeräusche entstehen. Man unterscheidet hier grob zwei Kategorien von Störungen: Additive Störungen, die sich als zum Sprachsignal additives Rauschen darstellen lassen, sowie multiplikative Störungen, die mittels Faltung (Multiplikation im Frequenzbereich) in das Signal eingehen.

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.

Wie bei der PLP-Analyse wird das Spektrum zunächst in den Bark-Bereich transformiert, mit einer Frequenzgruppen-Maskierungskurve gefaltet und in 1-Bark-Schritten neu abgetastet. Das so berechnete Bark-Spektrum wird nun nichtlinear mit Hilfe eines Logarithmus komprimiert. Hierdurch werden die multiplikativen Störungen (z.B. Mikrophon-Übertragungsfunktionen) in additive überführt. Die dann additiven Störungen können mit dem schon erwähnten Bandpass ausgefiltert werden. Anschließend erfolgt die von der PLP bekannte Höhenanhebung und die Berechnung der Kubikwurzel. Die Komprimierung wird nun durch Expansion mittels der e-Funktion rückgängig gemacht. Es erfolgt eine anschließende inverse Fourier-Transformation und die bekannte Berechnung der LPC-Koeffizienten.

In Experimenten konnte Hermansky zeigen, dass die RASTA-Analyse zwar robust gegenüber multiplikativen Störungen ist, nicht jedoch gegenüber additiven. Durch Einführung eines speziellen Parameters in der Kompressionsstufe konnte dieses Problem teilweise gelöst

werden (sog. Lin-Log-RASTA). Der Parameter kann je nach Verhältnis von Signal- zu Störenergie optimal gewählt werden. Details hierzu sind bei Hermansky et al. (1993) beschrieben.

7.1.4 Hidden-Markov-Modelle und neuronale Netze

Anhand der so bestimmten Merkmale soll nun eine Zuordnung zu Lauten oder Lautgruppen erfolgen. Hierzu haben sich statistische Methoden bewährt, die die zeitliche Abfolge von Merkmalen berücksichtigen. Die bekanntesten zwei Methoden, nämlich Hidden-Markov-Modelle und neuronale Netze, sollen hier kurz vorgestellt werden; Details finden sich in der u.a. Literatur.

Hidden-Markov-Modelle sind ein Standardwerkzeug zur Beschreibung und Modellierung von Zufallsprozessen. Der Prozess der Spracherzeugung hier wird als ein solcher Zufallsprozess angesehen: Kann die Erzeugung des Sprachsignals durch ein Modell beschrieben werden, welches eine Zuordnung zwischen (zu erkennenden) Symbolen und (beobachteten) Merkmalen herstellt, so kann dieses Modell anschließend zur Generierung von möglichen Merkmalssequenzen verwendet werden. 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.

Diese Maximierungsaufgabe kann mathematisch wie folgt beschrieben werden:

$$\arg \max_w P(W | A) = \arg \max_W P(A | W) \cdot P(W) \quad (7.1)$$

mit W der Sequenz von (zu erkennenden) Symbolen und A der Sequenz von Merkmalen.

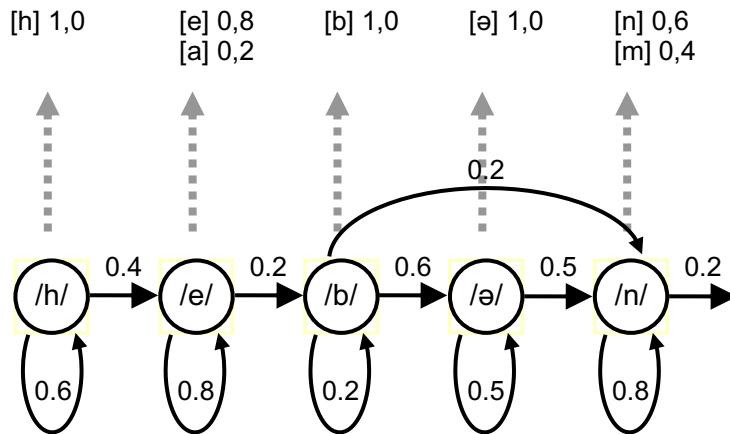
Hidden-Markov-Modelle sind zunächst Markov-Ketten, d.h. Ketten von Zuständen, die jeweils bestimmte Ausgaben generieren können. Zunächst sollen hier *Modelle mit diskreten Ausgaben* (Ausgabesymbolen) betrachtet werden. Beobachtet man die Markov-Kette an einer diskreten Folge von Zeitpunkten, so durchläuft man zu jedem Zeitpunkt genau einen Zustand und „generiert“ ein Ausgabesymbol. Die Übergänge zwischen den Zuständen werden durch sog. Übergangswahrscheinlichkeiten bestimmt. Darüber hinaus können weitere Einschränkungen für mögliche Folgezustände des aktuellen Zustands gegeben werden; bspw. verwendet man in der Sprachverarbeitung meist Links-Rechts-Modelle, d.h. bei einer Kette von Zuständen kann der Übergang nur von links nach rechts erfolgen (und nicht etwa zurück), oder spezieller Bakis-Modelle (bei denen maximal ein Zustand übersprungen werden darf).

Die unten stehende Abbildung zeigt ein Beispiel für eine einfache Markov-Kette. Diese besteht aus:

- 5 Zuständen
- 7 Symbolen
- Zustandsmenge $\{/h/, /e/, /b/, /ə/, /n/\}$
- Symbolmenge $\{[h], [e], [a], [b], [\theta], [n], [m]\}$

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

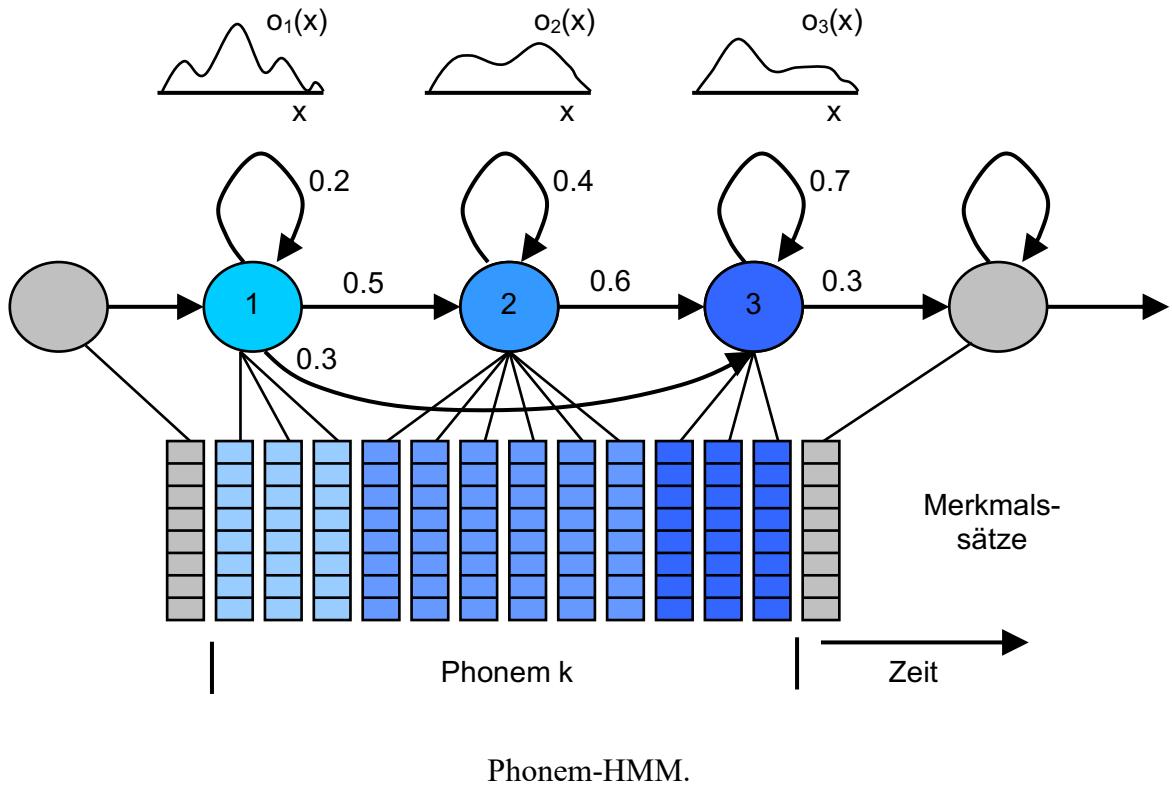
unterschiedliche Symbole emittieren kann. Man bezeichnet diese Wahrscheinlichkeiten als Emissionswahrscheinlichkeiten.



Beispiel eines einfachen Hidden-Markov-Modells für Sprache.

Durch die *nicht 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“. Das Hidden-Markov-Modell ist also in doppelter Hinsicht stochastisch: Nicht nur die Übergänge zwischen den Zuständen sind stochastisch geregelt, sondern auch die Ausgabesymbole, die ein Zustand emittiert.

Den Zuständen müssen nun Merkmale zugeordnet werden. Dies geschieht, indem die Zustände nicht wie bislang einzelne Symbole (z.B. Phoneme) emittieren, sondern Merkmale ähnlich denen, die aus dem Sprachsignal extrahiert werden können. Diese Ausgaben sind dann *kontinuierlich* (unbegrenzte Anzahl an Werten, die die Merkmalsvektoren annehmen können) und nicht – wie bislang – diskrete Symbole. Man bezeichnet dies als ein *Phonem-HMM*. 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. Die untenstehende Abbildung zeigt ein solches Phonem-HMM.



Wir betrachten nun die beobachtete (d.h. aus dem Sprachsignal berechnete) Folge von Merkmalsvektoren und nehmen an, dass sich das Markov-Modell im Startzustand 1 befindet. Es wird nun die Wahrscheinlichkeit berechnet, dass der erste Merkmalsvektor vom Zustand 1 erzeugt worden sein könnte (Emissionswahrscheinlichkeit). Diese wird nun mit der Übergangswahrscheinlichkeit zum nächsten Zustand multipliziert, z.B. (wie durch die Striche im unteren Teil der Abbildung angedeutet) wiederum in den Zustand 1 (Übergangswahrscheinlichkeit 0.2). Das Produkt wird nun mit der Emissionswahrscheinlichkeit des Merkmalsvektors 2 im Zustand 1 multipliziert, dann wiederum mit der Übergangswahrscheinlichkeit zum nächsten Zustand (hier wiederum Zustand 1), und so fort bis die Merkmalsvektoren abgearbeitet sind. Diese Prozedur muss nun für alle möglichen Abfolgen von Zuständen wiederholt werden. Als Maximum der Gesamtwahrscheinlichkeiten erhält man die wahrscheinlichste Kette von Zuständen; diese beschreiben dann die wahrscheinlichste Abfolge von Symbolen, d.h. die „erkannte“ Symbolkette.

Damit sich die Merkmalsvektoren adäquat modellieren lassen, muss das HMM zunächst trainiert werden; d.h. seine Emissions- und Übergangswahrscheinlichkeiten müssen so bestimmt werden, dass sie die zu erkennende Sprache (also Merkmalsvektoren und zugehörige Symbolketten) bestmöglich repräsentieren. Hierzu bedarf es zum einen etikettierter Daten (meist mehrere Zig oder Hundert Stunden Sprachdaten), zum anderen effizienter Algorithmen zum Training des HMMs. Hierzu können iterative Verfahren verwendet werden, z.B. der Baum-Welch-Algorithmus.

Sobald das HMM trainiert ist, kann es zur Erkennung verwendet werden. Hierbei muss – wie oben beschrieben – eine große Anzahl von möglichen Zustandsabfolgen bzgl. ihrer Gesamtwahrscheinlichkeit überprüft werden. Hierbei kann die Markov-Eigenschaft des

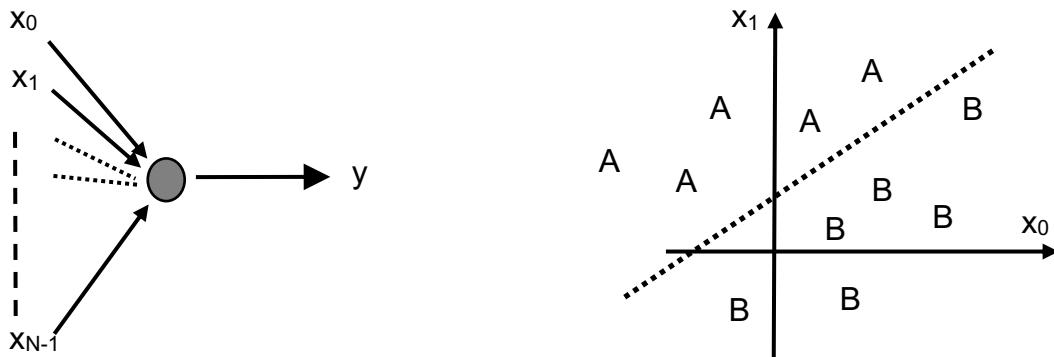
HMMs ausgenutzt werden, nämlich dass die Wahrscheinlichkeit, sich zu einem Zeitpunkt in einem bestimmten Zustand zu befinden, nur von der Wahrscheinlichkeit des vorhergehenden Zustands abhängt. Diese Eigenschaft erlaubt die Benutzung des Viterbi-Algorithmus, mit dessen Hilfe man die wahrscheinlichste Zustandsabfolge bestimmen kann, ohne alle möglichen Zustandsabfolgen zu berechnen.

Hidden-Markov-Modelle können sowohl zur Bestimmung der Phonemwahrscheinlichkeiten als auch zur Bestimmung der wahrscheinlichsten Folge von Phonemen/Wörtern benutzt werden. Die Phonemwahrscheinlichkeiten lassen sich aber auch mit einer anderen Art von Klassifizierern bestimmen, den sog. neuronalen Netzen. Eine zu diesem Zweck verbreitete Klasse von Netzen sind die Perzeptron-Netze, die im Folgenden kurz erläutert werden sollen.

Ein Perzeptron ist zunächst ein einfacher Klassifikator. Es verfügt über einen Eingangsvektor x (z.B. den beobachteten Merkmalsvektor), einen Gewichtsvektor w , und einen (eindimensionalen) Ausgangswert y . Letzterer berechnet sich aus dem Eingangsvektor x wie folgt:

$$y = f_h \left(\sum_{i=0}^{N-1} w_i \cdot x_i - \theta \right) = \begin{cases} +1 & \rightarrow \text{Klasse A} \\ -1 & \rightarrow \text{Klasse B} \end{cases} \quad (7.2)$$

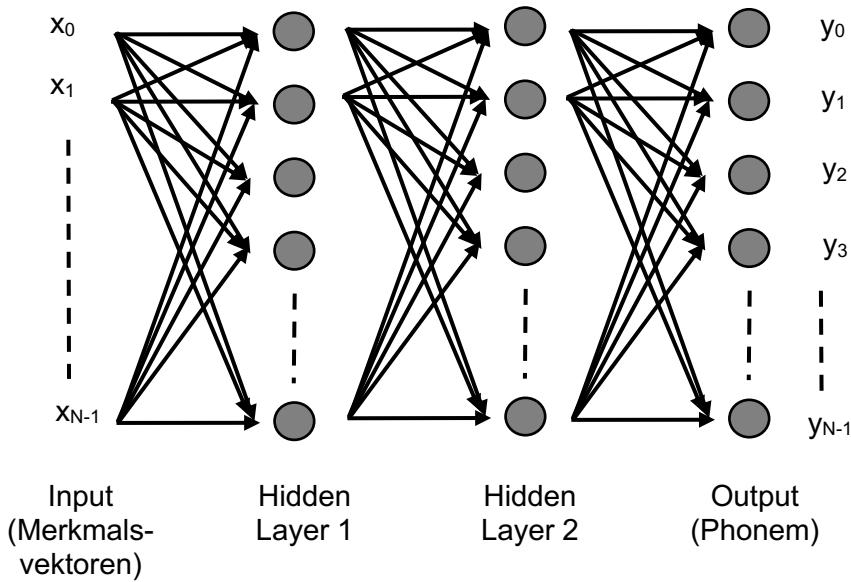
d.h. das Ausgangssignal kann nur zwei Werte (+1 oder -1) annehmen. Die Funktionsweise lässt sich graphisch wie folgt veranschaulichen.



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.

Nun sollen aber mehr als zwei Symbole klassifiziert werden. Zu diesem Zwecke werden mehrere Perzeptronen hintereinander geschaltet, sodass das Ausgangssignal eines Perzeptrons die Eingangsgröße zum folgenden Perzeptron liefert. Dabei werden mehrere Perzeptronen parallel in Schichten angeordnet; man spricht deshalb von einem Multi-Layer-Perzeptron-Netz. Die nachfolgende Abbildung zeigt die dabei entstehende Struktur.



Multi-Layer-Perzeptron-Netz.

Durch diese Struktur lassen sich also Eingangsvektoren in mehrere Ausgangsklassen klassifizieren. Dieses Prinzip wird bspw. zur Phonem-Klassifikation angewandt. Die Ausgänge können dann an den Dekodierer weitergegeben werden, der z.B. durch ein HMM gebildet werden kann.

7.1.5 Sprachmodelle

Das HMM beinhaltet schon gewisse Informationen über die Abfolge von Lauten, die in den Übergangswahrscheinlichkeiten kodiert sind. Für die kontinuierliche Spracherkennung ist es aber wünschenswert, Informationen über die Aufeinanderfolge von Wörtern beim Erkennungsprozess zu berücksichtigen. Hierzu sind zwei Ansätze verbreitet.

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. Diese Grammatik sollte nicht unbedingt mit der Standard-Grammatik des Deutschen übereinstimmen, sondern sollte an die Erkennungsaufgabe angepasst sein. Vor allem bei der Erkennung von Spontansprache ist nicht davon auszugehen, dass sich die Sprecher an die Standardgrammatik des Deutschen halten. Für spezielle Anwendungen (z.B. Diktieren von Arztbriefen oder juristischen Schriftstücken) reichen aber u.U. recht einfache, dabei aber hochgradig spezialisierte Grammatiken aus.

Eine solche Grammatik wird oft als sog. kontextfreie Grammatik formuliert. Folgendes Beispiel veranschaulicht den Aufbau:

```
start <sentence>;
<sentence>: <yes> | <no>;
<yes>: yes | yep | yes please ;
<no>: no | no thanks | no thank you ;
```

Hierbei bezeichnen die Ausdrücke in $\langle \rangle$ Bestandteile, die noch weiter zerlegt werden können, und alle anderen Ausdrücke Wörter des Vokabulars. Es ist offensichtlich, dass die Erzeugung

einer solchen Grammatik selbst für einen beschränkten Anwendungsfall recht aufwendig ist. Insbesondere müssen alle möglichen Benutzeräußerungen vorweggenommen werden.

Man versucht deshalb häufig, die Aufeinanderfolge von Wörtern statistisch zu beschreiben. 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). Für den Fall, dass eine bestimmte Aufeinanderfolge von Wörtern nicht in der Grammatik enthalten ist, kann die entsprechende Wahrscheinlichkeit aus den Auftretenswahrscheinlichkeiten der Einzelwörter berechnet werden. Man bezeichnet dies als ein n -gram *Backoff Language Model*.

7.1.6 Erkennungsleistungen

Es wurde bereits eingangs ausgeführt, dass die Erkennungsaufgabe von einer Vielzahl von Faktoren beeinflusst wird. Von diesen Faktoren hängt auch der Erfolg der Erkennung zu einem wesentlichen Teil ab. Deshalb ist es schwierig, die Leistungsfähigkeit von Spracherkennern direkt zu vergleichen.

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. Für jeden Satz werden nun die erkannte und die Referenz-Transkription verglichen. Dazu müssen beide Transkriptionen zunächst so einander zugeordnet werden, dass sich ein minimaler Fehler ergibt. Dieses *Alignment* wird z.B. mittels dynamic time warping erzeugt. Nach dem Alignment werden alle korrekt erkannten Wörter (c_w), vertauschten Wörter (s_w), gelöschten Wörter (d_w) und eingefügten Wörter (i_w) gezählt. Die Word Accuracy und Word Error Rate errechnen sich dann mit Hilfe der Gesamtzahl der Wörter W in der Referenz wie folgt:

$$\begin{aligned} WA &= 1 - \frac{s_w + i_w + d_w}{W} \\ &= 1 - WER \\ &= \frac{c_w - i_w}{c_w + s_w + d_w} \end{aligned}$$

und

$$WER = \frac{s_w + i_w + d_w}{W}$$

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. Auch wurden Maße vorgeschlagen, die sich auf die Auftretenshäufigkeit von Erkennungsfehlern in einem Satz beziehen, sowie solche, die nur einen bestimmten Teil des Vokabulars umfassen. Beispiele finden sich in Möller (2005).

In der folgenden Tabelle sind einige Erkennungsraten für verschiedene Erkennungsaufgaben angegeben, die aus der Literatur bekannt sind. Es zeigt sich, dass bei geringem Wortschatz und einem ungestörten Eingangssignals eine fast perfekte Erkennung möglich ist. Allerdings sinkt die Erkennungsleistung doch erheblich, wenn ein größerer oder komplexerer Wortschatz zu erkennen ist. Bei Spontansprache ist die Erkennungsleistung häufig noch sehr bescheiden.

Tabelle 1: Leistungsfähigkeit verschiedener Spracherkennner
(Literaturdaten aus verschiedenen Jahren).

Korpus	Sprechstil	Vokabular	Wortfehler-rate	Wortfehler-rate	Wortfehler-rate
kontinuierliche Zahlen	vorgel.	10	0,1%	< 0,3%	
Fluginform.-System	spontan	2.500	3%	2%	
Wall Street Journal	vorgel.	64.000	12%	7%	
Nachrichten	vorgel./spontan	64.000		30%	
Anrufe beim Dialogsystem	Konversation	10.000 / 1.957		50%	20,4%
			(1996)	(1997)	(2000)

Die Erkennungsleistung wird darüber hinaus sehr stark von den akustischen Umgebungsbedingungen beeinflusst. Besonders in akustisch widrigen Umgebungen (z.B. bei der Erkennung über das Mobiltelefon oder aus dem fahrenden Auto) ist es wichtig, Maßnahmen zur Verbesserung der Erkennungsleistung zu ergreifen. Folgende Lösungsmöglichkeiten sind denkbar:

- Aufzeichnung umfangreichen Trainingsmaterials, unter realistischen (akustischen) Bedingungen
- Verwendung robuster Merkmalssätze und/oder einer Signalvorverarbeitung, bei der Störgeräusche unterdrückt werden
- Aufspaltung des Frequenzbereiches und Benutzung paralleler Erkennner für die unterschiedlichen Teile des Spektrums (sog. *Multi-Stream-Spracherkennung*); diesem Ansatz liegt die Annahme zugrunde, dass die unterschiedlichen Teile des Spektrums unterschiedlich stark gestört sind, d.h. das eventuelle Störgeräusche nicht in allen Spektralbereichen stören

Diese Ansätze können natürlich auch kombiniert werden.

7.1.7 Literatur

Comerford R., Makhoul J., Schwartz R. (1997). The Voice of the Computer is Heard in the Land (and it Listens too!). IEEE Spectrum, Dez., 39-47.

Davis S.B., Mermelstein P. (1980). Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. IEEE Trans. ASSP 28, 357-366.

Hermansky H. (1994). RASTA Processing of Speech, IEEE Trans. Speech and Audio Processing 2, 578-589.

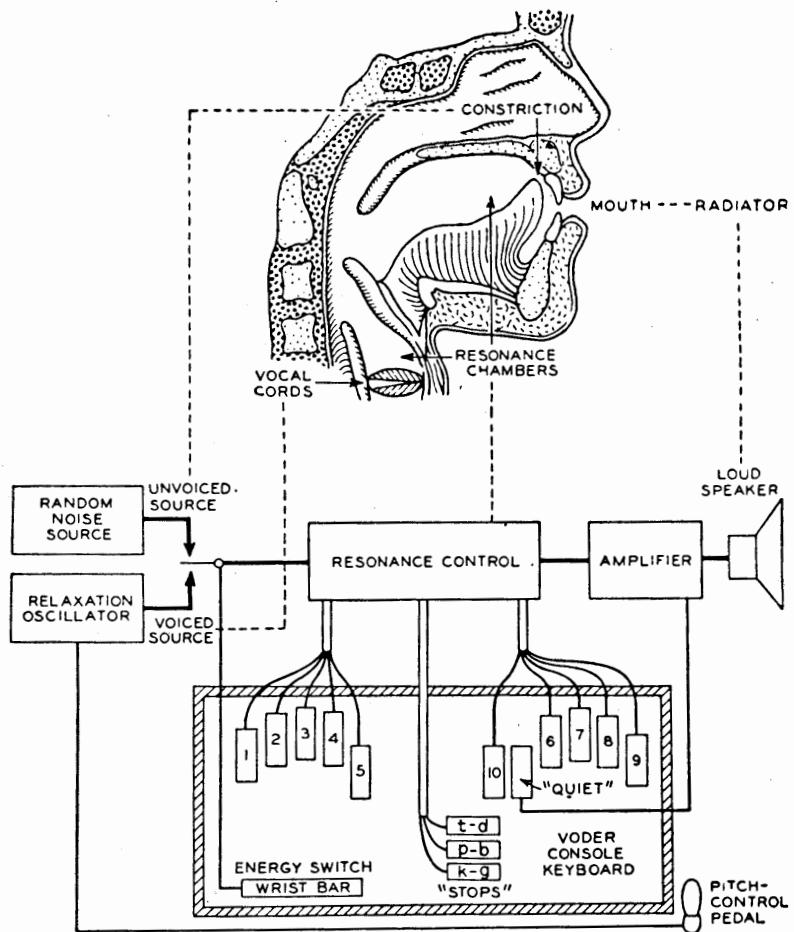
Hermansky, H., Morgan, N., Hirsch, H.-G. (1993). Recognition of Speech in Additive and Convolutional Noise Based on RASTA Spectral Processing. In: Proc. ICASSP 1993, 2, 83-86.

- Hermansky, H. (1990). Perceptual Linear Predictive (PLP) Analysis of Speech. *J. Acoust. Soc. Am.* 87(4), 1738-1752.
- Lee C., Soong F.K., Paliwal K.K. (1996). Automatic Speech and Speaker Recognition, Kluwer Academic Publ., Boston MA.
- Lippmann R.P. (1987). An Introduction to Computing with Neural Nets, *IEEE ASSP Mag.*, April 87, 4-22.
- Rabiner L.R., Juang B.H. (1986). An Introduction to Hidden Markov Models. *IEEE ASSP Magazine* 3(1), 4-16.
- Rabiner L.R., Juang B.H. (1993). Fundamentals of Speech Recognition. Prentice Hall, Englewood Cliffs.

7.2 Sprachsynthese

Die Aufgabe der Sprachsynthese besteht darin, aus auf Symbolebene vorliegendem Text ein Signal zu generieren, welches als Sprache wahrgenommen wird. Die dabei angewendeten Verfahren unterscheiden sich zum Teil erheblich in ihren Ansätzen und Zielen.

Historisch gesehen versuchte man zunächst, eine Maschine zu bauen, die die Funktion des menschlichen Sprechapparates nachbildet, jedoch vom Menschen gesteuert wird. Dem Menschen oblag also die Definition der Ansteuerparameter, während die Maschine die physikalische Erzeugung des Sprachsignals übernahm. Beispiele hierfür sind z.B. die Sprechmaschinen nach von Kempelen, vgl. Abschnitt 3.5. 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). Synthesesysteme nach Regeln mit Lautschrifteintrag wurden erstmals in den 50er Jahren entwickelt, und durch die Fortschritte in der Symbolverarbeitung wurden Ende der 70er Jahre die ersten Sprachsynthesen entwickelt, die eine komplette Generierung gesprochener Sprache aus Rechtschrift-Text erlauben.



Schematische Zeichnung des Voder (VOice DEMonstratoR),
nach Dudley und Tarnoczy (1950).

Eine vollständige Generierung auf Basis von Text ist nicht unbedingt immer notwendig, vor allem dann nicht, wenn die zu synthetisierenden Äußerungen vorher bekannt sind. Dies ist in vielen interessierenden Anwendungen der Fall, sodass voll synthetisierte Sprache bislang nur in Ausnahmefällen – nämlich da, wo es unabdingbar ist – eingesetzt wird. Der Hauptgrund hierfür ist die immer noch recht begrenzte Qualität synthetischer Sprache.

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*). Die Sprachsignale können als Signalform kodiert oder in parametrischer Form abgelegt sein (vgl. Kapitel 6). Die Signalabschnitte werden entweder unverändert abgespielt, oder sie können an den Segmentgrenzen noch manipuliert werden; bspw. ist eine Anpassung des Grundfrequenzverlaufes oft wünschenswert, um wahrnehmbare Sprünge in der Melodie abzumildern.

Man kann Sprachausgabesysteme nach ihren Leistungsmerkmalen in folgende Klassen einteilen (vgl. Blauert und Schaffert, 1985):

- Ansageautomaten
- Aussageautomaten
- Vorleseautomaten

Leistungsmerkmal eines *Ansageautomaten* ist die Stimmgenerierung, d.h. die Erzeugung eines Sprechschalls, z.B. aus Lautschriftketten oder anderen Parametern. Diese Automaten können nach unterschiedlichen Prinzipien arbeiten. Häufig werden dabei folgende Verfahren unterschieden:

- *Parametrische Verfahren*, bei denen der Sprechschall aus Parametern generiert wird. Diese Parameter beschreiben häufig die Funktionen oder die Anatomie des menschlichen Sprechapparates, bspw. Formanten, Anregungssignale, etc.
- *Nichtparametrische Verfahren*: Diesen liegt kein parametrisches Modell der Spracherzeugung zugrunde, sondern der Sprechschall wird aus Daten vorher aufgezeichneter Sprache zusammengesetzt (datenbasierte Synthese). Man bezeichnet ein solches Verfahren als Verkettungssynthese.

Aussageautomaten erzeugen Sprechschalle aufgrund eines logisch-semantischen Konzeptes. Man nennt sie deshalb auch Konzept-nach-Sprache- oder *Concept-to-Speech-Systeme (CTS)*. Neben der Generierung der Sprechschalle ist die Generierung der Aussage in symbolischer Form also ein weiterer wesentlicher Teil des Systems. Insbesondere ist die Generierung von prosodischer Information (Tonhöhe, Sprechrhythmus, Lautdauern, Betonung) wichtig. Hierzu können regel- und datenbasierte Verfahren eingesetzt werden.

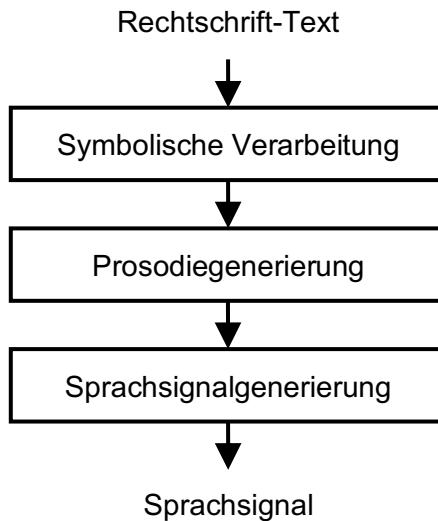
Vorleseautomaten erzeugen Sprechschalle direkt aus Rechtschrift-Text (*Text-to-Speech, TTS*). Es werden also die Lautschrift und die zur Prosodiegenerierung notwendigen Informationen direkt aus dem Text generiert. Dies ist mit Hilfe einer umfangreichen linguistischen Vorverarbeitung möglich, die wiederum auf Regeln oder Daten basieren kann.

Im Folgenden soll zunächst die Struktur des Allgemeinsten dieser Systeme, des Vorleseautomaten, vorgestellt werden (Abschnitt 7.2.1). Dann werden die Bestandteile symbolische Verarbeitung (7.2.2) und Prosodiegenerierung (7.2.3) erläutert. Daran schließt sich die Beschreibung einiger wichtiger Prinzipien zur eigentlichen Generierung des Sprachsignals (Ansageautomat) an (7.2.4). Hierbei kann ein parametrisches Modell der

Spracherzeugung zugrunde gelegt werden, es können Einheiten kurzer oder längerer Dauer verkettet werden (Verkettungssynthese oder Unit-Selection-Synthese), oder es können statistische Verfahren zur Generierung von Sprache (z.B. HMMs) eingesetzt werden.

7.2.1 Struktur eines Vorleseautomaten

Trotz recht unterschiedlicher Systemarchitekturen lassen sich Vorleseautomaten (TTS-Systeme) grob in drei Komponenten aufteilen. Die nachfolgende Abbildung veranschaulicht den Aufbau.



Schematischer Aufbau eines Vorleseautomaten.

Die Sprachsignalgenerierung ist für die eigentliche Synthese des Sprachsignals verantwortlich und wird daher auch als Synthesator bezeichnet. Sie übernimmt die Rolle des Sprechapparates beim Menschen und setzt die parametrisch vorliegenden Eingangsinformationen in ein Sprachsignal um. Solche Eingangsinformationen sind z.B. die Lautsymbolkette des zu synthetisierenden Signals, die Lautdauern, die Grundfrequenzkontur und evtl. noch Informationen zum Intensitätsverlauf.

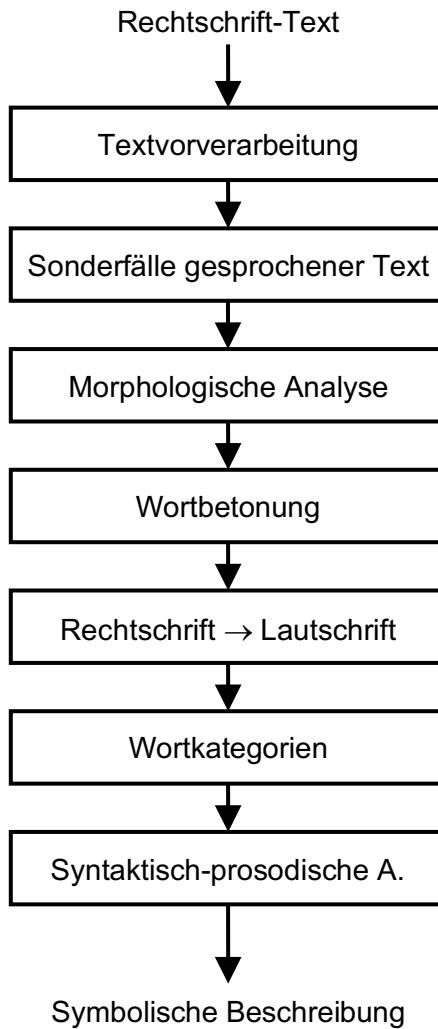
Die Eingangsinformationen müssen von der Prosodiegenerierung aus den von der symbolischen Verarbeitung bestimmten Daten generiert werden. Dazu werden die Lautsymbolketten sowie die Phrasengrenzen und die Akzente einer zu synthetisierenden Äußerung benötigt.

Diese Informationen werden wiederum von der symbolischen Verarbeitung bestimmt. Sie ersetzt damit sozusagen den Verstand oder das Gehirn beim Vorlesen. Die symbolische Verarbeitung analysiert den Text und setzt ihn in eine erweiterte symbolische Darstellung um.

Bei einem Concept-to-Speech-System (CTS) wird als Eingabe nicht ausformulierter Text, sondern eine abstrakte Darstellung des zu synthetisierenden Sachverhaltes verwendet. Hierdurch können viele Probleme, die bei der Ansteuerung durch reinen Rechtschrift-Text entstehen, vermieden werden und damit die Qualität der ausgegebenen Sprache erhöht werden. Dabei spielt es nur eine untergeordnete Rolle, auf welcher Ebene diese Informationen geliefert werden.

7.2.2 Symbolische Verarbeitung

Die einzelnen Ansätze und Schwerpunkte, die bei der symbolischen Verarbeitung verfolgt werden, können je nach Sprache mehr oder weniger stark variieren. Für das Deutsche können exemplarisch die in folgender Abbildung gezeigten Schritte unterschieden werden, die nachfolgend beschrieben sind.



Symbolische Verarbeitung von Rechtschrifttext.

Textvorverarbeitung:

Bevor eine lexikalische Analyse des Eingabetextes möglich ist, muss dieser erst in eine einheitliche Form gebracht werden. Hierzu müssen Sonderfälle im geschriebenen Text, bspw. Ziffern, Abkürzungen oder Nummerierungen, in eine sprachliche Form gebracht werden. Dies kann z.B. mit einem Wörterbuch geschehen, welches Abkürzungen enthält. Anhand der relevanten Interpunktionszeichen kann eine Unterteilung der Sätze (z.B. in Haupt- und Nebensätze) erfolgen.

Behandlung von Sonderfällen im gesprochenen Text:

Hierbei sind z.B. die Aussprache von Eigennamen oder fremdsprachlicher Wörter zu nennen. Eigennamen folgen oft nicht den Standard-Ausspracheregeln und erfordern somit eine

spezielle Behandlung. Diese kann aus Wörterbüchern oder aus gesonderten Regeln bestehen. Fremdsprachliche Wörter müssen zunächst im Text erkannt werden und können dann (bei einer multilingualen Synthese) nach den Regeln der Fremdsprache – oder nach noch zu definierenden Regeln einer „Zwischensprache“ – ausgesprochen werden.

Morphologische Analyse:

Um eine zuverlässige phonetische Transkription durchzuführen können einzelne Worte in Wortstämme, Präfixe, Suffixe, Flexionsendungen und Fugenelemente aufgeteilt werden. Die morphologische Struktur bestimmt die Wortbetonung und gibt Aufschluss über die syntaktische Struktur eines Wortes, die wichtig für die Satzbetonung ist. Zur morphologischen Analyse können regel- oder datenbasierte Verfahren verwendet werden, oder Kombinationen aus beiden.

Bestimmung der Wortbetonung:

Für die Generierung der richtigen Prosodie müssen die Betonungsstufen der Silben erkannt werden. Diese können aus Regeln abgeleitet oder einem Wörterbuch entnommen werden. Allerdings kann sich die Wortbetonung durch die Satzbetonung noch ändern. Die Wortbetonung ist u.U aber auch für die Rechtschrift-nach-Lautschrift-Umsetzung wichtig.

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. Darüber hinaus kann man sich an Standardregeln für die Aussprache des Deutschen halten, die jedoch nicht immer fehlerfrei sind. Ausnahmen kann man über ein Aussprachewörterbuch abfangen. Häufig verwendet man Kombinationen von regel- und datenbasierten Ansätzen.

Bestimmung von Wortkategorien:

Die verschiedenen Wortklassen können z.B. einem Wörterbuch entnommen werden. Wörter, die nicht im Wörterbuch enthalten sind, können anhand von Regeln klassifiziert werden. Dabei kann auch die Groß- und Kleinschreibung helfen, oder die Informationen der morphologischen Analyse

Bestimmung der syntaktischen und prosodischen Struktur:

Die Prosodiegenerierung benötigt zur Bestimmung der Satzintonation semantische, syntaktische und lexikalische Informationen. Da eine semantische Analyse normalerweise nicht durchgeführt werden kann beschränken sich Synthesesysteme auf die Analyse syntaktischer und lexikalischer Strukturen. Auf syntaktischer Ebene werden die Sätze nach den Regeln der deutschen Grammatik durchsucht und einzelne Wörter zu syntaktischen Objekten zusammengefasst. Auf prosodischer Ebene werden Phrasierung und Akzentuierung bestimmt. Eingangsinformationen hierzu sind z.B. die Interpunktionszeichen, die Wortbetonungen und die syntaktischen Objekte.

7.2.3 Prosodiegenerierung

Neben der textlichen Information stellt die Prosodie wichtige Informationen zur Interpretation einer Aussage bereit. So können manche mehrdeutigen Aussagen erst mit Hilfe der Prosodie richtig interpretiert werden, bspw. die Zeugenaussage „Er wollte den Radfahrer umfahren“ vs. „Er wollte den Radfahrer umfahren“. Auch können sowohl die Verständlichkeit als auch die Natürlichkeit synthetisierter Äußerungen stark unter einer fehlerhaften Prosodie leiden. Der menschliche Hörer verwendet prosodische Informationen nämlich auch, um den Verlust von

nicht verstandenen Teilinformationen (verschluckte oder gestörte Silben oder Wörter) auszugleichen.

Aus der symbolischen Verarbeitung stehen der Prosodiegenerierung lexikalische und syntaktische Informationen zu Verfügung. Hieraus müssen nun die akustischen Parameter Grundfrequenz, Laut- und Pausendauer sowie die Sprachsignalenergie bestimmt werden. Dabei kommen folgende Verfahren zum Einsatz:

- *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. Man unterscheidet hierbei Wort- und Phrasenakzente, die zunächst getrennt generiert und dann einer Grundlinie (z.B. leicht fallende Intonation) überlagert werden. Andere Ansätze beruhen auf der Übertragung von Kopiekonturen (Adriaens et al., 1991) oder auf einer LPC-Beschreibung der Grundfrequenzkontur (Mersdorf et al., 1997). Zur Lautdauersteuerung werden ebenfalls häufig regelbasierte Verfahren angewendet.
- *Datenbasierte und statistische 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 sein.

7.2.4 Sprachsignalgenerierung

Aus den Lautketten, den Grundfrequenz- und Intensitätsverläufen muss nun das Sprachsignal generiert werden. Die dabei verwendeten Ansätze kann man grob in

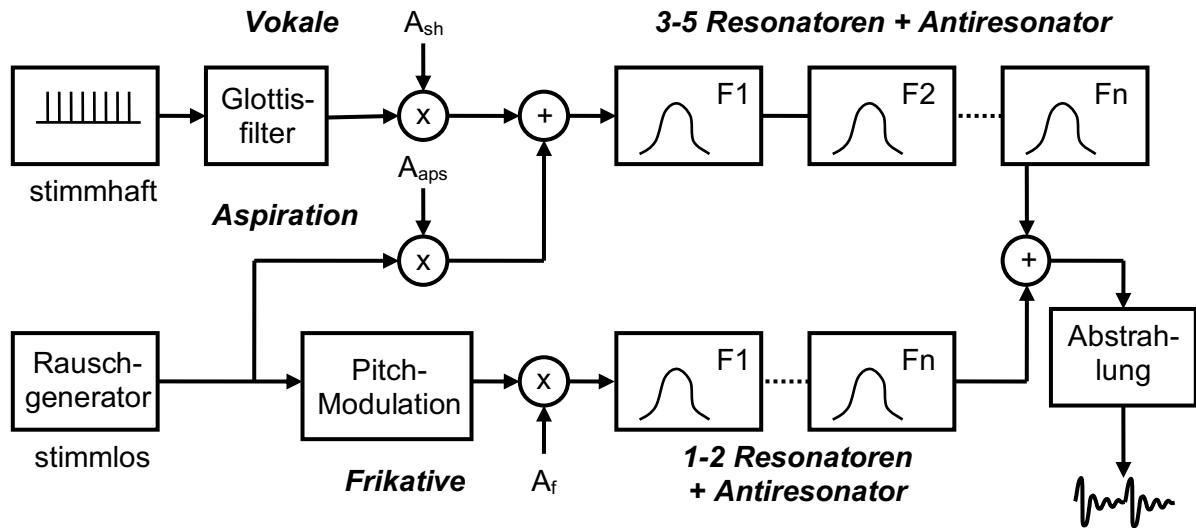
- regelbasierte vs. datenbasierte oder
- parametrische vs. nichtparametrische

Ansätze unterscheiden. Eine genauere Unterscheidung lässt vier Prinzipien erkennen, die heutzutage angewandt werden. Diese Prinzipien sind im Folgenden beschrieben

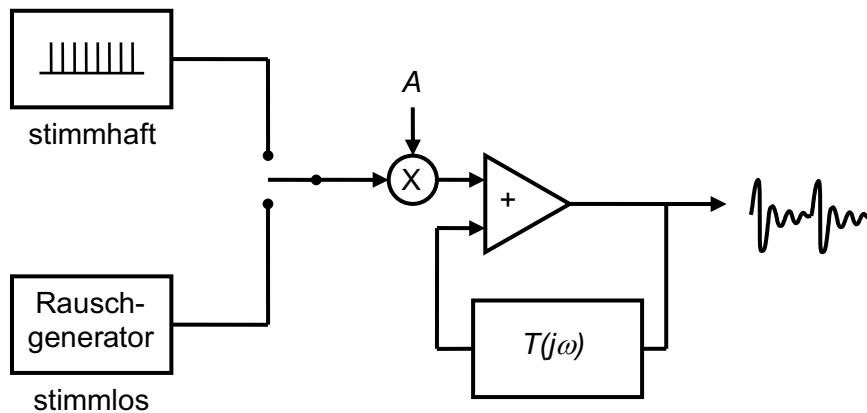
Parametrische Synthese:

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) sowie den in Kapitel 4 beschriebenen Verfahren zur Signalanalyse. Zwei bekannte Verfahren sollen hier kurz vorgestellt werden.

Zum einen kann man – wie schon Dudley – versuchen, Sprache aus Formantfiltern zu synthetisieren, die mit idealisierten Anregungssignalen angesteuert werden. Hierbei wird meist eine getrennte Generierung der Vokale, der Frikative und der aspirierten Laute durchgeführt (vgl. O'Shaughnessy, 2000). Vokale werden durch eine stimmhafte Anregung (Impulskamm) und durch einen längeren Vokaltrakt (3-5 Formanten) realisiert. Frikative werden durch eine Rauschanregung und einen kürzeren Vokaltrakt generiert. Bei den aspirierten Lauten wird die Rauschanregung durch den (längeren) Vokaltrakt der Vokale gefärbt. Die nachfolgende Abbildung zeigt die Prinzipschaltung einer solchen *Formantsynthese*. Die dabei benötigten Ansteuerungsparameter sind die Grundfrequenz, die Frequenzen und Bandbreiten der Formantfilter, sowie die Amplituden der Anregungssignale.

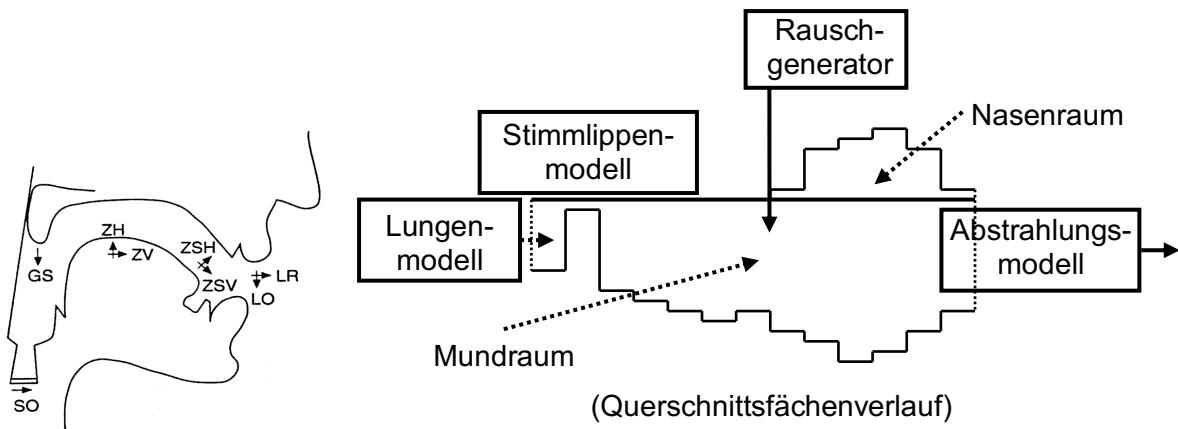


Eine andere Variante der parametrischen Synthese ist der *LPC-Synthetisator*. Dabei werden aus natürlicher Sprache zunächst LPC-Parameter bestimmt, mit deren Hilfe dann neu synthetisiert werden kann. Hierzu wird meist eine rein stimmhafte oder stimmlose Anregung gewählt, und mit Hilfe des Prädiktions-Synthesefilters (vgl. Abschnitt 3.5) das Sprachsignal synthetisiert. Benötigt werden also die LPC-Parameter, die Amplitude des Anregungssignals und eine Stimmhaft-Stimmlos-Entscheidung. Die LPC-Synthese ist einfacher als die Formantsynthese, da die LPC-Parameter vollautomatisch bestimmt werden können; allerdings sind die Variationsmöglichkeiten stark eingeschränkt, und somit lässt sich meist nur weniger natürlich klingende Sprache synthetisieren.



Neben diesen Verfahren, die von einem vereinfachten Modell der menschlichen Sprachproduktion Gebrauch machen, kann man aber auch versuchen, die genauen Bewegungen der Artikulationsorgane nachzubilden. Man bezeichnet dieses Verfahren als *artikulatorische Synthese*. Das Verfahren ist vor allem deshalb interessant, da es genauere

Rückschlüsse auf die menschliche Spracherzeugung zulässt als jedes andere Syntheseverfahren. Allerdings ist die Komplexität der Modellierung wie auch der Berechnung sehr hoch. Deshalb wird die artikulatorische Synthese hauptsächlich zu Forschungszwecken eingesetzt.



Vereinfachtes Schema der artikulatorischen Synthese (nach Kröger, 1996).

Die artikulatorische Synthese ist – wie die Formant- oder LPC-Synthese – eine parametrische Synthese, da ein parametrisches Modell der Spracherzeugung verwendet wird. Die Parameter sind aber direkt mit der Physik der menschlichen Spracherzeugung verbunden. Die folgende Tabelle zeigt Parameter, die hierfür verwendet werden können.

Tabelle: Parameter zur artikulatorischen Synthese (nach Kröger, 1996).

Abkürzung	Parameter
GS	Gaumensegelsenkung
LO	Lippenöffnung
LR	Lippenrundung
ZH	Zungenhebung
ZV	Zungenvorverlagerung
ZSH	Zungenspitzenhebung
ZSV	Zungenspitzenvorverlagerung
SO	Stimmlippenöffnung

Verkettungssynthese:

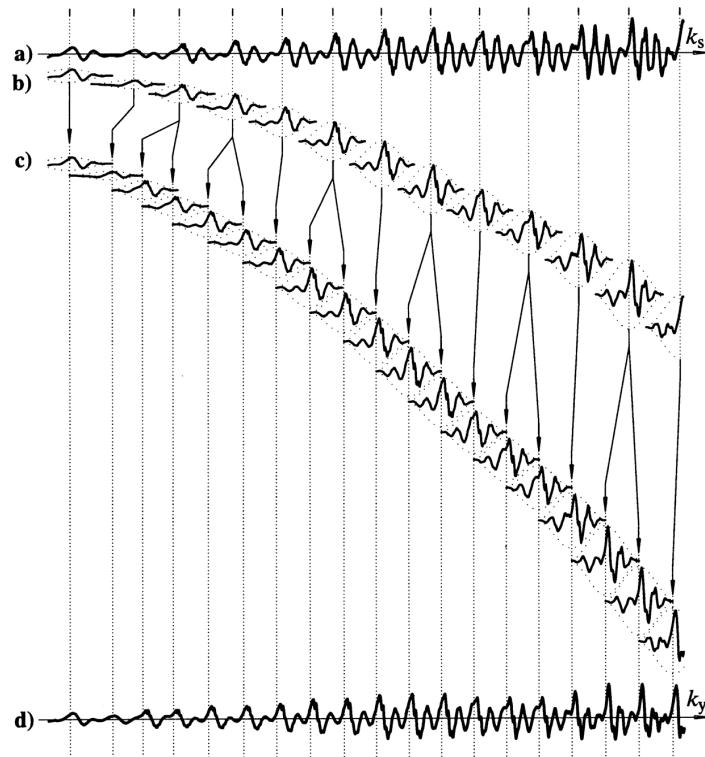
Die Verwendung von Modellen der menschlichen Spracherzeugung bei der Synthese hat den Nachteil, dass nicht alle Aspekte genügend genau modelliert werden können. Deshalb ist die damit erzielbare Qualität im Allgemeinen recht begrenzt. Daher versucht man, durch Verkettung einzelner Bausteine natürlich 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. Da sich Sprache vor allem an den Lautübergängen ändert und darin informationstragende Aspekte liegen, sollten Einheiten gewählt werden, die Lautübergänge schon enthalten. Umgekehrt wird die Anzahl der benötigten Einheiten, um

beliebigen Text zu synthetisieren, umso größer, je länger die Einheiten sind. Man verwendet deshalb meist Diphone, Halbsilben oder Silben.

Die Einheiten werden dann nach speziellen Verfahren miteinander verkettet. Hierbei kommt es darauf an, dass die Einheiten so verbunden werden, dass sich die Grundfrequenz, Amplitude und Dauer getrennt manipulieren lassen, ohne dass die spektrale Struktur (die den Klang beeinflusst) verändert wird. Dies ist möglich durch eine Verkettung synchron zur Grundfrequenz. Dazu hat sich das sog. *PSOLA-Verfahren (pitch-synchronous overlap-and-add)* bewährt, welches von Charpentier und Moulines 1989 entwickelt wurde.

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 herausfenstert, 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. Die untenstehende Abbildung zeigt das Verfahren.



Beispiel für die Wirkungsweise des PSOLA-Verfahrens. Verringerung der Grundperiodendauer um 1/3 (nach Vary et al., 1998, 478).

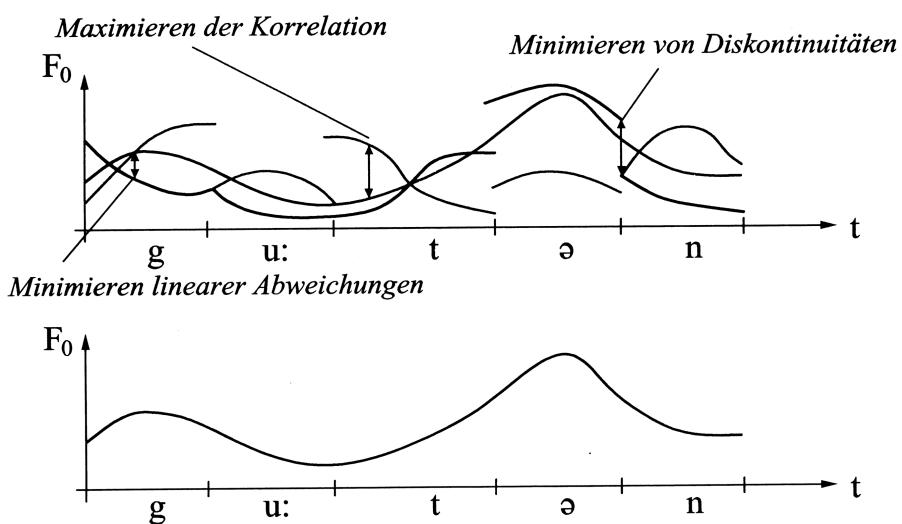
- a): Originalsignal; b) Elementarbausteine; c) 2-Perioden-Fenster nach Verschiebung; d): manipuliertes Ausgangssignal.

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. Dadurch klingt ein mittels PSOLA erzeugtes Sprachsignal meist viel natürlicher als ein parametrisch erzeugtes. Dennoch ergeben sich durch die Signalmanipulation wahrnehmbare Artefakte, die die Qualität nicht optimal werden lassen. Um diese zu umgehen wurden die Unit-Selection-Methoden entwickelt.

Unit-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 Aufgabe des Synthesizers ist es also, unter den vorhandenen Bausteinen diejenigen zur Verkettung auszuwählen, die möglichst gut zum zu synthetisierenden Text und auch untereinander gut zueinander „passen“. Dies ist in folgender Abbildung angedeutet.



Prinzip der Unit-Selection-Synthese (nach Hess, 2000).

Die Passgenauigkeit wird über eine Kostenfunktion definiert, und die eigentliche Syntheseaufgabe ist die Minimierung der Kosten. Dabei werden zwei Typen von Kosten unterschieden:

- *Einheitenkosten*: Diese ergeben sich aus der zu synthetisierenden Äußerung, d.h. wie gut die im Inventar vorhandenen Bausteine zu den zu synthetisierenden passen. Diese Kosten müssen online zur Laufzeit der Synthese bestimmt werden.

- *Verkettungskosten:* Diese ergeben sich aus der Passgenauigkeit der Bausteine des Inventars zueinander. Verkettungskosten können offline vor der eigentlichen Synthese bestimmt werden.

Zur Synthese mittels des Unit-Selection-Ansatzes muss zunächst ein prosodisch und phonemisch etikettierter Datensatz erzeugt werden. Zur phonetischen Etikettierung dieses Datensatzes können Methoden der Spracherkennung verwendet werden; die prosodische Etikettierung beschränkt sich meist auf einige grobe Kategorien auf Silbenbasis. Während das Inventar vor der eigentlichen Synthese etikettiert werden kann, muss der Eingabetext online, d.h. zur Synthese-Laufzeit, etikettiert werden.

Mit solchen Verfahren lässt sich Sprache synthetisieren, die natürlich gesprochener qualitativ recht nah kommen kann. Allerdings ist die Qualität sehr stark abhängig vom zu synthetisierenden Vokabular: Sind für die zu synthetisierende Äußerung zufällig gut zueinander passende Bausteine im Inventar (u.U. Halbsätze oder ganze Sätze), so ist die Qualität praktisch optimal (es ist in der Tat ja natürlich aufgezeichnete und unverändert wieder abgespielte Sprache); finden sich dagegen keine passenden Bausteine im Inventar, so können deutliche Sprünge auftreten, die perzeptiv sehr auffällig sein können.

HMM-basierte Synthese:

Die Unit-Selection-Synthese versucht, die zu synthetisierende Äußerung aus einem möglichst optimal passenden Satz von Bausteinen zusammenzusetzen. Diese Aufgabe erinnert an die der Spracherkennung, bei dem die erkannte Merkmalsfolge möglichst optimal (ähnlich) mit Hilfe eines Hidden-Markov-Modells generiert werden sollte; die Folge von Zuständen beschrieb dann das erkannte Wort.

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.

Derzeit wird noch viel an der HMM-basierten Synthese geforscht. Erste Ergebnisse zeigen jedoch, dass sich hiermit eine recht gute Sprachqualität erzielen lässt. Wie bei der Spracherkennung und der Unit-Selection-Synthese wird hier der Schwerpunkt auf eine möglichst optimale Abdeckung des zu erkennenden bzw. zu synthetisierenden Raumes durch das Trainingsmaterial gelegt. Wissen über den Prozess der menschlichen Spracherzeugung lässt sich so allerdings nur schwerlich gewinnen.

7.2.5 Literatur

- Adriaens, L.M.H. (1991). Ein Modell deutscher Intonation. Dissertation, TU Eindhoven.
- Blauert, J. (1994). Kommunikationsakustik II: Audiokommunikation und virtuelle Realität. Skriptum zur Vorlesung am Institut für Kommunikationsakustik, Ruhr-Universität, Bochum.
- Blauert, J., Schaffert, E. (1985). Automatische Sprachein- u. Ausgabe. Schriftenreihe der BundesAnst. F. Arbeitsschutz Fb 417, Dortmund.
- Dudley, H. (1939). The Vocoder. Bell Labs Record 17, 122-126.
- Dudley, H., Tarnoczy, T.H. (1950). The Speaking Machine of Wolfgang von Kempelen. J. Acoust. Soc. Am. 22(2), 151-166.
- Fujisaki, H. (1983). Dynamic Characteristics of Voice Fundamental Frequency in Speech and Singing. In: The Production of Speech, P.F. MacNeilage, ed., Springer, New York NY, 39-55.
- Hess, W. (1996). Maschinelle Sprachsynthese. Spektrum der Wissenschaft, Dez. 1996, 100-103.
- Kröger, B.J. (1996). Von Sprechbewegungen zum akustischen Signal: Artikulatorische Sprachsynthese. Spektrum der Wissenschaft, Dez. 1996, 105-107.
- Mersdorf, J. (2001). Sprecherspezifische Parametrisierung von Sprachgrundfrequenzverläufen: Analyse, Synthese und Evaluation. Dissertation, Institut für Kommunikationsakustik, Ruhr-Universität Bochum, Shaker Verlag, Aachen.
- O'Shaughnessy, D. (2000). Speech Synthesis. In: Speech Communication: Humans and Machines. IEEE Press, New York NY, 337-366.
- Vary, P., Heute, U., Hess, W. (1998). Sprachsynthese. In: Digitale Sprachsignalverarbeitung, B.G. Teubner, Stuttgart, 465-497.

7.3 Natürlichsprachliche Dialogsysteme

Sprachdialogsysteme erlauben eine natürlichsprachliche Interaktion zwischen Mensch und Maschine. Sie stellen dabei meist eine Schnittstelle her zwischen dem Menschen auf der einen Seite und einem Anwendungssystem (z.B. einer Datenbank) auf der anderen. Diese Schnittstelle muss zwei Arten von Informationen verwalten: Diejenigen, die mit dem Benutzer ausgetauscht werden (natürlichsprachliche Information) als eine Art *voice user interface* (VUI, analog zu GUI), und diejenigen, die mit dem Anwendungssystem ausgetauscht werden (z.B. SQL-basiert). Daneben gibt es ein weiteres Szenario, nämlich ein Sprachdialogsystem, das zwischen zwei kommunizierenden Menschen steht, bspw. als Übersetzungssystem. Die folgenden Betrachtungen beschränken sich auf das erste Szenario, auch wenn viele Vorgehensweisen für beide Szenarien gleich sind.

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.

Diese Aufgaben können i.a. nicht in einem Schritt gelöst werden, sondern erfordern mehrere Äußerungen auf beiden Seiten – vom Benutzer und vom System.

Je nach Komplexität der Interaktion können folgende Klassen von Dialogsystemen unterschieden werden:

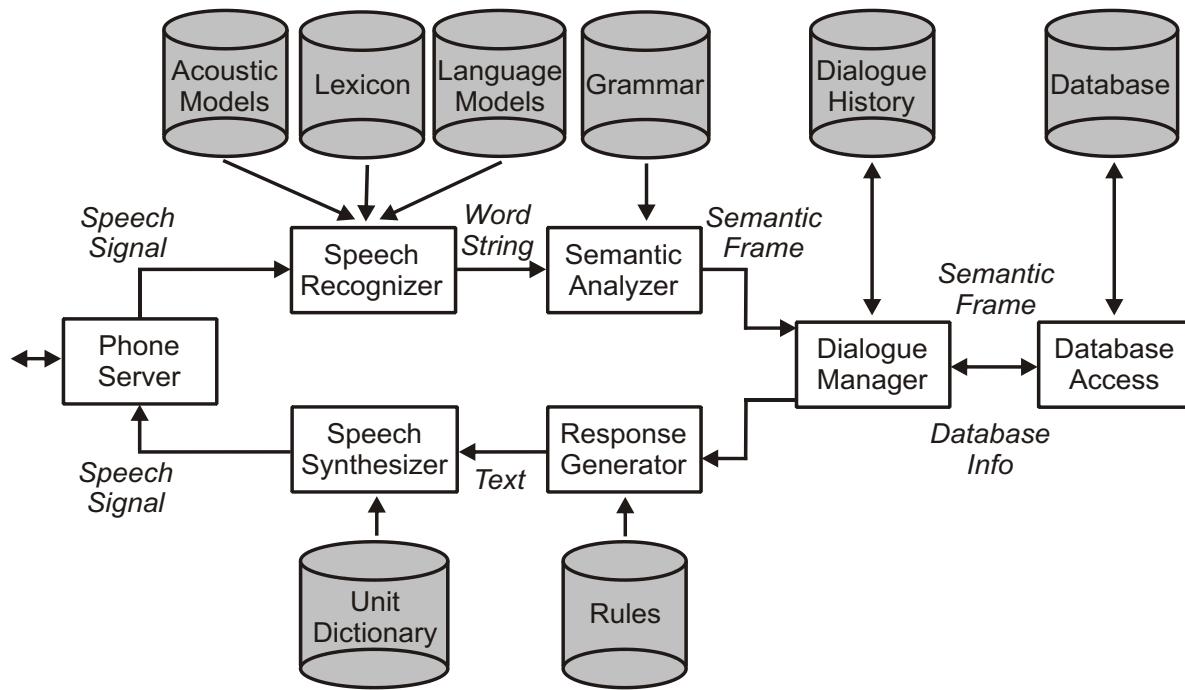
- *Kommandosysteme*: Diese sind durch eine direkte und deterministische Reaktion des Systems gekennzeichnet, d.h. jeder Äußerung des Benutzers entspricht genau eine Reaktion des Systems – ohne die Einbeziehung von Vorwissen. Beispiel: Das Drücken einer Taste erzeugt ein bestimmtes Symbol auf dem Computerbildschirm.
- *Menü-orientierte Systeme*: Diese bieten dem Benutzer an jeder Verzweigungsstelle ein explizites Menü an; dadurch wird der Dialog stark vom System bestimmt. Im Gegensatz zu Kommandosystemen können mehrere Äußerungen notwendig sein, um ein kommunikatives Ziel zu erreichen. Die Interaktion muss nicht rein natürlichsprachlich sein; bspw. fallen auch sog. *interactive voice response* (IVR)-Systeme, bei denen die Eingaben des Benutzers durch die Tastatur eines Telefons (Tonwahl, DTMF) erfolgt, in diese Klasse.
- *Sprachdialogsysteme*: Diese Systeme sollen im Folgenden behandelt werden. Sie verfügen über eine Spracherkennung, eine sprachverstehende Komponente, eine Dialogsteuerung, eine Schnittstelle zum Anwendungssystem, eine Komponente zur Generierung der Systemantwort, und die eigentliche Sprachausgabe.
- *Multimodale Dialogsysteme*: Diese Systeme benutzen neben der Sprache noch eine oder mehrere weitere Modalitäten, bspw. Eingabe über einen Touchscreen oder über Gestenerkennung, Ausgabe von Videos und Grafiken, etc. Näheres hierzu findet sich in Kapitel 8.

Sprachdialogsysteme werden heute meist über das Telefonnetz betrieben. Sie bieten Dienste an wie eine automatische Auskunft, eine Abfrage von Zugfahrplänen oder Flugzeiten, Informationen über touristische Ziele, Reservierung von Fahrkarten oder Hotelzimmern, etc. Neben dem Telefon ist aber natürlich auch eine direkte Interaktion zwischen Benutzer und System möglich, bspw. als „Kiosk“ auf einem Bahnhof, einem Flughafen oder einem Fremdenverkehrsbüro, oder als Navigationssystem in einem Kraftfahrzeug.

Solche Systeme können neben der gesprochenen Sprache auch weitere Modalitäten anbieten. Die Wahl der „richtigen“ *Modalität* wird teilweise durch die Informationen bestimmt, die zwischen Benutzer und System ausgetauscht werden können, und teilweise durch die Umgebung, in der die Interaktion stattfindet. So können lange Listen von Ortsnamen in einem Navigationssystem schlecht vorgelesen werden, ohne dass der Benutzer gelangweilt wird und die wichtigsten Informationen verpasst; diese sollten dann besser über ein Display angezeigt werden. Allerdings ist das nicht wünschenswert, wenn der Benutzer gleichzeitig Auto fährt.

7.3.1 Struktur eines Sprachdialogsystems

Die Funktionsweise eines Sprachdialogsystems lässt sich am besten an einer sequentiellen Struktur erläutern. Diese ist in folgender Abbildung gezeigt.

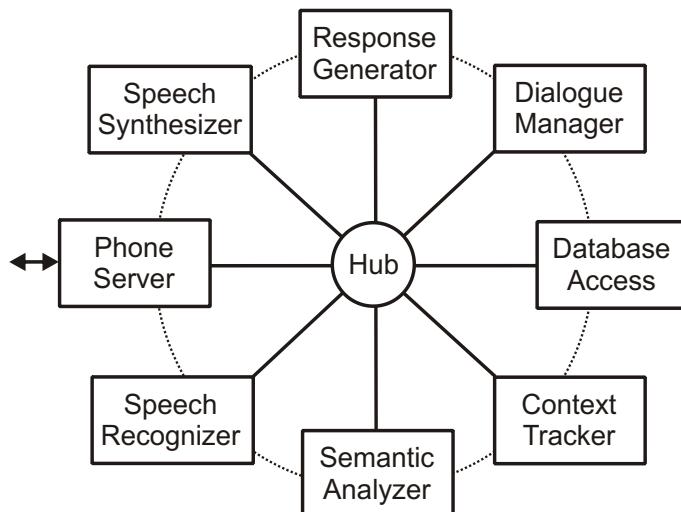


Sequentielle Struktur eines telefonbasierten Sprachdialogsystems,
ähnlich Lamel et al. (2000).

Die vom Benutzer kommende Sprache wird über das Telefon-Interface an den Spracherkenner weitergeleitet. Dieser generiert daraus orthographischen Text, wozu das akustische Modell, das Sprachmodell und das Vokabular (Lexikon) benötigt werden. Der orthographische Text wird anschließend einer sprachverstehenden Einheit zugeführt. Diese versucht, aus dem Text diejenigen Informationen zu extrahieren, die für den Dialog wichtig sind. Die Informationen werden z.B. als semantische Attribut-Wert-Paare (sog. slots) abgelegt. Sie dienen zur Steuerung des Dialogablaufes im Dialog-Manager.

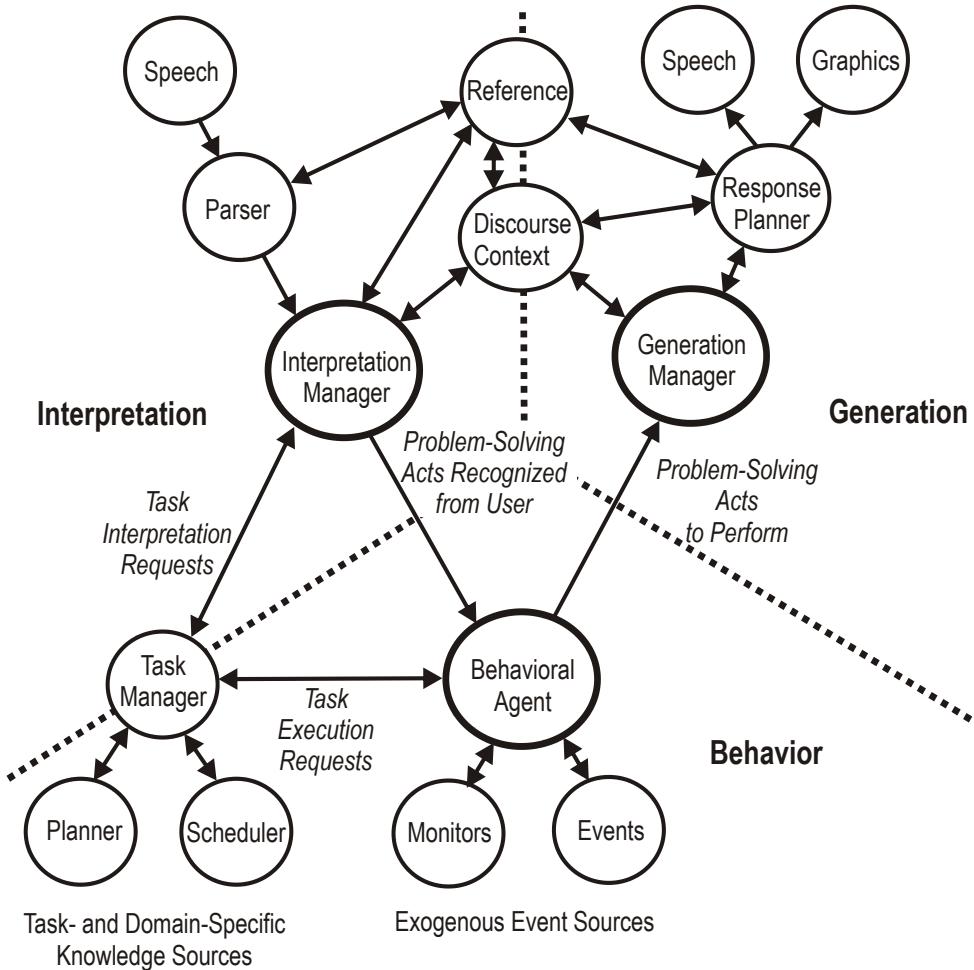
Der Dialog-Manager ist das Herzstück des Systems. Er bestimmt den Ablauf der Interaktion basierend auf den (interpretierten) Benutzeräußerungen und den Ergebnissen des Anwendungssystems (hier: Datenbank). Daraus werden adäquate Systemäußerungen generiert, bspw. um Rückfragen zu stellen, um Missverständnisse aufzuklären, oder um letztendlich die gewünschten aufgabenbezogenen Informationen an den Benutzer auszugeben. Die Antwort kann zunächst in Textform generiert und anschließend mit einer Sprachsynthese vorgelesen werden, oder es wird direkt auf vorher aufgezeichnete Äußerungen zurückgegriffen (canned speech). Das so erzeugte Sprachsignal wird über das Telefon-Interface an den Benutzer ausgegeben.

Dieser prinzipielle Ablauf kann auf unterschiedliche Arten implementiert werden. Eine ist die schon gezeigte sequentielle Struktur. Eine andere Struktur verwendet ein sog. *Hub* (engl. für Nabe) zur zentralen Verwaltung der Informationen. Auf dieses Hub können unterschiedliche Module (Telefon-Interface, Spracherkennung, Sprachverstehen, Kontext-Tracker, Schnittstelle zum Anwendungssystem, Dialogmanager, Antwortgenerierung und Sprachsynthese) jederzeit zugreifen und Informationen abfragen oder eintragen. Die Informationen können dabei in einer Art *blackboard* gespeichert werden. Die Struktur ist im Folgenden illustriert.



Hub-Struktur eines telefonbasierten Sprachdialogs, vgl. Seneff (1998) und Zue et al. (2000).

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. Die untenstehende Architektur des sog. TRIPS-Systems (Allen et al., 2001) implementiert diese Idee. Informationen werden zwischen den Modulen über einen zentralen Hub ausgetauscht, der die Informationen weiterleitet, mitschreibt und die Syntax überprüft.



Asynchrone TRIPS-Architektur, vgl. Allen et al. (2001).

Im Folgenden sollen die wichtigsten Komponenten des Sprachdialogsystems kurz vorgestellt werden. Dabei werden die Antwortgenerierung und die Sprachsynthese zusammengefasst, da eine Vollsynthese aus Rechtschrift-Text oft nicht notwendig ist und diese beiden Module deshalb in vielen Dialogsystemen in einem einzigen Sprachausgabemodul vereint sind.

7.3.2 Spracherkennung

Das Prinzip der automatischen Spracherkennung wurde schon in Abschnitt 6.1 beschrieben. Für ein Sprachdialogsystem, welches von einer prinzipiell unbegrenzten Zahl von Benutzern frequentiert wird (bspw. eine Fahrplanauskunft), ist es notwendig, einen sprecherunabhängigen Erkenner für begrenzten Wortschatz zu verwenden. Dieser beruht meist auf HMMs oder auf einer Kombination von neuronalen Netzen (zur Bestimmung der Phonemwahrscheinlichkeiten) und HMMs (zur Dekodierung). Die Merkmale sollten robust gegenüber typischen Störungen wie dem Einfluss des Telefonkanals, der Benutzung unterschiedlicher Telefon-Endgeräte (Handys, Freisprecher mit unterschiedlichen akustischen Eigenschaften) sowie gegenüber Hintergrundgeräuschen sein (Anruf aus einem lauten Büro, einer Bahnhofshalle, einem fahrenden Auto).

Zur Bestimmung des Vokabulars und des Sprachmodells bzw. der Grammatik ist es notwendig, Wissen über die spezielle Anwendung und die dabei vorkommenden Benutzeräußerzungen zu bekommen. Bei einem Dienst, der auch ohne Sprachdialogsystem

funktioniert (bspw. einer Bahnauskunft) kann dies durch Aufzeichnung typischer Mensch-zu-Mensch-Dialoge geschehen. Allerdings sollte dabei berücksichtigt werden, dass sich der Mensch in einer Interaktion mit einem System u.U. anders verhält, als er dies in der zwischenmenschlichen Kommunikation tut. Bspw. passen viele Menschen ihr Vokabular an, sprechen langsamer und deutlicher, oder sie sind gehemmt und sprechen nur ungern mit einem System. Deshalb ist es vorteilhaft, Vokabular und Sprachmodell aus Interaktionen mit einem prototypischen System zu sammeln. Dabei kann der fehlende Spracherkenner (und u.U. weitere fehlende Module) zunächst durch einen transkribierenden Helfer ersetzt werden; man bezeichnet ein solches Vorgehen als *Wizard-of-Oz-Experiment*.

Sprachdialogsysteme müssen in der Realität mit einer Reihe von Benutzeräußerungen zureckkommen, die in einem begrenzten Trainingskorpus zunächst nicht enthalten sind. Deshalb muss das Sprachmodell und/oder die Grammatik offen gegenüber Äußerungen sein, die nicht trainiert wurden. Dies kann z.B. durch Verwendung von Kategorien in Grammatiken geschehen, bspw. für Ortsnamen oder für Uhrzeiten.

Die Erkennungsleistung kann verbessert werden, wenn in jeder Dialogsituation (Dialogzustand, vgl. Abschnitt 7.3.4) ein spezielles Vokabular und Sprachmodell verwendet wird. Bspw. kann auf eine Frage des Systems, die mit „ja“ oder „nein“ zu beantworten ist, mit großer Wahrscheinlichkeit eine positive oder negative Rückmeldung des Benutzers erwartet werden. Allerdings kann es dann passieren, dass zusätzlich gelieferte Informationen nicht erkannt werden. Beispiel:

System: Sie möchten also nach München fahren?

Benutzer: Nein, nach Münster.

Wenn nun das Vokabular und das Sprachmodell nur auf positive oder negative Rückmeldungen begrenzt ist könnte der Dialogablauf unnötig verlängert werden.

7.3.3 Sprachverstehen

Obwohl eine Maschine natürlich nicht im kommunikativen Sinne „verstehen“ kann, so wird dieser Begriff im Folgenden trotzdem (in Anlehnung an das im Englischen verwendete *speech understanding*) benutzt, um die Interpretation der relevanten Informationen durch das System zu beschreiben.

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. Die Ausgabeform ist dabei häufig durch die Aufgabe vorgegeben, die mit Hilfe des Dialogsystems erledigt werden kann (bspw. Bahnauskunft). Diese kann z.B. mittels *Attribut-Wert-Paaren (slots)* beschrieben werden:

Abfahrtsort: Bochum Hbf

Zielort: Münster Hbf

Abfahrtstag: Sonntag, 1. Mai 2005

Abfahrtszeit: 11 Uhr vormittags

Die Identifikation dieser Informationen erfordert eine syntaktische Analyse (zur Bestimmung der Funktion der einzelnen Worte im Satz), eine semantische Analyse (zur Bestimmung der Bedeutung der Worte), sowie eine kontextuelle Analyse. Die syntaktische und semantische

Analyse kann z.B. mittels eines *Parser*s durchgeführt werden, der einen erkannten Satz nach seinen grammatischen Bestandteilen durchsucht, diese etikettiert, und u.U. weitere semantische Informationen (z.B. semantische Klassen) liefert. Die Ausgabe des Parsers kann dann direkt zur Zuordnung zwischen Attributen und Werten (Füllen der slots) genutzt werden.

Eine kontextuelle Analyse ist notwendig, um die aktuelle Benutzeräußerung im Kontext des gesamten Dialogs zu interpretieren, sowie um Weltwissen in die Interpretation einzubringen. Bspw. kann die Äußerung

Benutzer: Münster.

sich sowohl auf einen Abfahrtsort als auch auf einen Zielort beziehen, je nachdem, welche Frage des Systems voraus ging. Äußerungen wie

Benutzer: Nächsten Sonntag um drei Uhr.

müssen mit Hilfe eines Kalenders in eine Datumsangabe umgesetzt werden (das aktuelle Jahr wird z.B. vorausgesetzt), und die Angabe „3 Uhr“ wird wahrscheinlich zunächst als „15 Uhr nachmittags“ interpretiert, solange der Benutzer sich nicht gegenteilig äußert.

Aufgrund der Natur spontaner Sprache ist eine komplette grammatische 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. Allerdings können dabei Bezüge zwischen Wörtern nur sehr umständlich abgebildet werden (z.B. „von Bochum nach Münster“).

Die sequentielle Struktur, die zu Beginn dieses Kapitels aufgezeigt wurde, geht von komplett getrennten Modulen zur Spracherkennung und zum Sprachverständigen aus. Dies ist aber nicht unbedingt vorteilhaft. Bspw. kann ein Spracherkennender mehrere Hypothesen liefern, unter denen dann erst nach Ergebnis des Sprachverständigen – auf Basis einer Gesamtwahrscheinlichkeit von Erkennung und Verstehen – eine definitive Entscheidung getroffen wird.

7.3.4 Dialogmanagement

Obwohl die Interaktion zwischen Benutzer und System nicht unbedingt den Regeln eines zwischenmenschlichen Dialogs folgt, wird sie i.a. als Dialog bezeichnet. Ein Dialog gliedert sich in verschiedene Unter-Dialoge, welche jeweils eine bestimmte kommunikative Funktion haben: Unter-Dialoge zum Austausch einer aufgabenbezogenen Information sind i.a. stark von der jeweiligen Aufgabe abhängig, Unter-Dialoge zu allgemeinen Aspekten des Gesprächs (Begrüßung, Beendigung des Gesprächs) sind oft unabhängig von der speziellen Aufgabe.

Insbesondere muss das Dialogsystem auch zu *Meta-Kommunikation*, d.h. zu *Kommunikation über Kommunikation*, fähig sein. Meta-Kommunikation ist immer dann notwendig, wenn Missverständnisse auftreten und korrigiert werden müssen. Wegen der stark eingeschränkten Fähigkeiten eines Dialogsystems, Sprache zu erkennen und zu verstehen, ist Meta-Kommunikation eine der wichtigsten Fähigkeiten eines robusten Dialogmanagers.

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,

da sich der Austausch von Informationen meist über eine Anzahl von Äußerungen (von Benutzer und System) verteilt.

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.

Darüber hinaus stellt der Dialogmanager die Schnittstelle zum Anwendungssystem dar. Die Informationen, die das Anwendungssystem liefert, können ebenfalls zur Dialogsteuerung eingesetzt werden. Als Beispiel könnte der Dialogmanager zunächst nach Kriterien fragen, die den Suchraum einer Datenbank stark eingrenzen, oder er könnte unnötige Fragen weglassen (bspw. die Frage nach einem bestimmten Bahnhof, wenn eine Stadt nur einen Bahnhof hat).

Die geforderten Funktionen können auf unterschiedliche Arten implementiert werden. Churcher et al. (1997) unterscheiden drei grundlegende Strategien des Dialogmanagements:

- *Dialog-Grammatiken*: Dies ist ein einfacher Top-Down-Ansatz, der über einen Zustandsautomaten oder eine deklarative Grammatik implementiert werden kann. 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. Jedem Zustand kann ein eigenes Vokabular und eine eigene Grammatik zugeordnet werden. Dialog-Grammatiken sind relativ einfach zu implementieren und gut geeignet für stark strukturierte Aufgaben und Dialoge; sie stoßen jedoch bei komplexen Dialogen an ihre Grenzen.
- *Plan-basierte Ansätze*: Hierbei wird explizit versucht, einzelne aufgabenbezogene Dialogziele zu modellieren. Die Dialogziele können z.B. als Plan-Operatoren implementiert werden, welche den Dialogverlauf nach möglichen Zielen durchsuchen. Dabei ist es wichtig anzunehmen, dass der Benutzer und das System dieselben Ziele verfolgen, da der Dialogverlauf sonst in eine falsche Richtung führen kann. Plan-basierte Ansätze sind i.a. komplizierter als Dialog-Grammatiken, aber sie können auch komplexere Dialogphänomene berücksichtigen.
- *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. Dabei werden die Annahmen (beliefs) beider Gesprächspartner modelliert und abgeglichen; von Benutzer und System akzeptierte Ziele werden als gemeinsame Annahmen abgespeichert.

Obwohl plan-basierte und kollaborative Ansätze allgemeine Dialogphänomene besser modellieren können, sind die Chancen, dass der Dialog in eine unvorhergesehene Richtung läuft und nicht zum Ziel führt, doch deutlich größer als bei Dialog-Grammatiken.

Um den Dialogverlauf verfolgen zu können benutzt ein Dialogmanager eine Reihe von Speichern und Modellen (hier mit ihren englischen Bezeichnungen angegeben):

- *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.

Diese Speicher oder Modelle können in expliziter Form vorliegen, oder sie sind implizit im Dialogmanager implementiert. Bspw. sind bei einem Zustandsautomaten das domain model und das conversational model normalerweise in den Zuständen und ihren möglichen Übergängen festgelegt.

Eine Hauptaufgabe des conversation model ist die Steuerung der Initiative zwischen Benutzer und System, d.h. welcher Gesprächspartner als Nächster an der Reihe ist. Hierbei kann unterschieden werden zwischen

- *System-Initiative*, d.h. die Initiative bleibt beim System und die Aufgabe des Benutzers ist es, Fragen zu beantworten,
- *Benutzer-Initiative*, 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.

Zwischenmenschliche Dialoge lassen normalerweise eine gemischte Initiative zu, und deshalb wird versucht, dies auch bei Sprachdialogsystemen nachzubilden. Allerdings wird der Dialogablauf – je mehr Initiative vom System zum Benutzer übergeht – zunehmend unvorhersagbarer; damit steigen leider auch die Chancen, dass die Aufgabe nicht erfüllt werden kann und der Dialog ohne Erfolg endet. Deshalb sind Systeme, die die Initiative rein beim Benutzer belassen, bislang hauptsächlich zu Forschungszwecken implementiert worden.

Da die Erkennung und das Verständnis eines Systems immer begrenzt sind, ist es wichtig, dass das System Rückmeldungen darüber gibt, was es verstanden hat. Dies kann auf folgende Weisen geschehen:

1. *Explizit*: Der Benutzer wird explizit gefragt, ob eine neu verstandene Information richtig oder falsch ist. Bsp.:
Benutzer: Nach Münster.
System: Sie möchten also nach Münster fahren?
2. *Implizit*: Das System verbindet die zuletzt verstandene oder interpretierte Information mit einer neuen Frage. Bsp.:

Benutzer: Nach Münster.

System: Wann möchten Sie nach Münster Hauptbahnhof fahren?

Eine explizite Rückmeldung erhöht die Anzahl der Äußerungen unnötig und führt nicht zu einem „natürlichen“ (d.h. mit einem zwischenmenschlichen Dialog vergleichbaren) Dialogablauf. Allerdings besteht bei der impliziten Rückmeldung die Gefahr, dass der Benutzer implizit gegebene Informationen überhört und das System danach von der Korrektheit dieser (überhörten) Informationen ausgeht. Auch wird eine implizite Bestätigung kompliziert, wenn Fehler (u.U. mehrere in einer Äußerung) korrigiert werden müssen.

Der Dialogmanager muss nicht unbedingt statisch sein, sondern er kann seine Strategie (Initiative, Rückmeldung, Hilfestellungen) dynamisch anpassen. Die Anpassung kann entweder global für einen speziellen Benutzer geschehen (bspw. kürzere Rückmeldungen und Erläuterungen für erfahrenen Benutzer), oder lokal anhand von Ereignissen im Dialogverlauf (bspw. die Verwendung einer expliziten Bestätigung, wenn die Erkennungsrate niedrig ist).

7.3.5 Sprachausgabe

Dies umfasst die Formulierung der linguistischen Antwort und die Generierung des entsprechenden Sprachsignals.

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. Die Auswahl der Informationen sollte z.B. die Überlegung umfassen, welche Informationen primär wichtig sind, und welche vielleicht nur auf Nachfrage gegeben werden sollten. Auch macht es keinen Sinn, lange Listen von Ergebnissen (z.B. Zugverbindungen) vorzulesen, die sich der Benutzer nicht merken können wird. Hierbei müssen die kognitiven Grenzen der menschlichen Verarbeitung insgesamt wie auch die Grenzen den akustischen (Sprach-) Kanals beachtet 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.

Die Art der Generierung des Sprachsignals hängt stark von der Aufgabe ab. So kann eine Bahnauskunft zunächst mit vorgefertigten und vorher aufgezeichneten Segmenten von natürlicher Sprache auskommen, die ohne Signalmanipulation verkettet werden. Allerdings ist dies problematisch, wenn das System aktualisiert werden soll (z.B. neue Bahnhofs- und Zugbezeichnungen) und der ursprünglich verwendete Sprecher nicht mehr zur Verfügung steht. Bei Anwendungen mit prinzipiell unbegrenztem Wortschatz (z.B. E-Mail-Vorlesesystem) kann auf eine Vollsynthese nicht verzichtet werden. Hier ist u.U. auch eine multilinguale Synthese mit automatischer Fremdsprachenerkennung vorteilhaft.

7.3.6 Beispiele

In den vergangenen 15 Jahren wurden eine Reihe von Sprachdialogsystemen entwickelt und zum Teil auch kommerziell vermarktet. Die Systeme sind meist aufgabenorientiert, d.h. sie sind nur zur Lösung einer speziellen Aufgabe – in einer eng umgrenzten Domäne – gedacht.

In diesem eng umgrenzten Gebiet können recht gute Ergebnisse erzielt werden. Typische Aufgaben umfassen z.B. die Folgenden:

- Zug-, Flug- und Bus-Informationen
- Touristische Informationen (Hotels, Restaurants, Kino, Sehenswürdigkeiten)
- Wettervorhersage
- Telefonauskunft oder Rückwärts-Auskunft
- Portoauskunft
- Automatische Weiterverbindung (auch „How may I help you?“-Systeme)
- Messaging-Systeme
- E-Mail-Zugang über das Telefon
- Telefon-Banking
- Bestellsysteme
- Konferenzanmeldung
- Kleinanzeigen-Verwaltung
- Kooperatives Problem-Lösen (z.B. circuit fix-it shop)
- Umfragen (Census)
- Sprache-nach-Sprache-Übersetzungssysteme

Literaturreferenzen finden sich in Möller (2005).

Einige dieser Systeme können auch aus dem öffentlichen Netz angerufen werden, z.B.

- Scansoft Bahnauskunft: +49 (0)241 604020
- MIT-Wetterauskunft: +1 617 258 0300
- Deutsche Bahn Fahrplanauskunft: +49 (0)800 1507090
- Berti-Infoportal zur 1. Bundesliga: +49 (0)9131 610017
- Viasuisse Reiseinformationen: +41 900 400500

7.3.7 Literatur

- Allen, J., Ferguson, G., Stent, A. (2001). An Architecture for More Realistic Conversational Systems. In: Proc. Of Intelligent User Interfaces 2001 (IUI-2001), Beijing, 2, 118-121.
- Bernsen, N.O., Dybkjær, H., Dybkjær, L. (1998). Designing Interactive Speech Systems: From First Ideas to User Testing. Springer, Berlin.
- Churcher, G.E., Atwell, E.S., Souter, C. (1997). Dialogue Management Systems: A Survey and Overview. Report 97.06, School of Computer Studies, University of Leeds, Leeds.
- Lamel, L., Minker, W., Paroubek, P. (2000). Towards Best Practice in the Development and Evaluation of Speech Recognition Components of a Spoken Language Dialogue System. Natural Language Engineering 6(3-4), 305-322.
- McTear, M.F. (2002). Spoken Dialogue Technology: Enabling the Conversational Interface. ACM Computing Surveys 34(1), 90-169.
- McTear, M.F. (2004). Spoken Dialogue Technology: Toward the Conversational User Interface. Springer, London.
- Möller, S. (2005). Quality of Telephone-Based Spoken Dialogue Systems. Springer, New York NY.
- Seneff, S. (1998). Galaxy-II: A Reference Architecture for Conversational System Development. In: Proc. 5th Int. Conf. On Spoken Language Processing (ICSLP'98), Sydney, 3, 931-934.

Zue, V., Seneff, S., Glass, J.R., Polifroni, J., Pao, C., Hazen, T.J., Hetherington, L. (2000).
JUPITER: A Telephone-Based Conversational Interface for Weather Information. IEEE
Trans. Speech and Audio Proc. 8(1), 85-96.

8. Multimodale Dialogsysteme

Bislang haben wir uns auf rein sprachbasierte Systeme beschränkt. In der Tat bieten solche Systeme einige prinzipielle Vorteile gegenüber einer traditionellen Mensch-Computer-Interaktion mittels Maus und Tastatur. So ist Sprachinteraktion häufig schneller in der Informationsübermittlung, und sie ist expressiver – man kann also z.B. Ideen, Gefühle, Gemütszustände, Einstellungen und Informationen über den Sprecher sehr viel genauer übermitteln. Darüber hinaus lässt sich eine Interaktion mittels Sprache durch die effiziente Übertragung (vgl. Kapitel 6) von praktisch jedem Ort aus durchführen; man erreicht also praktisch die gesamte Bevölkerung des Globus, sofern sie Zugang zu einem Telefon hat. Außerdem belegt Sprache nur einen Interaktionskanal und lässt bspw. die Hände und Augen unbelastet; Sprache kann also gut als Interaktionsmodalität eingesetzt werden, wenn andere (parallele) Aufgaben erfüllt werden müssen, bspw. im Kraftfahrzeug.

Trotzdem ist eine rein sprachbasierte Interaktion eingeschränkt. 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. Insbesondere die Leistung von Spracherkennung und Sprachverstehen bleibt häufig noch sehr stark hinter der menschlichen Interpretationsleistung zurück. Aber auch das Dialogmanagement kann nicht alle Funktionen abdecken, die ein menschlicher Kommunikationspartner im Gespräch abdeckt. Der serielle, akustische Kanal hat auch prinzipielle Grenzen: Wird die Anzahl der zu übermittelnden Informationen zu groß (wie bspw. bei der Ausgabe verschiedener Zugverbindungen), so bietet sich eher eine strukturierte Darstellung mittels einer Liste an. Damit lässt sich die Information i. Allg. sehr viel schneller von Zuhörer (Leser) aufnehmen.

Ein multimodales System kann verschiedene Modalitäten kombinieren und somit eine fast optimale Interaktion zwischen Mensch und Maschine zulassen. Ähnlich wie ein Mensch z.B. auf einen Gegenstand zeigt und gleichzeitig „den möchte ich gern haben“ sagt, so kann durch Kombination unterschiedlicher Modalitäten Ein- und Ausgabe von Informationen effektiv, effizient und intuitiv gestaltet werden.

Dabei verstehen wir unter dem Begriff „Medium“ ein Kommunikationsmittel (Material oder Gerät), welches einen bestimmten physikalischen (z.B. akustischen, optischen) 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 (Fühlen; umfasst die taktile Wahrnehmung/ Oberflächensensibilität, die kinästhetische Wahrnehmung/ Tiefensensibilität, die Temperaturwahrnehmung sowie die Schmerzwahrnehmung). Unterschiedliche Modalitäten eignen sich unterschiedlich gut für verschiedene Zwecke der Informationsübermittlung; dies ergibt sich teilweise aus ihren prinzipiellen Eigenschaften, welche in Kapitel 8.1. angerissen werden. Darüber hinaus sind aber auch Konventionen zu beachten, wenn multimodale Dialogsysteme – als vierte Klasse der Dialogsysteme, welche bereits in Kapitel 7.3. eingeführt wurden – gestaltet werden sollen. In Kapitel 8.2. wird die prinzipielle Architektur solcher Systeme vorgestellt. Kapitel 8.3 bis 8.5. beschreiben Details derjenigen Komponenten solcher Systeme, die nicht bereits in Kapitel 7 erfasst wurden, die sich also nicht auf die Modalität „Intonation“ oder vereinfacht „Sprache“ beziehen. Kapitel

8.6. schließt mit ausgewählten Beispielen multimodaler Systeme. Die folgenden Ausführungen sind zumeist López-Cózar Delgado und Araki (2005) entnommen.

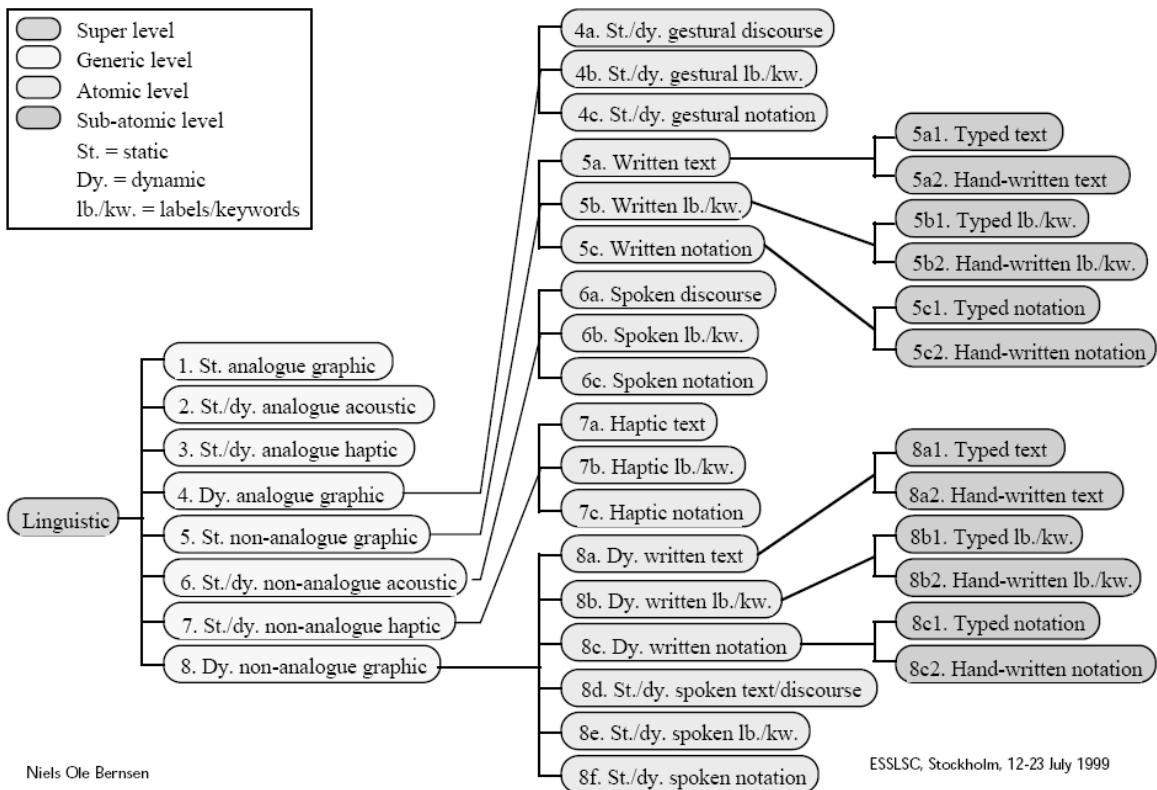
8.1 Eigenschaften von Modalitäten

Um multimodale Systeme gestalten und evaluieren zu können bedarf es Wissens über die Eigenschaften von Modalitäten, sowie über ihr Zusammenspiel. Obwohl bislang keine umfassende Theorie der Modalitäten verfügbar ist, so gibt es jedoch einige interessante Versuche, Eingabe- und Ausgabe-Modalitäten in Ihren Eigenschaften zu beschreiben, und dieses Wissen für das Design multimodaler Systeme – z.B. auch in Form von Tools – nutzbar zu machen.

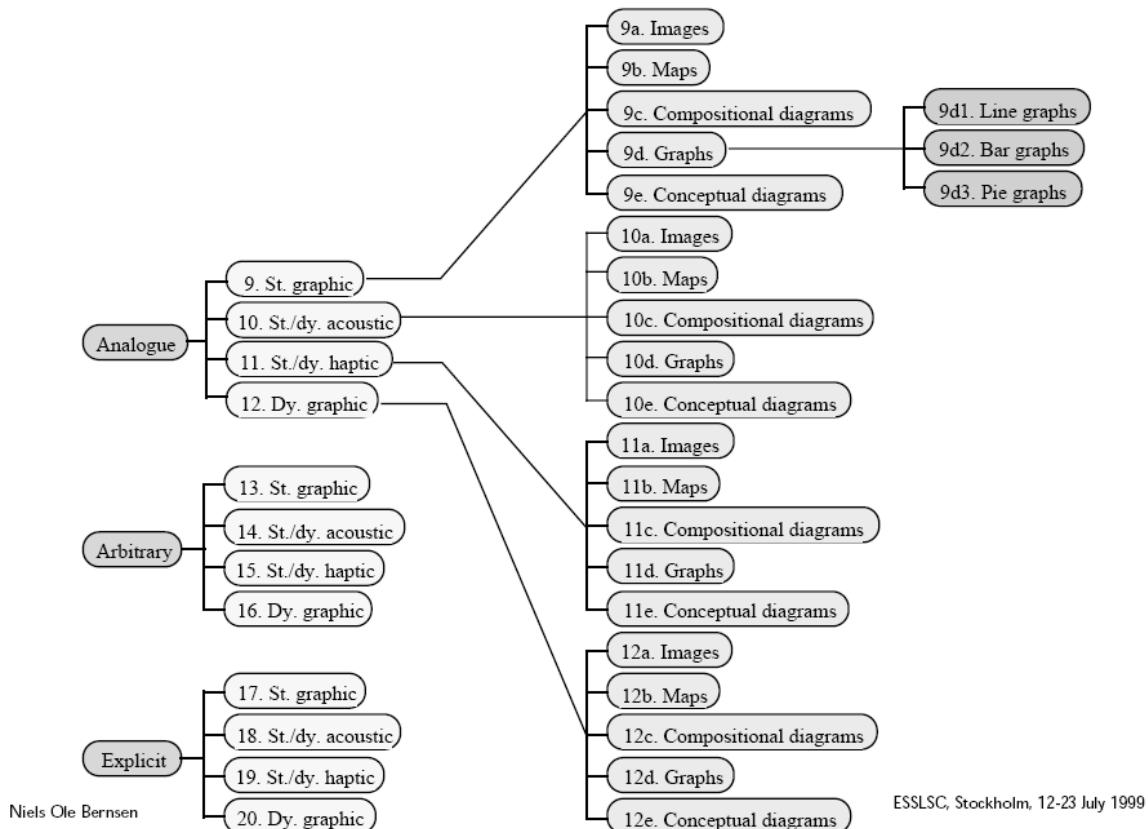
Zur Klassifikation von Ausgabe-Modalitäten verwendet Bernsen (1999) z.B. die 5 Merkmale

- Linguistisch vs. nicht-linguistisch: Linguistische Modalitäten beruhen auf einem syntaktisch-semantisch-pragmatischen System von Bedeutungen. Beispiele hierfür sind z.B. geschriebener Text oder gesprochene Sprache
- Analog vs. nicht-analog: Analoge Modalitäten beruhen auf einer Ähnlichkeit zwischen dem Bezeichner und dem Bezeichneten; man bezeichnet sie deshalb auch als ikonische Modalitäten. Beispiele für analoge Modalitäten sind Bilder und Diagramme.
- Arbiträr vs. nicht-arbiträr: Nicht-arbiträre Modalitäten beruhen auf einem bestehenden System von Bedeutungen, arbiträre nicht.
- Statisch vs. dynamisch: Statische Modalitäten können vom Benutzer prinzipiell in beliebiger Reihenfolge und für eine beliebige Dauer wahrgenommen werden, dynamische nicht.
- Klasse von Medien: Grafisch (visuell wahrnehmbar), akustisch (auditiv wahrnehmbar) oder haptisch.

Mit Hilfe dieser Merkmale ist man in der Lage, eine Vielzahl gebräuchlicher Modalitäten zu klassifizieren. Die folgenden beiden Abbildungen zeigen eine Klassifikation der linguistischen und der weiteren Modalitäten auf verschiedenen Ebenen.



Klassifikation der linguistischen Modalitäten. Aus Bernsen (1999).



Klassifikation der analogen, arbiträren und expliziten Modalitäten. Aus Bernsen (1999).

Eine ähnliche Taxonomie lässt sich auch für die Eingabeseite aufstellen. Hier ist die Unterscheidung statisch/dynamisch weniger angebracht, und die haptischen Modalitäten sind stärker differenziert.

Den so klassifizierten Modalitäten können verschiedene Eigenschaften zugeordnet werden. Diese Eigenschaften lassen bestimmte Modalitäten prinzipiell geeignet oder weniger geeignet für den Austausch von bestimmten Informationen erscheinen. In der folgenden Tabelle (aus Bernsen, 2002, unvollständig) sind einige Modalitäts-Eigenschaften aufgelistet, welche sich für sprachbasierte und multimodale Systeme als relevant herausgestellt haben.

No.	<i>Modality</i>	<i>Modality Property</i>
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.
...

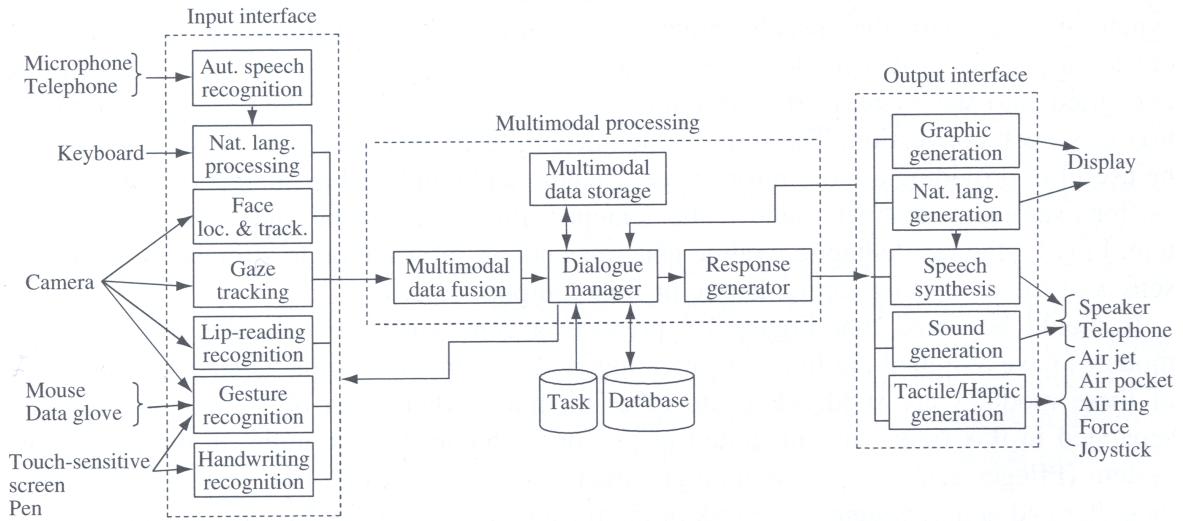
Diese Modalitäts-Prinzipien geben eine Hilfestellung, welche bekannten Modalitäten prinzipiell – unter Beachtung der Aufgabe und des Nutzungskontextes – geeignet sein könnten. Allerdings gelten sie nur für einzelne Modalitäten, nicht für Kombinationen derselben. Insbesondere in der Kombination steckt aber viel Potential.

8.2 Allgemeine Architektur eines multimodalen Dialogsystems

Ein multimodales Dialogsystem lässt eine Interaktion über mehrere verschiedene Modalitäten zu. Hierbei werden die vom Benutzer erhaltenen und die auszugebenden Informationen auf höheren Repräsentationsebenen zusammengeführt, sodass sich gegenseitige Abhängigkeiten ergeben. In diesem Punkt unterscheidet sich ein *multimodales* Dialogsystem von einem *multimedialen* System, bei dem diese Zusammenführung nicht stattfindet, sondern einfach parallele Modalitäten angeboten werden, die in keiner bestimmten Beziehung zueinander stehen.

In einem multimodalen Dialogsystem sind im Allgemeinen das Dialogmodell, das Aufgabenmodell, 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. Ein typisches Beispiel ist z.B. der Benutzer eines Naviagationssystems, welcher nacheinander auf zwei Punkte einer auf dem Touchscreen dargestellten Karte zeigt und sagt: „Wie komme ich von hier nach hier?“. Dabei werden die Informationen über die beiden (Start- und Ziel-) Punkte durch die Zeige-Geste auf dem Touchscreen eingegeben; die Aufgabe sowie die Zuordnung der Punkte zu Start und Ziel wird durch die Spracheingabe bestimmt, wobei es entscheidend ist, dass die Worte „hier“ synchron zu den Zeigegesten stattfinden, damit eine korrekte Interpretation – d.h. eine korrekte Zuordnung zu den Konzepten des Systems – geliefert werden kann. Auf der Ausgabeseite könnte ein multimodales Systems diese Eingabe z.B. mit einer Linie, die den Routenverlauf auf der Karte einzeichnet, sowie der gesprochenen Information „Der kürzeste Weg ist auf der Karte gelb markiert. Wenn Sie ‚Auto‘ sagen kann ich Ihnen alternativ auch eine für das Auto schnellste Version anzeigen.“ quittieren.

Unten stehende Abbildung zeigt eine allgemeine Architektur eines solchen multimodalen Systems. Es besteht prinzipiell aus drei Teilen: Der Eingabe-Schnittstelle, welche eine Reihe unterschiedlicher Eingabe-Modalitäten aufweisen kann; der multimodalen Verarbeitungseinheit, die u.U. auf ein externes Aufgabenmodell und/oder eine Datenbank zugreifen kann; sowie der Ausgabe-Schnittstelle, wiederum mit einer Reihe von Ausgabe-Modalitäten.



Allgemeine Architektur eines multimodalen Dialogsystems,
nach López-Cózar Delgado und Araki (2005).

Auf der Eingabeseite wurden Spracherkenner (*automatic speech recognition, ASR*) und die sprachverstehende Einheit bereits behandelt. Letztere könnte auch direkt Text als Eingabe akzeptieren und wird deshalb hier allgemein als Textverarbeitungs-Einheit (*natural language processing*) bezeichnet. Daneben finden sich Einheiten zur Lokalisierung und Verfolgung des Gesichtes im Kamerabild (*face tracking*), zur Blickbewegungserkennung (*gaze tracking* oder *eye tracking*), zur Erkennung der Lippen (*lip-reading recognition*), zur Erkennung von Gesten (*gesture recognition*), sowie zur Handschrift-Erkennung (*handwriting recognition*). Diese Modalitäten werden über unterschiedliche physikalische Medien bedient, bspw. über das Mikrofon (bei Sprache), über ein Keyboard (bei Texteingabe), über eine Kamera (bei Gesichts-, Blick-, Lippen- und Gestenerkennung), eine Maus oder einen Datenhandschuh (bei Gestenerkennung), oder durch ein berührungssensitive Oberfläche, die direkt durch Hautkontakt oder einen Stift (*stylus*) bedient wird.

Die genannten Eingabemodule stellen i. Allg. eine Vielzahl von Informationen bereit, welche anschließend korrekt interpretiert werden müssen. Dabei ist es wichtig, die einzelnen Informationskanäle nicht getrennt zu betrachten, sondern in ihrer zeitlichen und inhaltlichen Kombination. Nur so lassen sich die oben genannten Interaktionen realisieren. Die erhaltenen Informationen müssen also sinnvoll zusammengeführt werden; man nennt diesen Prozess Fusion (engl. *fusion*). Auf Basis der interpretierten Informationen muss – wie beim Sprachdialogsystem – der Dialogfluss gesteuert werden. Hierbei wird auf die Informationen, die durch die Eingabeschnittstellen bereitgestellt werden oder im Interaktionsverlauf bereits bereitgestellt wurden, sowie auf Informationen des Aufgabenmodells und/oder der Datenbasis zurückgegriffen. Informationen über den Dialogverlauf werden im multimodalen Datenspeicher (*multimodal data storage*) festgehalten. Der Dialogmanager entscheidet über den nächsten Dialogschritt, d.h. die „Antwort“ des Systems. Diese muss hier natürlich nicht unbedingt in Textform (durch die Textgenerierung, *natural language generation*) oder als Sprache (durch die Sprachsynthese, *speech synthesis*) ausgegeben werden. Zusätzlich oder alternativ können Grafiken, Bilder, Icons oder Videos generiert werden, oder auch ein animierter künstlicher Agent, ein sog. Avatar (hier in der Komponente *graphic generation* zusammengefasst), darüber hinaus auch nicht-sprachliche Audioausgaben, oder taktiler bzw. haptischer Ausgaben. Dazu werden neben einem Lautsprecher und einem Display Geräte zur Erzeugung von Kräften (bspw. über Luftstrom, elektrische oder magnetische Kraft, Bewegung von Massen, etc.) verwendet. 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.

Die Eingabe- und Ausgabemodule arbeiten nicht unbedingt unabhängig voneinander. In der Tat ist es vorteilhaft, wenn durch das Zusammenspiel verschiedener Modalitäten Vieldeutigkeiten aufgelöst werden. Bspw. kann der Dialogmanager aus dem Dialogzusammenhang ein bestimmtes Vokabular oder eine Grammatik in den Spracherkennern laden, oder aber verschiedene Modalitäten aktivieren oder deaktivieren. Umgekehrt kann auch durch das Ausgabemodul direktes Feedback zum Benutzer gegeben werden, noch während er die Eingabe bedient. Augenblickliches Feedback ist insbesondere bei haptischen Modalitäten sinnvoll.

Die beim Sprachdialogsystem vorhandenen Module wurden bereits in Kapitel 7.3. vorgestellt. In den folgenden drei Abschnitten beschränken wir uns deshalb auf die Erläuterung der darüber hinaus gehenden Module, zunächst auf der Eingabeseite (Kapitel 8.3), dann bzgl. der multimodalen Verarbeitung (Kapitel 8.4), und schließlich auf der Ausgabeseite (Kapitel 8.5). Eine Übersicht über existierende multimodale Systeme aus Forschung und Anwendung schließt die Ausführungen in Kapitel 8.6.

8.3 Multimodale Eingabe-Schnittstellen

Geräte für die multimodale Eingabe wurden bereits im vorangegangenen Kapitel kurz umrissen. Für die Interaktion mit Computern haben sich offensichtlich die Tastatur zur Eingabe größerer Informationsmengen und die Maus (bzw. Joystick, Trackball) zur Eingabe von Kommandos und Anweisungen als vorteilhaft erwiesen. Das sollte nicht darüber hinweg täuschen, dass diese Geräte meist nicht intuitiv sind, und ihre Bedienung zunächst erlernt werden muss, bis eine einigermaßen effiziente Benutzbarkeit gegeben ist. Je nach Benutzungskontext werden diese Geräte aber auch durch berührungssensitive Displays (*touchscreens*) ersetzt, bspw. bei mobilen Anwendungen. ipod und iphone sind dafür prominente Beispiele. In mobilen Geräten, insbesondere Mobiltelefonen, sind auch zunehmend Kameras eingebaut.

Die Funktionsweise von Spracherkennung und sprachverstehender Einheit wurde bereits in Kapitel 7.1 bzw. 7.3 beschrieben. 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. Die Lokalisation eines Gesichtes im Kamerabild wird durch eine Reihe von Einflussfaktoren erschwert: Die Positionierung und Neigung des Gesichtes zur Kamera variieren, ebenfalls die Beleuchtung, Kameraeinstellung, die Hautfarbe und der Gesichtsausdruck. Zusätzliche Merkmale wie Bärte, Kopfbedeckungen oder Brillen, oder auch andere im Raum befindliche Objekte decken Hautpartien ab. Dies führt zu einer großen Variabilität der Gesichtspartie im Kamerabild.

Zur Gesichtserkennung werden unterschiedliche Verfahren verwendet. Die vier wichtigsten sind:

- *Regelbasierter Ansatz*: Hierbei wird die Gesichtspartie durch einen Regelsatz beschrieben, der das „Wissen“ über das Aussehen eines Gesichtes umfasst. Dabei werden z.B. die relativen Positionen verschiedener Gesichtsmerkmale (Augen, Mund

Nase) über Regeln beschrieben. Der Algorithmus versucht zunächst, Kandidaten für diese Merkmale im Kamerabild zu finden, und dann anhand der Regeln zu überprüfen.

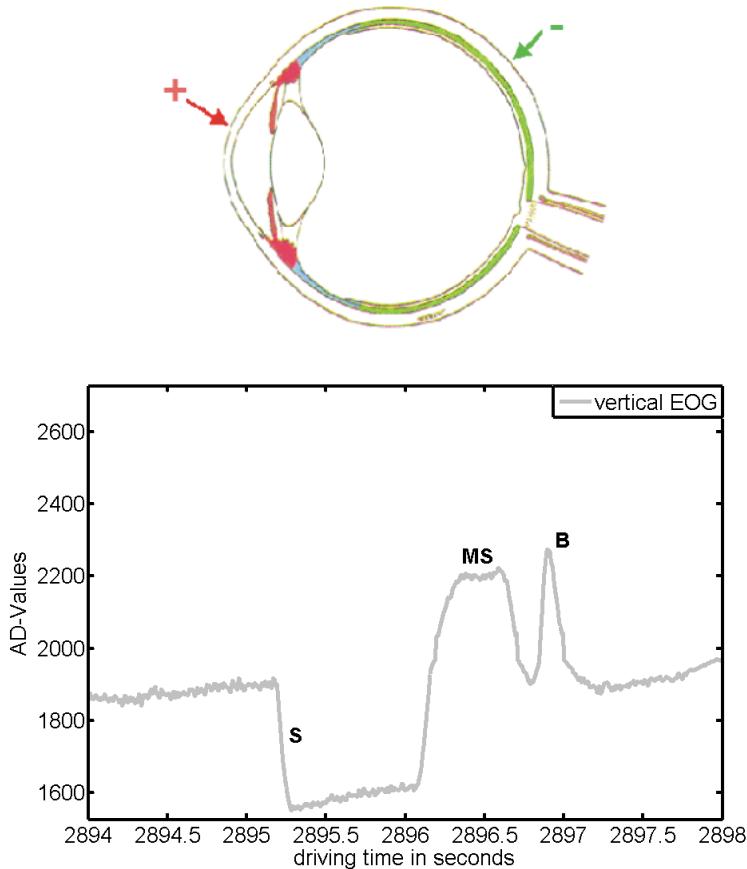
- *Invariante Merkmale*: Wie oben werden bei diesem Ansatz ebenfalls zunächst Merkmale extrahiert, z.B. durch Kantendetektion im Bild. Auf Basis dieser Merkmale wird dann ein statistischer Klassifizierer trainiert. Der Klassifizierer – wie auch die Merkmale – müssen robust gegen Änderungen der Position, Lichtverhältnisse, Deformationen etc. sein, um eine gute Erkennung zu ermöglichen.
- *Mustervergleich*: Wie zuletzt werden Muster aus dem Kamerabild extrahiert, entweder für einzelne Gesichtsmerkmale oder das gesamte Gesicht. Zwischen diesen Mustern und vorgeübten Prototypen werden anschließend Ähnlichkeiten (z.B. als Korrelationen) berechnet. Um die Muster robust zu machen kann man auf Teilmuster reduzieren oder Deformationen zulassen.
- *Farbe*: Interessanterweise hat sich die Farbe der menschlichen Haut als sehr hilfreich zur Gesichtserkennung (wie auch zur Erkennung von Handbewegungen und Gesten) erwiesen. Obwohl unterschiedliche Menschen in ihrer Hautfarbe variieren zeigt sich, dass die Unterschiede meist in der Luminanz (Helligkeit) und nicht in der Chrominanz (Farbart) liegen. Im Gegensatz zu den merkmalsbasierten Verfahren lässt sich die Farbinformation recht schnell gewinnen, und sie ist auch recht robust gegenüber Änderungen der Position und Beleuchtung. Bewegungen können ebenfalls schnell erfasst werden, denn es ist nur ein translatorisches Modell hierzu notwendig.

Diese Verfahren können natürlich auch kombiniert werden, um eine robuste Gesichtserkennung zu erzielen.

Sobald das Gesicht erkannt – und über einen längeren Zeitraum verfolgt, also „getrackt“ – ist, können Merkmale aus dem Gesicht extrahiert werden, welche mögliche Zusatzinformationen beinhalten. Dabei sind insbesondere die Erkennung der Augen (und daraus Blickrichtungen), der Nase und des Mundes (und daraus Lippeninformationen) interessant. Blickbewegungen spielen in der zwischenmenschlichen Kommunikation eine wichtige Rolle, bspw. beim Wechsel der Sprecher- / Hörerrolle. Durch die Blickrichtung wird der Aufmerksamkeitsfokus des Benutzers deutlich; dieser mag auf einem zu erkennenden Objekt liegen, oder auf einem Gesprächspartner (bei mehreren Personen im Raum). Ein abgewandter Blick mag darauf hindeuten, dass der Benutzer derzeit nicht mit dem System interagieren möchte. Die mit Hilfe von Blickbewegungen aufgezeichnete Information ist zumeist nicht vordergründig und beinhaltet normalerweise keine Kommandos; stattdessen ist sie gut zur Disambiguierung verwertbar, oder sie kann anstelle von Zeigegeräten (wie z.B einer Maus) verwendet werden.

Blickbewegungs-Detektoren arbeiten z.B. nach den folgenden Prinzipien:

- *Cornea-Reflex-Methode*: Beim Eintritt ins Auge wird ein Lichtstrahl teilweise an der Oberfläche der Hornhaut (Cornea) reflektiert. Zur Blickbewegungsmessung wird dazu ein künstlicher Lichtstrahl durch eine LED oder einen Laser erzeugt und dessen Reflexion gemessen. Hierzu ist allerdings zunächst ein aufwendiger Kalibrierungsprozess erforderlich.
- *Elektro-Okulogramme (EOG)*: Hierbei wird das elektrische Potenzial zwischen der Netzhaut und der Hornhaut oder um die Augenpartie herum gemessen. Die Funktionsweise und ein sich daraus ergebendes EOG ist in unten stehender Abbildung skizziert.



Blickbewegungsmessung mittels EOG. Oben: Prinzip; unten: Beispiel-EOG.
B=Blink (Lidschlag), S=vertikale Sakkade (Blicksprung), MS=Mikroschlaf (überlanger Lidschluss). Aus Schleicher et al. (2007).

- Kontaktlinsen:** Hierbei werden entweder teilweise verspiegelte Kontaktlinsen verwendet, deren Reflexionen dann von einer Kamera aufgezeichnet werden, oder es werden mit Spulen versehene Kontaktlinsen verwendet, deren Bewegung in einem Magnetfeld dann über die induzierte Spannung gemessen wird. Wie die Elektro-Okulogramme sind diese Methoden allerdings invasiv.

Darüber hinaus können die Blickrichtungen aus der relativen Position der Gesichtsmerkmale zueinander geschätzt werden; hierzu werden bspw. neuronale Netze verwandt. Blickbewegungs-Detektoren, die zur Mensch-Maschine-Interaktion eingesetzt werden sollen, werden entweder als Remote-Systeme oder als Head-Mounted-System konstruiert.

Neben den Augen ist vor allem die Erfassung des Mundes und seiner Sprechbewegungen ein wichtiges Hilfsmittel bei der multimodalen Interaktion. 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. Insbesondere Laute wie z.B. /d/ und /b/ bzw. /m/ und /n/ lassen sich auf rein akustischem Wege nur schlecht, mit Hilfe zusätzlicher visueller Information aber recht gut voneinander unterscheiden. Dabei wird zumeist ein viereckiger Bereich rund um die Mundpartie des Sprechers (auch inkl. des Kinns) ausgewertet, der zunächst mittels geeigneter Normierungen weitgehend unabhängig von Größe, Rotation und Lichteinfall gemacht wird. Aus dem normierten Bild können anschließend Merkmale

berechnet werden, die eine Identifikation von Visemen (den kleinsten bedeutungsunterscheidenden visuellen Korrespondenten der Phoneme) ermöglichen. Hierzu kann explizit die Form der Lippen extrahiert werden, oder es werden einzelne Punkte des Mundbereiches als Merkmale verwendet (bspw. Breite, Höhe und Fläche des von der Innenlinie der Lippen aufgespannten Bereiches).

Die anschließende Klassifikation der Merkmale erfolgt z.B. mit Hilfe von neuronalen Netzen. Dabei können die visuellen Informationen entweder auf Merkmalsebene (*feature fusion*) oder auf der Ebene der berechneten Wahrscheinlichkeiten (*decision fusion*) mit den entsprechenden akustischen Informationen (vgl. Kapitel 7.1) kombiniert werden. Bei der Fusion auf Merkmalsebene lassen sich natürlich vorkommende Asynchronitäten zwischen den akustischen und den visuellen Merkmalen leider nicht korrekt abbilden. Dies lässt sich durch Fusion der Wahrscheinlichkeiten lösen. Bspw. wurden gekoppelte HMMs für diesen Zweck entwickelt, welche die zeitliche Asynchronität explizit modellieren können. Man bezeichnet die Spracherkennung mittels akustischer und visueller Merkmale allgemein als *Audio-Visual Automatic Speech Recognition, AVASR*.

Neben dem Gesicht können aber auch durch andere Körperpartien Informationen ausgedrückt werden, bspw. durch Gesten. Im einfachsten Falle werden solche Gesten explizit mit einem speziellen Gerät erzeugt, bspw. bei der Maus oder einem Stylus, der über eine berührungssensitive Oberfläche (Touchscreen) bewegt wird. Solche Gesten sind einfach zu erkennen, da explizit Anfang und Ende einer Geste vom Benutzer bestimmt werden. Demgegenüber ist die Erkennung von allgemeinen Bewegungen der Arme, Beine oder des Rumpfes komplizierter, da die kontinuierlichen Bewegungen zunächst segmentiert und von irrelevanten Bewegungen getrennt werden müssen.

Zur Eingabe eignen sich intrusive (z.B. Datenhandschuhe) oder nicht-intrusive Geräte (z.B. mittels einer Kamera). Intrusive Geräte nehmen die Bewegungsinformationen direkt auf und verarbeiten sie meist auf unterschiedlichen Ebenen weiter; bspw. können bei einem Datenhandschuh zunächst die grobe Position und Orientierung der Hand und der Finger erfasst werden, danach feinere Bewegungen der Finger, und schließlich können daraus Gesten bestimmt werden. Bei nicht-intrusiven Geräten müssen zunächst die Positionen der Gestenerzeugenden Körperpartien bestimmt werden. Dies kann z.B. wie bei der Gesichtserkennung durch Farbinformationen (insbes. bei den Händen) geschehen.

Daraus lassen sich unterschiedliche Gesten bestimmen. Scherer und Ekman (1982) und McNeill (1992) unterscheiden fünf 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.

Zur eigentlichen Klassifikation wurden in der Vergangenheit eine Vielzahl unterschiedlicher Klassifikatoren eingesetzt. Hierzu zählen der Vergleich von Schablonen-Mustern, neuronale Netze, HMMs, Bayessche Klassifikatoren, Hauptkomponentenanalysen, sowie Kombinationen dieser Verfahren.

Die Erkennung von Handschrift hat in jüngerer Zeit wieder an Bedeutung gewonnen, insbesondere durch die Miniaturisierung mobiler Geräte, welche häufig keine Tastaturen besitzen, aber natürlich auch zur Eingabe von Schriften, die über eine Vielzahl unterschiedlicher Charaktere verfügen (z.B. Kanji). Hier kann eine Erkennung über den Touchscreen helfen. Darauf können entweder einzelne Buchstaben separat erkannt werden (*offline*, entspricht *optical character recognition*), oder es wird eine kontinuierliche Trajektorie als ganzes Wort erkannt (*online*). Im ersten Falle werden nur die räumlichen Informationen über die Linien ausgewertet; im zweiten Falle können auch zeitliche Informationen über die Produktion der Trajektoren verwendet werden.

In beiden Fällen müssen die entstehenden Eingangsinformationen klassifiziert werden. Dieses Problem ist dem der Spracherkennung sehr ähnlich, sowohl was die Variationsmöglichkeiten (Vokabular, Schreibstil, Sprecherabhängigkeit, Einzelwörter vs. kontinuierliche Sprache) als auch die dabei verwendeten Lösungen angeht. Meist werden also auch hier neuronale Netze oder HMMs als Klassifikatoren verwendet. Bei kontinuierlich geschriebener Sprache kommt z.B. noch ein statistisches Sprachmodell in Form eines n-grams hinzu.

8.4 Multimodale Verarbeitung

Die von den einzelnen Eingabe-Modalitäten bereitgestellten Informationen müssen nun gemeinsam weiterverarbeitet werden, um den Sinn der eingegebenen Informationen zu extrahieren und auf Basis dessen den nächsten Interaktionsschritt zu planen. Dazu müssen die Informationen zunächst fusioniert werden. Hierbei stellen sich drei Fragen:

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

Als Kriterium für die erste Frage wird meist die zeitliche Synchronizität herangezogen: 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. Der Nachteil dieses Verfahrens liegt darin, dass es sich nur schlecht auf zusätzliche Modalitäten erweitern lässt, dass es einen großen, spezifischen Trainingskorpus benötigt, und dass die rechnerische Verarbeitung aufwändig ist. Demgegenüber können bei einer Fusion auf semantischer Ebene zunächst die Einzelhypthesen mittels einzeln trainierter Erkenner bestimmt werden. Aus diesen Hypothesen kann dann eine umfassende Interpretation der Semantik der Eingangsinformationen abgeleitet werden. Dabei können recht einfach neue Modalitäten integriert werden, und es können auch Modalitäten, die unterschiedliche zeitliche Verarbeitungen implizieren, miteinander kombiniert werden.

Die extrahierten Informationen müssen in einem Format bereitgestellt werden, dass die konsekutive Aufnahme von Informationen widerspiegelt. Hierbei sollten neben den extrahierten Informationen auch die Zeit der Extraktion sowie die Konfidenz in die extrahierte Information beachtet werden. Auf Basis der zeitlichen Informationen können Zuordnungen unterschiedlicher Informationsquellen durchgeführt werden, selbst wenn die

Verarbeitungszeiten unterschiedlich sind. Auf Basis der Konfidenzen können widersprüchliche Informationen im Sinne einer besten Gesamt-Erkennungsrate aufgelöst werden. Üblicherweise werden wie bei Sprachdialogsystemen semantische Rahmen (Attribut-Wert-Paare) verwendet, die aber u.U. noch geschachtelt und erweitert werden können, oder sogenannte Melting Pots, die multimodale Informationen mit Zeitstempeln einkapseln und für die Fusion bereit stellen. Verschiedene Markup-Sprachen stehen hierzu zur Verfügung, wie bspw. XML, M3L, MIAMM, MMIL, MUST, MXML, etc.

Die Aufgaben des Dialogmanagers gleichen denen im Fall einer Sprachdialogsystems. Allerdings kann ihre Lösung im multimodalen Fall aufwändiger sein als im unimodalen (Sprach-) Fall. Bspw. kann durch die Vielzahl von Modalitäten die Konfidenz der Eingabeinformationen stark variieren. In diesem Falle bietet es sich an, die entsprechende Modalität, durch die eine Information erhalten wurde, mit zu erfassen, um daraus Schlüsse im Falle von Missverständnissen abzuleiten. Zum Beispiel könnte der Benutzer in Falle von Spracherkennungsfehlern aufgefordert werden, die entsprechende Information lieber über den Touchscreen einzugeben. Multimodale Dialogsysteme werden häufig für komplexere Aufgaben eingesetzt als sprachbasierte Systeme, was sich auch in der Komplexität des Dialogmanagements widerspiegelt. Zusätzlich zur Systemeingabe muss der Dialogmanager auch berücksichtigen, in welcher Modalität die vorangegangene Information vom System ausgegeben wurde. Bspw. könnte der Dialogmanager dann die nächste von ihm benötigte Information ebenfalls in dieser Modalität anfordern, und den nächsten Systemprompt dementsprechend ausrichten. Generell führt eine Vielzahl von angebotenen Modalitäten natürlich auch zu einer Vielzahl von Fehlermöglichkeiten, mit denen ein multimodaler Dialogmanager umgehen müssen muss. Daher wurden spezielle Architekturen für multimodale Dialogmanager entwickelt. Details und Literaturverweise finden sich z.B. in López-Cózar und Araki (2005).

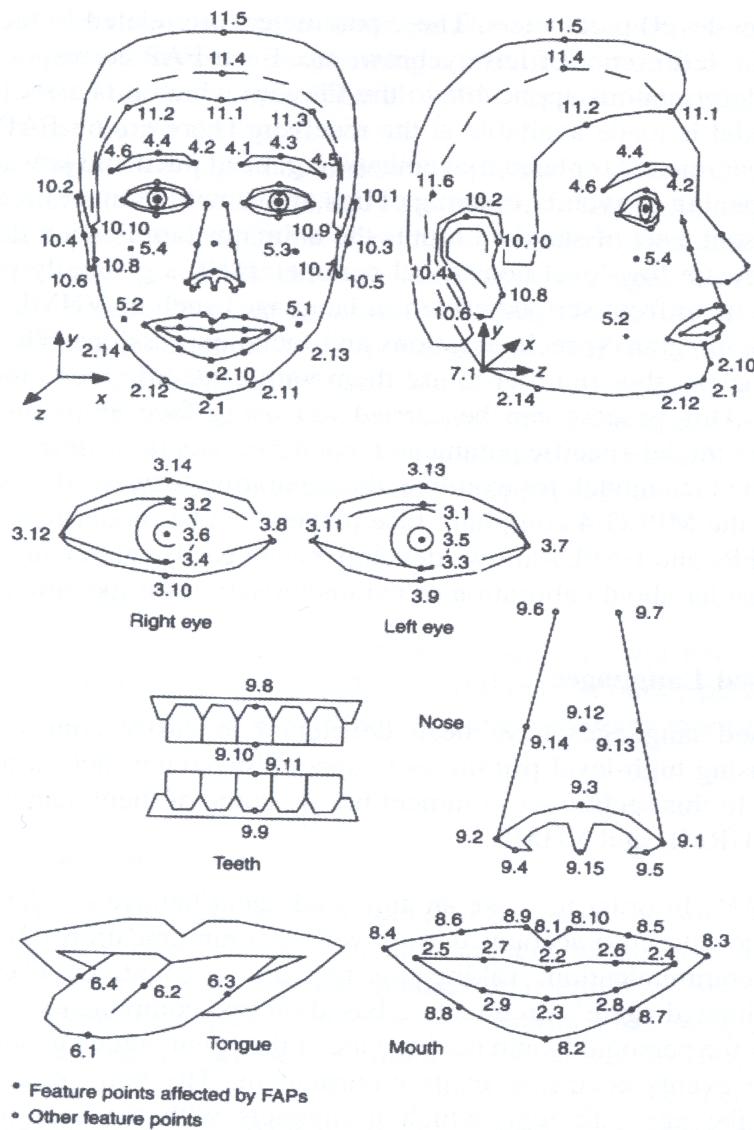
Sobald der Dialogmanager den nächsten Interaktionsschritt geplant hat müssen die auszugebenden Informationen bestimmt werden. Dies ist die Aufgabe des Response-Generation-Modules. Insbesondere müssen die auszugebenden Informationen auf die zur Verfügung stehenden Modalitäten aufgeteilt werden (Fission). Dabei können natürlich die eingangs in Kapitel 8.1. beschriebenen Modalitäts-Eigenschaften zu Rate gezogen werden. Häufig bietet es sich an, parallel mehrere Modalitäten zu verwenden, die sich optimalerweise aufeinander beziehen sollten. Bspw. könnte als Ergebnis einer Navigationsaufgabe eine Karte auf dem Display angezeigt werden, während eine Stimme den Benutzer darüber informiert, dass er über einfaches Berühren einzelner Punkte der Route Zusatzinformationen bekommen kann. Auch Referenzen der Modalitäten aufeinander („Die schnellste Route finden Sie oben in der Liste“) sind zulässig, verlangen jedoch, dass die internen Repräsentationen innerhalb der Modalitäten zugänglich sind. Um bestimmte Informationen hervorzuheben bieten sich darüber hinaus Kombinationen von Modalitäten an, die bewusst Redundanzen erzeugen.

8.5 Multimodale Ausgabe-Schnittstellen

Die Module zur Text- und Spacherzeugung sind bereits aus Kapitel 7 bekannt. Grafiken lassen sich leicht nach Regeln aus bestimmten Prototypen (Listen, Balkendiagramme, Kuchendiagramme, etc.) erzeugen. Positionsinformationen lassen sich sehr gut auf Karten darstellen, in denen die gewünschten Informationen dann markiert werden können. Wird eine solche Karte oder auch eine Liste auf einem Touchscreen dargestellt ergibt sich eine optimale Kombination von Eingabe- und Ausgabemodalitäten, die die Effizienz und Intuitivität des Dialoges deutlich steigern kann.

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.

ECAs beruhen entweder auf einer Animation vorher aufgezeichneter Bilder oder auf einem parametrisch angesteuerten Modell. Letztere verwenden Parameter zur Beschreibung der Körper- und Gesichtsform (Nase, Mund, etc.) sowie zur Beschreibung des Ausdrucks (Bewegungen von Augenbrauen, Mund, Lidern, etc.). Die Animation besteht in einer Variation der Parameter über der Zeit, wodurch Bewegungen des animierten Agenten erzeugt werden. Dafür können z.B. die Punkte des in MPEG-4 beschriebenen Standardes zur Kodierung von Multimedia-Informationen verwendet werden, vgl. untenstehende Abbildung.



Merkmalspunkte des MPEG-4-Gesichtsmodells, nach ISO/IEC IS 14496/2 Visual (1999).

Zur Beschreibung des Gesichtsausdrucks werden zumeist Parameter des *Facial Action Coding System (FACS)* verwendet. Diese bestehen aus einer begrenzten Zahl an sog. *Action Units*, aus denen sich Mimik und Gesichtsausdrücke erzeugen lassen. In einfachen Fällen wird nur die Gesichtsoberfläche durch diese Parameter bestimmt. Komplexere physiologische Modelle beschreiben neben der Hautoberfläche auch die Struktur der Anatomie und die korrespondierenden Muskel-Aktivitäten. Sie sind rechenintensiver und arbeiten deshalb nicht immer in Realzeit, lassen dafür aber genauere Modellierungen – und mitunter auch realistischere Animationen – zu.

Basic expression	Set of Action Units
Surprise	AU1, 2, 5, 15, 16, 20, 26
Fear	AU1, 2, 4, 5, 7, 15, 20, 25
Disgust	AU4, 9, 10, 17
Anger	AU2, 4, 5, 10, 20, 24
Happiness	AU6, 11, 12, 25
Sadness	AU1, 4, 7, 15
AU1 Inner Brow Raiser	AU30 Jaw Sideways
AU2 Outer Brow Raiser	AU31 Jaw Clencher
AU4 Brow Lowerer	AU32 Lip Bite
AUS Upper Lid Raiser	AU33 Check Blow
A116 Cheek Raiser	AU34 Check Puff
AU7 Lid Tightener	AU35 Check Suck
AU8 Lips toward each other	AU36 Tongue Bulge
AU9 Nose Wrinkler	AU37 Lip Wipe
AU 10 Upper Lip Raiser	AU38 Nostril Dilator
AU1 1 Nasolabial Deepener	AU39 Nostril Compressor
AU12 Lip Corner Puller	AU41 Lip Droop
AU13 Cheek buffer	AU42 Slit
AU14 Dimpler	AU43 Eyes Closed
AU15 Lip Comer Depressor	AU Squint
AU16 Lower Lip Depressor	AU45 Blink
AU17 Chin Raiser	AU46 Wink
AU18 Lip Puckerer	AU51 Head turn left
AU19 Tongue Show	AU52 Head turn right
AU20 Lip Stretcher	A1J53 Head up
AU21 Neck Tightener	AU54 Head down
AU22 Lip Funneler	AU55 Head tilt left
AU23 Lip Tightener	AU56 Head tilt right
A1124 Lip Pressor	AU57 Head Forward
AU25 Lips Part	AU58 Head back
AU26 Jaw Drop	AU61 Eyes turn left
AU27 Mouth Stretch	AU62 Eyes turn right
AU28 Lip Suck	AU63 Eyes up
AU29 Jaw Thrust	AU64 Eyes clown
	AU65 Wall-eye
	AU66 Cross-eye

Facial Action Coding Scheme (FACS) mit Emotionen und zugehörigen Action Units (AUs).
Nach López-Cózar Delgado und Araki (2005).

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 Ressourcen zur Verarbeitung, und sie sind meist sehr intuitiv und in verschiedenen Sprach- und Kulturgemeinden gleich. Durch *Auditory Eyecons* kann dem Nutzer sehr einfach und schnell eine Rückbestätigung gegeben werden; insbesondere bei taktiler Eingabe hat sich dies bewährt. Darüber hinaus ist hier natürlich auch taktiles Feedback wünschenswert. Dies kann z.B. pneumatisch (mittels Luftdruck), vibrotaktile (z.B. durch Ausnutzung des piezoelektrischen Effektes), elektrotaktile (Anregung mittels Elektroden) oder funktional-neuromuskulär (direkte Anregung des neuro-

muskulären Systems) gegeben werden; insbesondere vibrotaktile Interfaces erfreuen sich wachsenden Interesses.

8.6 Systembeispiele

Seit Mitte der 1990er Jahre wurde eine Reihe von multimodalen Dialogsystemen entwickelt, von denen es jedoch bislang nur sehr wenige zur Anwendungsreife gebracht haben. Insbesondere sind zu nennen:

- MASK-Kiosk für Zuginformationen (Frankreich)
- AdApt-System zur Information über Appartments in Stockholm (KTH & Telia Research, Schweden)
- August für touristische Informationen (KTH, Schweden)
- Jeanie zur Terminabsprache (CMU, USA)
- Matis für Zuginformationen (Nijmegen, Niederlande)
- MUST für touristische Informationen
- Olga zur Steuerung von Mikrowellenherden (KTH, Schweden)
- Rea für Immobilieninformationen (MIT Media Lab, USA)
- SmartKom für Hausgeräte und öffentliche Informationen (DFKI, Deutschland)
- Waxholm für touristische Informationen (KTH, Schweden)

Referenzen zu diesen Systemen finden sich bei López-Cózar und Araki (2005).

8.7 Literatur

Bernsen, N. O. (2002). Multimodality in Language and Speech Systems - From Theory to Design Support Tool. In Granström, B., House, D., and Karlsson, I. (Eds.): *Multimodality in Language and Speech Systems*, Dordrecht, Kluwer Academic Publishers, 93-148.

Bernsen, N.O. (1999). Multimodality in Language and Speech Systems - From Theory to Design Support Tool. Invited course, 7th European Summer School on Language and Speech Communication (ESSLSC), Stockholm, Luly 1999.

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought*. The University of Chicago Press.

Scherer, K., Ekman, P. (1982). *Handbook of Methods in Nonverbal Behavior Research*. Cambridge University Press.

Schleicher, R., Galley, N., Briest, S. & Galley, L. (2007). Blinks and saccades as indicators of fatigue in sleepiness warners: looking tired? Accepted for publication in *Ergonomics*.