J.-W. Goethe Universität Frankfurt

Fachbereich Informatik

Lehrstuhl für praktische Informatik VSFT

Dr.R.Brause

Automatische Spracherkennung

Vorlesung SS 1987

Inhalt

1.0 Spracherkennung - Möglichkeiten und Grenzen 2

2.0 Das Sprachsignal 6

2.1 Erzeugung des Sprachsignals 6
Stimmhafte Laute - Stimmlose Laute
Artikulationsorte der Konsonanten
Artikulationsarten der Konsonanten
Artikulationsorte der Vokale
Artikulationsarten der Vokale

2.2 Physikalische Charakterisierung 10
Pitch und Formanten
Die Übertragungsfunktion des Sprachtrakts

2.3 Phonetische Charakterisierung 16
Phon, Phonem, Allophon
Morph, Morphem, Allomorph
Wort und Lexem
Suprasegmentale Merkmale

2.4 Verständlichkeit 20
Sprachgütemessungen
Logatom-Tests
Reimtests
Ergebnisse

3.0 Kodierung und Sythese des Sprachsignals 24

3.1 Digitale Kodierung 24
Analog-Digitalkonvertierung
Signal-Geräusch-Verhältnis
Kompandierung und 13-Segment-PCM-Kennlinie

3.2 Lineare Prädiktion 28
Delta-Modulation, Adaptive Delta-Modulation
Differenz-PCM
Prädiktor-Koeffizienten
PARCOR-Koeffizienten
Lineare Filter und Inverse Filterung
ADPCM-System
Intervall-Fenster

3.3 Parametrische Kodierung 38
Kanalvokoder und LPC-Vokoder
Formantenanalyse
Pitchanalyse
Center Clipping
SIFT-Verfahren
Cepstrum
Multi-Puls-Anregung

3.4 Hardware zur Spracherkennung und Synthese 45
Sprachanalyse
Sprachsynthese
Phonemübergänge

4.0 Spracherkennungsalgorithmen und Systeme 49

4.1 Einzelworterkennung 49
Zeitnormalisierung
Dynamic Time Warping (DTW)
Hidden Markov Models (HMM)
4.2 Erkennung fließend gesprochener Sprache 58
statistischer Ansatz
4.3 Syntaxgesteuerte Satzerkennung 60
Reguläre, kontextfreie Grammatik
stochastische Grammatiken
Der Kartesische Produkt-Algorithmus

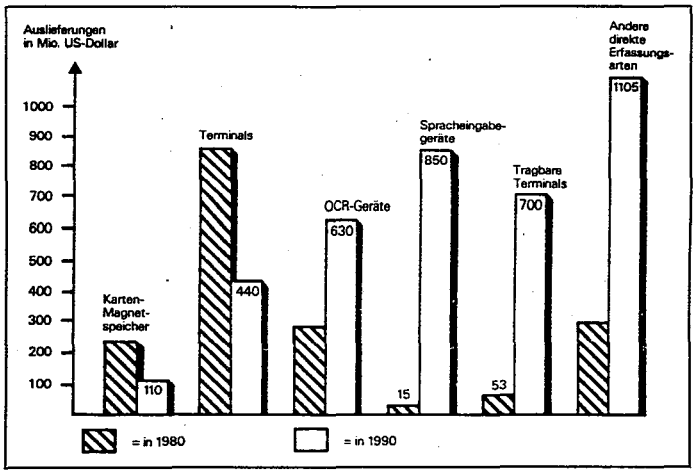
5.0 Hören 65
5.1 Physiologie des Höhrapparats 65
5.2 Die Funktionm des Innenohrs 69
5.3 Experimentalpsychologische Hörerkenntnisse 72
Subjektive Hörwahrnehmung
Verarbeitungsmodelle und Maskierungsexperimente
Missing Fundamental
Kategoriale Wahrnehmung

6.0 Ein Modell für Spracherkennen 77
6.1 Modellierung des menschlichen Spracherkennens 77
6.2 Ein Modell für "Klassifizieren und Ergänzen" 79
Die Feinstruktur des Großhirns
6.3 Das Kreuzkorrellations-Matrixmodell 80
Das lineare Matrixmodell
Assoziative Speicherung
Das nichtlineare Matrixmodell
Die Annahme konstanter Aktivität
Der Hamming-Abstand zweier Muster
Die Klassifizierungsoperation
Die Ergänzung lückenhafter Daten

Literatur 87

1.0 Spracherkennung - Möglichkeiten und Grenzen

Viele Menschen haben Schwierigkeiten, die Automaten im täglichen Leben (Fahrstuhl, Bankautomat, Textverarbeitungssystem etc.) richtig zu bedienen. Deshalb erhoffen sich viele Organisatoren und Firmen von der Spracherkennung und Sprachausgabe eine "menschengerechtere" Ein- und Ausgabemöglichkeit ("Mensch-Maschine-Schnittstelle") für alle computergestützten Systeme und prognostizieren den Sprachsystemen einen großen Markterfolg.
In den letzten Jahren hat nun die Realisierung kleiner, preisgünstiger Sprachsysteme große Fortschritte gemacht.
Die folgende Abbildung zeigt eine Prognose der International Data Corporation (IDC) für den amerikanischen Markt.



Trend zur automatischen direkten Datenerfassung in den USA

Quelle: IDC

Für Westeuropa werden von Frost&Sullivan (London) 1986 folgende Zahlen genannt:

Büroautomation	1985	0.5	Mio $
	1994	55	Mio $
Industrie	1994	50	Mio $
Gesamtmarkt	1985	25	Mio $ (BRD 5 Mio $)
	1994	600	Mio $

Wie man sieht, wird dem Markt "Sprache und Computer" ein starkes Wachstum prognostiziert, so daß die Industrie aus gutem Grund an diesem Thema interessiert ist.

Was können nun die Benutzer von dem Medium "Spracheingabe" erwarten ?
Zweifelsohne ist die Sprache nicht in allen Anwendungsfällen das schnellere und flexiblere Eingabemedium. Beispielsweise lassen sich Textkorrekturen an gespeicherten Texten leichter mit Zeigeinstrumenten (Maus, Lichtgriffel) am Bildschirm vornehmen als Sprachsteuerung. Manche Praktiker sind sogar der Meinung, daß bei allen Anwendungen, bei denen der Benutzer die Hände für Tasten frei hat, Spracheingabe nicht angebracht sei. Überzeugend sei die Spracheingabe dementsprechend nur in folgenden Anwendungsfällen:

- Alle Tätigkeiten, bei denen die Hände nicht frei sind, aber Daten ermittelt werden:

  - **Lager und Transportsysteme**
    Bsp: Paketsortierung

  - **Qualitätskontrolle**
    Bsp: versch. Meßgeräte ohne Rechner-Interface,
    Ablesen von Werten auf Bauteilen, etc.

  - **Behinderten-Hilfen**
    Bsp: sprachgesteuerter Rollstuhl und Auto
    Sprachausgabe bei Thermometer und Uhr

  - **Telefon-Wahl**
    Bsp: Rufnummernspeicherung mit Nahmen (SELund AEG)

- Alle Tätigkeiten, bei denen der Blick nicht verändert werden kann, aber Daten bzw. Kontrollinformation einegegeben werden muß:
  - **Arbeiten am Mikroskop**: Laboruntersuchungen,
    sprachsteuerung für neurolog. Operationen

  - **Arbeiten am Meßgeräten**

- Alle Tätigkeiten, die über Telefon abgewickelt werden können:

  - **Automatische Auskunfts-** und Reservierungssysteme
    Bsp: Flug, Zug, Hotel, Theaterreservierungen

  - **Warenbestellung** per Telefon (Grossisten etc.)

- Sprachgesteuerte **Schreibmaschine**

- Und wie überall: **Militärische Anwendungen**

Inwieweit kann nun der jetzige Stand der Technik diesen potentiellen Anwendungen genügen?

Betrachten wir dazu die Fähigkeiten und Leistungsmerkmale der auf dem Markt befindlichen Geräte. Die Mehrzahl der zur Zeit produzierten Geräte erkennt keine vollständig gesprochenen Sätze, sondern nur einzelne Worte, die von anderen Worten durch eine kurze Pause getrennt sein müssen. Bei dieser "Einzelworterkennung" wird das gesprochene Wort mit einer Liste von zuvor eingegebenen Worten verglichen. Es wird auf das Wort aus der Liste, das dem fraglichen Wort am ähnlichsten ist, entschieden. In Abbildung 1.0b ist ein solches System zur Einzelworterkennung gezeigt.

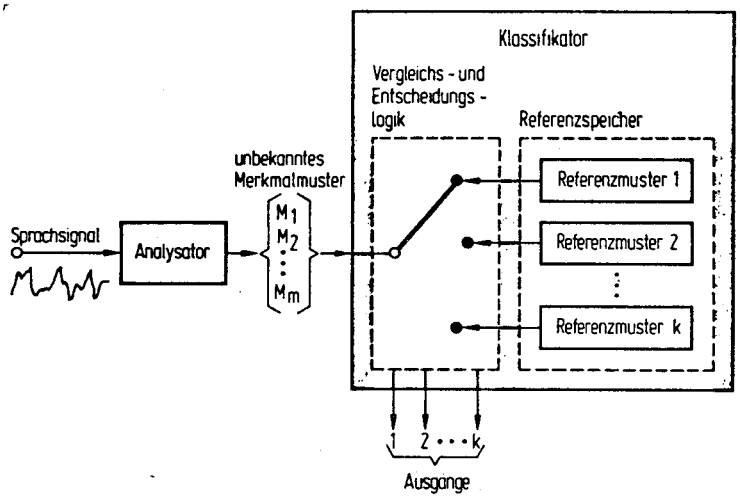
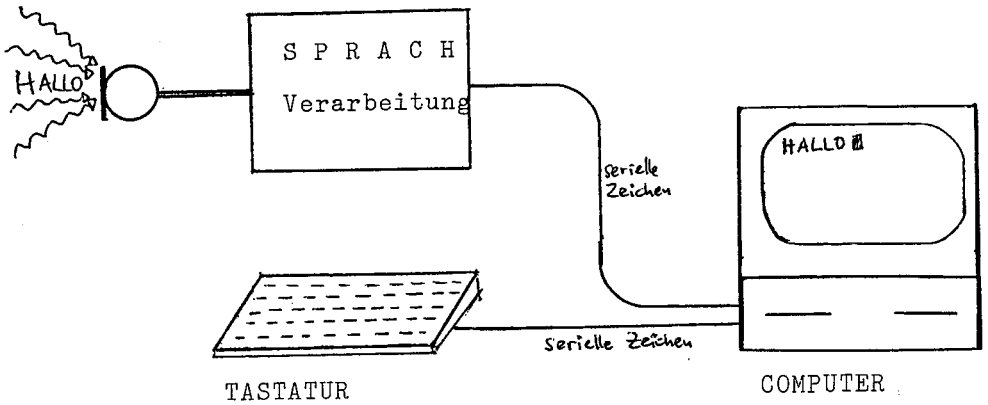


Bild 1. Allgemeine Darstellung eines Spracherkennungsautomaten

Da es für die Wiedererkennung vollkommen unerheblich ist, was vorher als Referenzmuster für die Liste vorgesprochen worden ist, müßte diese Methode eigentlich "Geräuscherkennung" heißen.
Der Vorteil dieses Verfahrens liegt in der hohen Erfolgswahrscheinlichkeit (ca 98%), egal ob Dialekt, Akzent oder fremdsprachliche Worte verwendet werden.
Die Nachteile sind im Wesentlichen

- Abhängigkeit der Wiedererkennung vom Sprecher
- Erkennung von Sätzen nur bei eingeschobenen künstlichen Pausen
- Empfindlichkeit gegenüber Störgeräuschen

Systeme mit Einzelworterkennung haben typischerweise folgende Anschluß-Konfiguration:



Die Spracheingabe ist parallel zu einem Keyboard angeschlossen und produziert bei jedem erkannten Wort eine dazu gespeicherte, vom Benutzer spezifizierte ASCII-Zeichenkette. Damit besteht für das Betriebssystem des Computers kein Unterschied, ob die Zeichen eingetippt oder gesprochen wurden, was natürlich die Weterverarbeitung mit Standardsoftware ungemein erleichtert.
Obwohl in der Bürokommunikation ein Vokabular von ca. 10000 Worten nötig ist, besitzen handelsübliche Systeme meist nur die Möglichkeit, bis zu 250 Worte zu speichern. Sie sind damit nicht für die Bürokommunikation brauchbar und können nur in den genannten Spezialanwendungen (Behindertenhilfen, Mikroskopsteuerung etc) eingesetzt werden. Diese Beschränkung ist zwar zur Zeit noch ein Preis-Leistungsproblem, aber es zeichnet sich ab, daß

dies durch die genannten Nachteile auch eine prinzipielle Obergrenze ist. Nimmt ein gutwilliger, fortschrittsgläubiger Anwender noch hin, ein Mikrophon am Kopf tragen zu müssen, immer gleichartig zu sprechen und künstliche Pausen zwischen den Worten zu machen, so verliert er doch leicht die Geduld, wenn zum Training 250 Worte mehrmals (Mittelwertbildung!) vorgesprochen werden müssen. Aus der Erfahrung heraus, daß andere Leute ihn auch verstehen, ohne daß er stundenlang Wortlisten mit ihnen übt, wird er allerdings das wochenlange Vorsprechen von Wortlisten ablehnen. Hat er mehrere Geräte verschiedener Hersteller, dauert das Ganze natürlich etwas länger...
Die Grundlagen der Einzelworterkennung sind schon seit über 10 Jahren bekannt und haben sich kaum weiterentwickelt. Heutige Forschungsbemühungen versuchen deshalb immer noch, die Nachteile der Einzelworterkennung zu vermeiden, allerdings bisher ohne einen bahnbrechenden Erfolg. Die Erfolgsquote der sprecherunabhängigen Worterkennung ist mit ca. 60%-80% noch zu gering, da erfahrungsgemäß die Benutzer bei einem Mißerfolg von mehr als 5% solche Systeme ablehnen.
Der Traum einer sprachgesteuerten Schreibmaschine beispielsweise dürfte damit erstmal in die Zukunft verschoben sein.

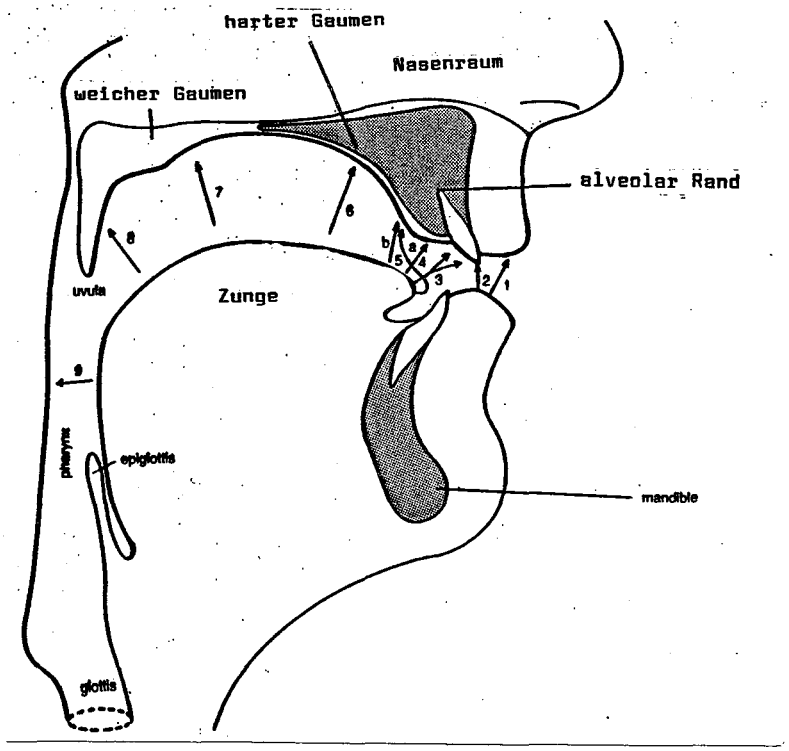
Im Gegensatz zur Spracherkennung ist die Synthese von Sprache (Sprachausgabe) weitgehend unproblematisch. Die Simulation der Charakteristika des menschlichen Sprachtrakts durch digitale Filter auf Chip-Basis ist ziemlich weit fortgeschritten, so daß auf eine Anwendung der Spracheingabe zur Zeit ca. 20 Systeme der Sprachausgabe kommen. Mit der Stagnation der sprecherunabhängigen, kontinuierlichen Spracherkennung werden deshalb in den aktuellen Marktschätzungen der Anteil der Spracheingabesysteme zugunsten der Sprachausgabe nach unten korrigiert.

2.0 Sprache

Im Unterschied zu primären, natürlichen Reizen wie Sehen und das Hören von Geräuschen, ist Sprache ein Kunstprodukt: von Menschen für Menschen bestimmt. Was die Sprache von Geräuschen unterscheidet ist also subjektiv, sehr kultur- und zeitabhängig. Im Unterschied zum Sehen ist also das Sprachverstehen nicht genetisch bedingt und muß von jedem Menschen erst gelernt werden.
Wodurch ist nun Sprache charakterisiert? Wie läßt sie sich beschreiben?
Zweifelsohne läßt sich Sprache physikalisch als Frequenzgemisch beschreiben, wobei die relativen Amplituden, die Phasenverschiebungen sowie die Gesamtamplitude zeitabhängig ist. Wollen wir auf Basis dieser physikalischen Charakterisierung eine Wiedererkennung von Worten versuchen, so ist dies nicht möglich, da auch der selbe Sprecher das selbe Wort niemals physikalisch exakt reproduzieren kann. Es gibt immer Abweichungen in der Intonation, der zeitlichen Reihenfolge der Lautmuster und dergleichen mehr. Außerdem würde diese Analyse sehr lang dauern, da sowohl das Extrahieren aller Zeitfunktionen als auch der Vergleich mit anderen, bereits gespeicherten Worten ziemlich viel Zeit in Anspruch nimmt.
Stattdessen ist es sicher günstig, nur 'typische', charakteristische Merkmale bekannter Sprachelemente zu vergleichen. Das große Problem ist aber nun: Welches sind die 'typischen' Merkmale? Da Sprache von Menschen gemacht wird, sind die Merkmale sicher sehr subjektiv. Andererseits lassen sie sich auch durch Selbstbeobachtung erschließen: Was tue ich, um diese oder jene Laute zu erzeugen?
Spracherkennung ist deshalb auch ein Wissensgebiet, zu dem eine Vielzahl von Wissenschaften beitragen: Physik, Physiologie, Neurologie, Phonetik, Linguistik, etc. In den folgenden Abschnitten soll nun kurz geschildert werden, was die einzelnen Wissensgebiete zur Erschließung der typischen Sprachmerkmale aussagen können.

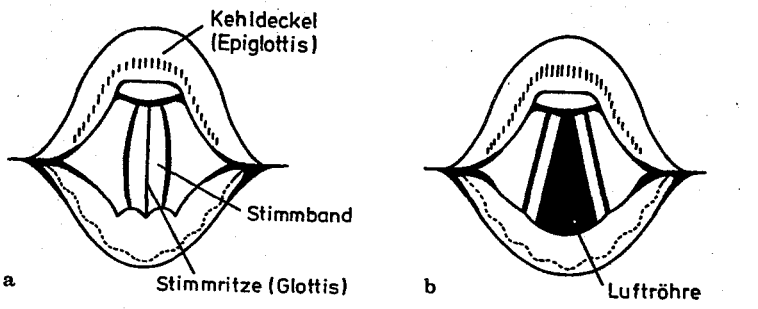
2.1 Erzeugung des Sprachsignals

Energetische Ursache des Sprachsignals ist zweifelsohne unsere Lunge, die Luft durch den Sprachtrakt (Hals, Mund, Nase) presst. Dabei gibt es zwei verschiedene Arten, Sprachlaute zu erzeugen: durch Anregung der Stimmbänder ("stimmhafte" Laute), die schwingen und damit einen Ton erzeugen oder aber nur durch das einfache Rauschen der bewegten Luft im Sprachtrakt ("stimmlose" Laute).
Betrachten wir dies näher anhand der Abbildung 2.1a.



Stimmhafte Laute

Zur Erzeugung des stimmhaften Tons wird der Stimmbandmuskel zusammengezogen (s. Abb 2.1b) und ähnlich wie bei einer Polsterpfeife (s. Abb. 2.1c) die Luft durch den schmalen Spalt gepreßt.



bei der Stimmbildung eines vokalischen Lautes,

beim Atmen

Abb. 2.1b

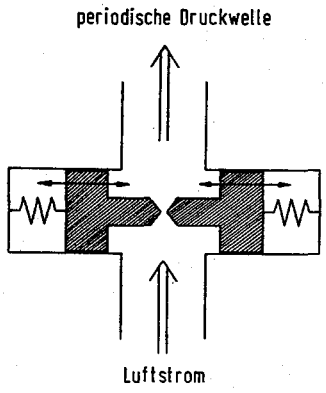


Abb.2.1c

Hat sich die Stimmritze (Glottis) geöffnet und der Überdruck wurde abgebaut, so schließt sich die Stimmritze wieder, bedingt durch die Muskelkontraktion und den Bernoulli-Effekt, der bei der erhöhten Luftgeschwindigkeit einen Druckabfall an der Glottis bewirkt. Ist die Glottis wieder geschlossen, so baut sich erneut ein Druck auf und der gesamte Vorgang wiederholt sich periodisch. Da die Muskelspannung nicht konstant ist, variiert die Periode leicht, so daß man nur noch von einer quasi-periodischen Bewegung sprechen kann. Ähnlich wie bei einem Blasinstrument hängt der tatsächlich hörbare Ton von einer Reihe weiterer Faktoren ab, wie beispielsweise der Öffnungen (Nasenraum, Mundraum) und der Länge des Sprachtrakts (Resonanzlänge). Einen reinen Ton hört man dabei allerdings nur bei den Vokalen und bei Gesang;normalerweise kommt

eine zusätzliche Beeinflussung des Tons durch die Bewegung der Zunge, der Zähne und der Lippen hinzu.

Stimmlose Laute
Die stimmlosen Laute entstehen durch Reibung und Turbulenzen der ausströmenden Luft an Verengungen (Rachen-, Zungen-, Zahnstellungen) sowie der dadurch angestoßenen Resonanzschwingungen im Sprachtrakt.

Artikulationsorte der Konsonanten
Je nachdem, an welcher Stelle im Sprachtrakt die Artikulation der stimmhaften und stimmlosen Laute erfolgt, resultieren verschiedene Laute. Da der genaue Artikulationsort eher ein gelernter Mittelwert und weniger eine streng definierte Mundstellung ist, weichen die verschiedenen Beschreibungssysteme je nach Muttersprache des Authors voneinander ab. Betrachten wir ein deutsches System für Konsonanten (Mitlaute) mit Hilfe der Abbildung 2.1a. Beginnen wir bei den Zähnen:

Bezeichnung	beteil. Organe	Beispiel
1) bilabial	Unterlippe - Oberlippe	[p]
2) labiodental	Unterlippe - obere Schneidezähne	[f]
3) dental	Zunge - obere Schneidezähne	[s]
4) alveolar	Zunge - obere Zahnseite	[d]
5) palatal	Zunge - harter Gaumen	[ch]
6) velar	Zunge - weicher Gaumen	[k]
7) uvular	Zunge - Zäpfchen	Zäpfchen-r
8) glottal	Stimmritze	[h]

In Abbildung 2.1d ist dazu die Stellung der Zunge bei verschiedenen Konsonanten gezeigt.

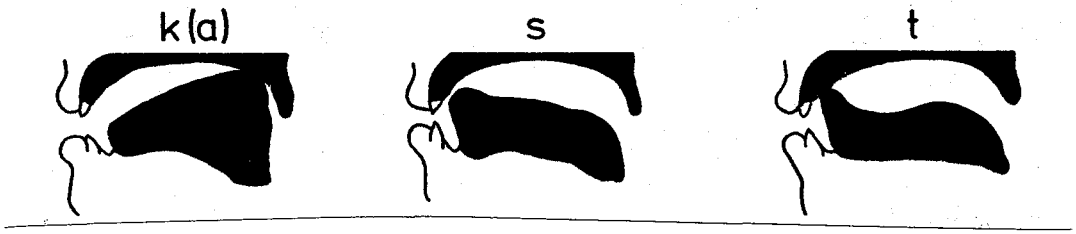


Abb.2.1d Stellung der Zunge bei verschiedenen Konsonanten

Artikulationsarten für Konsonanten
Zusätzlich zur Lauterzeugung (stimmhaft/stimmlos) und dem Ort der Lautbeeinflussung gibt es weitere Merkmale, die die Erzeugung von Konsonanten charakterisieren. Diese sind

Bezeichnung	Merkmal	Beispiel
- Verschlußlaute (Explosivlaute)	Verschließen des Luftstroms und plötzl. Freigabe	[b] , [d] , [p] , [t]
- Reibelaute (Frikative)	Einengung des Luftstroms	[j] , [f]
- Nasale	Verschließen der Mundhöhle	[m] , [n]
- Laterale	Luftstrom seitlich an der Zunge vorbei	[l]
- Vibranten	Zunge bzw Zäpfchen schwingen	Zungen-r, Zäpfchen-r

Im Prinzip kann jeder Artikulationsort mit jeder Artikulationsart kombiniert werden; tatsächlich ist dies aber nur eingeschränkt möglich, wie in der folgenden Abbildung 2.1e gezeigt ist.

Artikulationsort

Artikulationsart		bi-labial	labio-dental	dental	alveolar	palatal	velar	uvular	glottal
Verschlußlaute	sth.	b			d		g		
	stl.	p			t		k		
Reibelaute	sth.		v	z		j			
	stl.		f	s	ʃ	ç	x		h
Nasale		m			n		ŋ		
Laterale					l				
Intermittierende					r			R	

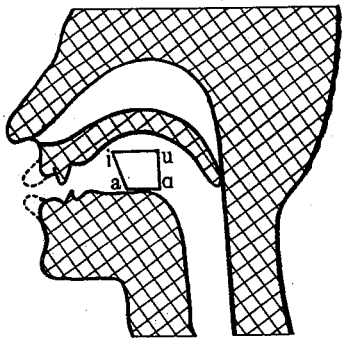
Umschrift

ŋ	wie dt.	la*ng*	R	wie dt.	Zäpfchen-r
ʃ	dt.	*Sch*iff	s	dt.	Ha*ß*
ç	dt.	i*ch*	z	dt.	Rose
x	dt.	Ba*ch*	v	dt.	*W*all
r	dt.	Zungen-r			

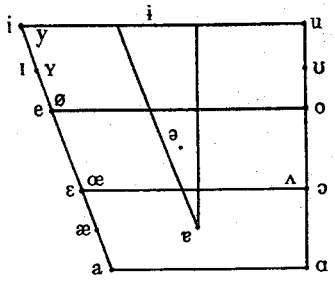
Abb.2.1e Bildungsort und -art der Konsonanten

Artikulationsorte für Vokale

Vokale werden grundsätzlich stimmhaft erzeugt. Die Bildung von Vokalen ist noch ungenauer beschreibbar als die der Konsonanten. Prinzipiell ist der Abstand des höchsten Punkts der Zunge (Zungenrücken) vom Gaumen und der Abstand Zungenrücken-Zähne wichtig. Dies schlägt sich in dem sog. Vokalviereck nieder.



Zungenstellung bei der Artikulation,



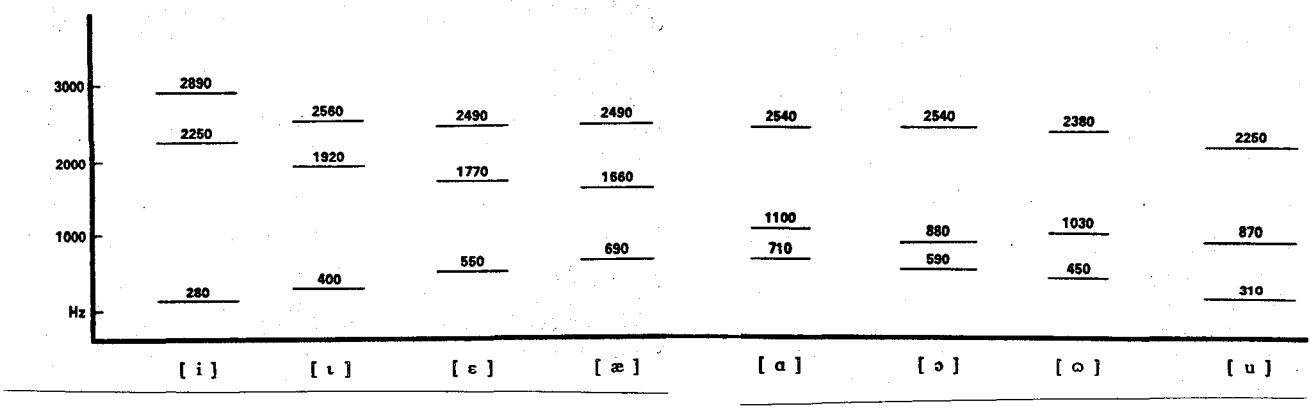
Vokalviereck

Artikulationsarten für Vokale
Die Artikulation erfolgt durch die

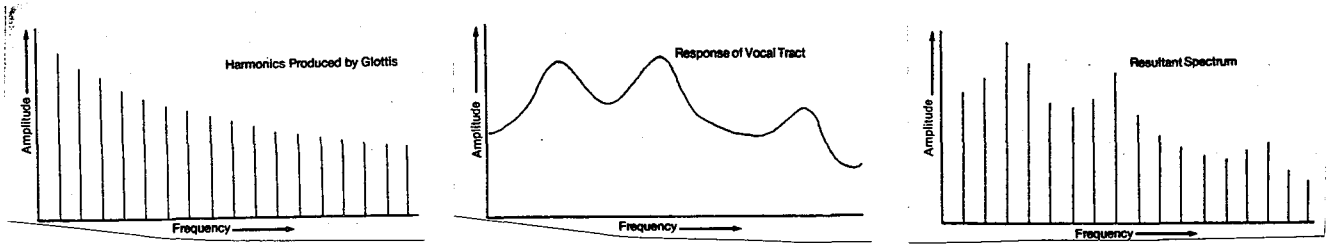
- Mundstellung	offen, halboffen, fast geschlossen	[a] , [ə] , [o] , [u]
- Lippenstellung	gerundet, ungerundet	[u] , [e]
- Diphtonge	Übergang zwischen zwei Vokalen	[au] , [ai]
- Nasale	Umleitung der Luft durch die Nase	[ɛ̃]

2.2 Physikalische Charakterisierung

Eine Charakterisierung der Lauterzeugung bildet zwar ein einfacheres System als die resultierenden Klanggebilde, für die Sprachanalyse ist sie aber nur bedingt nutzbar, da man z.B. Artikulationsorte nicht direkt hören kann. Deshalb wollen wir im Folgenden das physikalische Schallereignis "Sprache" etwas näher betrachten.
Wie wir im vorigen Abschnitt gesehen haben, wird bei stimmhaften Lauten mit der Stimmritze ein Ton erzeugt. Dieser Ton hat als Grundfrequenz normalerweise 80Hz (tiefe Männerstimme) bis 350Hz (Kinderstimme). Durch den nicht-linearen Resonanzraum (Reflexionen an Kanten und Wölbungen) des Sprachtrakts ergeben sich Oberwellen, die aber durch die Quasi-Periodizität nicht ein Vielfaches der Grundfrequenz haben, sondern in ihren Frequenzen verschoben sind. Abbildung 2.2a zeigt dies für verschiedene Vokale.



Die Obertöne haben außerdem verschiedene, relative Amplituden, je nach Artikulation. Dies ist in Abb. 2.2b illustriert, in dem das ursprüngliche, exponentiell gedämpfte Obertonspektrum, die Dämpfungscharakteristik einer bestimmten Artikulation sowie das resultierende Spektrum zu sehen ist.



Die Obertöne werden dabei als "Formanten" bezeichnet; der Grundton heißt "Pitch". Wie am Beispiel des Vokales /i/ in der Abb.2.2c zu sehen ist, kann die Dämpfungscharakteristik nicht einfach berechnet werden, da sich der Querschnitt des Vokaltrakts mit steigendem Abstand von der tonerzeugenden Glottis stark ändert.

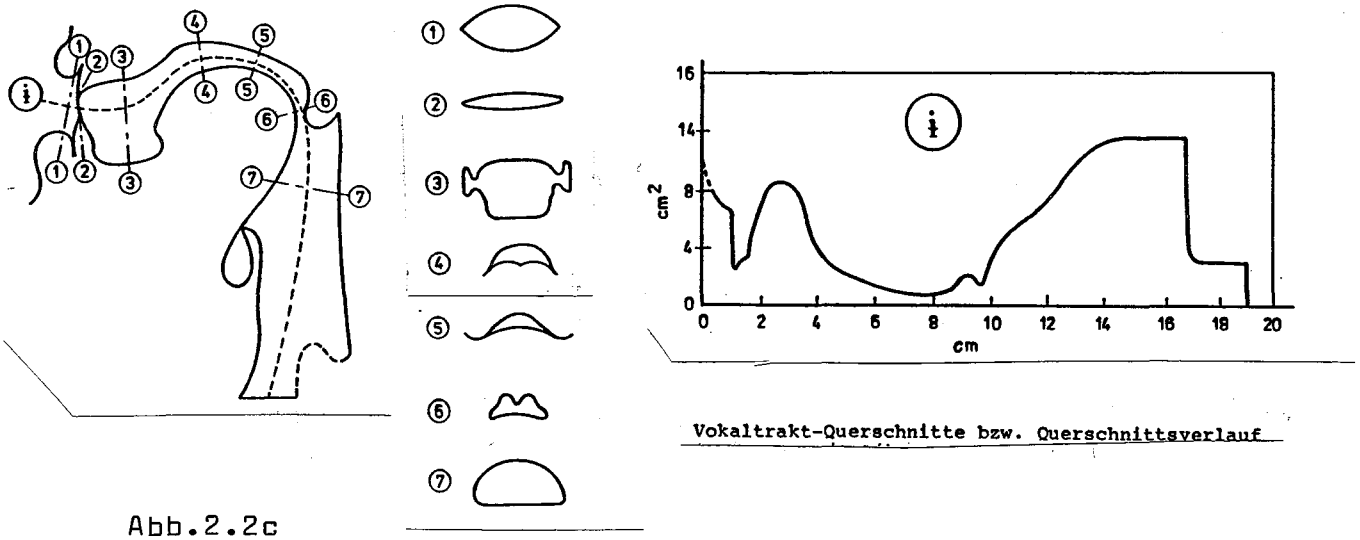


Abb.2.2c

Die Modellierung dieser Spracherzeugung betrachtet meist nur die mittlere Fläche und das dazu gehörende Röhrenstück in einem Segment des Vokaltrakts. Die Verkettung dieser Röhrenstücke bildet ein durchaus brauchbares Modell des Vokaltrakts bei

einem bestimmten Vokal.

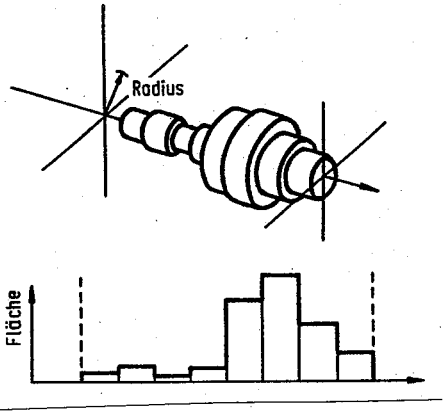


Abb.2.2d Röhrenmodell des Vokaltrakts

Bei den Nasalen muß natürlich ein zweites Röhrenstück parallel geschaltet werden.

Nachrichtentechnisch erhält man die Übertragungsfunktion als Faltung der Anregung (Obertöne) s(t) mit der komplexen Übertragungsfunktion h(t) des Artikulationstrakts.Für die Verbindung der Fouriertransformierten S(f) der Anregung s(t) und der Filterfunktion H(f) der Artikulation h(t) gilt bekanntermaßen die einfache Multiplikation, um das Spektrum der Gesamtübertragung zu erhalten, wie in Abb. 2.2e illustriert.

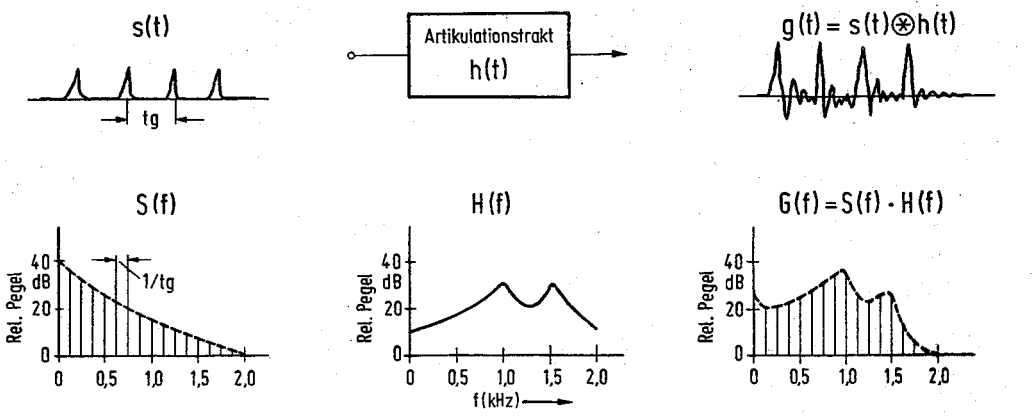
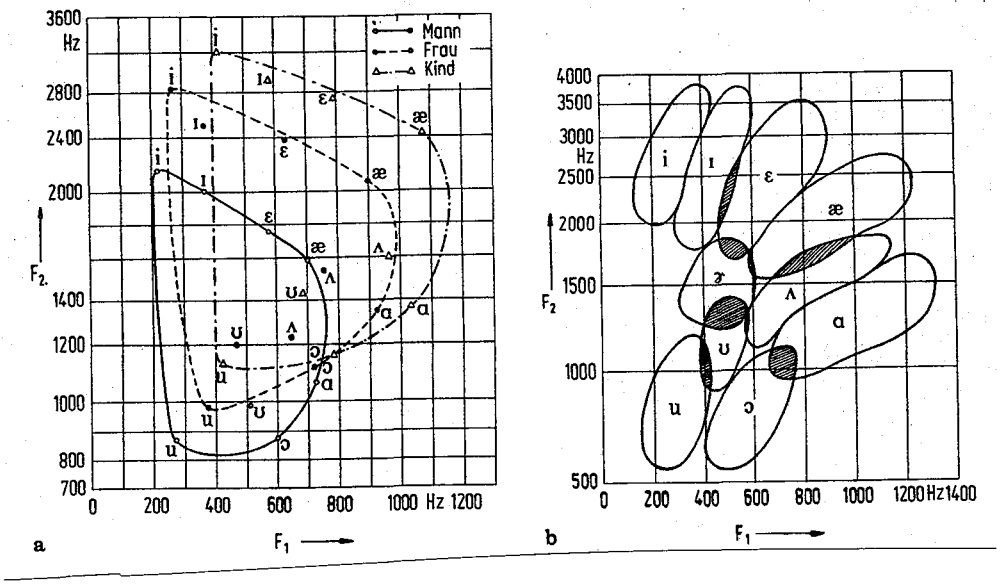


Abb. 2.2e Übertragungscharakteristik des Sprachtrakts

Bei den Nasalen kann auch eine Interferenz auftreten (Dämpfung bei nahe beieinander liegenden Resonanzen), was durch die Einführung von "Antiformanten" modelliert werden kann.

Die einzelnen Töne können also durch die Lage und Energie der einzelnen Formanten gekennzeichnet werden. Diese Merkmale sind allerdings stark sprecherabhängig. Abb.2.2f zeigt die Formantentupel (F_1,F_2) der amerikanischen Vokale von drei verschiedenen Personen sowie daneben den allgemeinen Streubereich der einzelnen Vokale.



Diese Merkmale sind aber zusätzlich zeitabhängig. Den Amplitudenverlauf und das Spektrogramm zweier Vokale 'a' und 'u' sind in Abb.2.2g gezeigt.

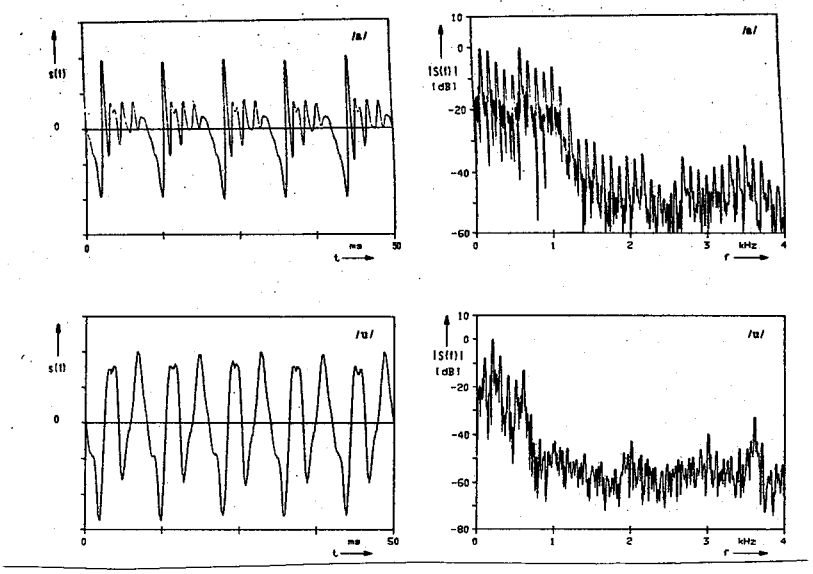
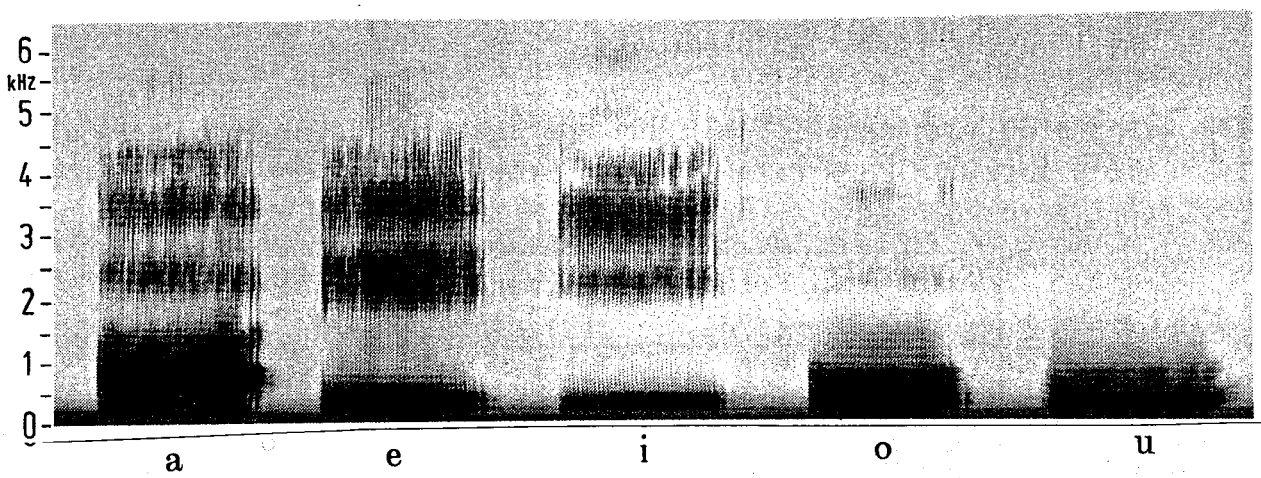


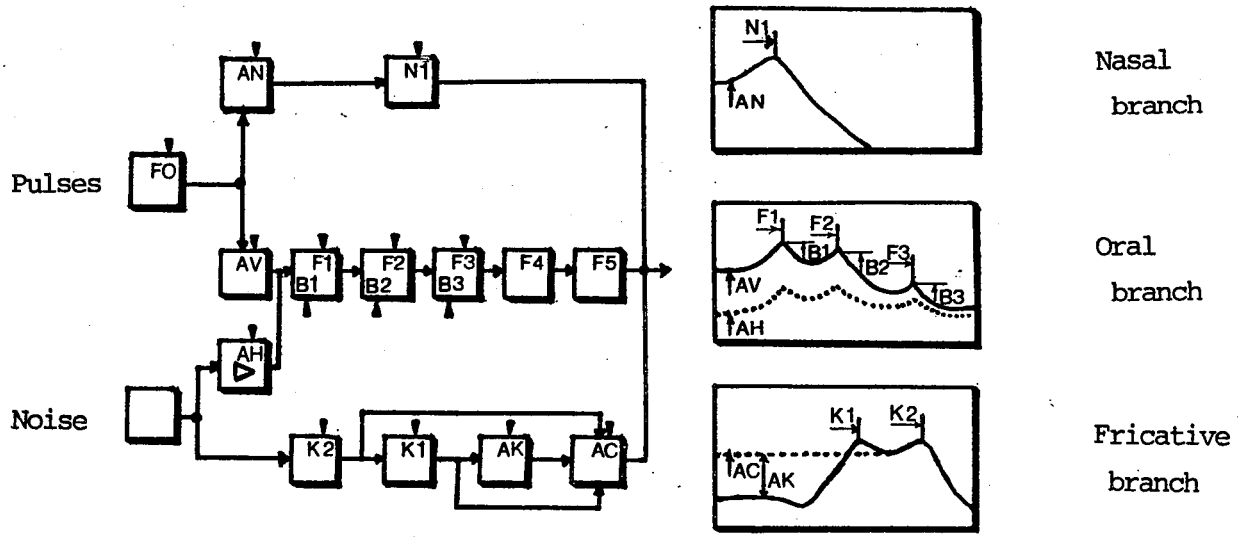
Abb. 2.2g

Markiert man die Intensität |S(f)| nicht durch einen Punkt der |S(f)|-f Ebene, sondern durch eine entsprechende Schwärzung auf der Frequenzachse, so lassen sich durch Nebeneinanderstellen der Schwärzungen an verschiedenen Zeitpunkten der zeitliche Verlauf der Intensitäten verschiedener Frequenzkomponenten (z.B. der Formanten!) verfolgen. Diese Form der spektralen, zeitabhängigen Darstellung heißt "visible-speech" Diagramm und ist in der Sprachforschung sehr beliebt. In Abb. 2.2h ist das Diagramm der 5 wichtigsten Vokale gezeigt.



. 'Visible Speech'-Diagramme (Spektrogramme) der wichtigsten Vokale

Welches sind nun die physikalischen Charakteristika der Sprache? Eines der bekanntesten Systeme, Sprache zu charakterisieren, stammt von FANT (1960). Im Folgenden soll das Blockdiagramm zur Sprachgeneration und die Erläuterung der verwendeten Parameter wiedergegeben werden.



Symbol und Bezeichnung	Physiolog.Basis	Akust.Beschreibung
1. A_V abs.Amplitude des Sprachsignals	Amplitude der Stimmbandschwingung	Intensität der per. Schallwellen
2. F_0 Pitch	Frequenz der Stimmbandschwingung	Grundfrequenz der per.Schallwellen
3. B_1 rel.Amplitude 4. F_1 Frequenz des 1.Formanten	1. Oberton des Vokaltrakts	Amplitude Frequenz der 1.Gruppe der Obertöne

5. B_2	rel.Amplitude	2. Oberton	Amplitude
6. F_2	Frequenz des 2.Formanten	des Vokaltrakts	Frequenz der 2.Gruppe der Obertöne
7. B_3	rel.Amplitude	3. Oberton	Amplitude
8. F_3	Frequenz des 3.Formanten	des Vokaltrakts	Frequenz der 3.Gruppe der Obertöne
9. A_k	abs.Amplitude des frikat. Geräuschs	Art der frikat. Verengung	abs.Intensität der aperiod.Komponenten
10. K_1	Frequenz des frikat. Geräuschs	Art der frikat. Verengung	Grundfrequenz der aperiod.Komponenten
11. A_c	rel.Amplitude des 2.frikat. Geräuschs	Art der frikat. Verengung	rel.Intensität der 2.frikat.Kompon.
12. K_2	Frequenz des 2.frikat. Geräuschs	Art der frikat. Verengung	Grundfrequenz der 2.frikat.Kompon.
13. A_H	Amplitude des Luftausstoßes	Höhlenreibung	aperiod.Komponent. der Formanten
14. A_N	Amplitude und	Resonanz des	Intensität und
15. N_1	Frequenz der nasalen Formanten	Nasentrakts	Frequenz der nasalen Resonanz

2.3 Phonetische Charakterisierung

Schon seit ziemlich langer Zeit versuchen Sprachwissenschaftler, den Sprechakt (cf.Phonetik) und das Gesprochene (cf.Phonologie) durch Normen und Regeln zu beschreiben. Dabei haben sich für Laute, die von Menschen als Teil der Sprache erkannt werden, verschiedene Klassifikationssysteme herausgebildet. Selbst physikalisch vollkommen verschiedene Laute, wie sie beispielsweise der Vokal [y] (aus dem frz."lune") darstellt, wenn er jeweils von einer Frau und einem Mann ausgesprochen wird (s.Abb.2.3a), werden von trainierten Phonetikern als "bis ins phonetische Detail gleich" bewertet.

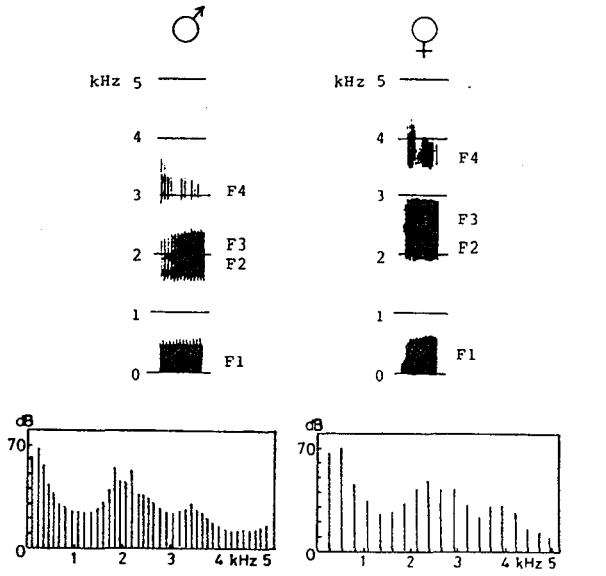


Abb. 2.3a Geschlechtsspezifische Sprachspektren von [y]

Im Unterschied zu den spracherkennenden Maschinen, die mit der Verarbeitung dieser geschlechtsspezifischen Sprachunterschiede ziemlich viele Probleme haben, scheinen Menschen aus den beiden so verschiedenen Sprachlauten eine Invariante herauszufiltern, die in beiden Fällen eindeutig **einen** Vokal [y] identifiziert. Diese Invarianten sind sehr unterschiedlich und damit typisch für die jeweilige Sprache. Beispielsweise sind die Vokalabstufungen einiger asiatischer Sprachen für uns unhörbar, da wir gelernt haben, sie als "unwesentliche Variationen" zu behandeln. Andererseits bilden die für uns wichtigen Konsonantenunterschiede (wie 'r' und 'l') nur unwesentliche Variationen in diesen Sprachen. Experimentelle, entwicklungspsychologische Arbeiten deuten darauf hin, daß diese Fähigkeiten bereits im Säuglingsalter erweorben werden.
Betrachten wir nun die gängisten phonetischen Charakterisierungen dieser kleinsten lautlichen Einheiten.

Phone
Die kleinsten unterscheidbaren Lauteinheiten werden **Phone** genannt. Ein Phon unterscheidet sich durch verschiedene Eigenschaften von den anderen Phonen:

- Klangfarbe (z.B.Verschiedene Vokale)
- zeitliche Länge (z.B. [a] in Nase und Tasse)
- Betonungsstärke (z.B. das erste [a] in Nasa)
- Tonhöhe durch Variation der Stimmlage

Es wird geschätzt, daß es in der deutschen Sprache über 40.000 Phone gibt.

Phonem
Verursacht die Änderung von Lauteinheiten auch einen Bedeutungsunterschied (wie bei Tanne und Tenne), so werden sie **Phoneme** genannt.

Allophone
Da das selbe Wort-je nach Kontext- verschieden ausgesprochen wird, ist ein Phonem meist nur der willkürlich ausgewählte Repräsentant einer ganzen Gruppe von Phonen, die als **Allophone** bezeichnet werden. Phoneme werden üblicherweise in schrägen Klammern // geschrieben, Allophone dagegen in eckigen Klammern [ ].
Ist die Wahl des Allophons vom Kontext bestimmt, so heißen sie **stellungsbedingte Allophone.**
Beispiel: [ç] steht nach /i/, /e/, /y/ ("**ich**","Bl**ech**","K**üche**"),
[x] dagegen nach /a/,/u/, /o/ ("W**acht**","W**ucht**,"W**oche**")
Vertauschen wir die Allophone probeweise, so sehen wir sofort an den "unmöglich klingenden" Worten, daß die Allophone stellungsbedingt sind.
Lassen sich die Allophone frei wählen (Beispiel: Zungen-r und Zäpfchen-r), so heißen sie **freie Allophone.**
Da die Aussprache eines Phonems meist vom restlichen Kontext abhängt, gibt es sehr viele Allophone. Beispielsweise ist die Aussprache des [k] verschieden in "Kasten", "Keller", "Kiste", "Kohle", "Kuhle", "klein", "wacker", etc. Insgesamt gibt es über 100 Allophone von [k].

Um trotz der Vielzahl der Lautvarianten eine einheitliche internationale Lautschrift zu erhalten, wurden die wichtigsten Phone der europäischen Sprachen zusammengestellt, die dann als Kombinationen die sprachtypischen Phoneme jeder Sprache darstellen können. Zweifelsohne ist dies bei der Zahl der existierenden Allophone nur eine erste, grobe Näherung. Für die deutsche Sprache ist nach der Norm der API bzw. IPA folgende phonetische Charakterisierung festgelegt:

a) Konsonanten

b	Bad	ba:t	n̩	roden	'ro:dn̩
ç	nicht	nɪç t	ŋ	Ding	'dɪŋ
d	dann	dan	p	plus	plus
d͡ʒ	George	d͡ʒɔ:d͡ʒ	p͡f	Pfand	p͡fant
f	falsch	fa'lʃ	r	rund	rʊnt
g	gut	gu:t	s	Last	last
h	heute	'hɔ͡ytə	ʃ	schon	ʃo:n
j	jetzt	je͡tst	t	Tanne	'tanə
k	kein	ka͡in	t͡s	Zahn	t͡sa:n
l	lang	laŋ	t͡ʃ	Klatsch	klat͡ʃ
l̩	Kittel	'kɪtl̩	v	wann	van
m	Mann	man	x	ach	ax
m̩	großem	'gro:sm̩	z	Vase	'va:zə
n	nach	na:x	3	Gage	ga:ʒə
			ˈ	Beamte	b 'ˈąmtə

Konsonanten für fremdsprachliche Aussprache

ð	father	'fɑ:ðə
θ	south	'saʊθ
ɲ	Cognac	kɔ'ŋak

b) Vokale

a	was	vas	o:	so	zo:
a:	Bahn	'ba:n	o̯	loyal	lo̯a'ja:l
ɐ	Meter	me:tɐ	ɔ	sonst	zɔnst
ɐ̯	Uhr	u:ɐ̯	ø	Ödem	ø'de:m
a̯i	Leim	la̯im	ø:	Röhre	'rø:rə
a̯u	Haus	ha̯us	œ	Töpfer	'tœp͡fɐ
e	Schwerin	ʃve'ri:n	ɔ̯y	neu	nɔ̯y
e:	geben	'ge:bn̩	u	zuviel	ʦu'fi:l
ɛ	West	vɛst	u:	Buße	'bu:sə
ɛ:	während	'vɛ:rənt	u̯	aktuell	ak'tu̯ɛl
ɛ	helfen	'hɛlfn̩	ʊ	Wunde	'vʊndə
i	Zigarre	ʦi'garə	y	Synagoge	zyna'go:gə
i:	ihm	i:m	y:	Typ	ty:p
i̯	Studie	'ʃtudi̯ə	ʏ	Etui	e'tʏi:
ɪ	Bitte	'bɪtə	ʏ	Syntax	'zʏntaks
o	Morast	mo'rast			

Vokale für fremdsprachliche Aussprache

ɑ	father	'fɑðə	ɛ̃:	Timbre	tɛ̃:br
æ	man(engl.)	mæn	ɨ	Gromyko	gra'mɨkɐ
ã	Chantal	ʃã'tal	õ	Fondue	fõ'dy:
ã:	Gourmand	gʊr'mã:	õ:	Chanson	ʃã'sõ:
ʌ	but	bʌt	œ̃	Parfum	par'fœ̃:
ɛ̃	timbrieren	tɛ̃'bri:rən			

Ergänzungen

ˀ	Knacklaut (Stimmritzen-Verschlußlaut, z.B. beacht! [bə 'ˀaxt].
:	Längezeichen; der vorhergehende Vokal wird lang gesprochen.
'	Hauptbetonung; steht unmittelbar vor der hauptbetonten Silbe.
~	nasale Vokale, z.B. [ã,ɛ̃].
̩	kennzeichnet silbischen Konsonant, z.B. kleben ['kle:bn̩].
̯	kennzeichnet unsilbischen Vokal, z.B. Statue ['ʃta:tu̯ə].

Abb.2.3b Phonetische Charakterisierung der deutschen Sprache

Auch die Folgen von Lauten tragen Bezeichnungen.

Silbe
Lesen wir ein rhythmisches Gedicht, so fällt uns auf, wenn die "Zahl der Silben" dabei nicht stimmt. Intuitiv haben wir dabei ein Gefühl für das, was mit **"Silbe"** gemeint ist. Verzichten wir auf die umstrittene exakte Definition und begnügen uns mit der intuitiven, so läßt sich beobachten, daß Silben meist die Form "Konsonant-Vokal-Konsonant" (CVC) haben. Der mittlere Vokal wird als **Silbengipfel** bezeichnet; die Laute davor bzw. danach als **Halbsilben.** Dies wird in manchen Spracherkennungssystemen verwendet, in denen versucht wird, Spracherkennung durch Erkennung von Halbsilben durchzuführen.

Morph
Betrachten wir das Wort "be-kleid-et". Die Vorsilbe "be" deutet auf eine Assoziation von "kleid" zu jemandem hin; die Endung "et" ist eine Vergangenheitsform.
Das Wort läßt sich also in drei Teile (Lautfolgen) unterteilen, die jeweils für sich eine Bedeutung haben. Diese Lautfolgen werden als **Morphe** bezeichnet.

Morphem
Verschiedene Morphe mit gleicher Bedeutung lassen sich zu einer Menge zusammenfassen, von denen eines -als Repräsentant ausgewählt- als **Morphem** bezeichnet wird.

Allomorphe
Analog zu den Allophonen heißen die Varianten eines Morphems **Allomorphe.** Auch hier gibt es wieder stellungsbedingte und freie Allomorphe.

Lexem
Steht ein Morphem für sich allein und läßt sich nicht weiter aufspalten, so heißt es ein **Wort** oder **Lexem.** Beispiel: "Hut".

Suprasegmentale Merkmale
Die große Zahl von Phonen und Allophonen basiert, wie erwähnt, im wesentlichen auf der Beeinflussung der Laute durch ihren Kontext. Die Ursache dafür sind Masse und Impuls der Zunge und der anderen am Sprechakt beteiligten, beweglichen Teile des Sprachapparats. Bei einer schnellen Sequenz von Phonemen können die nötigen Impulsänderungen (Zunge,..) und Stellungsänderungen (Lippen, Zähne,..) nicht in so schneller Folge durchgeführt werden. Deshalb kann jedes einzelne Phonem nicht klar ausgeprägt, sondern nur mehr oder weniger angenähert ausgesprochen werden. Wir tendieren dazu, vorzugsweise nur Phonemfolgen zu verwenden, die "nicht viel Mühe" machen, d.h. bei denen Impuls und Lage der beteiligten Organe möglichst wenig geändert werden muß. Dies führt dazu, daß eine Mund-, Zungen- oder Lippenstellung, die erst für ein späteres Phonem gebraucht wird, bereits vorher eingenommen wird (Bsp.: [k] -Allophone) oder aber nach dem Phonem weiterhin beibehalten wird und so andere Phoneme "färbt" (**Koartikulation**). Eine andere Möglichkeit besteht darin, Worte zusammenzuziehen und störende Phoneme einfach wegzulassen oder sie durch zusätzliche Phoneme zu "entschärfen".

Beispiel:frz. "le" und "la" werden vor einem Vokal zu "l'"; im Plural wird von "les" das 's' zusätzlich gesprochen.

Ein extremes Beispiel für diese Sprachmutation ist das Walisische, das je nach Kontext ein Wort mit einem anderen Anfangsbuchstaben anfangen läßt. Dies verleiht dieser Sprache einen sehr melodischen Aspekt. Beispiel dafür ist der Stadtname 'Bangor', der auch 'ym Mangor' (in Bangor) und 'i Fangor' (nach Bangor) ausgesprochen wird.
Da diese Sprachmutationen strengen Regeln unterworfen sind, sind die Ursprungsbuchstaben rekonstruierbar und die Spracherkennung dieser Teile ist nicht so problematisch, wie man denken könnte.
Weitere Arten, Phoneme zu beeinflussen, sind **Intonation** und **Betonung.** Frühere Sprachsythesizer, bei denen dies fehlte, prägten damit den Begriff der "Roboterstimme".
Als abschließendes Beispiel für die phonemübergreifende Beeinflussung ist in der folgenden Abbildung das Spektrogramm des Satzes "the girl was watching the fat man in the park" zu sehen.

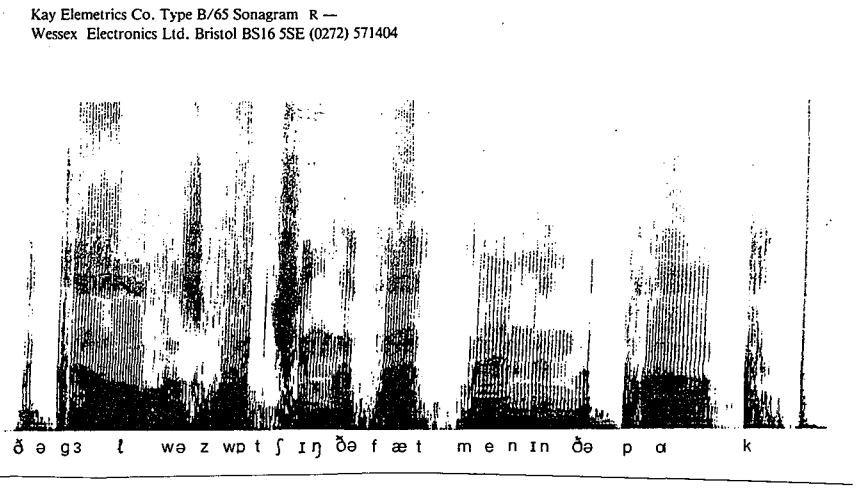


Abb.2.3c Zusammenziehen von Worten im Satz

Man beachte, daß in diesem zusammenhängenden Satz die Minima der Intensität nicht immer Wortgrenzen andeuten und andererseits Wortgrenzen nicht immer Minima haben ("wa_tching", "pa_rk"). Dies zeigt eines der dringendsten, ungelösten Probleme der aktuellen Forschung zur Erkennung kontinuierlich gesprochener Sprache.

2.4 Verständlichkeit

Will man beurteilen, wie gut man sich über eine neue Telefonleitung unterhalten kann, so reicht es nicht aus, die physikalisch-technischen Daten der elektrischen Signalübertragung zu beurteilen, Vielmehr muß man prüfen, ob die für die Wiedererkennung der im vorigen Abschnitt beschriebenen, bedeutungsunterscheidenden, lautlichen Einheiten benötigten physikalischen Signalanteile übertragen worden sind. Wie wir gesehen haben, ist aber diese physikalische Charakterisierung wiederum sehr problematisch, so daß sich es als bestes Meßinstrument bewährt hat, einen Menschen an das andere Ende der Leitung zu setzen und das subjektive Erkennen von gesprochener Sprache zu registrieren. Aus der Veränderung der Verständlichkeit beim Verändern der physikalischen Parameter der Übertragung läßt sich angeben, wie die Übertragungsverstärker möglichst kostengünstig gebaut werden können, ohne die Sprachverständlichkeit allzusehr zu verschlechtern. Weitere Erkenntnisse, die uns in diesem Zusammenhang eher interessieren, ist eine erhoffte Charakterisierung des Sprachverstehens durch die physikalischen Parameter.

Sprachgütemessungen
Als wichtigstes Verfahren der subjektiven Sprachgütemessungen haben sich die **Verständlichkeitstests** durchgesetzt. Dabei kann man zwei Gruppen unterscheiden: Zum einen die Verständlichkeit von **Worten** und **Sätzen**, zum anderen von **Silben** und **Lauten.** Grundsätzlich müssen bei diesen Verfahren, da ja die Verständlichkeit von der Wiedererkennung von Phonemen abhängt, das verwendete Lautmaterial (laute, Silben, Worte, Sätze) so selektiert werden, daß es phonetisch gemäß der Auftrittswahrscheinlichkeit der Phoneme in der deutschen Sprache ausgewogen ist. In der folgenden Abbildung ist dies von einem Testsatz gezeigt.

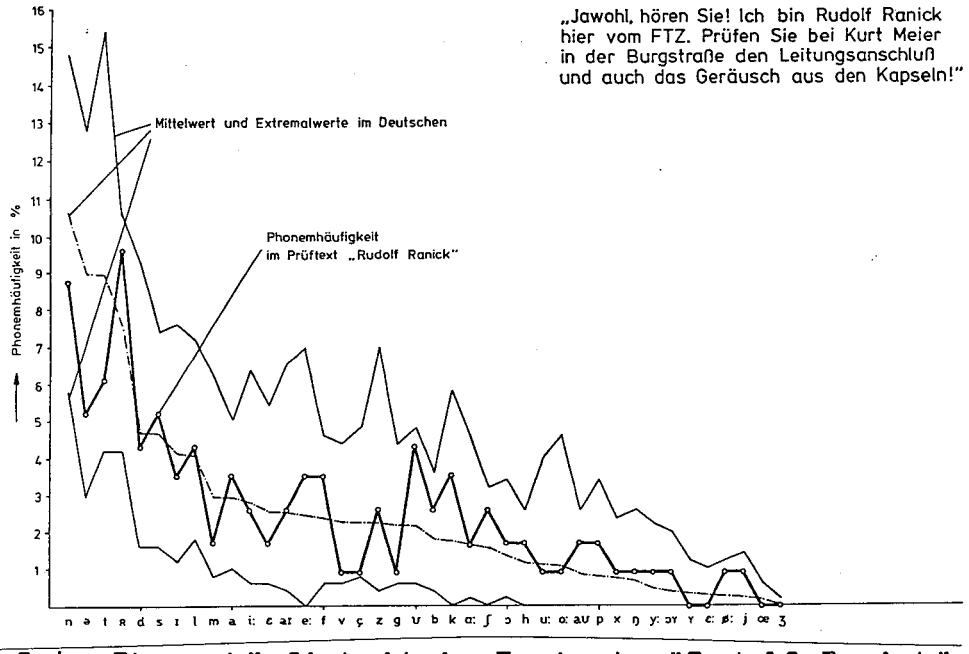
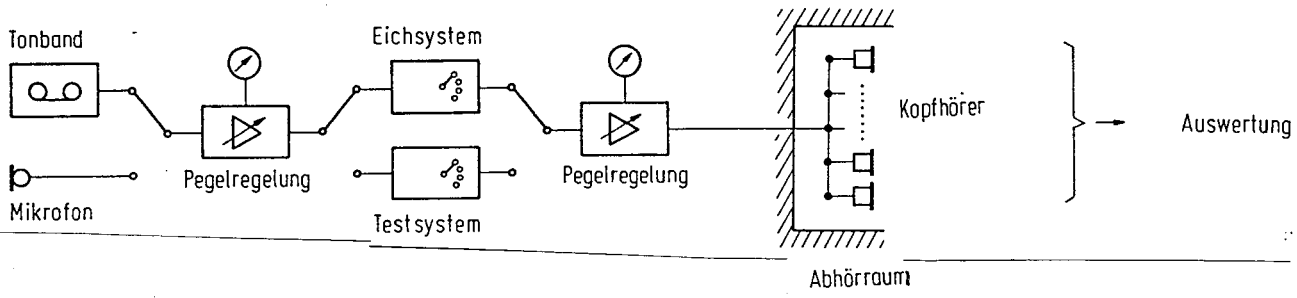


Abb.2.4a Phonemhäufigkeit im Testsatz "Rudolf Ranick"

Logatom-Tests

Da bei den Satz- und Wortverständlichkeitsmessungen fehlende, gestörte Laute, Silben und Wörter von Menschen ergänzt werden können, ist diese Art von Messungen meist für eine differenzierte Aussage über Übertragungscharakteristiken nur schlecht zu gebrauchen. Deshalb werden heutzutage in der Nachrichtentechnik meist nur Silben- und Lautverständlichkeitsmessungen durchgeführt. Hier läßt sich das Kontext-Hören dadurch unterbinden, daß künstliche, sinnlose Silben (**Logatome**) verwendet werden. Diese sind in der CVC-Form (s.2.3) genormt und werden ebenfalls phonetisch der verwendeten Sprache angepaßt. In Abb.2.4b ist ein Blockschaltbild der Testanordnung sowie eine Tabelle zur Logatombildung zu sehen.



Logatom-Aufbauelemente nach Schneider

50 Anlaute			50 Vokale	50 Auslaute	
b (2)	k (2)	schl	a (10)	b (2)	rs (2)
bl	kl	schr	e (10)	d (2)	s (3)
br	kr	sl	i (10)	f (2)	sch (3)
d (3)	l (2)	sp	o (10)	ft (2)	st (2)
dr	m (2)	st	u (10)	g (3)	t (3)
f	n (2)	str		k (3)	tsch(2)
fl	p	t (3)		l (3)	z (2)
fr	pl	tr		lf (3)	
g	pr	w (2)		m (3)	
gl	ps	z		n (3)	
gr	r (2)	zw		ng (2)	
h	s (2)			p (2)	
j	sch(2)			r (3)	

Wird ein Element mehr als einmal verwendet, so ist die Anzahl angegeben.

Abb.2.4b Testanordung und Logatomtabelle

Jedes der 50 Logatome wird aus drei Lauten (Anlaut,Inlaut, Auslaut) gebildet; die Kombination wird zufällig gewählt. Beispiele für Logatome sind "bab", "blef", "kruz", "glutsch", etc. Dabei werden die aus 8 Listen à 50 Logatome gebildeten 400 Logatome zur Hälfte (jedes 2.Wort) über das Testsystem und über das Eichsystem mit bekannter Verständlichkeit geschickt. Damit lassen sich individuelle Faktoren (Unaufmerksamkeit, Ermüdung etc) kompensieren.

Reimtests
Eine andere Form, die Phoneme darzubieten, ist der Reimtest. Dabei werden Worte oder Logatome beim Sender vorgesprochen. Gleichzeitig erscheint beim akustischen Empfänger auf einem Bildschirm eine Gruppe von 6 geschriebenen Worten, die sich in nur wenigen Phonemen vom übermittelten Wort unterscheiden. Die Aufgabe besteht nun darin, von den sich reimenden Worten das Richtige mit einem Knopfdruck zu kennzeichnen. Die folgende Abbildung zeigt Beispiele dazu.

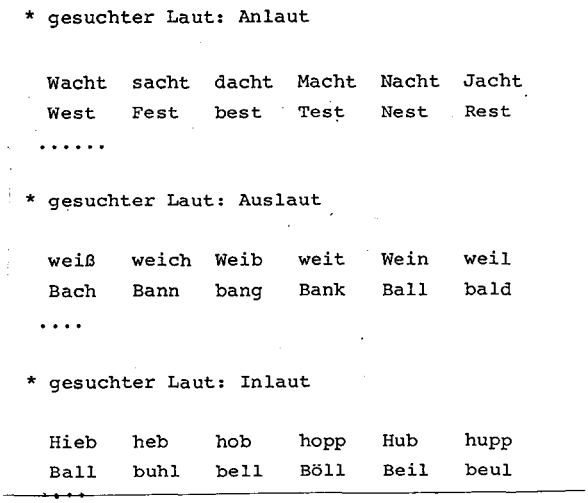


Abb. 2.4c Beispiele für Reimtests

Der Vorteil dieses Verfahrens liegt in dem geringeren Versuchsaufwand um einen Meßwert der Verständlickeit bei festen physikalischen Parametern zu ermitteln. Mußten beim Logatom-Test noch mit 4 Versuchspersonen insgesamt 1600 Logatome abgehört, aufgeschrieben und ausgewertet werden, so reduziert sich der Aufwand bei 100 Reimgruppen und 12 Testpersonen auf insgesamt 1200 Worte, die per Knopfdruck automatisch protokolliert und per Rechner ausgewertet werden können. Berücksichtigt man noch bei der Auswertung, daß auch eine Wahrscheinlichkeit dafür besteht, das gesuchte Wort beim Reimtest richtig zu raten, so sind die Ergebnisse genauso gesichert wie beim Logatom-Test. Der einzige Nachteil der Reimtests liegt darin, daß die Gütefunktion erst bei schlechten Übertragungsbedingungen deutlich absinkt, dann aber stärker als beim Logatom-Testverfahren.

Ergebnisse
Die Auswertung der Meßergebnisse hängt stark von der vorgegebenen Fragestellung ab. Ist beispielsweise gefragt, wieviel Aufwand bei den Übertragungsverstärkern zur Unterdrückung von Störgeräuschen getrieben werden muß, um eine gute Verständlichkeit zu garantieren, so ergibt sich folgende Meßkurvenschar:

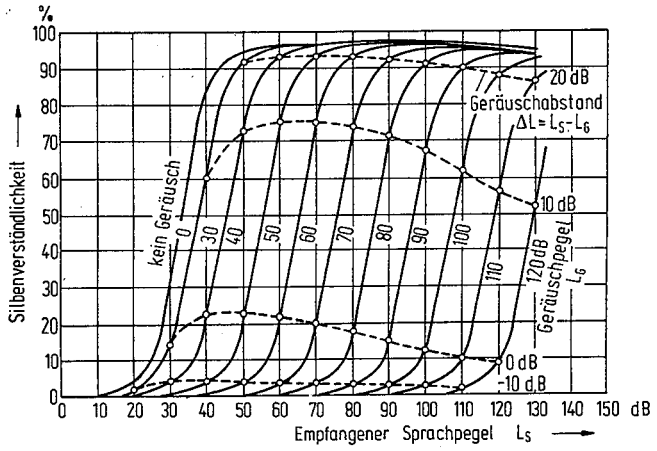


Abb.2.4d Verständlichkeit bei verschiedenen Rauschpegeln

Wie man der Kurve entnimmt, ist ein Störgeräuschabstand von 20dB bei allen Lautstärken ausreichend, um eine Verständlichkeit von 90% zu garantieren.
Wird andererseits als physikalischer Parameter beispielsweise die oberste oder unterste, übertragbare Frequenz (Grenzfrequenz und Bandbreite) gewählt, so ergeben sich die beiden Kurven in Abb. 2.4e.

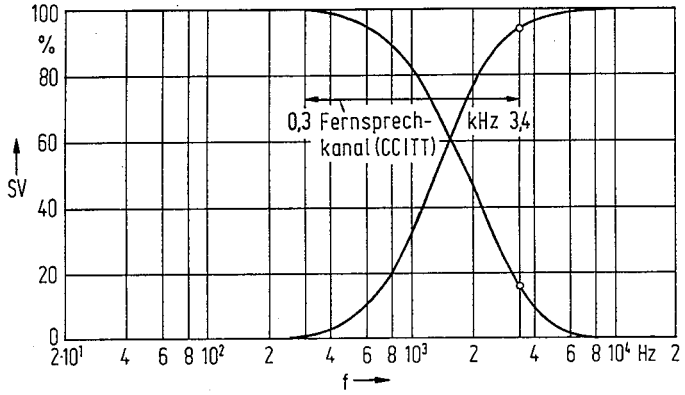


Abb.2.4e Mindest-Bandbreite der Sprache

Aus dem Diagramm ist ersichtlich, daß für eine verständliche Telefonunterhaltung ausreichend ist, Frequenzen von 300 Hz bis zu 3,4 KHz zu übertragen, so daß sich dies als Fernsprechnorm durchgesetzt hat. Die stärkere Beschneidung bei den hohen Frequenzen garantiert auch eine Rauschunterdrückung, da das thermisch bedingte Rauschen erst ab 5 KHz stärker auftritt.

3.0 Kodierung und Synthese des Sprachsignals

In den vorigen Abschnitten beschäftigten wir uns mit der Frage, wie das Sprachsignal charakterisiert werden kann, um durch eine automatische Erkennung der charakteristischen Merkmale eine automatische Spracherkennung zu ermöglichen. Dieses Problem läßt sich auch mit den Mitteln der Nachrichtentechnik anders formulieren: Wie ist die Sprachinformation im Sprachsignal kodiert?
Ist diese Frage beantwortet, so reduziert sich das Problem der Spracherkennung auf das Problem der Kodierung und Dekodierung des Sprachsignals. Da auch im menschlichen Hörsystem (s. Kap.5) eine Umkodierung der Schalleindrücke vorgenommen wird, können wir sicher einiges für die Spracherkennung lernen, wenn wir die Erkenntnisse und Verfahren der Nachrichtentechnik beim Kodieren und Dekodieren von Sprache betrachten.
Dazu vergleichen wir die haupsätzlichen Möglichkeiten, sprachliche Information zu übertragen:

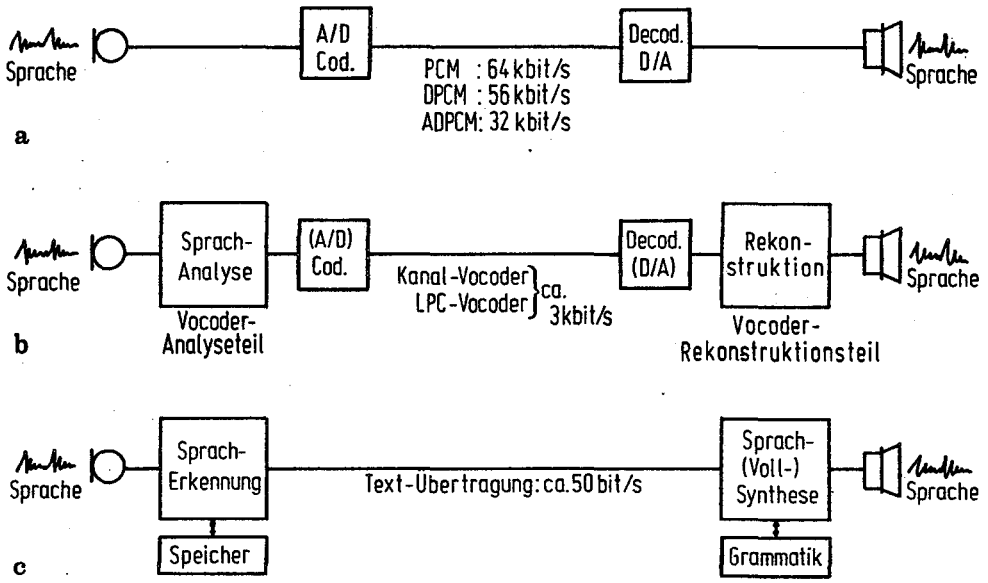


Abb.3.0a Übertragungssysteme für Sprache

Wie man sieht, gibt es drei verschiedene Grundverfahren zur Sprachübertragung:
a) die digitale Kodierung des physikalischen Signals
b) die Extraktion und Kodierung von quantitativen Sprachmerkmalen
c) die vollständige Spracherkennung und Übertragung des reinen Texts

3.1 Digitale Kodierung

Ähnlich wie ein Sprachsignal, das trotz störender Umweltgeräusche beim Zuhörer("Empfänger") noch verständlich sein soll, soll auch in der Nachrichtentechnik ein physikalisches Signal trotz widriger Umwelteinflüsse ("Störrauschen") möglichst ungestört beim Empfänger ankommen. Für dieses elemtare Problem gibt es viele Lösungsansätze. Einer der erfolgversprechendsten ist die digitale Kodierung. Worum geht es bei diesem Verfahren?
Betrachten wir dazu die folgende Abbildung.

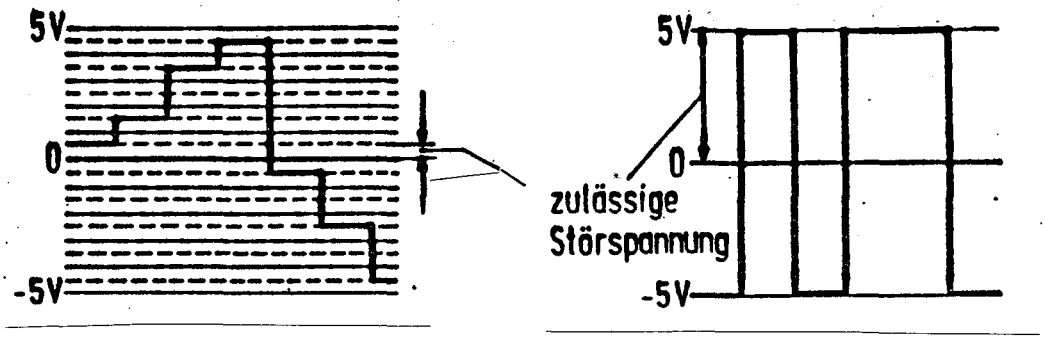


Abb.3.1a Störabstand zweier werte-diskreter Signale

Links ist ein Signal zu sehen, das im Intervall [-5V, +5V] 10
Werte annehmen kann. Übertragen wir anstelle eines wertekontinuierlichen Signals nur diese diskreten Werte, so kann eine Störung oder Verfälschung vom Empfänger korrigiert werden, insofern sie nicht größer als 0.5 Volt ist: Da nur die wenigen, diskreten Werte erlaubt sind, sind alle davon abweichenden Werte gestört und können durch Verschieben zum nächstliegenden, erlaubten Wert hin korrigiert werden. Eine Diskretisierung der Werte erlaubt also eine Fehlerkorrektur.
Verringern wir die Zahl der möglichen Werte, beispielsweise bis auf zwei wie im rechten Teil der Abb.3.1a, so erhöht sich zweifelsohne die Möglichkeit, Störungen zu erkennen und zu korrigieren. Andererseits aber ist das resultierende Signal dem ursprünglichen Signal immer unähnlicher. Eine Abhilfe dafür schafft eine andere Idee: Übertragen wir nämlich das gewünschte Signal nicht zeitkontinuierlich, sondern nur als eine Folge von Amplitudenwerten von aufeinanderfolgenden Zeitpunkten, so läßt sich aus dieser Folge ebenfalls beim Empfänger das ursprüngliche Signal näherungsweise aufbauen. Die kontinuierlichen, reellen Amplitudenwerte lassen sich nun ebenfalls diskretisieren, so daß das ursprüngliche Signal durch eine Folge von zeitdiskreten und wertediskreten Abtastpunkten (Zahlen) übertragen wird.

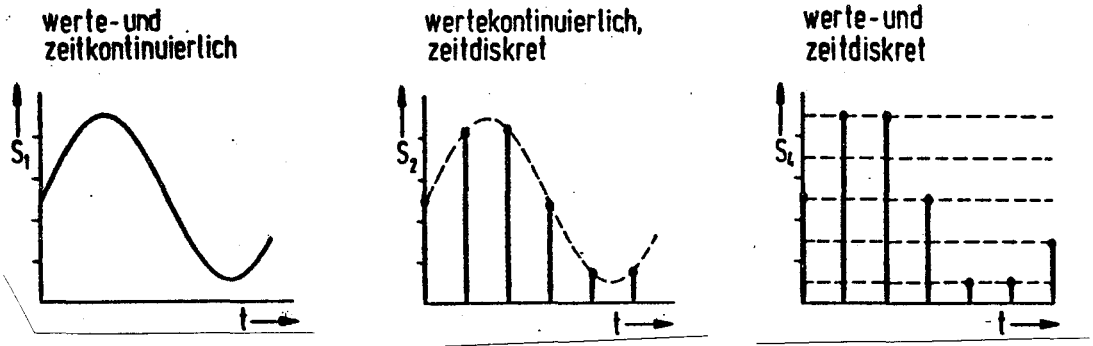


Abb.3.1b Zeit- und Wertediskretisierung eines Signals

Jede Zahl läßt sich binär kodieren, so daß die Folge der binär kodierten Signalwerte physikalisch ein zeitkontinuierliches Signal mit zwei Werten bildet, das einen hohen Störabstand aufweist.
Der beschriebene Kodierungsprozeß heißt **Analog-Digital Konvertierung** (A/D) und ist ein beliebtes Hilfsmittel, analoge Meßgrößen zur Speicherung und rechnergestützten Verarbeitung aufzubereiten. Bei der Sprachübertragung findet sie bei der direkten und parametrischen Übertragung von Sprachsignalen Anwendung(s. Abb.3.0a).

Allerdings gibt es bei der digitalen Kodierung auch Probleme. Lassen wir die Abtastprobleme außerhalb unserer Betrachtung, so verbleibt ein wichtiges Problem: die ausreichende Abschirmung von Störeinflüssen.
Betrachten wir dieses Problem genauer. Analysieren wir unser empfangenes Signal, so sehen wir, daß die Gesamtintensität (Quadrat der Amplitude) aus zwei Teilen zusammengesetzt ist: das ursprüngliche Signal P_s und das überlagerte Störgeräusch P_q, das beispielsweise von der Signalverzerrung durch zu grobe Quantisierung herrührt. Der Quotient aus beiden Anteilen wird logarithmisch in der Einheit "Dezibel"(dB) gemessen und wird als das **Signal-Geräusch-Verhältnis** (SNR) bezeichnet:

$$SNR = 10 \lg P_s/P_q \qquad [dB]$$

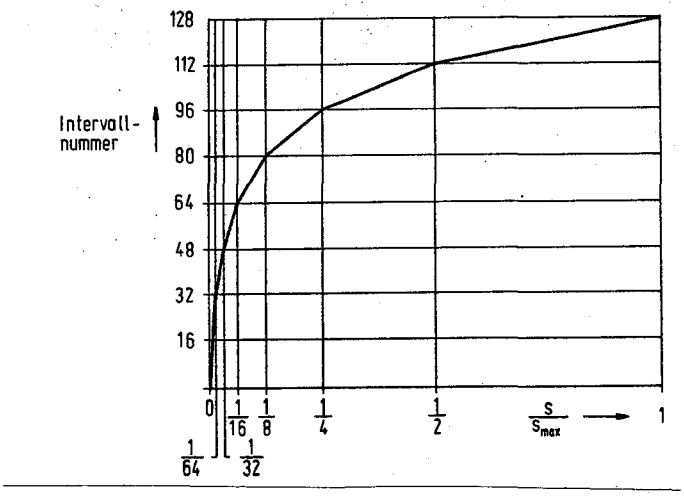
Sind Signal und Störgeräusch gleich, so ist das SNR Null; das Signal "geht im Störgeräusch unter".
Das Gütekriterium SNR hilft uns, Maßnahmen zur Anhebung der Übertragungsqualität quantitativ zu beurteilen. Beispielsweise läßt sich errechnen, daß der A/D Konverter, um eine Übersteuerung zu vermeiden, den 4-fachen Bereich der effektiven Sprachamplitude umfassen sollte. Verdoppelt man bei einem solchen Konverter die Anzahl der Quantisierungsstufen (Erhöhung der Kodierung um 1 Bit), so verbessert sich das SNR durch geringeren Quantisierungsfehler um 6 dB. Legt man dies zugrunde, so müßte ein 8 Bit A/D Wandler einen Störabstand von 40 dB haben und damit eine gute Sprachverständlichkeit aufweisen. Tatsächlich aber benötigt man 12 Bit.
Woran liegt das ?
Beobachten wir die tatsächlich übertragenen, digitalen Werte, so stellen wir schnell fest, daß meist nur niedrige Werte übertragen werden; die gesamte Aussteuerungsbreite wird nur selten genutzt. Diese Häufung bei niedrigen Amplitudenwerten und wenigen Quantisierungsstufen beschert uns einen größeren Quantisierungsfehler, als wir erwartet haben. Würden wir die niedrigen Amplitudenwerte beim Sender gleichmäßig erhöhen und beim Empfänger entsprechend erniedrigen, so müßte das SNR amplitudenunabhängiger werden. Diese Maßnahme der **Kompandierung** (Kompression) der Signale wird tatsächlich angewendet; Beispiel ist das bekannte analoge Dolby-Verfahren bei Kassetten-Recordern, wobei allerdings das Störgeräusch thermische Ursachen hat. Analoge Verfahren der Kompandierung haben allerdings ein Problem: die zur Kompression benutzte Kennlinie ist invers bei der analogen Dekompression meist nicht exakt realisiert, da Verstärkungsfaktoren und Offsetspannungen thermisch nicht stabil, sind. Besser ist es, bei der Digitalisierung nicht die kleinen Amplituden zu erhöhen, sondern die Zahl der Quantisierungsstufen in diesem Bereich zu vergrößern. Verringert man dazu noch entsprechend die Zahl der Quantisierungsstufen bei den großen Amplituden, so resultiert eine **ungleichmäßige Quantisierung.** Hat diese nicht-lineare Quantisierung logarithmischen Verlauf, so erhält man ein pegelunabhängiges, konstantes SNR. Deshalb wird beim europäischen Puls-Code-Modulationsverfahren (PCM) die sogenannte **A-Kennlinie** bei der Transformation y=f(x) des Eingangspegels x zum Ausgangspegel y verwendet:

$$y = \frac{1+\ln(Ax)}{1+\ln(A)} \qquad 1/A \leq x \leq 1 \qquad \text{(log. Teil)}$$

und $A := 87{,}56$

$$y = \frac{Ax}{1+\ln(A)} \qquad 0 \leq x \leq 1/A \qquad \text{(linearer Teil)}$$

die bei x=1/A ineinander übergehen.
Die logarithmische Kennlinie läßt sich durch lineare Stücke nicht nur in der Nähe des Nullpunkts annähern, sondern durch ein Polygonzug im gesamten Bereich. Beispiel dafür ist die **13-Segment-Kennlinie**, die sich bei der Digitalisierung bewährt hat.



Segment	Codewort
6	1111XXXX
5	1110XXXX
4	1101XXXX
3	1100XXXX
2	1011XXXX
1	1010XXXX
0	1001XXXX
	1000XXXX

Abb. 3.1c Der obere Teil der 13-Segment-Kennlinie

In der Abbildung ist links die Kennlinie der Kompandierung gezeigt, rechts die entsprechende Kodierung in den Segmenten. Im Unterschied zu den Segmenten 1..6 mit jeweils 4 Bit Quantisierung sind im Segment Null um den Nullpunkt jeweils 5 Bit im positiven wie auch im negativen Amplitudenbereich vorgesehen, also insgesamt 6 Bit einer linearen Quantisierung in einem Bereich 1/64 = 2^{-6} der Gesamtaussteuerung. Damit wird im Segement Null so quantisiert wie es eine 12-Bit Auflösung im Gesamtbereich tun würde. Abbildung 3.1d zeigt das Resultat.

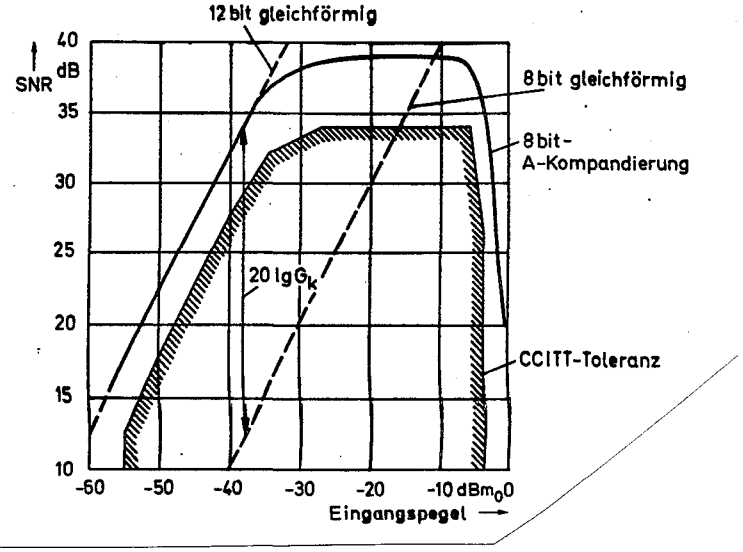


Abb.3.1d Störabstand der 13-Segment Kennlinie

Man sieht, daß nach dem linearen Teil im Segnment Null (-55 bis -

35 dB) die logarithmische Quantisierung einen ungefähr konstanten SNR bewirkt, um ca. 24 dB erhöht gegenüber einer reinen 8 Bit Auflösung. Bei sehr starken Amplituden (-5 bis 0 dB) macht sich rasch die Signalverzerrung durch Übersteuerung bemerkbar.

3.2 Lineare Prädiktion

Bei der digitalen Kodierung ist zwar ein guter Störabstand garantiert, dieser Vorteil wird aber durch einen gegenüber dem Orignalsignal (bis 20KHz) erhöhtem Übert ragungsaufwand (64 KBit/sec, also mind. 64KHz) wieder wettgemacht. Die Frage ist nun: Läßt sich die Kodierung vereinfachen und damit der Übertragungsaufwand senken?
Betrachten wir ein normales Sprachsignal, so stellen wir fest, daß die Folge der Signalwerte nicht völlig zufällig und beliebig ist, sondern von Wert zu Wert sich meist nur wenig ändert. Nutzen wir diese Redundanz aus, so müßte es möglich sein, Sprache ökonomischer zu kodieren als beim log. PCM-Verfahren des vorigen Abschnitts.
Ein erster Ansatz könnte davon ausgehen, daß wir nur konstante Änderungen der Signalwerte betrachten. Wird das Sprachsignal größer, so übertragen wir Einheitssignale für positive Sprachsignaländerungen (+1); nimmt das Sprachsignal ab, so senden wir Signale für negative Änderungen (-1). Abbildung 3.2a zeigt den Verlauf des Signales $s(t_n)$ und seiner Annäherung $\hat{s}(t_n)$ durch eine sich konstant ändernde Funktion (Treppenfunktion).

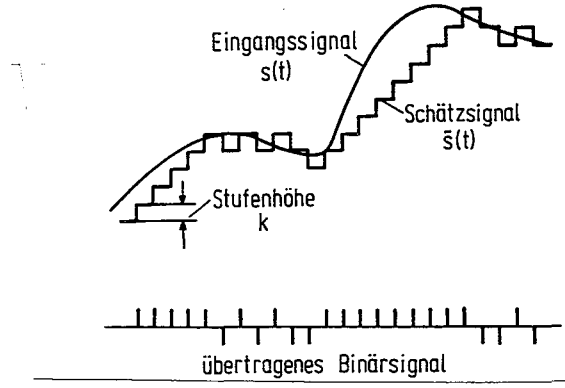


Abb 3.2a Schätzung durch eine konstante Änderung

Dieses Verfahren heißt **Deltamodulation** (DM) und benutzt nur relativ einfache Hardware. In Abb.3.2b sind die Blockschaltbilder von Sender und Empfänger gezeigt. In beiden Systemen wird aus dem Kodesignal für positive oder negative Änderung durch Integration der geschätzte Sprachsignalwert $\hat{s}(t_n)$ gewonnen.

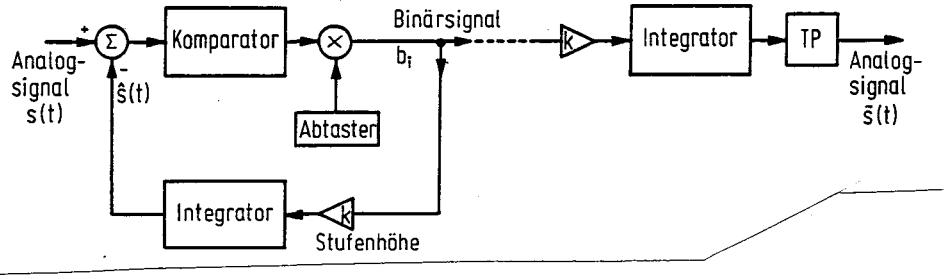


Abb. 3.2b Blockschaltbild eines Systems zur Deltamodulation

Beim Sender wird das Vorzeichen +1 oder -1 des Kodesignals $b(t_n)$ mittels eines einfachen Komparators bestimmt:

Mit $$d_n := s(t_n) - \hat{s}(t_n)$$

ist $$b(t_n) := \begin{cases} +1 & d_n > 0 \\ -1 & d_n \leq 0 \end{cases}$$

Eie lineare Annäherung an das Sprachsignal führt bei schnellen Signaländerungen leicht zu Fehlschätzungen von $\hat{s}(t_n)$. Abhilfe dafür schafft die Überlegung, das Kodesignal nicht nur von einem, dem letzten Wert von s(t) abhängen zu lassen, sondern von mehreren vorhergehenden Werten. Werden bei schnellen Amplitudenänderungen mehrere Male hintereinander positive Inkremente nötig, so soll auch die Größe der Inkremente ansteigen, um sich s(t) schneller anzunähern. Kehrt sich das Vorzeichen der Inkremente um, so wird die Größe der Inkremente wieder reduziert:

$$b(t_i) = b(t_{i-1}) = \begin{cases} b(t_{i-2}) = +1 \\ INC^{i-2} = k_1 \end{cases} \Rightarrow \quad INC := k_2$$

$$\begin{cases} b(t_i) = +1 \\ INC^i = k_2 \end{cases} \Rightarrow \quad INC := k_3$$

$$\begin{cases} b(t_i) = -1 \\ INC^i = k_3 \end{cases} \Rightarrow \quad INC := k_2$$

$$\begin{cases} b(t_i) = -1 \\ INC^i = k_2 \end{cases} \Rightarrow \quad INC := k_1$$

$$k_1 < k_2 < k_3$$

Dieses modifizierte Deltaverfahren, das seine Inkremente an das aktuelle Signal anpaßt, heißt **adaptive Deltamodulation (ADM).** Die Abbildung 3.2c zeigt davon eine typische Kodierungssequenz

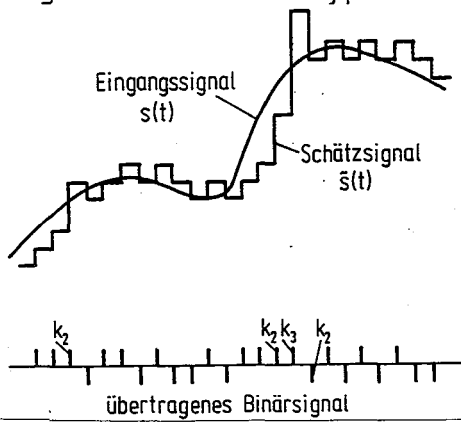


Abb. 3.2c Näherung und Kodierung beim ADM-Verfahren

Vergleichen wir die Leistung der verschiedenen Verfahren (Signal-

Rauschverhältnis) bei verschiedenen Abtastraten in Abbildung 3.2d

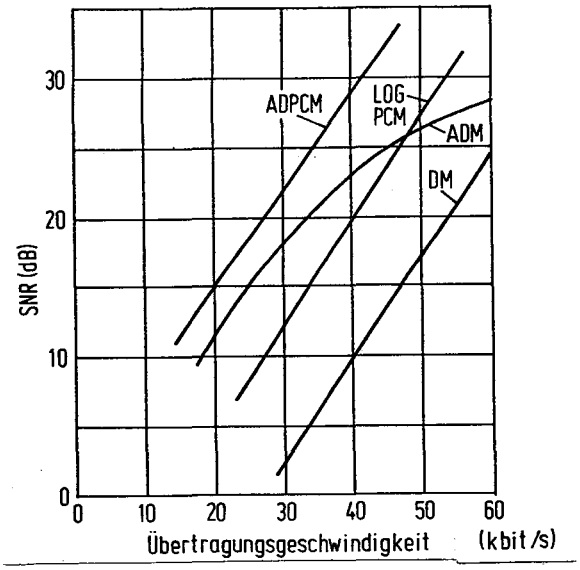


Abb. 3.2d Vergleich der verschiedenen Kodierungsverfahren

so fällt uns auf, das es noch ein besseres Verfahren gibt mit der Bezeichnung ADPCM. Worum handelt es sich dabei?
Zweifelsohne ließe sich das ADM-Verfahren verbessern, wenn man anstelle der festen Inkremente variable Inkremente benutzt, diese kodiert und den Kode zusätzlich übermittelt. Nehmen wir außerdem zusätzlich zum vorigen Wert $\hat{s}(t_{n-1})$ noch weitere, frühere Werte $\hat{s}(t_{n-2})$ .. $\hat{s}(t_{n-p})$ für die Berechnung der Schätzung $\hat{s}(t_n)$ hinzu, so ist die digital zu kodierende Differenz mit der Notation $s(t_i) = s_i$

$$d_n := s_n - \sum_{i=1}^{p} a_i \hat{s}_{n-i} \qquad (3.2a)$$

Wird die Differenz im PCM-Kode übermittelt, so heißt das Verfahren **Differenz-PCM (DPCM)** Verfahren. In Abbildung 3.2e ist ein solches System gezeigt.

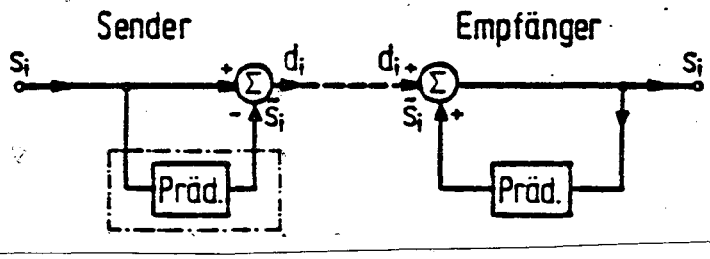


Abb.3.2e Blockschema eines DPCM-Systems

Ein solches lineares Netzwerk läßt sich auch modular aufbauen, so daß leicht durch eine Aneinanderreihung in Chip-Form oder VLSI-Design-Makro eine schrittweise Verbesserung der Schätzung und damit auch ein besserer Rauschabstand (s. Abb.3.2d) möglich ist. Eine solche Umformung ist in Abb. 3.2f gezeigt.

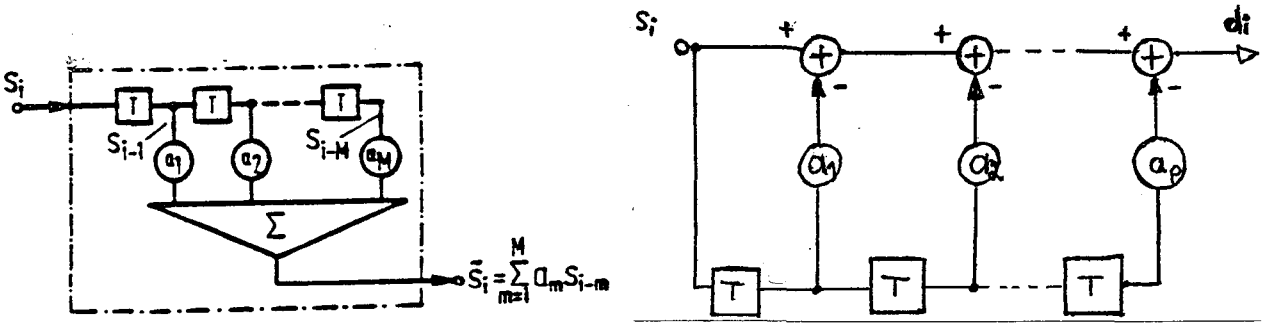


Abb.3.2f Prädiktorfilter aus 3.2e und modulare Systemversion

Das oben beschriebene Verfahren, jeden Wert aus den p vorhergehenden Werten zu schätzen, heißt **Prädiktion.** Im Unterschied zu dem von Kolmogorov in den vierziger Jahren angegebenen Verfahren mit allgemeinen Polynomen reicht es normalerweise aus, die Schätzung durch eine lineare Annäherung zu erreichen. Diese **lineare Prädiktion (LPC)** ist ein allgemeines Verfahren und wird gern in verschiedenen Aufgabenstellungen verwendet.
Die p **Prädiktorkoeffizienten** a_i sind konstant und müssen für das vorliegende Sprachmaterial optimal ermittelt werden. Dazu wird die über die Zeit gemittelte, quadrarische Differenz (mittlerer quadratischer Fehler) minimisiert

$$\overline{d_i^2} \overset{!}{=} \min \qquad (3.2b)$$

Leiten wir diese Gleichung nach dem Parameter a_i ab und setzen die Ableitung null, so erhalten wir die optimalen Werte als Lösung der resultierenden Gleichungen. Konnte man noch die Summe aus Gleichung (3.2a) als Produkt zweier Vektoren $\mathbf{a} = (a_1,..,a_p)$ und $\mathbf{s} = (s_{n-1},..,s_{n-p})$ schreiben, so führt uns die p-dimensionale Ableitung von (3.2b) zu einem linearen Gleichungssystem

$$\mathbf{S\,a} = \mathbf{c} \qquad (3.2c)$$

mit $\mathbf{c} := (S_{10},S_{20},..,S_{p0})$ und $S_{jk} = \overline{s_{n-j}\,s_{n-k}}$.
Die zeitliche Mittelung der Matrixkoeffizienten läßt sich im Limes auf sämtliche Werte und damit aufs unendliche Intervall ausdehenen, so daß der Lösungsvektor **a** das Optimum bezüglich des gesamten, über das System übertragenen Sprachmaterials darstellt. Die Matrixkoeffizienetn sind dann nur noch vom relativen zeitlichen Abstand und nicht mehr vom absoluten Zeitpunkt n der betrachteten Abtastpunkte abhängig. Mit z.B. n=k ist

$$S_{jk} = \overline{s_{k-j}s_0} = S_{k-j,0} = c_{|j-k|}$$

Die pxp Matrix **S** nimmt damit eine sehr symmetrische Form an; alle Elemente der Hauptdiagonalen sowie der Nebendiagonalen sind jeweils gleich. Diese Matrix-Form heißt "Toeplitz-Form" und kann sehr effizient rekursiv gelöst werden (Levinson Rekursion). Die n-te Stufe lautet

$$a_n^{(n)} = \frac{c_n - \sum_{K=1}^{n-1} c_{|n-k|}\, a_k^{(n-1)}}{c_0 - \sum_{K=1}^{n-1} c_k a_k^{(n-1)}} \qquad (3.2d)$$

mit $1 \leq k \leq n-1$ und $2 \leq n \leq p$, wobei

$$a_k^{(n)} = a_k^{(n-1)} - a_n^{(n)}\, a_{n-k}^{(n-1)}$$

ist.
Bei Kenntnis der Koeffizienten von **c** lassen sich die Prädiktorkoeffizienten **a** errechnen.

Bei diesem Verfahren kann man nun einige interessante Nebenergebnisse beobachten.

Partielle Korrelation (PARCOR-Koeffizienten)
Die positive (oder positiv-semidefinite), symmetrische Matrix **S** aus (3.2c) läßt sich auch als Produkt zweier Dreiecksmatrizen schreiben.
Damit wird (3.2c) zu

$$\mathbf{q} = \mathbf{L}^T \cdot \mathbf{a}$$
$$\mathbf{L} \cdot \mathbf{q} = \mathbf{c}$$

mit $\mathbf{q} = (q_1, .., q_p)$. Dann lassen sich die sog. **PARCOR-Koeffizienten** definieren als

$$r_m = \frac{q_m}{(c_0 - \sum_{i=1}^{m-1} q_i^2)^{1/2}}$$

wobei $r_m \leq 1$ ist.
Es läßt sich zeigen, daß diese Koeffizienten den Wert einer partiellen Autokorrellation widergeben, d.h. ein Maß für die Ähnlichkeit von zwei Signalwerten s_n und s_{n-m} darstellen, bei dem der Einfluß der dazwischenliegenden Werte $s_{n-1}, .., s_{n-(m-1)}$ eleminiert worden ist.
Erweitern wir das betrachtete Intervall wieder auf die gesamte unendliche Folge von Signalwerten, so läßt es sich zeigen, daß dann r_m und $a_m^{(m)}$ aus Gleichung (3.2d) ineinander übergehen und die einen Koeffizienten mittels der anderen ausgerechnet werden können.

Lineare Filter
In Abb.3.2e ist gezeigt, wie das Eingangssignal s(t) in dem System der Linearen Prädiktion geformt wird und als Ausgangssignal das System wieder verläßt. Betrachtet man das umformende System als allgemeinen Filter, so stellt sich die

Frage: was ist die Übertragungsfunktion dieses Systems?
Das Differenzsignal, das aus dem abgetasteten Eingangssignal $s(t_n) =: s_n$ gewonnen wird, ist

$$d(n) := s_n - \hat{s}_n = s_n - \sum_{i=1}^{N} a_i \, s_{n-i} \qquad (3.2a)$$

Wollen wir die Frequenzabhängigkeit des Differenzsignals feststellen, so erreichen wir dies durch eine diskrete Fouriertransformation mit der Frequenzvariablen k

$$Q(k) = \sum_{n=1}^{N} d(n) \, e^{-in2\pi k/N} \qquad k=1,..,N$$

Da die Fouriertransformation nur für unendliche Intervalle gilt, muß das betrachtete Intervall periodisch zeitlich davor und danach fortgesetzt werden.
Die komplexe Variable

$$z_n := e^{-i2\pi k/N}$$

hat den Betrag $|z_n|=1$; der Punkt z_n liegt also auf dem Einheitskreis und ist durch die Frequenz k/N gekennzeichnet. Die Transformationsgleichung mit dieser verallgemeinerten Frequenz z lautet somit

$$Q(z) = \sum_{n=1}^{N} d(n) \, z^{-n}$$

und wird als **z-Transformation** von d(n) bezeichnet.
Die **Rücktransformation** dieser verallgemeinerten Fouriertransformation ist

$$s(n) = \frac{1}{2\pi i} \oint Q(z) \, z^{n-1} dz$$

Die z-Transformation ist - ebenso wie die Fouriertransformation - über das unendliche Intervall definiert. Da die unendliche Reihe nicht für alle Werte von z konvergiert, kann die Rücktransformation, und damit die Integrationslinie, nur im Bereich konvergenter z (Konvergenzgebiet) erfolgen.
Rechnen wir nun die z- Transformierte von d(n) aus, so ist die Übertragungsfunktion

$$Q(z) = 1 - \sum_{i=1}^{N} a_i \, z^{-i}$$

Die Übertragungsfunktion H(z) der Empfangsstation soll den Einfluß des Senders rückgängig machen, so daß

$$Q(z) \cdot H(z) \stackrel{!}{=} \text{const} =: G$$

gelten soll.
Damit ist

$$H(z) = \frac{G}{Q(z)} = G \cdot \left(1 - \sum_{i=1}^{N} a_i \, z^{-i}\right)^{-1} \qquad (3.2b)$$

mit dem konstanten Verstärkungsfaktor G.
Welche Eigenschaften hat nun der Filter der Empfangsstation?
Zweifelsohne gibt es N Nullstellen $z_1,..,z_N$ der Funktion Q(z). Diese Nullstellen ("Resonanzen") äußern sich als Unendlichkeitsstellen ("Pole") der Funktion H(z) ("All-Pol-Filter").
Interessanterweise läßt es sich zeigen, daß die Übertragungsfunktion eines All-Pol-Filters äquivalent der des Röhrenmodells aus Abbildung 2.2d ist, wobei das Modell aus N verlustfreien Röhrenstücken gleicher Länge gebildet wird. Da das Signal d(n) der Rest des Signals s(t) ist, der nicht durch den linearen Filter und damit durch das Röhrenmodell erzeugt werden kann, ist d(n) prinzipiell mit dem Signal der Stimmbänder identisch, bis auf die Anteile, die durch die Nicht-linearitäten des Sprachtrakts erzeugt werden.
Da Q(z) somit die Filterwirkung des Sprachtrakts für das anregende Signal invertiert, wird der Q(z) realisierende Filter auch **inverser Filter** genannt; der Vorgang selbst heißt **Inverse Filterung.**
Es läßt sich weiterhin zeigen, daß die Quotienten der Querschnittsflächen A_i der Röhrenstücke mit den PARCOR-Koeffizienten r_i durch

$$\frac{A_{i+1}}{A_i} = \frac{1 - r_i}{1 + r_i}$$

verbunden sind.
Da auch die Reflexion der Schallwellen an den Röhrenstücken durch die Reflexionskoeffizienten u_i beschrieben werden können,

$$\mu_i = \frac{A_i - A_{i+1}}{A_i + A_{i+1}}$$

kann man anstelle der Prädiktorkoeffizienten a_i nicht nur die PARCOR-Koeffizienten r_i, sondern auch die entsprechenden Querschnittsflächen A_i oder die Reflexionskoeffizienten μ_i des korrespondierenden Röhrenmodells für die Sprachbeschreibung und Sprachübertragung benutzen.

Das ADPCM-System
Zweifelsohne sind die beim LPC-System für das gesamte Sprachmaterial ermittelten Prädiktorkoeffizienten bei wechselnden Sprachcharakteristika nicht optimal. Besser ist es, für jedes Sprachintervall, für das die Sprachverhältnisse konstant sind (beispielsweise während eines Phonems (20-50 ms)) die Prädiktorkoeffizienten jeweils extra zu ermitteln. Die Folge von Prädiktorkoeffizienten-Vektoren $\mathbf{a} = (a_1,..,a_p)$ müssen dann zusätzlich zum Differenzsignal d(n) zum Empfänger übermittelt werden, um einen Gleichlauf der beiden LPC-Systeme zu erreichen. Die Abbildung 3.2g zeigt ein solches System.

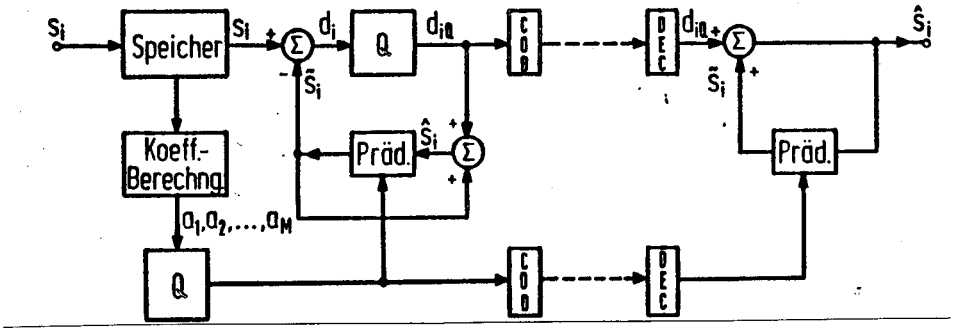


Abb 3.2g Blockschaltbild eines ADPCM-Systems

Wollen wir für dieses **Adaptive Differenz- PCM System (ADPCM)** die jeweiligen Prädiktor-Koeffizienten bestimmen, so ergeben sich gegenüber dem DPCM-System einige Schwierigkeiten. Da die Intervalle nicht mehr unendlich sind, sondern begrenzt, ist auch die zeitliche Mittelung abhängig davon, ob Randpunkte betrachtet werden oder nicht. Zum einen ergeben sich damit unerwünschte Randeffekte, zum anderen ist die Matrix **S** in Gleichung (3.2c) nicht mehr symmetrisch, und hat keine Toeplitz-Form mehr. Die Lösung ist auch nicht mehr als einfache Rekursion hinzuschreiben, sondern ist komplizierter, beispielsweise mit dem Gaß-Jordan Algorithmus zu finden. Darüberhinaus können die so errechneten Prädiktorkoeffizienten zu instabilen Filtern führen.

Intervall-Fenster
Eine Lösung aus diesem Problemkreis bietet die **Fenster-Technik.** Anstelle für die Berechnung ein Intervall aus einer Folge von Werten zu betrachten, werden bei der jeweiligen Berechnung alle Werte außerhalb des Intervalls als Null angesehen **(Rechteckfenster).** Um die dann aber starken Randeffekte abzuschwächen, nimmt man anstelle des Rechteckfensters ein sog. **Hamming-Fenster** wie es in Abb. 3.2h gezeigt ist.

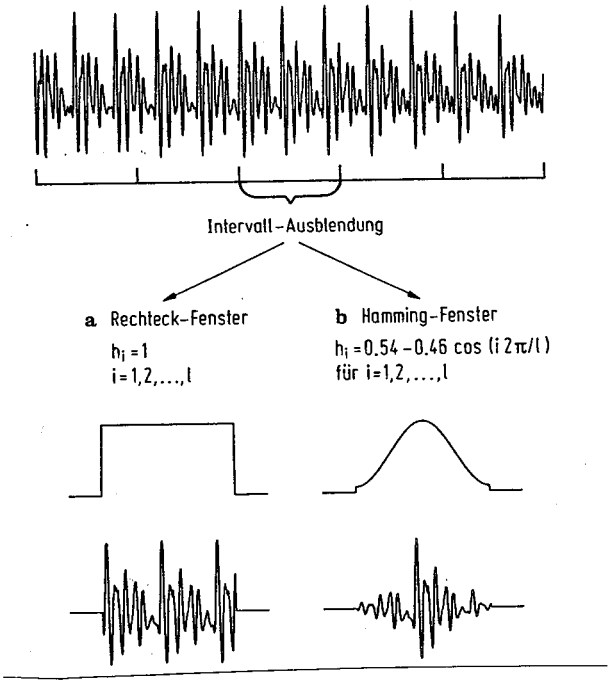


Abb. 3.2h Rechteck- und Hamming-Fenster

Dann aber können die Prädiktorkoeffizienten wieder wie im Falle der unendlichen Intervalle berechnet werden und die Gleichungen (3.2d) sind wieder gültig.
Für die Wahl der zeitlichen Breite des verwendeten Fensters gelten verschiedene Überlegungen. Betrachten wir dazu das Spektrum eines Sprachfensters. Wählen wir ein Rechteck-Fenster, so wird durch den höheren Anteil an hohen, von den Nicht-Linearitäten des Sprachtrakts herrührenden Frequenzen das resultierende Spektrum weniger regelmäßig als beim Hamming-Fenster (Abb. 3.2i).

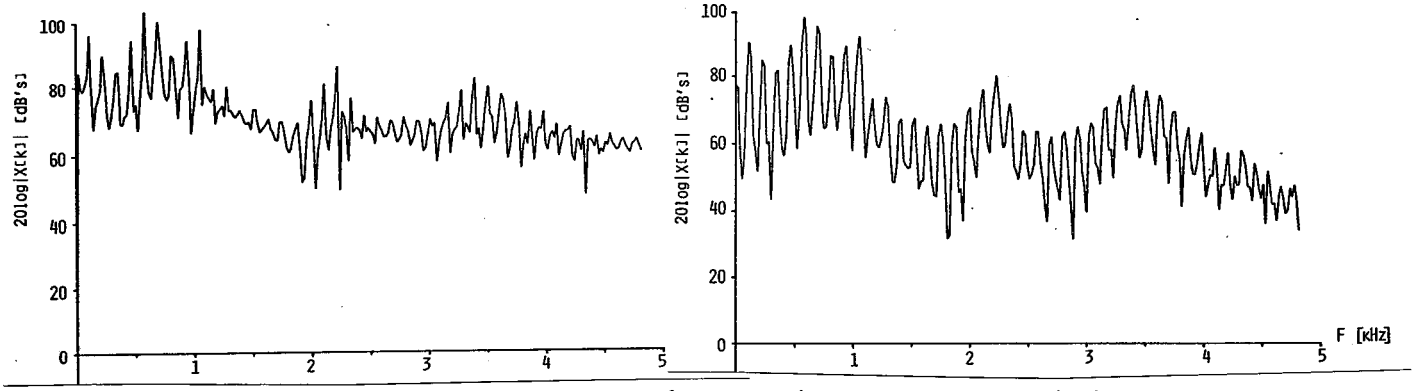


Abb. 3.2i Kurzzeit-Spektren (51,2ms) des Vokals /a/ beim a) Rechteckfenster und b) Hammingfenster

Wählen wir die Fenstergröße - und damit den Sprachabschnitt - zu groß, so beschreibt der daraus erhaltene Mittelwert die Sprachdynamik nicht mehr ausreichend. Wählen wir die Fenstergröße zu klein, so ist keine Periodik mehr erkennbar. Abb. 3.2j zeigt die Kurzzeit-Spektren vom selben Vokal /a/ wie in Abb.3.2i im kleinen Intervall von 6.4ms.

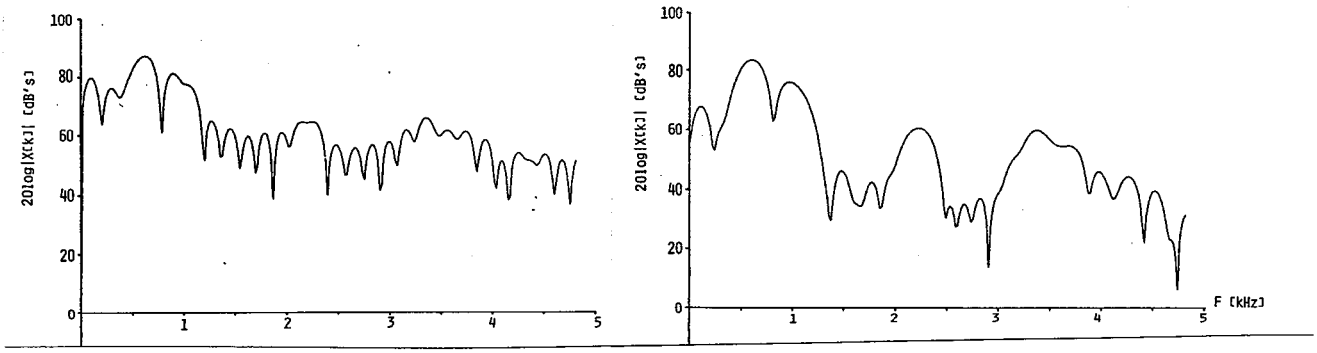


Abb. 3.2j Kurzzeit-Spektren (6,4ms) des Vokals /a/

Abschließend ist zur Verdeutlichung der Funktion des ADPCM-Systems das Orignal-Sprachsignal, das mit einem Hamming-Fenster gewichtete Signal und das resultierende Differenz-Signal in Abb. 3.2k gezeigt.

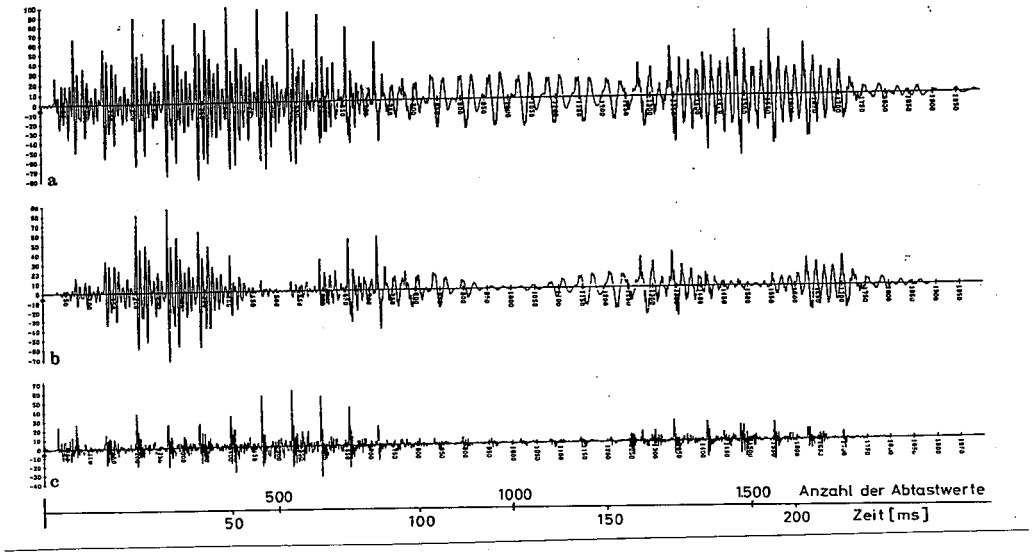


Abb. 3.2k Signalformen im ADPCM-System

3.3 Parametrische Kodierung

Eine direkte, digitale Kodierung des Sprachsignals benötigt trotz einiger Verfahren, vorhandene Redundanz im Sprachsignal bei der Kodierung auszunutzen, noch verhältnismäßig viel Bandbreite bei der Übertragung. Nun wissen wir aber, daß wir bei der Sprachgenerierung (s.Abschnitte 2.1 und 2.3) keine besonders schnelle Mund- und Zungenbewegungen machen können. Nehmen wir an, daß ein Phon minimal 10 ms dauert, so ist es für eine Sprachübertragung nur nötig, alle 10 ms oder mit einer Frequenz von 100Hz die für ein Phon charakteristische Information zu übertragen. Dies ist zweifelsohne eine wesentlich niedrigere Übertragungsfrequenz; das **parametrische Kodierungsverfahren** ist deshalb eine interessante Alternative zur digitalen Kodierung aus Abb. 3.0a. Entscheidend für die Güte der übertragenen Sprache ist dabei die Zahl und Art der Parameter, die die Sprache kodieren sollen.

Kanalvokoder

Eigentlich würde es reichen, die Amplitude der wichtigsten, also der ersten drei Formanten zu übertragen. Da aber sowohl Amplitude als auch Frequenz der Formanten stark wechseln (s.Abb.2.2a), werden für eine differenzierte Analyse nicht nur die Signalintensitäten in den Frequenzbereichen der am häufigsten auftretenden Formanten gemessen, sondern in einem dichten Raster über dem gesamten Frequenzbereich. Der relative Signalanteil von jedem der (je nach Ausführung) 10-20 Frequenzbereiche wird über einen eigenen Kanal (meist im Multiplexverfahren) übertragen.
Zusätzlich zu dem Anteil der Formanten, also dem Resonanzverhalten des Sprachtrakts, wird aus dem Sprachsignal die Information über das anregende Signal gewonnen: stimmlos (sl), stimmhaft(sh), und im zweiten Fall, die Grundfrequenz (Pitch) F_0 der stimmhaften Anregung. In Abb. 3.3a ist ein solcher Kanalvokoder gezeigt.

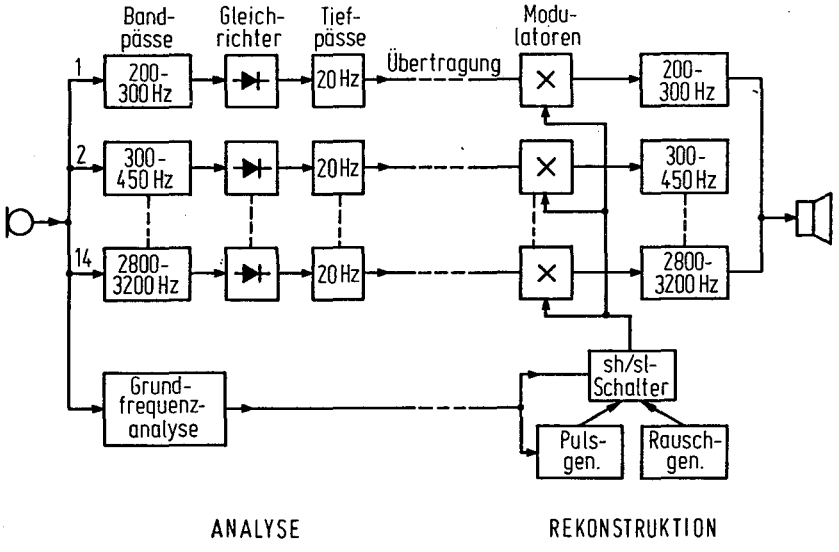


Abb. 3.3a Blockschema eines Kanalvokoders

Die Frequenzanalyse wird heutzutage meist über digitale Filter durchgeführt, d.h. durch Fast-Fouriertransformation eines kurzen Zeitabschnitts und Aufteilung des errechneten Spektrums in Frequenzbereiche. Über Größe und Form des betrachteten Signalfensters gilt dabei das Gleiche wie in Abschnitt 3.2.

LPC-Vokoder

Eine Alternative zu Darstellung der Filterwirkung des Sprachtrakts durch die Analyse und Synthese der Formanten mittels der Fouriertransformation ist die Darstellung durch einen linearen Filter (Abschnitt 3.2). Die den Filter beschreibenden Parameter sind die Prädiktionskoeffizienten; das System heißt **linearer Prädiktionsvokoder (LPC-Vokoder)** und ist in der Sprachqualität einem reinen Kanalvokoder überlegen. Der LPC-Vokoder ist dem ADPCM-Kodierer aus Abschnitt 3.2 sehr ähnlich; in beiden werden aus dem Sprachsignal die Prädiktorkoeffizienten $a_1,..,a_n$ errechnet. In Abb. 3.3b ist der Senderteil eines solchen Vokoders zu sehen.

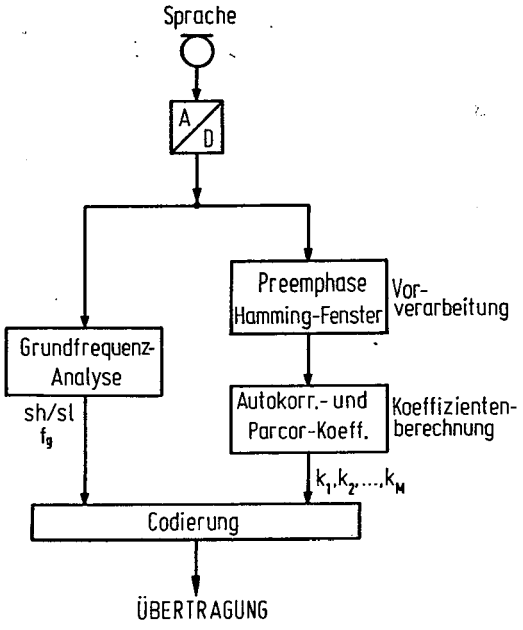


Abb.3.3b Blockschema eines LPC-Vokoders (Sendeteil)

Der tiefere Unterschied zwischen beiden Systemen liegt in der Erzeugung des Differenzsignals d_n. Beim ADPCM-Verfahren ist d_n die Differenz wischen dem Sprachsignal und den Anteilen des Sprachsignals, die von einem linearen Filter (s.3.2) "verursacht" worden sein können. Diese Differenz ist aber gerade das anregende Signal der Stimmbänder, das selbst nicht vom Filter erzeugt sein kann. Beim LPC-Vokoder wird nun anstelle des tatsächlichen anregenden Signals als vereinfachte Version die stimmhaft-stimmlos Entscheidung getroffen und die Pitchfrequenz bestimmt. Diese einfachen Parameter lassen sich mit bedeutend weniger Aufwand kodieren: Machte das Differenzsignal beim ADPCM-Verfahren bei mind. 8KHz Abtastfrequenz und 4 Bit Auflösung eine Datenrate von 32Kbit/sec, so läßt sich der gesamte Parametersatz des LPC-Vokoders mit 60-70 Bit kodieren, so daß bei einem Abtastintervall von 20ms, also 50Hz, eine Bitrate von ca. 3KHz resultiert. Dies ist gerade der Zehnte Teil des ADPCM-Verfahrens!
Wie beim ADPCM-Kodierer lassen sich auch beim LPC-Vokoder anstelle der Prädiktorkoeffizienten die PARCOR-Koeffizienten mit der Obergrenze von 1, oder, wie wir in 3.2 gesehen haben, auch die Reflexionskoeffizienten des Röhrenmodells verwenden.

Formantenanalyse

Eine weitere Eigenschaft des linearen Filters ist geeignet, um eine Formantenanalyse bei Vokalen vorzunehmen. Aus Abschnitt 3.2

wissen wir, daß die Übertragungsfunktion H(z) des Filters p Pole bei den komplexen Werten $z_1,..,z_p$ aufweist. Da

$$z_k = e^{(a_k+i\omega_k)T} \qquad \omega_k = 2\pi f_k$$

gilt, lassen sich aus den Polen z_k leicht die korrespondierenden (Resonanz-) Frequenzen f_k ermitteln. Dies sind aber gerade die gesuchten Resonanzfrequenzen bzw. Formanten F_k des Sprachtrakts! Abbildung 3.3c zeigt das Visible-Speech-Diagramm und die errechneten Formantfrequenzen des Satzes "Why do I owe you a letter". Wie man sieht, stimmt die Analyse mit der Realität für Vokale ganz gut überein; problematischer ist dies allerdings bei kurzen, dynamikreichen Ereignissen ("Stop"-Laute /t/, /p/,..,) und bei Lauten mit großem stimmlosem Anteil.

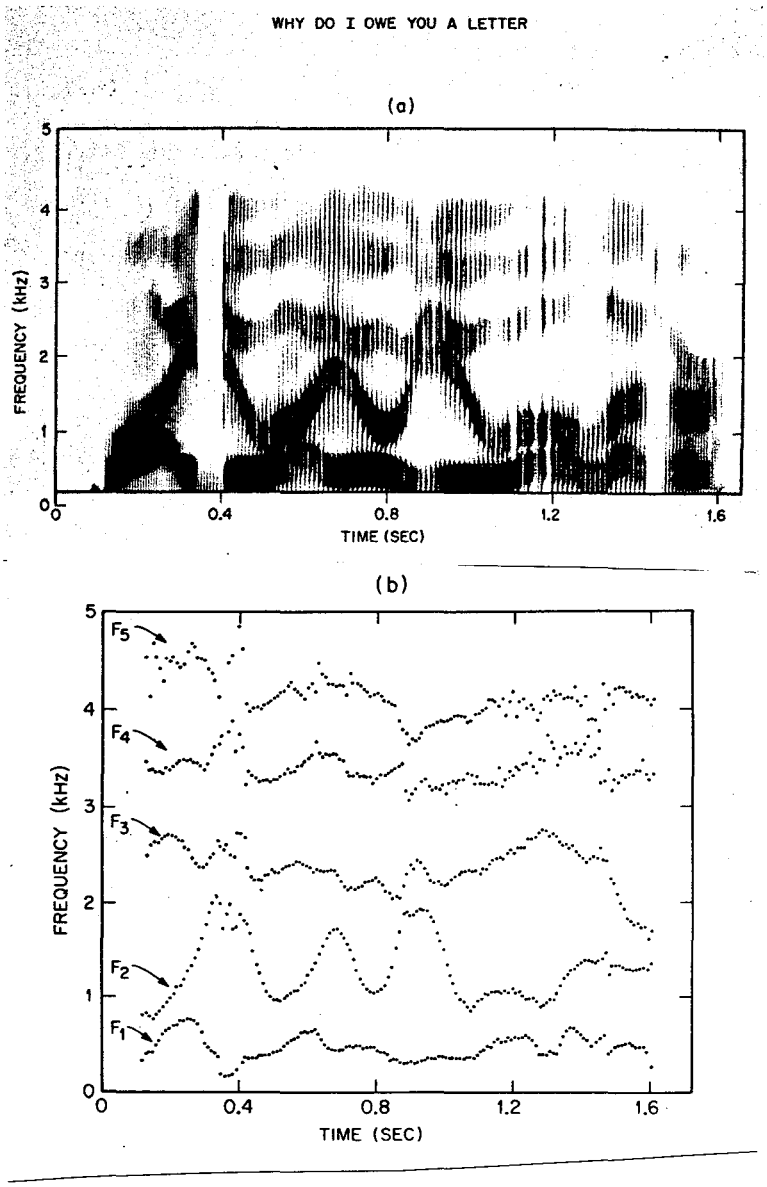


Abb 3.3c tatsächliche und errechnete Formantfrequenzen

Pitchfrequenzanalyse
Die Annäherung an das exakte Sprachsignal mit Hilfe der Formantenanalyse (Kanalvokoder) oder der linearen Prädiktion (LPC-Vokoder) ist nur durch die genaue Kenntnis des anregenden Signals, der stimmhaft/stimmlos Entscheidung und der Ermittlung der Pitchfrequenz möglich. Das tatsächliche Sprachsignal stellt allerdings einige Hindernisse für die Pitchfrequenzanalyse bereit. Dies sind beispielsweise

a) Das Signal der Stimmbänder ist nicht streng periodisch, sondern hat eine "variable" Frequenz.

b) Eine reine stimmlose oder stimmhafte Anregung ist meist nicht gegeben, sondern es liegt eine Mischanregung vor.

c) Ist das Sprachsignal auf Telefonbandbreite(300Hz-3,4KHz) begrenzt, so ist meist die Grundfrequenz direkt nicht mehr enthalten, sondern nur noch ihre Oberwellen.

d) Tritt ein starker stimmhaft/stimmlos-Wechsel mitten in einem Analyseintervall auf, so kann er evtl. erst im nächsten Intervall erkannt werden. Dies führt zu fehlerhafter Kodierung und damit zu Sprachverzerrungen bei der Wiedergabe.

e) Umgebungsgeräusche können das Analyseergebnis erheblich beeinflussen, da ja das Modell nur für **eine** Lautquelle gilt.

Aus der Fülle der in der Literatur beschriebenen Verfahren zur Pitchfrequenzanalyse sollen hier nur einige soweit beschrieben werden, um die wichtigsten Gedanken dieser Technik vorzustellen.

Center-Clipping
Betrachten wir das Sprachsignal in Abb. 3.2k oben, so zeigt sich das anregende Grundsignal in den besonders scharfen Signal-Peaks des Sprachsignals. Filtern wir das Sprachsignal bei fester Schwelle über einen Schmitt-Trigger (Schwellwert-Komparator), so erhalten wir ein normiertes Signal $\tilde{s}(t)$, siehe die folgende Abb.3.3d.

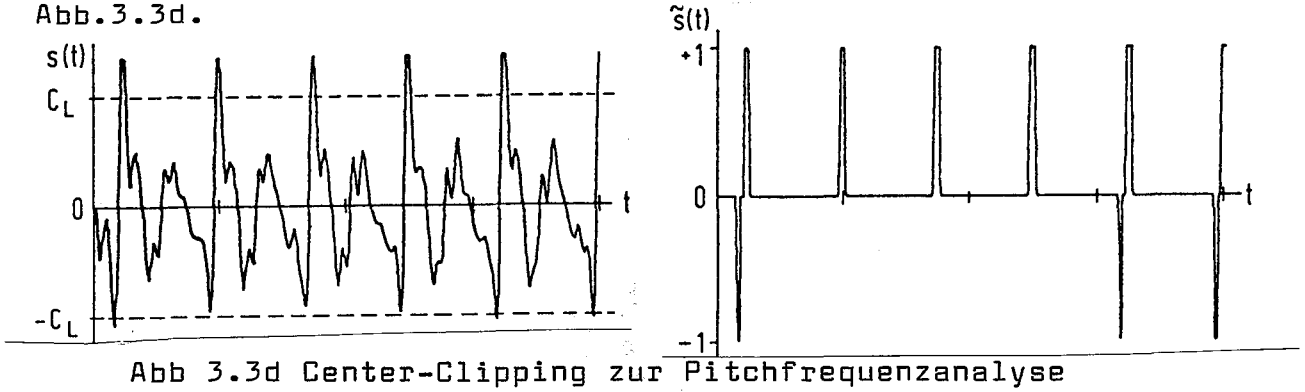


Abb 3.3d Center-Clipping zur Pitchfrequenzanalyse

Die Schwelle C_L des i-ten Intervalls wird dabei neu aus den Maxima max_{i-1} des vorhergehenden und des nachfolgenden Intervalls max_{i+1} gebildet:

$$C_L := k \cdot \min(max_{i-1}, max_{i+1}) \qquad \text{z.B. } k=0.8$$

Die Grundfrequenz wird über eine Kurzzeit-Autokorrellation

ermittelt. Dazu werden die Koeffizienten

$$R_m = \sum_{n=1}^{N-m} \tilde{s}_n \cdot \tilde{s}_{n+m}$$

gebildet.

Da hier die Signalwerte im Abstand m miteinander verglichen werden, reicht es, diese Summation (keine Multiplikation, da $\tilde{s}$ nur aus (-1,0,+1)) bei m=0 für die Grundernergie und ab m=20 bis m=160 (bei N=240) durchzuführen. Bei einer Abtastfrequenz von 8KHz entspricht dies gerade einer Periode von 20*125µs = 2,5ms $\hat{=}$ 400Hz bis 160*125µs = 20ms $\hat{=}$ 50Hz.
Die Entscheidung, ob "stimmhaft" oder "stimmlos" wird über das Maximum der Koeffizienten getroffen:

$$\max_m R_m \begin{cases} < aR_0 & \text{stimmlos} \\ \geq aR_0 & \text{stimmhaft mit } F_0 = (m*125\mu s)^{-1} \end{cases}$$

In Abb.3.3e ist dies an zwei Beispielen gezeigt.

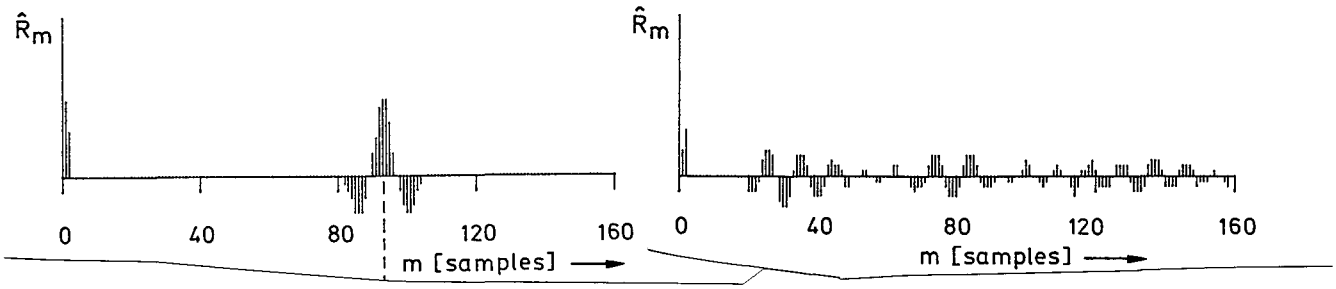


Abb.3.3e Autokorrellation eines stimmhaften und eines stimmlosen Intervalls

SIFT-Verfahren
Das Grundprinziep der "simplifizierten, inversen Filter-Technik" (SIFT) ist das adaptive Differenzenverfahren aus Abschnitt 3.2. Wie wir aus 3.2 wissen, ist das resultierende Differenzsignal identisch mit dem einen linearen Filter anregenden Sprachsignal. Aus dem Differenzsignal (s.Abb.3.2k) läßt sich nun beispielsweise über Autokorrellation wesentlich einfacher die Grundfrequenz bestimmen als aus dem orignalen Sprachsignal.

Cepstrum-Verfahren
Eine andere Möglichkeit als das Sift-Verfahren, durch inverse Filterung den Einfluß der Formanten zu beseitigen, ist die Anwendung der Fourieranalyse. Die Obertöne des Grundtons erscheinen im Spektrum G(f) als zusätzliche Amplitudenspitzen. Der Unterschied im Quadrat der Amplitude (der Intensität) lassen sich durch Logarithmieren zusätzlich verdeutlichen. In Abb.3.3f ist ein Beispiel gezeigt.

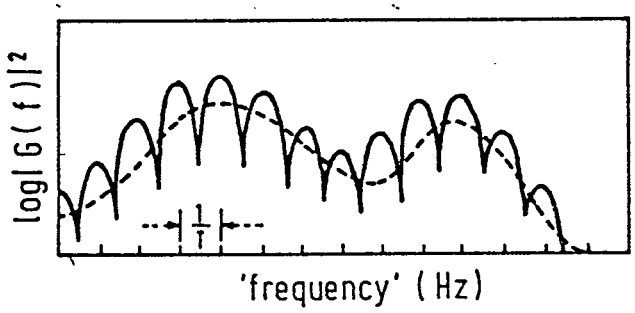


Abb.3.3f logarithmische Spektrumsintensitäten

Dabei erhalten wir aber noch einen interessanten Nebeneffekt. Sei G(f) die Fouriertransformierte der Übertragungsfunktion von Stimmband S(f) und Sprachtrakt H(f) (vgl.Abschnitt 2.2)

$$G(f) = S(f)*H(f)$$

Dann ist das logarithmierte Quadrat

$$\log(|G(f)|^2) = \log(|S(f)|^2 \, |H(f)|^2) = \log(|S(f)|^2) + \log(|H(f)|^2)$$

und entspricht einer Aufteilung in Sprachanteil und Filteranteil

$$c_G(f) = c_S(f) + c_H(f)$$

Wenden wir auf diesen Ausdruck die Rücktransformation an (was bei der nicht-logarithmischen Form nach dem Wiener-Kinchin-Theorem einer Autokorrellationsfunktion entspricht), so erhalten wir die Aufteilung der Sprache in Grundsignal und Sprachtraktverformung diesmal im Ortsraum

$$c_G(T) = c_S(T) + c_H(T)$$

Zur Erhöhung der Grundsignalspitze kann man diesen Ausdruck noch quadrieren. Dies Größe heißt **Cepstrum**, da sie aus dem inversen Spec-trum hergeleitet wurde. Entsprechend wird auch die Zeitfunk-tion als "inverse fre-que-ncy" oder **quefrency** bezeichnet. In Abb.3.3g ist ein solches Cepstrum gezeigt; wie auch in Abb.3.3f bedeuten niedrige quefrency-Werte hohe Frequenzen.

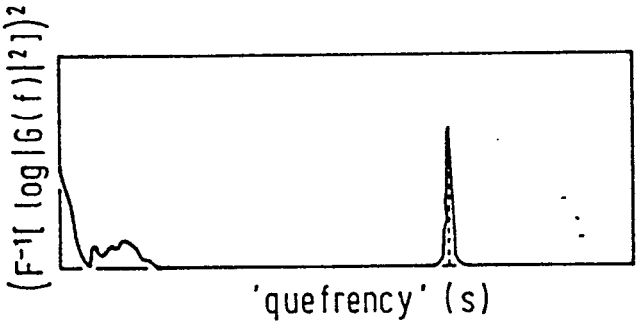


Abb.3.3g Cepstrum eines Sprachsignals

Die Entscheidung 'stimmhaft/stimmlos' wird wieder über den Schwellwert des Maximums getroffen.
Wird das Maximum (Grundfrequenz-Peak) entfernt und die

verbleibende Funktion Fourier-transformiert, so zeigt die so erreichte **homomorphe Filterung** nur noch den spektralen Verlauf des Sprachtrakt-Filters.

Multi-Puls-Anregung

Die bisherigen Verfahren zur Pitchfrequenzbestimmung bilden durch die Annahme einer reinen Grundfrequenz und einer diskreten 'stimmhaft-stimmlos' Entscheidung nur eine grobe Annäherung zur Lösung der anfangs genannten Probleme a) bis d). Anstelle einer Entscheidung zwischen regelmäßigen Signalimpulsen bei stimmhaften Lauten und unregelmäßigen Impulsen ("Rauschen") bei stimmlosen Lauten versucht die Analyse der **Multi-Puls-Anregung**, die Sprachanregung innerhalb eines betrachteten Intervalls durch eine solche Folge von Impulsen zu simulieren, daß das resultierende, durch einen LPC-Filter geleitete Signal dem tatsächlichen Sprachsignal möglichst nahe kommt.
Bei diesem Verfahren müssen also zusätzlich zu den Prädiktorkoeffizienten, die als Filterkoeffizienten die Einhüllende des Spektrums bestimmen, zur Beschreibung der Feinstruktur des Spektrums die Amplitude und der Abstand der anregenden Impulse bestimmt werden. Als Maß für die Güte der Näherung dient dabei wieder der quadratische Fehler, der aber noch zusätzlich spektral gewichtet wird (Höhenabsenkung), um das geringere menschliche Empfindungsvermögen für Abweichungen bei hohen Frequnzen zu berücksichtigen.In Abb.3.3h ist das Verfahrensschema gezeigt.

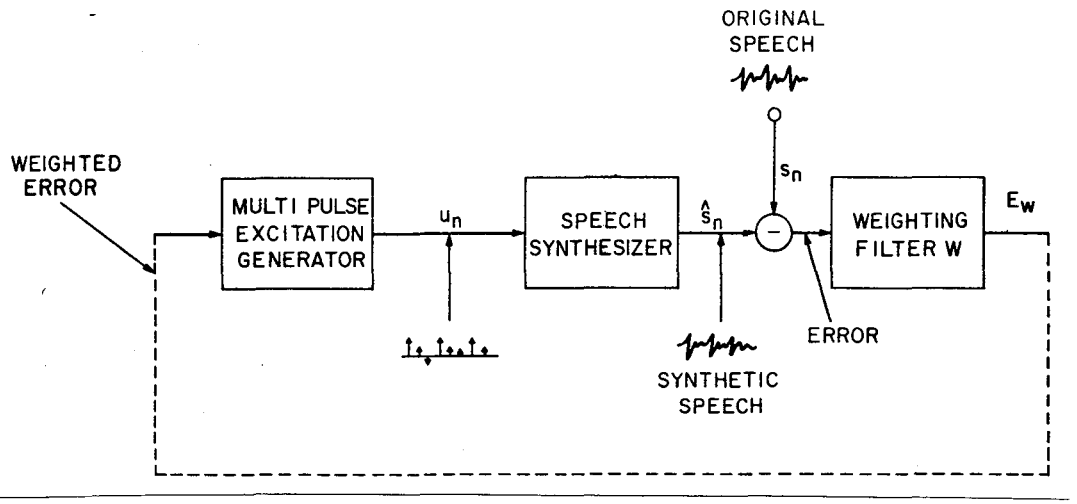


Abb.3.3h Multi-Puls Anregungs-Analyse

Aus dem resultirenden Fehler werden dann Amplitude und Abstand der nötigen Impulse bestimmt. Normalerweise werden in 10ms nicht mehr als 8 Impulse für eine gute Sprachqualität benötigt. Der Verbesserungsprozeß ist in Abb. 3.3i gezeigt.

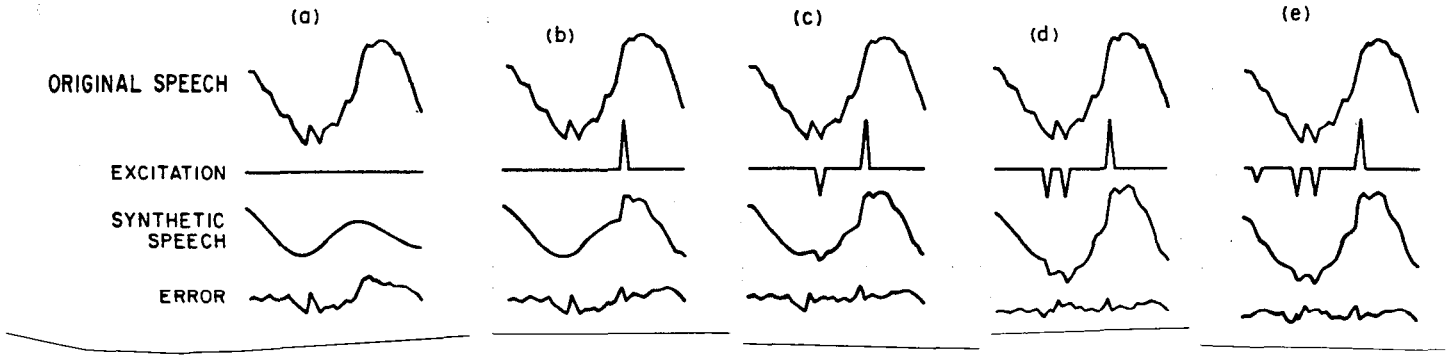


Abb.3.3i Einfügen von Impulsen ins Anregungssignal und der resultierende Fehler

3.4 Hardware zur Sprachanalyse und Synthese

Die Analyse und Synthese des Sprachsignals geschieht mit den heutigen Hardwarebausteinen nach zwei verschiedenen Prinzipien: entweder wird die Sprache als physikalisches Signal kodiert (Signalformkodierung), wobei Verfahren der vorigen Abschnitte 3.1 und 3.2 Verwendung finden, oder es wird die Sprache parametrisch kodiert (z.B.LPC-Kodierung, s.Abschnitt 3.3). Bei gleicher Bitrate ist das LPC-Verfahren der Signalformkodierung in der Sprachqualität überlegen.
Bei beiden Verfahren werden die Phoneme, Worte oder Sätze bei einem aufwendigen Entwicklungssystem eingesprochen und, wie oben beschrieben, analysiert. Die daraus resultierenden Parametersätze (ca 1Kbit/sec) wwerden sodann in ROMs gespeichert und extern oder intern einem Sprachchip beigegeben.

Sprachanalyse
In den vorigen Abschnitten wurden verschiedenen Verfahren zur Sprachanalyse vorgestellt. Diese Verfahren mußten dabei entweder von teuren Großrechnern oder aber Off-Line (nicht synchron zum Sprechen) durchgeführt werden. Abhilfe schaffte der Einsatz von schnellen Bit-Slice-Prozessoren wie AMD 2900. Eine solche Konfiguration ist in Bild 3.4a gezeigt.

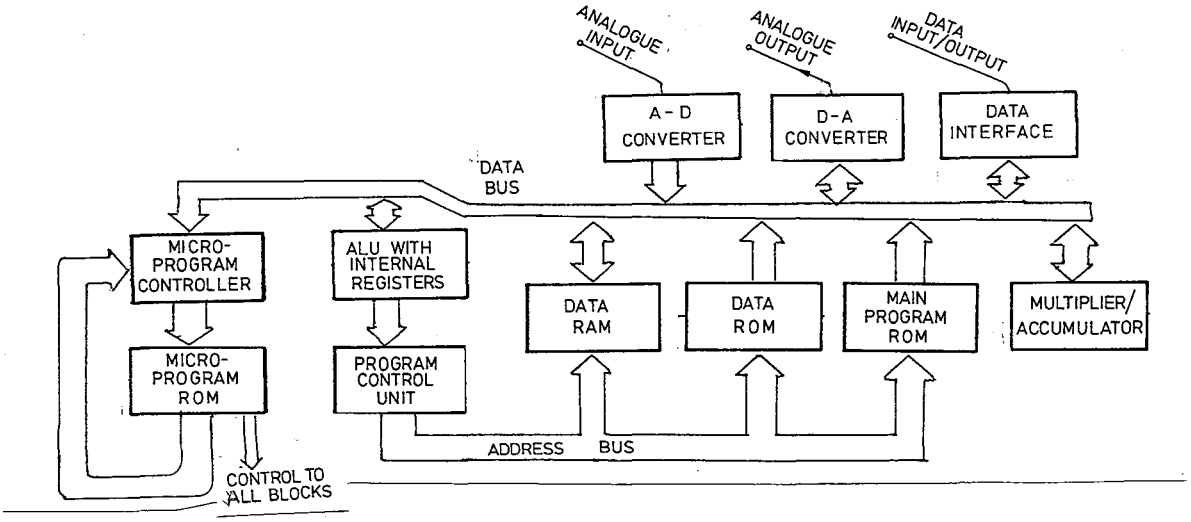


Abb.3.4a Bitslice-Mikroprozessor-Architektur

Leider verbraucht aber ein solcher Bit-Slice-Prozessor ziemlich viel Leistung. Bessere Ergebnisse kamen von speziellen, nur für die Signalanalyse konzipierten Ein-Chip-Coomputern wie z.B. NEC 7720 (ähnliche Architektur wie in Abb.3.4a) mit 250µs Zykluszeit und, als einer der ersten, dem Signalprozessor TMS 320 von Texas Instruments in Abb. 3.4b.

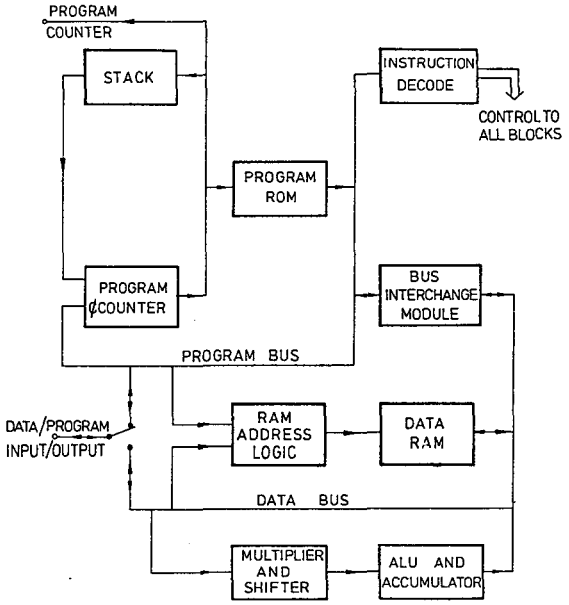


Abb. 3.4b Architektur des Signalprozessor TMS 320

Inzwischen gibt es eine Fülle von Prozessoren verschiedenster Hersteller, die sich durch kurze Zykluszeiten und die Fähigkeit auszeichnen, möglichst komplexe Algorithmen durch möglichst viele, voneinander unabhängige Einheiten auf dem Chip parallel arbeiten zu lassen.

Sprachsynthese

Die Chips zur Sprachsynthese realisieren im Prinzip den Empfangsteil der in Abb.3.0a gezeigten Übertragungssysteme, wobei allerdings anstelle der gesendeten Parameter diese bereits fest für jedes Wort bzw. Phonem im ROM gespeichert sind. In Abb.3.4c ist als Beispiel der Aufbau eines solchen Sprachsynthesechips der Fa. Hitachi gezeigt, der nach dem LPC-Kodierungsverfahren arbeitet.

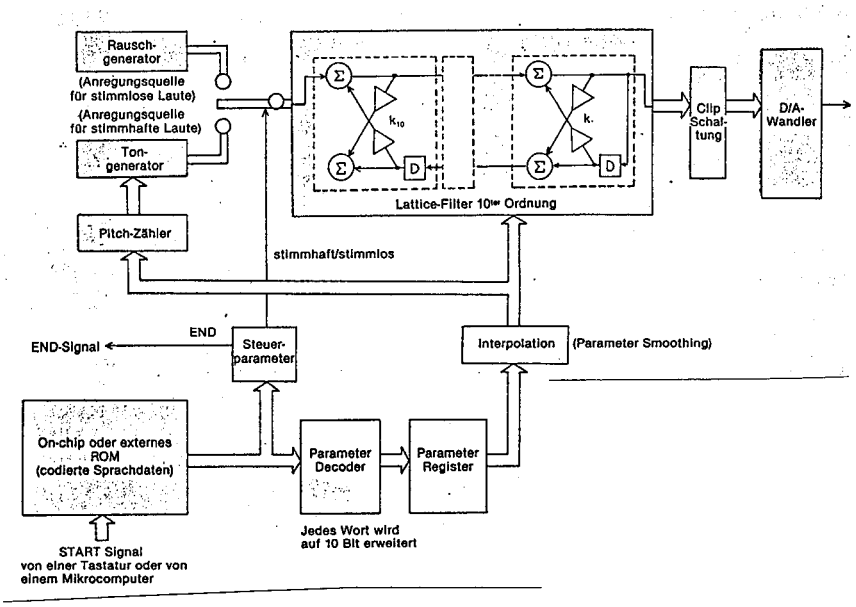


Abb. 3.4c Aufbau eines Sprachsynthese-Chips

Das betrachtete Intervall, in dem die 10 PARCOR-Koefizienten des 10 -poligen Filters konstant bleiben, ist wahlweise 10ms oder 20ms und benötigt 48 Bits zur Kodierung. der Filter ist dabei modular als Verkettung gleichartiger Operationen gestaltet.
Zum Ansteuern des Synthese-Chips wird meist ein Mikroprozessor benutzt, der Ereignisse erkennt und den Kontext berücksichtigt, in dem die Sprachausgabe eingesetzt wird. Beispielsweise kann man zur preiswerten Sprachausgabe mittlerer Qualität von Texten eine Übersetzung in Phoneme durchführen lassen und dann den Synthese-Chip mit der Sequenz der Phonem-Codes ansteuern lassen. Abbildung 3.4d zeigt eine Kodetabelle für einen Sprachchip der Fa. Votrax und in Abb. 3.4e ist ein solches System der Fa. Texas Instruments gezeigt.

Phoneme Code	Phoneme Symbol	Duration (ms)	Example Word
00	EH3	59	jacket
01	EH2	71	enlist
02	EH1	121	heavy
03	PA0	47	no sound
04	DT	47	butter
05	A2	71	made
06	A1	103	made
07	ZH	90	azure
08	AH2	71	honest
09	I3	55	inhibit
0A	I2	80	inhibit
0B	I1	121	inhibit
0C	M	103	mat
0D	N	80	sun
0E	B	71	bag
0F	V	71	van
10	CH	71	chip
11	SH	121	shop
12	Z	71	zoo
13	AW1	146	lawful
14	NG	121	thing
15	AH1	146	father
16	OO1	103	looking
17	OO	185	book
18	L	103	land
19	K	80	trick
1A	J	47	judge
1B	H	71	hello
1C	G	71	get
1D	F	103	fast
1E	D	55	paid
1F	S	90	pass

Phoneme Code	Phoneme Symbol	Duration (ms)	Example Word
20	A	185	day
21	AY	65	day
22	Y1	80	yard
23	UH3	47	mission
24	AH	250	mop
25	P	103	past
26	O	185	cold
27	I	185	pin
28	U	185	move
29	Y	103	any
2A	T	71	tap
2B	R	90	red
2C	E	185	meet
2D	W	80	win
2E	AE	185	dad
2F	AE1	103	after
30	AW2	90	salty
31	UH2	71	about
32	UH1	103	uncle
33	UH	185	cup
34	O2	80	for
35	O1	121	aboard
36	IU	59	you
37	U1	90	you
38	THV	80	the
39	TH	71	thin
3A	ER	146	bird
3B	EH	185	get
3C	E1	121	be
3D	AW	250	call
3E	PA1	185	no sound
3F	STOP	47	no sound

Phonemsymbole des SC-01

Abb. 3.4d Phonemtabelle des Synthes-Chips der Fa. Votrax

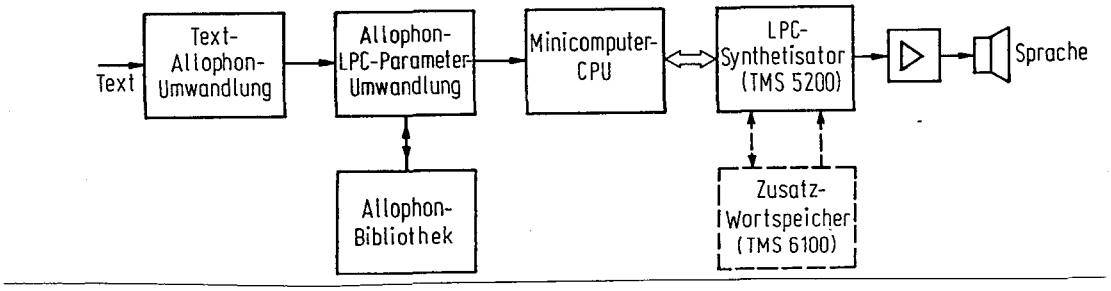


Abb.3.3e Sprachsynthese-System

Bei diesen Sprach-Chips muß man allerdings dabei beachten, daß die Allophone meist für die amerikanische Sprache eingegeben sind und deutsche Sprache deshalb meist einen amerikanischen Klang aufweist.

Phonemübergänge

Setzt man die Allophone einfach nebeneinander, so führt dies zu schwer verständlicher Sprache. Die Interpolation der die Allophone charakterisierenden Parameter ist aber auch nicht ohne weiteres möglich. Interpoliert man die Prädiktor-Koeffizienten, so führt dies zu unerwarteten Klangänderungen und teilweise zu Instabilitäten. Dieses Problem entfällt zwar bei den PARCOR-Koeffizienten, falls sie kleiner gleich eins sind, aber auch da sind die Interpolationen stark hörbar. Nur bei den korrespondierenden Querschnittsparametern des Röhrenmodells erhält man bessere Resultate; der direkte Zusammenhang zwischen den modifizierten Röhren-Querschnittsflächen und den Änderungen bei der Sprachgenerierung (Koartikulation!) ist leicht einsehbar.

4.0 Spracherkennungs-Algorithmen und Systeme

In der Einleitung im ersten Kapitel wurde bereits dargelegt, daß die Spracherkennung zur Zeit bei der sprecherabhängigen Einzelworterkennung gut funktioniert, dagegen bei der vom Sprecher unabhängigen Einzelworterkennung und bei dem Erkennen von fließend, ohne künstliche Pausen gesprochener Sprache noch viele Forschungsanstrengungen nötig sind. darüber hinaus sind Systeme, die aus den erkannten Lauten bzw. Worten syntaktisch und semantisch richtige Sätze generieren, praktisch nur in vereinzelten Ansätzen vorhanden. Aber betrachten wir die Situation genauer.

4.1 Isolierte Einzelworterkennung

Der grundlegende Ablauf bei der Erkennung von isoliert (mit Pausen) gesprochenen, einzelnen Worten wurde bereits in Kapitel 1 gezeigt. In Abb.4.1a ist nun ein realistischeres, detaillierteres Schema gezeigt. Es hat im Prinzip zwei Modi: In der Lernphase wird die Verbindung der lautlichen Charakterisierung eines Worts mit der entsprechenden Buchstabenkette (ASCII-String) gelernt; in der Erkennungsphase wird die lautliche Charakterisierung des unbekannten Worts mit der Charakterisierung eines bekannten Worts aus der Liste der bekannten Worte (Vokabular) verglichen und auf das "ähnlichste Wort" entschieden.

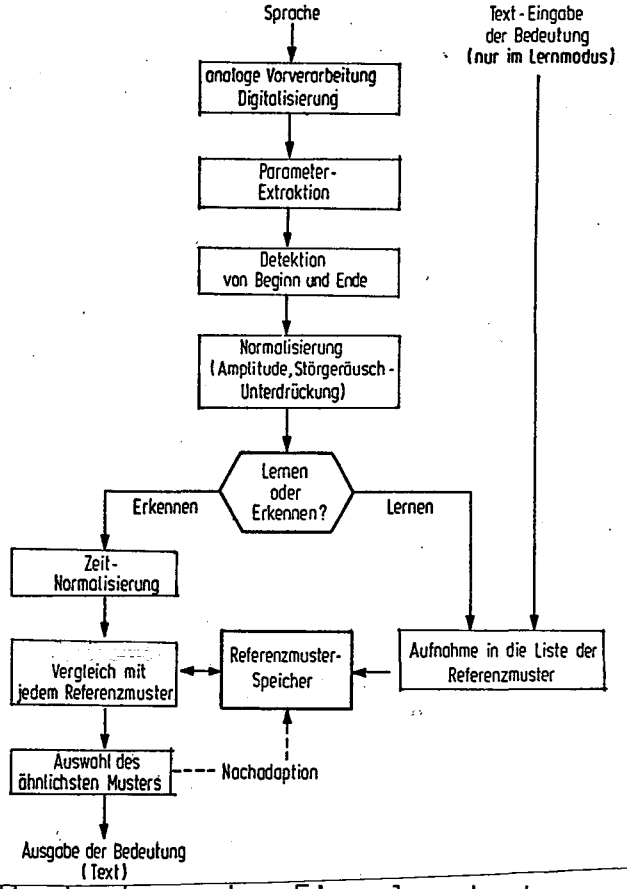


Abb.4.1a Blockschema der Einzelworterkennung

Welche Parameter werden zur Charakterisierung der Worte verwendet?
Legt man Wortmaterial zugrunde, das die selben Worte von den selben Sprechern enthält, und ermittelt für jede Parameterart die

Fehlerrate, so erhält man folgende, nach abnehmender Fehlerrate geordnete Liste:

a) Autokorrellations-Koeffizienten R_m
b) Prädiktor-Koeffizienten a_m
c) PARCOR-Koeffizienten r_m
d) äquidistante Werte des inversen Filters Q(z)
e) äquidistante Werte des homomorphen Filters
f) die ersten Werte des Cepstrums C(quefrency)

Die Parameter e) und f) unterscheiden sich von allen anderen nicht nur in der geringeren Fehlerrate, sondern durch einen von den häufigen Fouriertransformationen bedingten stark erhöten Rechenaufwand. Da sich die Fehlerrate von d) nicht wesentlich unterscheidet (ca 4%), bietet sich die inverse Filterung als günstigstes Parametrisierungsverfahren an.

Zeitnormalisierung
Leider lassen sich die Folgen $\mathbf{a}^1,..,\mathbf{a}^N$ und $\mathbf{b}^1,..,\mathbf{b}^M$ der Parametervektoren, die ein bekanntes Wort a und ein unbekanntes Wort b beschreiben, nicht direkt miteinander vergleichen, da das gleiche Wort vom selben Sprecher niemals physikalisch identisch ausgesprochen wird. Insbesondere weisen beide Worte eine unterschiedliche zeitliche Länge auf. Um beide Worte trotzdem miteinander vergleichen zu können, müssen beide auf die gleiche Länge gebracht werden. Es ist dabei unwichtig, ob man sämtliche Worte der Referenzliste auf die Länge des unbekannten Worts bringt oder ob das unbekannte Wort bei jedem Vergleich auf die zeitliche Länge des Referenzwortes gebracht wird.
Eine lineare Transformation (Strecken,Stauchen) bringt allerdings nur sehr unbefriedigende Resultate, da beim Wiederholen eines Wortes sich meist die Zeitlängen der Vokale ändern, nicht aber die der Konsonanten.

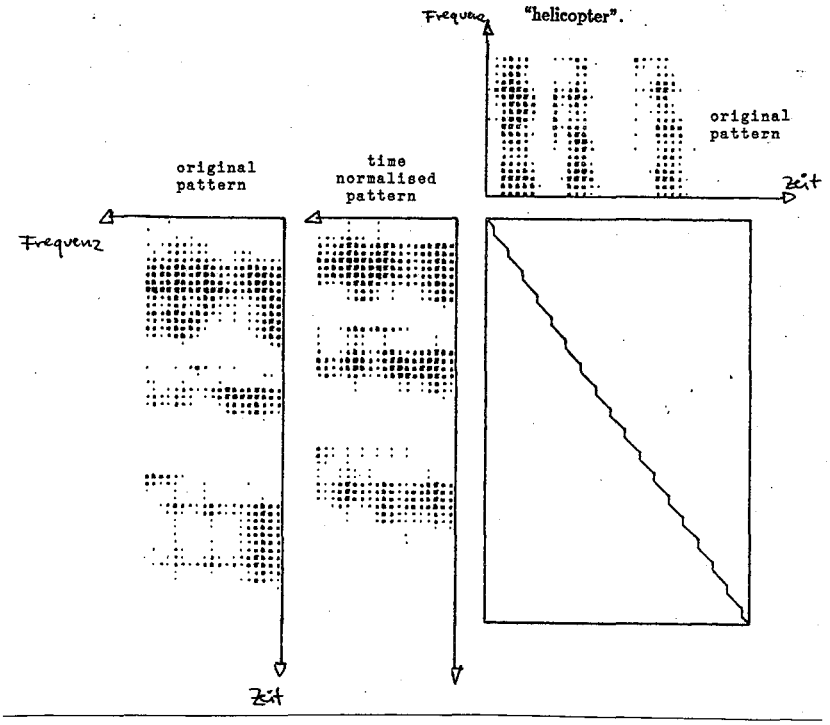


Abb.4.1b lineare Zeitransformation eines Testworts

In der obigen Abb.4.1b ist die lineare Transformation des visible-speech-Diagramms für das Wort "helicopter" gezeigt. Vergleicht man mit dem Auge die Diagramme von Referenzwort und dem gestreckten Testwort, so sieht man, daß die Übereinstimmung bei einer anderen, nicht-linearen Transformation besser sein könnte. In Abb.4.1c ist das Ergebnis einer besseren, nicht-linearen Anpassung zu sehen.

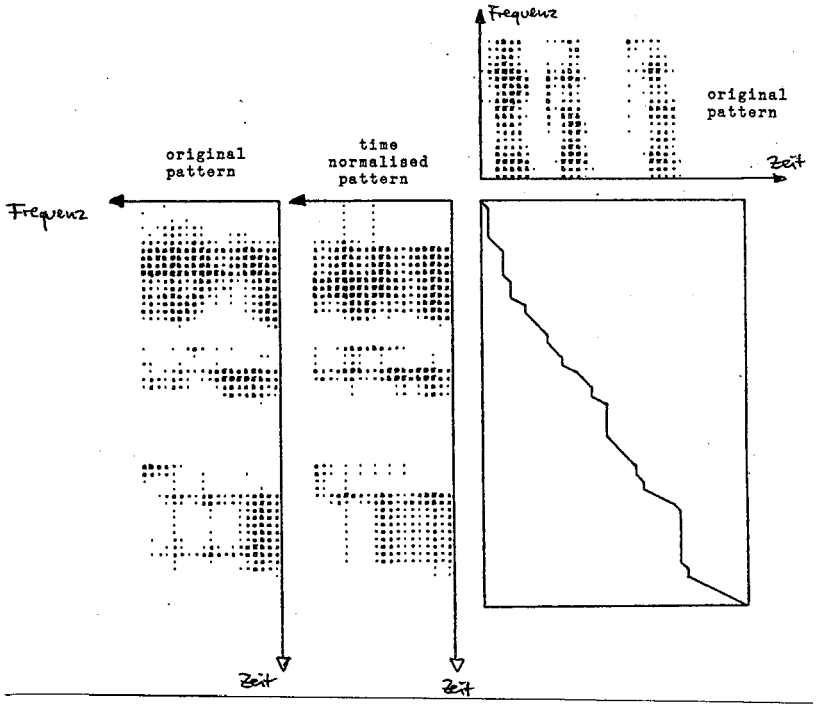


Abb.4.1c nicht-lineare Zeitanpassung durch dynamische Programmierung

Für die nicht-lineare Zeitnormalisierung werden hauptsächlich zwei Verfahren benutzt: die dynamische Zeitanpassung (Dynamic Time Warping) und die Analyse durch verborgene Makkov-Modelle (Hidden Markov Models).

Dynamic Time Warping (DTW)
Eine Folge von N (multi-dimensionalen) Vektoren $\mathbf{a}^1,..,\mathbf{a}^N$ sollen mit einer anderen Folge von M Vektoren $\mathbf{b}^1,..,\mathbf{b}^M$ verglichen werden. Ein mögliches Maß an Unterschied ist der Euklidische Abstand

$$L_{ij} := L(\mathbf{a}^i,\mathbf{b}^j) := |\mathbf{a}^i-\mathbf{b}^j|$$

In Abb.4.1d ist die gesamte Matrix (L_{ij}) von zwei Folgen $\mathbf{a}^1,..,\mathbf{a}^5$ und $\mathbf{b}^1,..,\mathbf{b}^4$ von Mustervektoren zu sehen. Die numerischen Werte a_1 werden dabei durch die visuellen Muster symbolisiert.

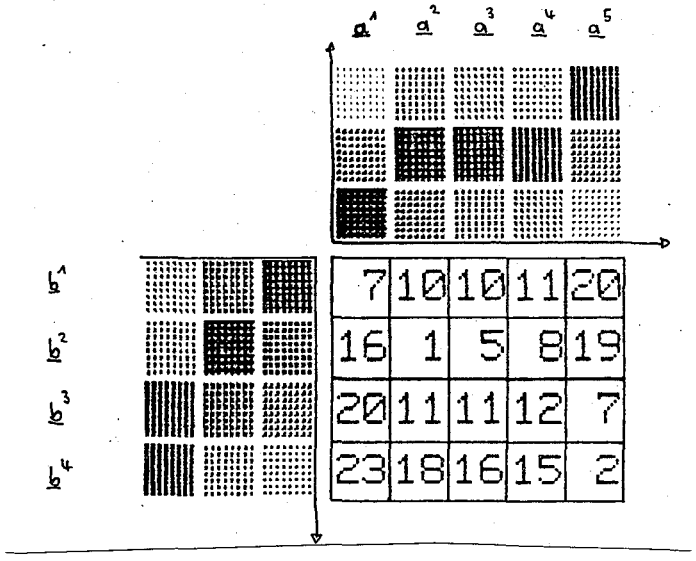


Abb.4.1d Fehlerfunktionsmatrix von zwei Folgen von Mustervektoren a und b

Die Aufgabe der Zeitanpassung besteht nun darin, eine solche Zuordnung zwischen den Vektoren a^i und b^j zu finden, daß der Gesamtfehler R_w minimal wird:

$$R_w \overset{!}{=} \sum_{(i,j)\ \text{aus}\ w} L_{ij} \overset{!}{=} \min$$

Der Zuordnungsvorschrift entspricht ein Weg w durch die Matrix L, dessen Anfang bei L_{11} und das Ende bei L_{NM} liegen und dessen Summe minimal ist.
Sehen wir uns Abb.4.1d an, so könnte eine initiale Idee darin liegen, als nächsten Punkt des Weges jeweils den Nachbarn L_{ij} mit dem niedrigsten Fehler zu suchen. Dies beschert uns für das Beispiel aus Abb.4.1d ein R_w von 30, was nicht optimal ist. Der Grund dafür liegt darin, daß bei der Suche nach dem R_{min} keine größeren Wegalternativen untersucht werden. Die Folge von lokalen Minima sichert eben kein globales Minimum wie es in Abb.4.1d durch den Weg 7-1-5-12-2 mit R_{min}=27 gegeben ist.
Eine Lösung dieses Problems bietet das Verfahren der dynamischen Programmierung. Dazu wird eine neue Matrix **R** errechnet, bei der jeder Wert eines Matrixelements R_{ij} durch die Summe aller Werte ersetzt wird, die auf dem optimalen Weg w zu diesem Matrixelement in L anfallen, plus der Wert L_{ij} selbst:

$$R_{ij} := R_w + L_{ij} \qquad w \text{ ist von } L_{11} \text{ bis } L_{ij} \text{ und ist optimal}$$

Der optimale Weg wird iterativ als lokale Entscheidung über die Nachbarschaft bestimmt. Betrachten wir dazu Abb. 4.1d. Nehmen wir an, daß für den Weg w die Indizes i und j nur zunehmen oder konstant bleiben dürfen, so lassen sich die Werte der Zeile R_{1j} und der Spalte R_{i1} direkt fortlaufend bestimmen. Als nächstes kann die Zeile R_{2j} komplettiert werden, indem für jedes L_{ij} das Minimum aus der Nachbarschaft gesucht und aufsummiert wird:

$$R_{ij} = \min(R_{i,j-1}, R_{i-1,j-1}, R_{i-1,j}) + L_{ij} \qquad (4.1b)$$

In Abbildung 4.1e ist dies für R_{34} gezeigt.

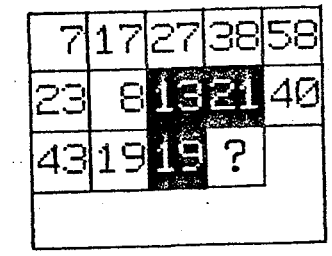


Abb.4.1e Errechnen der optimalen Summen-Fehlermatrix

So läßt sich Spalte für Spalte und Zeile für Zeile die Matrix **R** errechnen. Die Gesamtheit aller möglichen Wege, errechnet durch die lokalen Nachbarschaftsentscheidungen, ergibt durch die Restriktion der Indizes einen Baum:

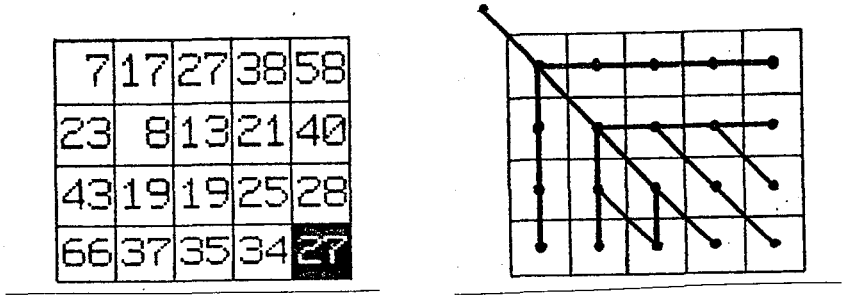


Abb.4.1f vollständige Summenfehlermatrix und der korrellierende Entscheidungsbaum

Der minimale Gesamtfehler ist somit

$$R_{min} = R_{NM}$$

und der optimale Weg derjenige, der bei (N,M) endet. Da es sich um eine Baumstruktur handelt, ist w damit eindeutig bestimmt.

Bei größeren Matrizen ist allerdings der Rechenaufwand zur Bestimmung von **R** ziemlich hoch. Die Abbildung 4.1g verdeutlicht dies für zwei visible-speech Diagramme des Satze "Joe took father'shoe bench out".In der Abbildung links hat die Matrix **R** bei kleinen Komponentenwerten einen großen Schwärzungsgrad; rechts ist der Entscheidungsbaum dargestellt.

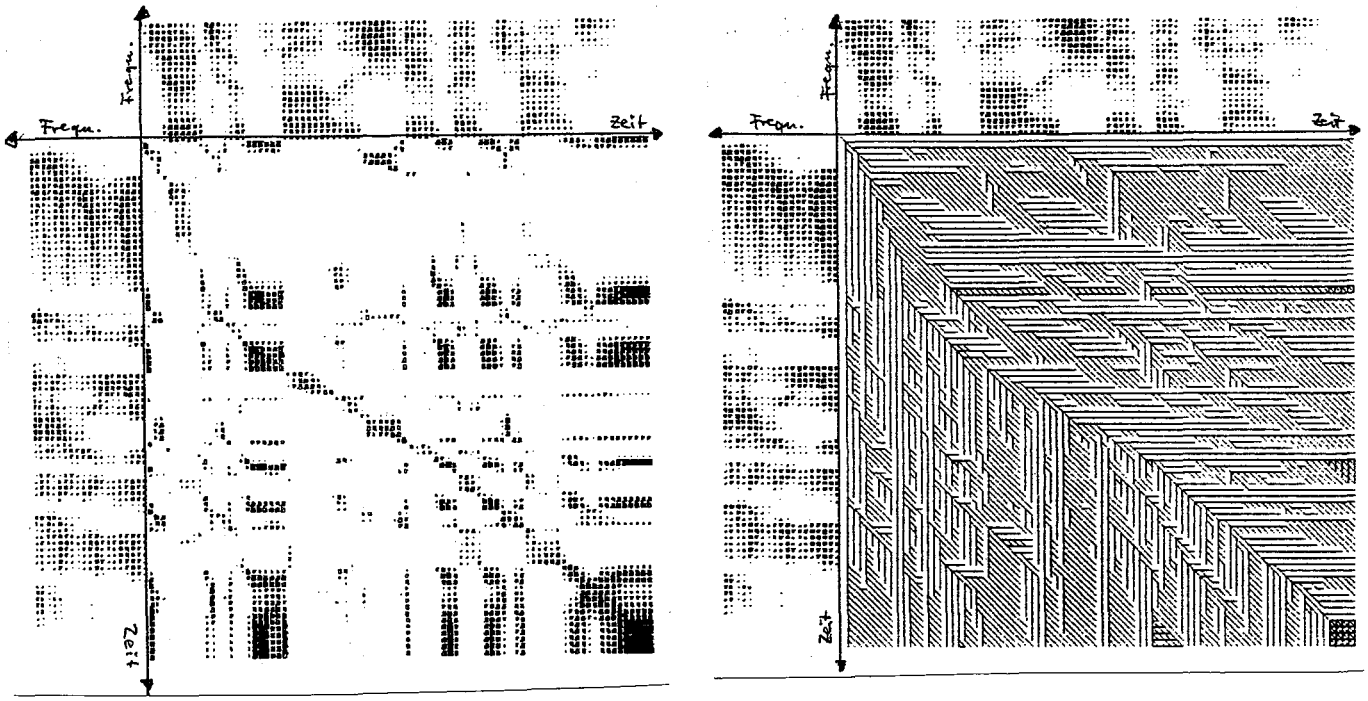


Abb.4.1g Fehlermatrix und Entscheidungsbaum zweier Spektrogramme zweier Sätze

Wie man sieht, ist es eigentlich für dieses Beispiel nicht nötig, die gesamte Matrix **R** auszurechnen; es würde reichen, sich auf einen Bereich um die Hauptdiagonale zu beschränken. Deshalb werden meist Restriktionen für die Nachbarschaftsauswahl getroffen, um den Rechenaufwand zu senken.
Betrachten wir nur Fälle, bei denen $2N \geq M \geq N/2$ gilt, ist folgende Modifikationen für die Nachbarschaft zweier Matrixelemente des optimalen Weges erfolgreich:

- Der Index i, i=1..N ändert sich jeweils um 1
- Der Index j, j=1..M darf sich nur um 0,1 oder 2 ändern
- War die vorherige Änderung von j Null, so darf j nur noch um 1 oder 2 geändert werden

Die Relation $M \geq N/2$ spiegelt sich in dem nicht erlaubten, konstanten Index für j wider; die Relation $2N \geq M$ entspricht der maximalen Indexänderung von j um 2.
Die obigen drei Regeln lassen sich als Produktionsregeln formulieren und legen zwei Geraden fest, die durch den Punkt (1,1) gehen. Die eine Gerade hat die Steigung -1/2, die andere die Steigung -2. Alle erlaubten Wege müssen sich in dem Gebiet zwischen den beiden Geraden befinden. Da dies entsprechend auch für den Endpunkt (N,M) gilt, bildet die betrachtete Fläche anstelle eines Rechtecks ein Parallelogramm, das wesentlich weniger Punkte enthält als das Rechteck vorher, siehe Abb.4.1h.

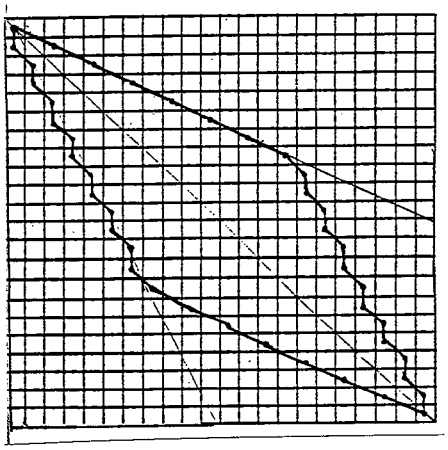


Abb 4.1h Restriktion des optimalen Weges durch Indexrestriktion

Abgesehen von Produktionsregeln für die Nachbar-Indizes läßt sich der Rechenaufwand auch durch Randbedingungen wie "maximale Zeitverschiebung" (Max. Abweichung von der Diagonalen) oder durch "maximalen Fehler" (Abbruch der Wegalternative, wenn der Summenfehler $R_{i,j}$ größer als ein Maximalwert wird).
Alle Varianten des DTW-Verfahrens haben den Vorteil, daß durch die nicht-lineare Transformation das Problem der linearen Transformation, genau den Anfang und das Ende des Wortes zu finden, automatisch gelöst wird.

Obwohl das DTW-Verfahren einen großen Durchbruch bei der Einzelworterkennung bedeutet, geht doch dabei Information verloren, die zur Worterkennung gebraucht werden könnte. Beispielsweise lassen sich keine Worte mehr richtig erkennen, die sich nur in der Länge von Vokalen unterscheiden, wie "Ball" und "Baal". Die Variation der gesprochenen Worte läßt sich leichter mit einem anderen Verfahren modellieren: der" Hidden Markov Modelle".
Zuvor aber sei noch ein Beispiel eines DTW-Systems angeführt, um die typischen Eigenschaften eines solchen Systems tabellarisch im Überblick zu präsentieren.

Beispiel: System Votan V5000

Vorverarbeitung:	16 Kanal Filter, kodiert in 2kbit/s
Segmentation:	500 ms Pause
Lernphase:	500 sec pro neues Wort, max 256 Worte
Erkennung:	Dynamische Zeittransformation DTW
Fehlerrate:	0,55% Fehler beim DSTI Datensatz
Kosten:	5000$

Hidden Markov Modelle (HMM)
Betrachten wir die Sprachproduktion vom phonetischen Standpunkt aus, so läßt sich jedes Wort als Folge von Phonemen auffassen, die sich in der konkret gesprochenen Sprache als Allophone ausbilden. Vom statistischen Standpunkt aus ist also ein Wort durch eine Folge von Zuständen (Phoneme) gekennzeichnet, wobei jeweils ein Ausgabesymbol (Allophon) zufällig gewählt wird. Hängt die Übergangswahrscheinlichkeit von einem Zustand zum nächsten nicht von den vorher durchlaufenen Zuständen("Vorgeschichte") ab, so wird die Zustandsmenge mit ihren Übergängen als **Markov-Modell** bezeichnet. In Abb. 4.1i ist ein solches Modell abgebildet.

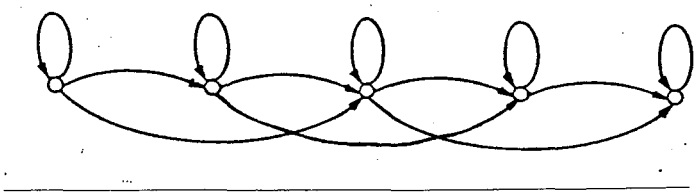


Abb. 4.1i Markov Modell mit 5 Zuständen

In der obigen Abbildung gibt es keine Übergänge zurück auf vorherige Zustände. Das Markov Modell wird deshalb auch als Links-Rechts Markov Modell bezeichnet. Da man nicht die Zustände, sondern nur die zufälligen Ausgabesymbole beobachten kann, heißt dieses Modell "verborgenes Markov Modell" (**Hidden Markov Model**). Ein verborgenes Markov Modell H sei durch k Zustände und l mögliche Ausgabesymbole gekennzeichnet. Tragen wir die Übergangswahrscheinlichkeiten von Zustand i zum Zustand j in der Matrix $(A_{ij}),i=1..k,j=1..k$ ein und die Wahrscheinlichkeit, im Zustand i das n-te Ausgabesymbol zu generieren in der Matrix (B_{in}), i=1..k, n=1..l, so ist mit den Wahrscheinlichkeiten a_i, initial im Zustand i zu sein, das Markov Modell H charakterisiert durch

$$H = (\mathbf{a}, \mathbf{A}, \mathbf{B})$$

Angenommen, eine bestimmte Sequenz $\mathbf{b}^1,..,\mathbf{b}^N$ von Ausgabesymbolen werden beobachtet. Dann läßt sich bei gegebenen Werten für **a**, **A** und B die Auftrittswahrscheinlichkeit $P_H(\mathbf{b}^1,..,\mathbf{b}^N)$ dafür ausrechnen. Da jedes Wort der Referenzliste durch einen unterschiedlichen Satz von Modellparametern gekennzeichnet ist, wird bei der Mustererkennung dasjenige Wort ausgesucht, das die Auftrittswahrscheinlichkeit maximimiert:

$$P_{H_{max}}(\mathbf{b}^1,..,\mathbf{b}^N) = \max_i P_{H_i}(\mathbf{b}^1,..,\mathbf{b}^N)$$

Wie werden bei diesem Modell in der Lernphase die Parameter ermittelt?
Da man die Zustände nicht direkt beobachten kann und auch die Ausgabesymbole nur zufällig erscheinen, scheint dies ein fast vergebliches Unterfangen zu sein.
Glücklicherweise ist dies nicht der Fall, wie uns der bekannte **Baum-Welch Algorithmus** zeigt. Bei diesem Algorithmus, der hier nur kurz angedeutet werden kann, werden iterativ für jedes Wort die Übergangswahrscheinlichkeiten B_{ij} neu mit Hilfe der beobachteten Folge $\mathbf{b}^1,..,\mathbf{b}^N$ des bekannten Wortes geschätzt. Dazu werden anfangs die A_{ij} aus der Zahl der bekannten Übergänge von i zu j, geteilt durch die Gesamtzahl der möglichen Übergänge von i, bestimmt; analog dazu auch die B_{ij}. In einem zweiten Durchgang wird sodann mittels dynamischer Programmierung das Gleiche in umgekehrter Richtung vorgenommen, so daß die Kombination von beiden Abschnitten als "Forward-Backward Algorithmus" bezeichnet wird.
Der Vorteil der HMM- Ansätze liegt darin, daß jedes Wort mittels eines Grundmusters beschrieben wird, das aber mit bestimmten Wahrscheinlichkeiten Abweichungen erfahren kann. Damit ist beispielsweise nicht von vornherein festgelegt, wieviele Parametervektoren $\mathbf{b}^1,\mathbf{b}^{i+1},..$ einem Vokal zugeordnet sind, sondern dies kondensiert sich beim Training heraus durch die Übergangswahr-

scheinlichkeit, in dem selben Zustand zu bleiben.
Dies ist auch der Grund, warum die Segmentationsprobleme (Wortanfang und -ende) auch bei diesem Verfahren keine Rolle spielen.
Die Fehlerquote der HMM-Systeme liegt ähnlich günstig wie beim DTW-Verfahren (4%).
Das HMM-Verfahren unterscheidet sich ziemlich stark von dem DTW-Verfahren in den Rechenzeitanforderungen. In der Lernphase benötigt das HMM-System durch die bessere Anpassung an das Sprachmaterial und den Sprecher eine längere Rechenzeit als das DTW-System; dagegen ist bei der Worterkennung die Rechenzeit des HMM-Systems um den Faktor 10 geringer, was sich in den Anforderungen an die Hardware der Erkennungssysteme positiv auswirkt.

Beispiel: Software-Produkt "Mark II" von Dragon Systems für 6502 und 8086 Prozessoren

Vorverarbeitung:	8 Bit A/D
Lernphase:	normal. 4 Wiederholungen pro Wort, max 96 Worte
Erkennung:	Hidden Markov Modell
Fehlerrate:	0,7% Fehler beim DSTI Datensatz
Kosten:	10$

4.2 Erkennung fließender Sprache

Bei der Erkennung fließend gesprochener Sprache muß zusätzlich zur Worterkennung auch die Segmentierung der Worte vorgenommen werden. Es ist sicher günstig, auch dies nach den Kriterium des kleinsten Fehlers vorzunehmen. Eine der einfachsten Algorithmen daz stammt von Bridle und Brown (1979), der effektiv nur eine Erweiterung unseres bekannten DTW-Algorithmus von 4.2 darstellt. In Abb. 4.2a ist links die isolierte Worterkennung des Wortes "one" mittels DTW gezeigt; rechts soll nun die Folge "1-1-2-1-3" wiedererkannt werden.

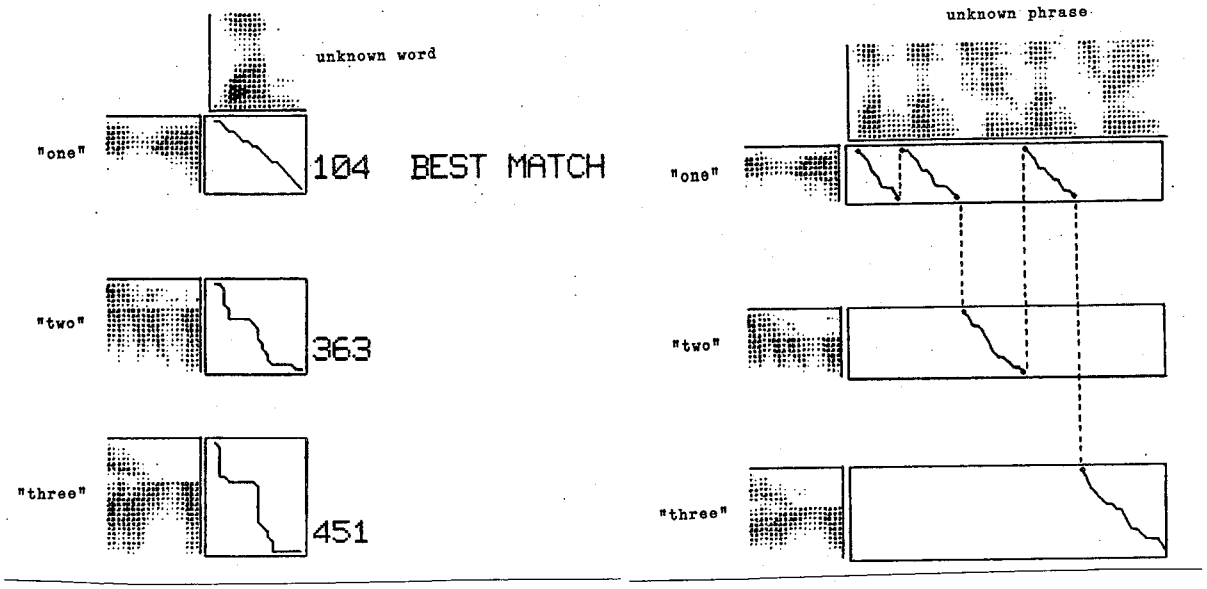


Abb.4.2a isolierte und verbundene Worterkennung mittels DTW

Wie man sieht , muß hier der optimale Weg über mehrere Ränder der Matrizen L(1),L(2),.., hinweg gehen können. Um den kleinsten Summenfehler in der Nachbarschaft auch in den Randbezirken der Matrizen zu finden, müssen die Zeilen der Matrizen oben ergänzt werden durch jeweils den günstigsten Wert der unteren Zeile aller möglichen Matrizen, die der Weg vorher durchlaufen haben könnte.

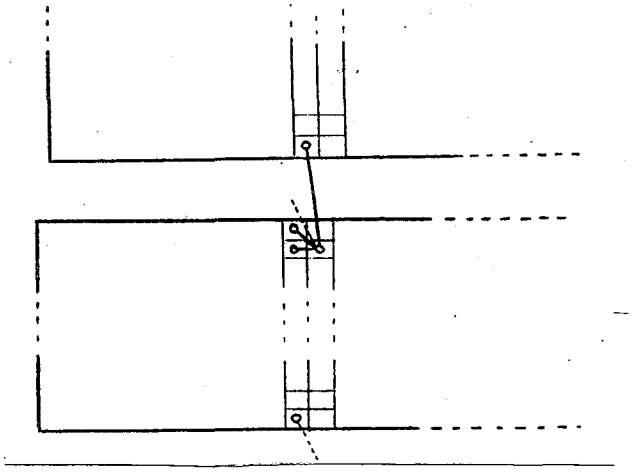


Abb 4.2b Randergänzung zur Wegsuche über mehrere Matrizen

Abgesehen von weiteren Verfahren, mittels modifizierter DTW-Verfahren fließende Sprache zu erkennen, gibt es noch den grundlegend anderen Ansatz der statistischen Betrachtungsweise.

Statistischer Ansatz

Bei der statistischen Beschreibung wird jeder Folge von Worten $\omega_1,..,\omega_k$ eine Auftrittswahrscheinlichkeit $P(\omega_1,..,\omega_k)$ zugeordnet. Diese Auftrittswahrscheinlichkeit kann man prinzipiell aus maschinenlesbar gespeicherten Texten erschließen. Dabei gibt es aber einige Schwierigkeiten, mit denen sich das Team von Jelinek (IBM,USA) konfrontiert sah. Zum einen gibt es keinen Text, indem alle beliebigen Wortkombinationen vorkommen, so daß die Wortketten auf drei Worte $\omega_1,\omega_2,\omega_3$ beschränkt wurden. Aber auch die Anzahl möglicher Tripel aus 5000 erkennbaren Worten ist immer noch zu groß: $1.25*10^{11}$!. Deshalb mußten spezielle Überlegungen zu Hilfe genommen werden, um über die Anzahl der in einem großen Datenmaterial vorhandenen Tripel, in denen ω_1, ω_2 oder ω_3 enthalten ist, indirekt auf die Auftrittswahrscheinlichkeit der nichtexistenten Tripel zu schließen. Damit wird auch der Ansatz fraglich, aus den vom Benutzer eingesprochenen Texten adaptiv die Auftrittswahrscheinlickeit der Wortketten zu bestimmen: Der Benutzer gibt niemals genug Text ein!
Eine mögliche Lösung aus diesem Dilemma besteht darin, anstelle auf Folgen ganzer Worte sich auf die Folgen der Grundformen (Wortstamm, Lexem) der Worte zu beschränken und die Variation (Morpheme, sprecherabhängige Modifikationen, u.ä.) als spezielle Manifestation dieses Wortes zu betrachten. Wollen wir wie im vorigen Abschnitt 4.1 die zeitliche Variation der gesprochenen modellieren, so läßt sich das Modell der verborgenen Markov-Modelle auch für die Erkennung fließend gesprochener Sprache verwenden. Dazu ermöglichen wir bei k Worten des Satzes den Übergang zwischen den verschiedenen, möglichen Worten. Abbildung 4.2c zeigt das Zustandsübergangsdiagramm dieses Modells.

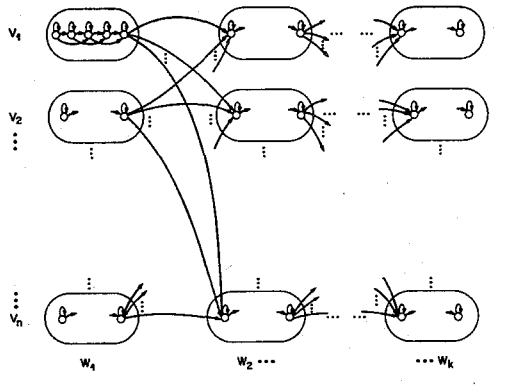


Abb. 4.2 Ein HMM für Wortketten

Die Sprechpausen zwischen den Worten werden dabei durch spezielle Ausgabesymbole der Anfangs- und Endzustände modelliert, die "lautlos" bedeuten.

Zur Erkennung des gesprochenen Satzes gibt es verschiedene Wege. Eine Möglichkeit besteht darin, diejenige Wortkette zu suchen, die die Verbundwahrscheinlichkeit der Wortkette mit der beobachteten Folge von Ausgabesymbolen (Sprach-Parametervektoren) maximiert. Dies läßt sich beispielsweise durch den **Viterbi-Algorithmus** vornehmen, der auf der uns bekannten dynamischen Programmierung beruht.

4.3 Syntaxgesteuerte Satzerkennung

Die deterministischen und stochastischen Techniken aus Abschnitt 4.2 berücksichtigen beim Vergleich des zu erkennenden Satzes mit allen möglichen Sätzen nicht die Redundanz, die in der menschlichen Sprache durch die Syntax gegeben ist. Ein großes Problem der SPracherkennung besteht gerade darin, daß die Menschen unvollständige oder fehlerhafte Sprache produzieren, weil sie aus der Erfahrung wissen, daß durch die bekannten Syntaxstrukturen eine Ergänzung oder Korrektur durch die Höhrer leicht möglich ist und deshalb eine perfekte Sprache überhaupt nicht nötig ist.
Beispielsweise zeigten Versuche eine Fehlerrate von 20% beim Erkennen von einzeln gesprochenen Buchstaben. Werden aber Namen buchstabiert, die aus einem Telefonbuch mit 18000 Einträgen stammen und deshalb impliziet Restriktionen an die Buchstabenfolge der Namen stellen, so sinkt die Fehlerrate auf 4%.
Ein anderes Beispiel sind die Arbeiten von Shipman und Zue (1982), die jedem Wort eines Lexikons von 20000 Worten eine Charakterisierung von 6 allgemeinen, sehr unspezifischen phonetischen Merkmalen zuordneten. Alle Worte mit einer gleichen Charakterisierung wurden in die gleiche Klasse eingeordnet. Es ergab sich eine mittlere Klassengröße von 2 Worten; selbst die größte Klasse umfaßte nur 1% des Lexikons (200 Worte).
Eine ähnlichgute Charakterisierung erfolgt, wenn die Worte nur durch die betonten Silben beschrieben werden und die unbetonten unbeachtet bleiben.
Selbstverständlich läßt sich durch diese phonetische Redundanz eine mehrstufige Worterkennung durchführen, die nur wenige phonetische Merkmale benutzt. Leider ist aber das Verfahren sehr empfindlich gegenüber Fehlern in den Anfangsstufen. Allerdings lassen sich phonetisch gleich klingende Worte so leichter unterscheiden oder vokabularabhängig kodieren.

Reguläre, kontextfreie Grammatik

Als ersten Ansatz, nicht beliebige Ketten zuzulassen, kann man die Registrierung aller erlaubten Wortketten ansehen. Die Menge L aller erlaubten Wortketten bildet somit eine Restriktion bei der Spracherkennung. Die erkannten Wortketten müssen zusätzlich mit Hilfe dieser Nebenbedingung geprüft werden, und, wenn sie nicht enthalten sind, durch alternative Wortkandidaten modifiziert und erneut geprüft werden. Ist das Vokabular groß, so macht es Schwierigkeiten, sowohl alle Sätze des Benutzers zu speichern,

als auch diese für jeden erkennbaren Satz zu durchsuchen. Vielfach sind die Sätze in dieser Form auch noch nicht gesprochen worden.
Einfacher ist es, die Menge alller möglichen Sätze durch die Angabe der Regeln zu beschreiben, mit deen sie generiert werden können. Traditionell wird gemäß der Chomsky-Hierarchie die Grammatik G als ein 4-Tupel charakterisiert:

$$G = (V_N, V_T, S, P)$$

mit

V_N= Menge der "nicht-terminalen" Symbole.
Bsp:"Verb","Subjekt",..

V_T= Menge der "terminalen" Symbole.
Bsp:"machen","Ich",..

S = ein Element aus V_N("Startsymbol")

P = Produktionsregeln der Form $(V_N \cup V_T)^* \rightarrow (V_N \cup V_T)^*$, wobei die *-Operation die beliebige Verkettung der Elemente andeutet

Die einfachsten Grammatiken sind die **reguläre Grammatik**, die durch die Produktionsregeln

$A \rightarrow aB$ $\quad$ A,B aus V_N, a aus V_T
$A \rightarrow a$

gekennzeichnet ist, sowie die **kontextfreie Grammatik** mit

$A \rightarrow \beta$ $\quad$ β aus $(V_N \cup V_T)^*$

Kennt man die zugrunde liegende Grammatik, so wird die Entscheidung, ob die vorliegende Wortkette syntaktisch richtig ist, durch einen sog. **Parser** getroffen.
In Abb.4.3a ist ein kurzer Ausschnitt aus einer kontextfreien Grammatik, der sog. ATN-Grammatik, abgebildet.

```
S -> NP Vint
S -> NP Vtr NP
S -> Vind NP NP
S -> NP AUX Vint
S -> NP AUX Vtr NP
S -> AUX Vind NP NP
S -> AUX NP Vint
S -> AUX Vtr NP
S -> AUX NP Vind NP NP
S -> NP Vint PP
S -> NP Vtr NP NP
S -> NP Vind NP NP NP
        .
        .
        .
S -> NP Vint PP PP
        .
        .
        .
S -> NP Vint PP PP PP
        .
        .
```

NP stands for Noun Phrase,
PP for Prepositional Phrase,
Vtr for a transitive verb,
Vint for an intransitive verb,
Vind for a verb that takes indirect objects,
AUX for an auxilliary verb such as "is" or "does"

Abb.4.3a Modellierung der englischen Syntax mittels ATN

stochastische Grammatiken

Eine Möglichkeit, eine Folge von Worten zu erkennen, liegt in der Auftrittswahrscheinlichkeit dieser Wortkette. In Abschnitt 4.2 wurde unter Anwendung der verborgenen Markov-Modelle die Wortkette mit der maximalen Auftrittswahrscheinlichkeit gesucht. Dies wird durch die sequentielle Auswahl der Einzelworte erreicht, die bei jedem neu erkannten Wort das Kriterium

$$Pr(\text{Wort}\,|\,\text{Wortkette}) \cdot Pr(\text{Wortkette}) \stackrel{!}{=} \max$$

erfüllen muß. Beachten wir die einschränkenden Nebenbedingungen einer möglichen, festen Wortfolge aus L, so ist es nötig, alle Kandidaten für die zu erkennende Wortkette, geordnet nach ihrer Auftrittswahrscheinlichkeit, darauf zu untersuchen, ob sie in L enthalten sind. Man erhält somit die wahrscheinlichste Wortkette aus L. Die Berechnung der Wortkette aus N Worten mit der maximalsten Wahrscheinlichkeit ist ziemlich rechenzeitaufwendig. Zur Einschränkung kann man den Algorithmus für Kartesische Produkte verwenden:

Der Kartesische-Produkt-Algorithmus
Die Aufgabe lautet, aus einem Vokabular von N Worten eine Wortkette von k Worten zusammenzustellen, die maximale Auftrittswahrscheinlichkeit besitzt und in der Menge L der möglichen Sätze enthalten ist.
Dier Auswahl aus den N^k Möglichkeiten läßt sich sukzessive vornehmen:

a) Zuerst wird (wie oben) sequentiell mit j=1..k jeweils das Wort w_m gesucht, das

$$Pr(w_m^j | \text{Wortkette}) = \max_i Pr(w_i | \text{Wortkette}) \qquad i=1..N$$

maximiert. Damit ist die Wortkette $w^1..w^k$ mit maximaler Auftrittswahrscheinlichkeit gefunden. Für jedes Wort w^j wurde somit eine Liste mit den für diesen Platz wahrscheinlichsten Worten gebildet.

b) Prüfe, ob die Wortkette aus L ist. Wenn ja, STOP.

c) Wähle für das j-te Wort w^j das nächste in der Liste der wahrscheinlichsten Worte. Wired dies unabhängig voneinander für alle k Worte der Wortkette durchgeführt, so resultieren k Variationen der Wortkette mit k Auftrittswahrscheinlichkeiten. Sortiere diese in eine Liste der wahrscheinlichsten Wortketten und wähle die mit der größten Auftrittswahrscheinlichkeit als nächsten Kandidaten.

d) Gehe zu b).

Dieser Algorithmus iteriert solange, bis ein geeignete Wortkette gefunden ist.
Leider ist die Abfrage, ob eine Wortkette zur Menge der als legal registrierten Sätze gehört, bei großen Werten von N (großem Vokabular) viel zu zeitaufwendig und unpraktisch. Wie bereits im vorigen Abschnitt beschrieben, kann L bei geeigneten Strukturen auch implizit durch eine Grammatik aufgebaut werden. Der einfachste Fall ist wieder die reguläre, kontextfreie Grammatik. Der Vorgang des Erkennens der Wortkette kann man nun mit der Anfrage an einen probabilistischen Parser verbinden, ob die Wortkette Teil der generierbaren Wortketten ist. Dabei werden nicht alle möglichen Wortketten ausprobiert, sondern es wird wieder sukzessive von j=1 bis k für jedes Wort dasjenige ausgewählt, was in der Grammatik erlaubt ist und eine hohe Auftrittswahrscheinlichkeit hat. Nimmt man wie beim Kartesischen Produkt-Algorithmus dasjenige mit der höchsten Auftrittswahrscheinlichkeit, so garantiert das aber noch nicht die optimale Wahl, da dies "lokale" Optimierung nicht auch automatisch ein "globales" Optimum bedeutet. Besser ist es, wie in 4.1 beschrieben, das Verfahren der dynamischen Programmierung zu verwenden.

Die bisher beschriebenen Ansätze versuchen, durch Übergänge zwischen den verschiedenen Schichten der Spracherkennung (Worterkennung, Satzerkennung) die Fehlerrate zu vermindern. Am Besten ist es aber sicher, die Nebenbedingungen auf verschiedenen Ebenen der Spracherkennung gleichzeitig zu benutzen.
Beim **Harpy-System** (Carnegie-Mellon Universität,USA) wird beispielsweise versucht, grammatische, lexikalische und phonetische Information in einem einzigen Netzwerk für mögliche Phonemsequenzen darzustellen, wobei jeder Weg durch dieses Netzwerk einem möglichen Satz entspricht. Die Aufgabe lautet dann, den optimalen Weg im Netzwerk zu finden.
Ein anderer, bekannter Ansatz wird vom **Hearsey II - System**

gewählt, das versucht, jeden Aspekt von einem einzigen "Agenten" (Manager, Prozess) bearbeiten zu lassen. Die Rohversion des Satzes, die auf einem von allen Agenten erreichbaren temprären Speicherplatz steht ("Blackboard"), wird von allen Agenten nach ihren speziellen Kriterien (Restriktionen) untersucht und korrigiert, bis Übereinstimmung zwischen den Agenten hergestellt ist.

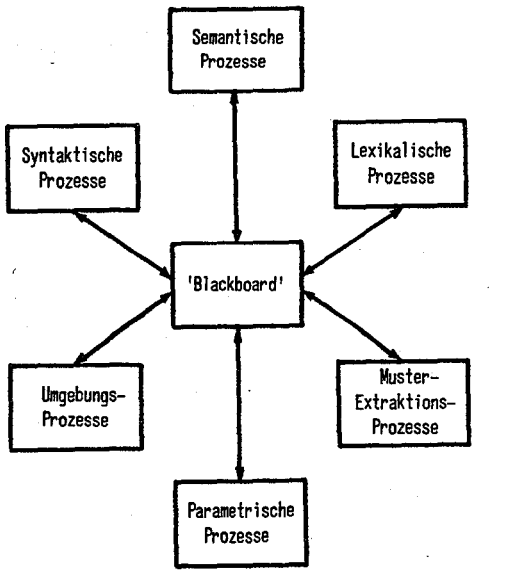


Abb.4.3 Hearsey-II System

5.0 Hören

Im folgenden Kapitel werden wir uns mit den Erkenntnissen über das menschliche hören befassen, um besser verstehen zu können, nach welchen Prinzipien - im Unterschied zu den bisher vorgestellten, elektronischen Systemen - das menschliche Erkennen von Sprache funktioniert. Das Ziel, das wir dabei im Auge haben, ist nicht, eine vollkommene Kopie des menschlichen Hör- und Erkenntnisapparates zu bauen, sondern aus den Konstruktionsdetails des menschlichen Hörapparates die wesentlichen Elemente herauszuziehen, um beim Bau eines Apparates zur automatischen Spracherkennung die in den vorhergehenden Kapiteln erläuterten Probleme der Erkennung von sprecherunabhängigen, fließenden Sprache überwinden zu können.
Betrachten wir dazu zuerst die anatomisch-elektrisch meßbaren Charakteristiken des Hörapparates.

5.1 Physiologie des Hörapparats

Der Weg vom äußeren Ohr bis zum Gehörorgan sei anhand der folgenden Abb. 5.1a beschrieben.

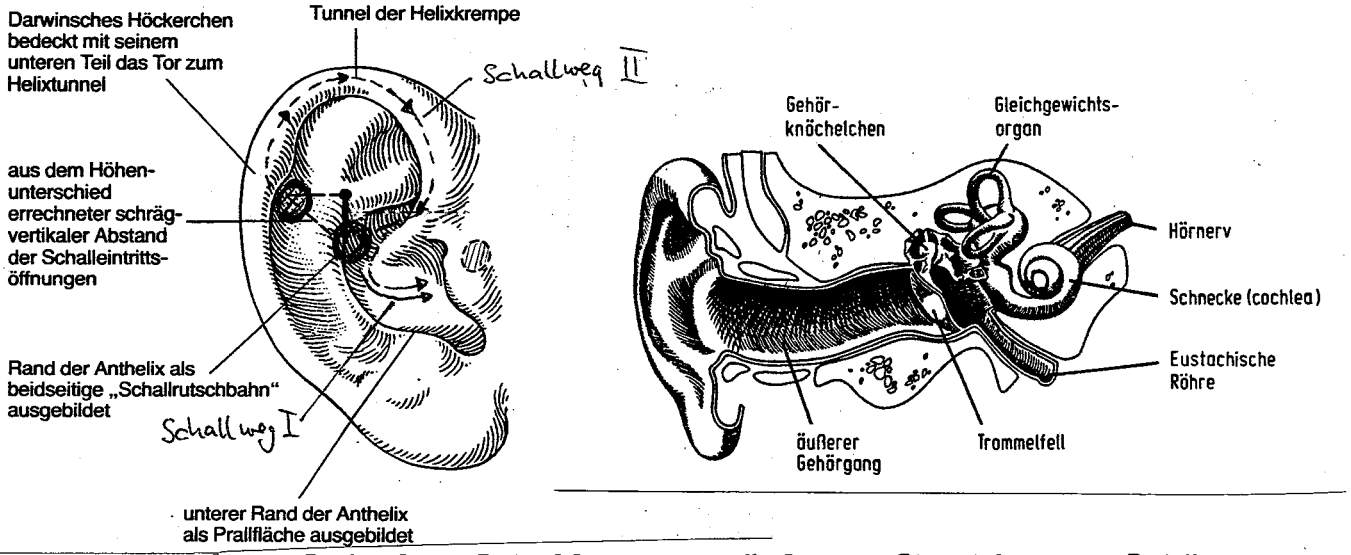


Abb. 5.1a Der Schallweg vom äußeren Ohr bis zum Gehörnerv

Der Schall wird zuerst vom äußeren Ohr links im Bild gesammelt und in den äußeren Gehörgang eingespeist. Dabei tritt durch die besondere Form des Ohres ein Effekt ein, der erst kürzlich aufgedeckt wurde.
Ähnlich wie beim Stereo-Sehen läßt sich auch durch Vergleich der Schalleindrücke beider Ohren beim Stereo-Hören eine Schallquelle auf einer horizontalen Linie lokalisieren. Aber auch in der Vertikalen ist dies möglich, was erst durch die besondere Form der Ohrmuscheln erklärt werden kann. Der Schall im Innenohr setzt sich, wie experimentell nachgewiesen wurde, aus dem Schallanteil zusammen, der direkt in den Gehörgang eintritt, als auch einem Anteil, der umgeleitet über den Gehörgang eintritt, als auch einem Anteil, der umgeleitet über den ca. 6,6 cm langen Umweg am Ohrrand entlang (Schallweg II in Abb. 5.1a) zeitlich versetzt in den Gehörgang eintritt. Da der zweite Schallweg oberhalb der Gehörgangeinmündung beginnt, ändert sich das Schall-Summensignal je nach vertikalem Abstrahlort. Dies ermöglicht uns eine Schallokalisierung, als ob wir 4 Ohren hätten.

Nach der Weiterleitung in den äußeren Gehörgang, der eine unscharf ausgeprägte Resonanz von 2-6 KHz hat, trifft der Schall auf das Trommelfell und bringt es zum Schwingen. Die durch die Luft hervorgerufenen Membranschwingungen werden durch die mechanische Hebelwirkung von drei **Gehörknöchelchen** ("Amboß", "Hammer" und "Steigbügel") auf das eigentliche, flüssigkeitsgefüllte Hörorgan übertragen.
Die Gehörknöchelchen erfüllen dabei zwei Aufgaben.
Zum einen bewirken sie eine Impedanztransformation von Luftschall auf Flüssigkeitsschall; die Schwingungen des Trommelfells werden auf einen 17-fach kleineren Teil ("ovales Fenster") des Gehörorgans mit dem 22-fachen Druck übertragen. Damit wächst die nutzbare Schallenergie von 5% auf 60%; vom Trommelfell werden nur noch 40% des Schalls reflektiert.
Zum anderen wird eine Schutzfunktion bei großen Schallintensitäten aktiviert: die Knöchelchen können durch Muskeln festgehalten werden (Reflex) und die Bewegungsrichtung des Steigbügels kann sich in eine Bewegung schräg zum ovalen Fenster hin ändern.
Für das Hören ist am Gehörorgan im wesentlichen ein schneckenartig gewundener Teil zuständig: die **Cochlea.** Die Cochlea oder Schnecke ist vollständig im Knochen eingebettet. Dadurch gelangt außer dem durch die Gehörknöchelchen übertragenen Luftschall auch Körperschall über die Knochen direkt in die Cochlea. Da der Luftschall auf diesem Wege um 50-60 dB gedämpft wird, spielt der Schallweg (bis auf die Wahrnehmung der eigenen Stimme) keine besondere Rolle. Die Cochlea hat ca 2 1/2 Windungen und ist abgewickelt ungefähr 32 mm lang. Schneidet man sie quer durch, so sieht man, daß sie der Länge nach durch Membranen in drei parallele Kammern (**scala vestibuli, scala media** und **scala tympani**) aufgeteilt ist.

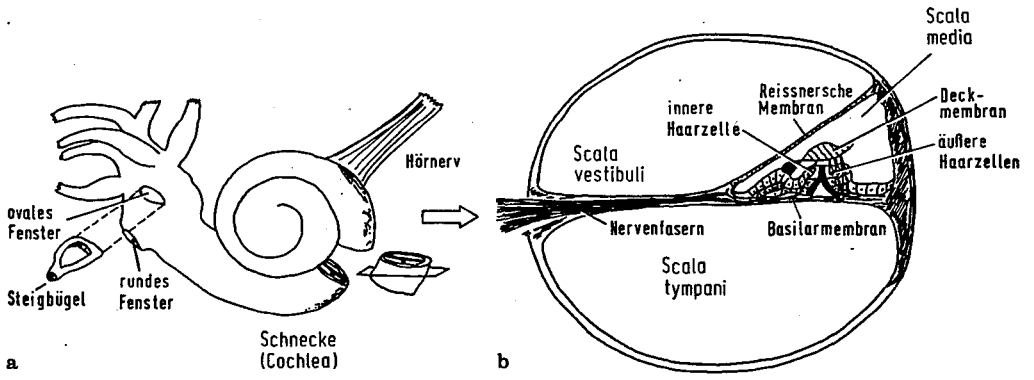


Abb. 5.1b Struktur des Gehörorgans

Der Schall wird vom Steigbügel auf die dünne Membran des ovalen Fensters übertragen, läuft durch die Windungen bis ans Ende der Cochlea und dort durch ein Loch ("**Helicotrema**") zwischen den beiden Hauptkammern von der scala vestibuli zur scala tympani. Zum Druckausgleich befindet sich noch ein mit einer Membran verschlossenes "rundes Fenster" in der scala tympani, analog zum ovalen Fenster. Da bei der scala media die Membran zur scala vestibuli akustisch (aber nicht elektrisch!) unwirksam ist, läßt sich die Cochlea als zwei-Kammern-System modellieren, das im Wesentlichen durch eine Membran, die **Basilarmembran,** getrennt ist.

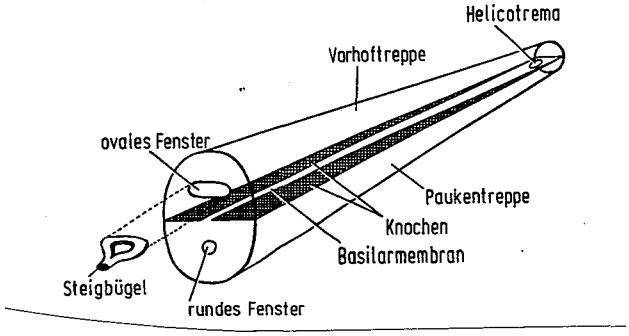


Abb.5.1c Modell der Cochlea

Die Basilarmembran ist schmal und straff gespannt am ovalen Fenster und verbreitert sich mit zunehmender Entfernung. Bei dem Helicotrema wird sie am breitesten und schlaffsten.
Wie wird nun die Schallerregung, die vom Steigbügel in die Cochlea übertragen wird, in die Nervenimpulse der Hörnerven verwandelt?
Betrachten wir dazu die Anatomie in Abb.5.1d genauer.

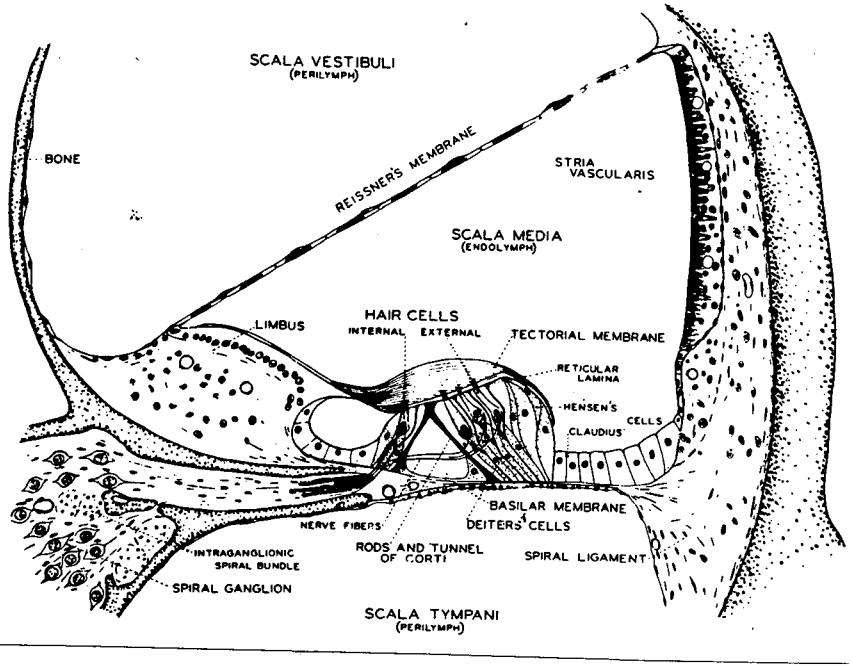
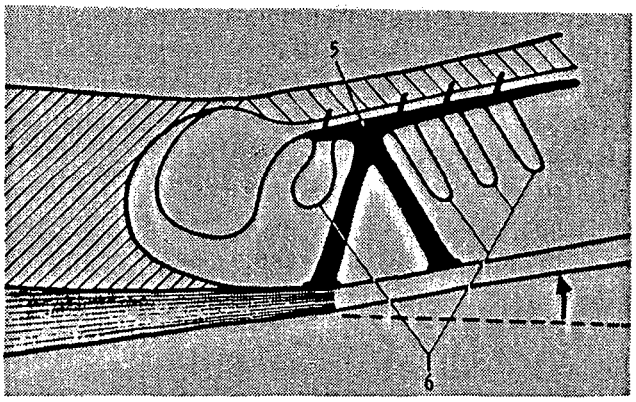
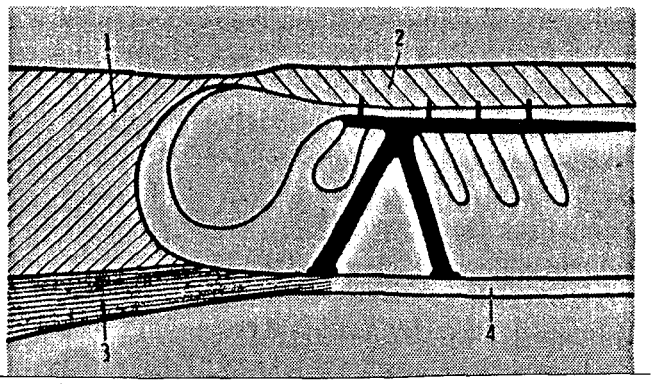


Abb. 5.1d Basilarmembran mit Limbus und Haarzellen

Die Basilarmembran ist mit einer Schicht von Zellen besetzt, in der periodisch eine besondere Sorte von Zellen vorhanden sind, die kurze Härchen besitzen. Diese Härchen sind auf der einen Seite über die Zellen fest mit der Basilarmembran verbunden und liegen auf der anderen Seite lose auf einer Membran ("**Tektorialmembran**") auf, die sie bedeckt. Die Basilarmembran ist auf beiden Seiten der Kammer befestigt, die Tektorialmembran dagegen nur auf einer, dem "Wickelzentrum" der Schnecke zugelegenen Seite am sog. "**Limbus**". Bewegt sich nun die Basilarmembran unter der Einwirkung der Schallwellen, so verschieben sich die beiden Membranen, und damit auch die Haare bezüglich der Tektorialmembran, gegeneinander.



Bewegung des Corti-Organes und der Tektorialmembran um gegeneinander versetzte Achsen, die zu einer Scherbewegung im Bereich der Sinneshaare führt



1 Limbus, 2 Membrana tectoria, 3 Knochen, 4 Basilarmembran, 5 Membrana reticularis, 6 Haarzellen

Abb.5.1e Erregungsmechanik der Haarzellen

Die Bewegung dabei ist in sehr kleinen Dimensionen. Führt ein Schallpegel nahe der Hörschwelle (20 uPa) noch zu einer Auslenkung des Trommelfells von 10^{-9} cm (Wasserstoffatom: 10^{-8} cm Durchmesser), so wird die Basilarmembran um 10^{-11} bis 10^{-9} cm ausgelenkt bei einer Länge der Haare von 10^{-4} cm und einem Abstand von 10^{-2} cm von der Basilarmembran. Damit sind die Haare ca 1 Million mal größer als die Schwingungsamplitude!
Die Haarzellen sind in zwei verschiedene Gruppen eingeteilt. Die in Abb.5.1d links dichter am Limbus (und damit am "Wickelzentrum" der Schnecke) sitzenden Haarzellen werden **innere Haarzellen** genannt; die in drei Reihen mit runden Haarbüscheln weiter rechts lokalisierten Haarzellen werden als **äußere Haarzellen** bezeichnet. Neurologisch unterscheiden sich die beiden Gruppen ziemlich stark. Von den 30.000 Nervenfasern, die den Hörnerv bilden und deren Zellkerne in der inneren Cochlea-Wandung gelagert sind, werden 95% (28.500) von den inneren 3000 Haarzellen, dagegen nur 5% (1.500) von den 13.000 äußeren Haarzellen innerviert. Jede Faser wird dabei von vielen Sinneszellen beeinflußt, andererseits beeinflußt auch jede Sinneszelle viele Fasern. Zusätzlich zu den ableitenden (**afferenten**) Neuronenverbindungen gibt es noch Verbindungen, die die Sinneszellen selbst beeinflussen (**efferente** Kontakte) und hauptsächlich aus dem Gehirnstamm kommen. Bei den Verbindungen kann man zwischen **radialen** (von "innen" nach "außen") und **zirkulären** (entlang der Reihen der Haarzellen) unterscheiden.
Die Hauptmerkmale der inneren Haarzellen sind

- viele radiale afferente Kontakte
- viele ziruläre efferente Kontakte

und der äußeren Haarzellen

- wenige afferente zirkuläre Fasern (ca 20 Kontakte pro Faser)
- viele efferente radiale Kontakte (ca 40.000)

Da die affrenten und efferenten Fasern zu den gleichen, weiterverarbeitenden Gehirnteilen (Olivenkern-Komplex) führen, liegt hier ein dynamisches, rückgekoppeltes System vor, dessen Funktion aber noch ziemlich unklar ist.

5.2 Die Funktion des Innenohres

Was weiß man von der Funktion der Cochlea?
Über die Funktion des Hörmechanismus gab und gibt es verschiedene Theorien. Helmholtz (1862) vermutete, daß jedem Ort der sich verbreiternden Basilarmembran eine Frequenz zugeordnet sei, ähnlich einem Xylophon oder einer Reihe von Resonatoren (z.B. Stimmgabeln), die bei unterschiedlichen Frequenzen in Schwingung verstzt werden (**Resonanztheorie**). Diese Theorie enthielt allerdings Widersprüche. Nahm man stark gedämpfte Resonatoren an, so ließ sich damit die gute Zeitauflösung (Wahrnehmung kurzer Impulse) erklären. Allerdings steht dies im krassen Gegensatz zu der Notwendigkeit, schwach gedämpfte Resonantoren annehmen zu müssen, um die guten frequenzauflösenden Eigenschaften (0.1% bei 1KHz) des Gehörs zu erklären.
Eine andere Hypothese, die **Telefontheorie** von Rutherford (1886), vermutete, daß im Innenohr das Schallsignal ähnlich einem Mikrofon direkt in Nervenimpulse umgewandelt wird. Die neurologischen Erkenntnisse des frühen 20.Jahrhunderts zeigten allerdings, daß die maximale Entladungsfrequenz eines Neurons bei 1000Hz liegt, also bedeutend unter der maximal hörbaren Frequenz. Umso überraschender waren die Beobachtungen, die 1930 von Wever und Bray gemacht wurden. Sie verbanden den gesamten Hörnerv einer Katze mit einem Verstärker und Lautsprecher und konnten die ins Ohr gesprochene Stimme einwandfrei hörbar machen. Daraufhin wurden Theorien entwickelt (**Volley-Theorie**), die das Ergebnis als Überlagerung von Nervenimpulsen verschiedener Fasern deuten. Tatsächlich können ja zwei Neuronen, die jeweils mit 1Khz feuern, zusammen überlagert eine Frequenz von 2KHz ergeben. Nachfolgende Unersuchungen zeigten aber, daß dieses dem mechanischen Schallsignal streng analoge und ohne Schwellwert beobachtbare Signal (**Mikrophonpotential**) von den äußeren Haarzellen stammt, die bei der Bewegung einen veränderlichen Widerstand zwischen dem elektrisch positiven Potential der scala media und der neutralen scala tympani darstellen.
Das Mikrophonpotential kann man an verschiedenen Teile der Cochlea ableiten und wird heutzutage klinisch von Ohrenärzten benutzt, um die Funktion des gesamten mechanischen Apparats vom äußeren Ohr bis zur Basilarmembran zu testen.
Die heutigen Theorien über die Funktion der Basilarmembran stammen im wesentlichen von dem ungarischen Naturforscher Bekesy, der in 50-jähriger Arbeit das Innenohr direkt unterm Mikroskop untersuchte. Dazu entfernte er unter Wasser an Leichenpräperaten ein Stück Knochen aus dem Schädel, um die Cochlea freizulegen, öffnete die Cochlea und plazierte Silberkristalle auf der Basilarmembran. Er ersetzte das ovale Fenster durch eine Gummimembran, die mit einem Lautsprecher gekoppelt war. Dann verschloß er das Loch im Knochen mit einer Glasplatte. Wurde nun der Lautsprecher erregt, so konnte er mit einem Lichtmikroskop für Unterwasserbeobachtung bei stroboskopischer Beleuchtung die Auslenkung der Basilarmembran sehen. Was er beobachtete, war folgendes: Durch die Erregung erschien eine Welle auf der Basilarmembran, wurde größer mit wachsender Entfernung vom ovalen Fenster, erreichte ein Maximum und erstarb ziemlich rasch auf dem weiteren Weg. Diese spezielle Welle, deren Form ziemlich unabhängig von der Form des 2-Kammersystems ist, wurde **Wanderwelle** genannt. In Abb. 5.2a ist ein Momentanzustand gezeigt.

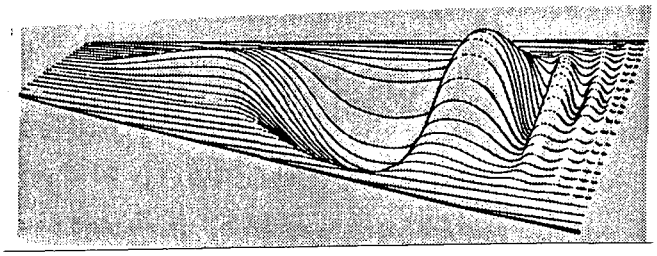


Abb. 5.2a Momemtanzustand der Wanderwelle

Das Maximum der Wanderwelle verschiebt sich dabei je nach Frequenz im Sinne der Resonanztheorie. Die dabei zugrunde liegende Mathematik ist allerdings kompliziertwer als bei der Resonanztheorie.
Allerdings vermag auch diese Theorie nicht die nicht-linearen Zusammenhänge zwischen mechanischer Auslenkung und neuronaler Frequenzselektivität erklären. In Abb.5.2b sind mehrere Abstimmkurven eingetragen, die man dadurch erhält, daß man die mechanische Auslenkung eines Punktes der Basilarmembran (gestrichelte Linie) und die Neuronenaktivität an dem zugeordneten Hörnerv (durchgezogene Linie) durch Anzapfung einer Faser bei verschiedenen Frequenzen mißt. Experimentell beobachtet man die Stelle der Basilarmembran, an der die Wanderwelle bei der "Resonanzfrequenz" des angezapften Neurons maximal wird und mißt die Schallintensität, um die mechanische Aulenkung bzw. die neuronale Aktivität bei verschiedenen Frequenzen konstant zu halten.

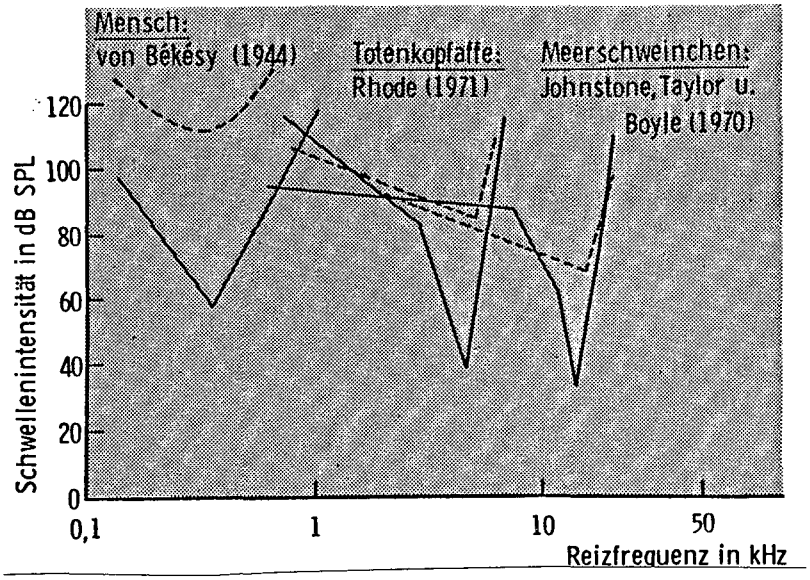


Abb.5.2b mechanische und neuronale Abstimmkurven

Wie man in der obigen Abbildung sieht, besteht in der neuronalen Frequenzselektivität eine Nicht-Linearität, die auch neuere Untersuchungen der mechanischen Auslenkung der Basilarmembran mittels Laserlicht, Mößbauereffekt und Kapazitäts-methoden nicht erklären können. Es besteht die starke Vermutung, daß die erwähnten neuronalen Verschaltungen und Rückkopplungsmechanismen dafür verantwortlich sind. Allerdings gibt es bisher noch keine Theorie, die die Unsymmetrie zwischen den afferenten Ableitungen und der Zahl der äußeren und inneren Haarzellen konsistent erklären kann.

- Eine Vermutung vergleicht die äußeren Haarzellen mit den Stäbchenzellen der Augenretina, die in der Retina durch Summierung der Reize (viele Zellen an einem Nerven) das Auge dämmerungsempfindlich macht, und schreibt ihnen die gute Empfindlichkeit des Hörens zu . Demgegenüber sollen die inneren Haarzellen unempfindlicher, aber (entsprechend der Farbtüchtigkeit der Zäpfchenzelle) bessere Frequenzselektivität aufweisen.
  Diese Vermutung wird aber nicht durch den Befund bei Einzelableitungen der Nervenfasern gedeckt. Alle angezapften Fasern mit guter Frequenzselektivität zeigten diese auch bis zur Hörschwelle; wurden weniger selektive Fasern gefunden, so waren sie auch unempfindlicher.

- Andere Hypothesen besagen, daß die gute Frequenzselektivität durch eine Wechselwirkung der inneren und äußeren Haarzellen hervorgebracht wird. Experimente, bei denen chemische Substanzen die äußeren Haarzellen für eine gewisse Zeit in der Funktion ausschalten, zeigen auch eine reversible Veränderung der Abstimmkurven aus Abb.5.2b.
  Leider aber konnten bisher in der Cochlea keinerlei synaptische Wechselwirkungen zwischen beiden Haarzellen-Systemen nachgewiesen werden.

- Eine weitere Vermutung sieht die inneren Haarzellen zuständig für die Frequenzanalyse, die äußeren für die Zeit- und Phasencodierung.
  Diese Vermutung erklärt aber damit nicht das 95:5 Verhältnis der afferenten Fasern.

Zusammenfassend kann man sagen, daß es bisher zwar viel Klarheit über die Anatomie, aber nicht über die Funktion der Cochlea gibt.

Abschließend soll noch kurz der weitere Verlauf der Hörnerven in Abb. 5.2c skizziert werden.

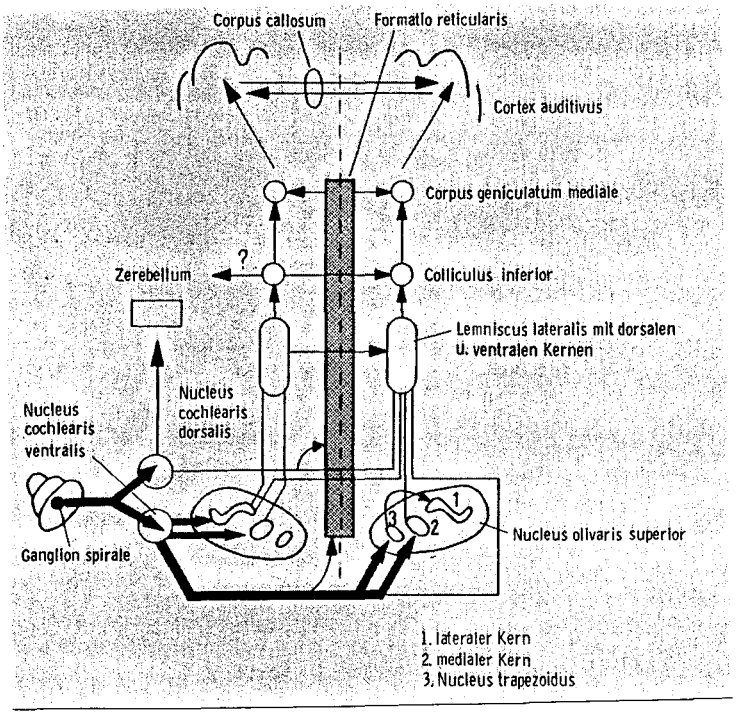


Abb. 5.2c Schema der auditiven Verarbeitung im Gehirnstamm

Über den Olivenkern-Komplex, einem neuronalen Verschaltungsteil vor dem eigentlichen Hörzentrum des Gehirns, werden die Signale beider Ohren verglichen und die Stereo-Information ausgefiltert. Auf dem Bild oben ist dabei nur eine Seite gezeigt; die andere Cochlea muß ergänzt werden.

5.3 experimentalpsychologische Hörerkenntnisse

Da Sprache in der uns bekannten Form etwas typisch menschliches ist (Ausanhme: Delphine, Wale etc), verbieten sich aus ethischen Gründen weiterführende neurologische Untersuchungen. Stattdessen versucht man, durch einen geschickten Versuchsaufbau und Befragung der Versuchspersonen etwas genaueres über die menschlichen Hörmechanismen zu erfahren. Das Hören wird dabei in möglichst kleine, von den Experimenten her begründbare Einzelschritte zerlegt und die Eigenschaften dieser Stufen experimentell verifiziert.

Subjektive Hörwahrnehmung

Die einfachste Versuchsanordnung testet die Frequenzempfindlichkeit bei verschiedenen Schallpegeln. Dazu erhält man auf einem Ohr einen Ton mit einer Frequenz und einer bestimmten Lautstärke und muß nun die Lautstärke eines anderen Tons einer anderen Frequenz auf dem anderen (oder gleichen) Ohr so justieren, daß beide Töne die gleiche subjektive Lautstärke besitzen. Bezieht man die relativen Lautstärken auf eine Bezugsfrequenz, beispielsweise 1000 Hz, so ergeben sich als Linien gleicher Lautstärkeempfindung die Isophonen in Abb.5.3a.

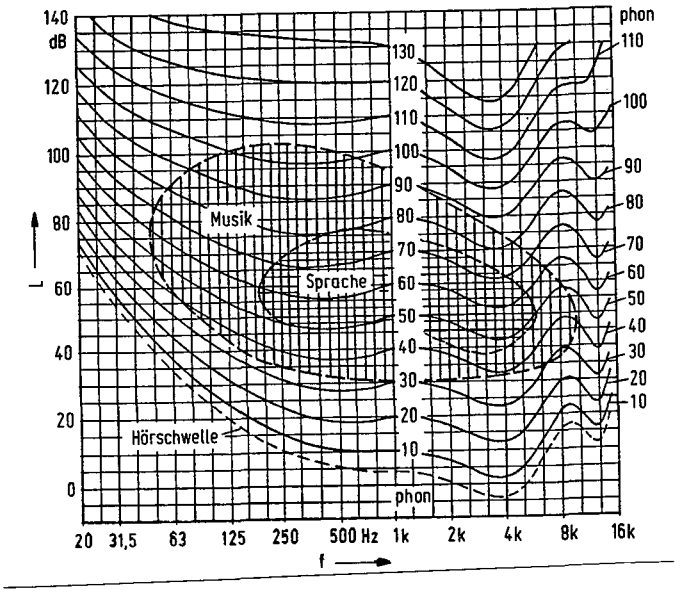


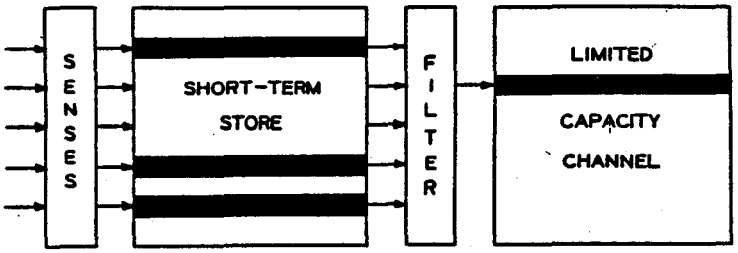
Abb.5.3a Hörfeld nach DIN

Wie wir aus dem Diagramm entnehmen können (Eichung in Dezibel!), gilt auch für die auditiven Sinneswahrnehmung das Weber-Fechnersche Gesetz, wonach die subjektive Wahrnehmung proportional zum Logarithmus des Reizes ist. Damit besitzt das Hörsystem eine Proportion, die gut mit dem Resultat in Kapitel 3.1 übereinstimmt, wonach der Rauschabstand der Signale nur bei einer

logarithmischen Kodierung konstant ist. Das Hörsystem verhält sich also hierbei "optimal" im Sinne der Nachrichtentheorie.

Verarbeitungsmodelle

Als einer der ersten Autoren modellierte Broadbent (1958) das Hörsystem als paralleles, informationsverarbeitendes System.



Flow diagram of two successive stages of processing in Broadbent's (1958) model.

Abb.5.3b Blockdiagramm der Verarbeitungsstufen von Broadbent

Typisch an diesem Modell ist die große Parallelität der Einzelsensorinformation, die nach einem Filterprozeß auf wenige, "wichtige" Information reduziert wird.
Alternativ dazu ist sequentielle Modell der auditiven Informationsverarbeitung von Massaro in Abb.5.3c aufgeführt:

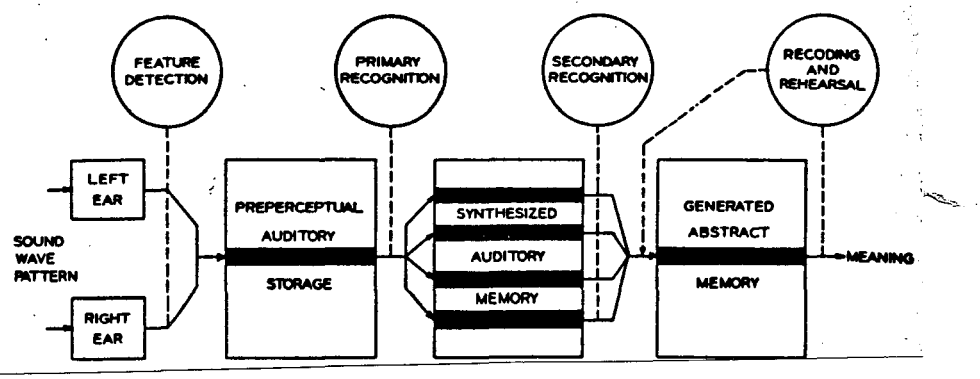


Abb.5.3c System mit begrenzter Kanalkapazität

Bei diesem Modell wird (im Unterschied zum ersten) nicht für jede Informationsart ein eigener Kanal vorgesehen, sondern die verschiedenen auditiven Reize werden auf einem Kanal mit begrenzter Kapazität vorverarbeitet und erst bei höheren Stufen parallel analysiert.
Die Entscheidung zwischen den beiden sich gegenseitig ausschließenden Modellen läßt sich gut mittels sog. **Maskierungsexperimente** treffen. Hierbei werden Reize dargeboten, deren Wiedererkennung durch andere Reize (**Maske**), die vorher (**forward masking**) oder hinterher (**backward masking**) dargeboten werden, gestört werden kann. Für das Hören wurden Testtöne und Maskierungstöne mit jeweils anderer Frequenz und Darbietungsfolge ausgewählt. In Abb.5.3d ist das Ergebnis zu sehen.

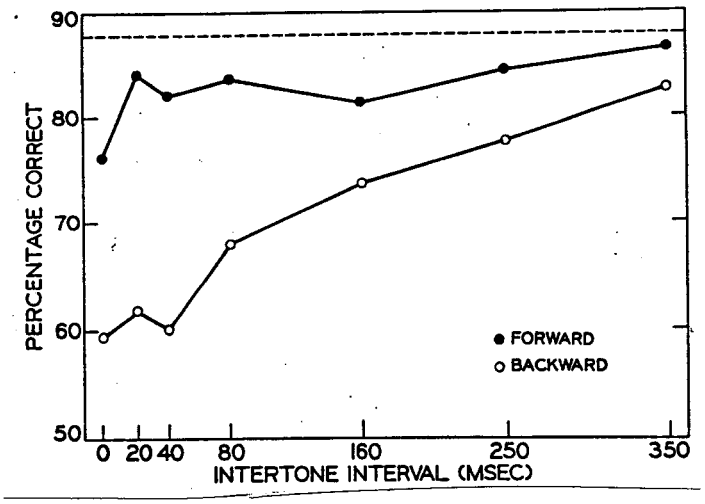


Abb.5.3d Maskierungsexperimente von Massaro

Wie man sieht, wird die Erkennung der Töne durch das Erscheinen der Maskierungstöne erheblich gestört; umgekehrt dagegen ist dies kaum der Fall. Diese Experimente sind mit der Broadbent-Hypothese kaum zu vereinen, so daß man eine begrenzte, sequentielle Verarbeitung schon bei den frühen Stufen annehmen muß.

Missing fundamental

Ein weiteres Phänomen ist das Hören nichtexistenter Töne. Bietet man einen Schalleindruck dar, der aus Obertönen(z.B. 730,1095,1460 und 1825 Hz des Grundtons 365 Hz) gemischt ist, ohne dabei die Grundfrequenz zu enthalten, so hört man trotzdem die Grundfrequenz mit. Das Hörsystem scheint die Grundfrequenz auf ziemlich früher Stufe zu ergänzen, wie man an den Brainstem-Potentialen der EEG-Ableitungen zeigen kann. Während man die Wahrnehmung eines reinen 365Hz-Tons durch überlagertes Schmalbandrauschen unterdrücken kann, ist dieses im Verbund mit den Oberwellen nicht möglich. Damit scheiden rein akustisch-physikalische Phänomene aus.
Das Phänomen der fehlenden Grundfrequenz macht man sich schon seit langer Zeit beim Glockenbau zu Nutze. Durch Abstimmung einer Glocke auf die verschiedenen Obertöne kann man eine wesentlich schwerere, größere und damit auch teurere Glocke mit der niedrigen Grundfrequenz vortäuschen.

Kategoriale Wahrnehmung

Das Phänomen der Ergänzung geht weitgehend einher mit einem anderen Phänomen, der Kategorisierung der wahrgenommenen Sprachlaute. Liberman (1959) versuchte mit seinen Mitarbeitern, die Grundelemente der Sprache durch Variation synthetisch erzeugter Bruchstücke zu ermitteln. Sie versuchten beispielsweise, einen /d/-Laut aus /di/ und /du/- Lauten zu isolieren. Erstaunlicherweise wollte ihnen dies nicht gelingen; entweder wurde der Laut als /d/ + Vokal wahrgenommen, oder aber als sprachuntypisches Geräusch. Auch eine systematische Variation eines Stop-Konsonanten wurde nicht bemerkt, sondern die Folge der kontinuierlich veränderlichen Laute wurden von den Versuchspersonen scharf in Abschnitte getrennt und als typische /b/, /d/ und /g/-Laute erkannt. Zwischen den Variationen innerhalb eines Abschnitts (einer Klasse) wurde nicht weiter unterschieden. In Abb.5.3e ist diese plötzliche Änderung in der Wahrnehmung anhand der beiden Silben PAH und BAH gezeigt, die durch kontinuierliche Variation

des zeitlichen Einsatzpunktes des Vokals /a/ ineinander überführt werden können.

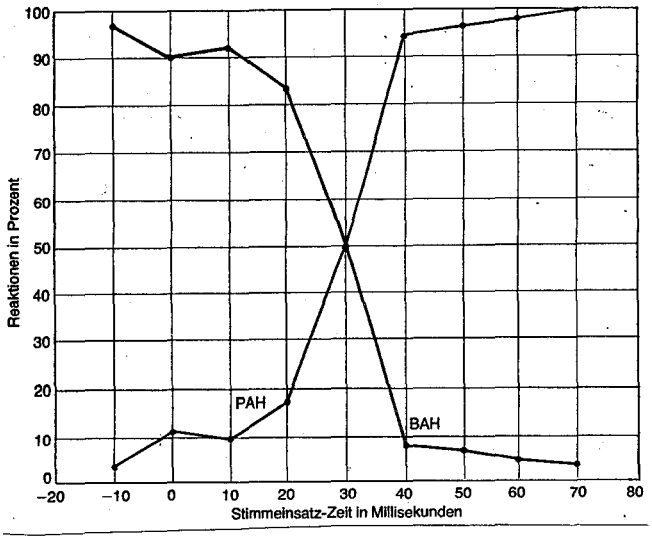
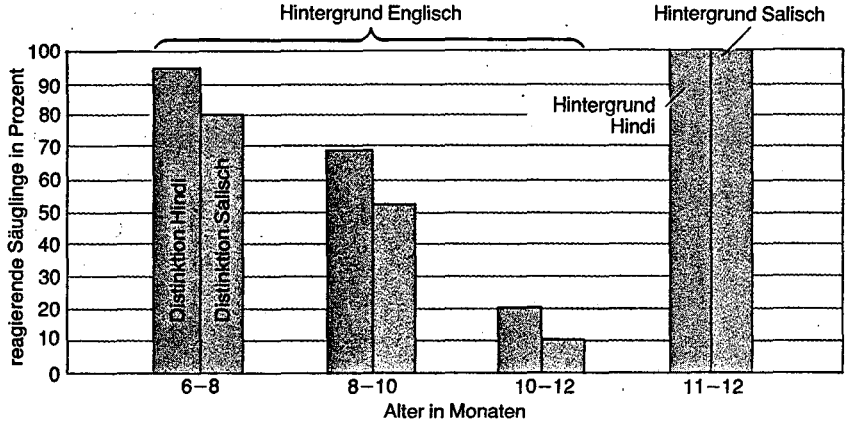


Abb.5.3e Kategoriale Wahrnehmung

Wie andere Experimente an Babys zeigen, kann man die Kategoriale Wahrnehmung schon bei wenigen Wochen alten Babys nachweisen. Dabei verliert sich mit fortlaufender Anpassung der Kinder an die Laute der Muttersprache die Fähigkeit, lautliche Unterschiede, die zwar in anderen, aber nicht in der Muttersprache eine Rolle spielen, überhaupt wahrzunehmen. In Abb.5.3f ist dies für die Wahrnehmung von nicht-englischen Sprachlauten gezeigt.



Die Fähigkeit, sprachliche Distinktionen wahrzunehmen, die dem Englischen fremd sind, bildet sich zurück, wenn sie nicht genutzt wird; dies zeigen die Reaktionen von Kindern mit englischsprachigem Hintergrund. Janet F. Werker von der Dalhousie-University und Richard C. Tees von der University of British Columbia testeten Kinder verschiedener Altersgruppen; sie stellten fest, daß der Anteil derjenigen, die auf konsonantische Unterschiede in Hindi und Salisch (einer nordamerikanischen Indianersprache) reagierten, mit dem Alter schnell absank. Einjährige Hindi- und Salisch-Kinder hingegen bleiben weiterhin fähig, die bedeutungsunterscheidenden sprachlichen Einheiten ihrer jeweiligen Muttersprachen zu erkennen.

Wie kann man die kategoriale Wahrnehmung erklären?

- Eine Theorie (**Motor-Theorie** von Liberman) vermutet, daß Sprachverstehen und Sprachgenerierung zwei verschiedene Modi ein-und desselben Mechanismus darstellen. Alle Laute, die man nicht gelernt hat zu sprechen, können deshalb auch nicht erkannt werden.
  Diese Theorie erklärt aber nicht, warum trotzdem von Geburt an stumme, aber nicht taube Menschen Sprache

verstehen können.

- Eine andere Hypothese, die **Analyse-durch-Synthese**, vermutet eine Sprachproduktion parallel zum Erkennungsprozeß. Durch Vergleich zwischen den möglichen Produktionen und den tatsächlich Gehörtem wird erkannt, was tatsächlich (vgl.4.2 und 4.3!) gesprochen wurde.

Es besteht Einigkeit darüber, daß keine Theorie alle vorkommenden Phänomene erklären kann; nur an welcher Stelle welche Aspekte dominieren ist unbekannt.

6.0 Ein Modell für Spracherkennen

In den bisherigen Kapiteln wurden die Versuche geschildert, Sprache anhand von bestimmten Charakteristiken zu beschreiben und zu erkennen. Dies begann mit den physiologischen, physikalischen und phonetischen Charkterisierungen, ging über die Frage, was das Erkennen von Sprache behindern kann und erstreckte sich über die verschiedenen Verfahren, Sprache zu kodieren sowie kodierte Worte und Sätze wiederzuerkennen.
In diesem Kapitel soll nunder Versuch unternommen werden, zwischen den Bruchstücken des Wissens eine Verbindung herzustellen, auch wenn diese nicht immer konsistent ist.

6.1 Modellierung des menschlichen Spracherkennens

Über die psychologischen, semantischen und syntaktischen Aspekte des menschlichen Spracherkennens gibt es viele Untersuchungen. Für unser Modell soll aber nur ein kleiner Teil der Ergebnisse herangezogen werden, um die wichtigsten Proportionen des menschlichen Spracherkennens zu modellieren.
Dies sind folgende:

- Spracherkennen geht in mehreren Stufen vor sich. Auf jeder Stufe (Laut, Wort, Satz) werden fehlende Teile kontextrichtig ergänzt (vgl 2.4 Verständlichkeitstests); es findet also eine **Ergänzungsoperation** statt.
  Die Ergänzung gilt auch für gleichzeitig auftretende Töne (5.3 missing fundamental).

- Auf der Lautebene gibt es eine Kategoriale Wahrnehmung (s.5.3), d.h.Variationen des gleichen Lauts sollen, sofern sie nicht zu stark abweichen, als ein-und derselbe Laut empfunden. Dies entspricht einer **Klassifizierung.**
  Da auf Wort- und Satzebene ebenfalls Variationen überhört werden, sofern diese unwesentlich sind, und dies in der Extrapolation auch für die Sätze selbst gelten, wenn der Inhalt extrahiert wird, so scheint die Klassifizierung (oder **Mustererkennung**) auch auf höheren Stufen zu gelten.

- Experimentalpsychologische Untersuchungen zeigten, daß die Laut-, Wort- und Satzerkennung gleichzeitig abläuft und sich gegenseitig beeinflußt (vgl. 4.3). Die Syntax- und Semantikerkennung wird also nicht erst nach der Worterkennubng vorgenommen, sondern beeinflußt diese auch.

In Abbildung 6.1a ist nun ein Verarbeitungsschema gezeigt, das diese Forderungen berücksichtigt.
Der Informationsfluß ist durch Doppelpfeile angedeutet; Zeitverzögerungsstufen sind mit "delay" gekennzeichnet.
Einerseits ist die Spracherkennung darin sequentiell in Stufen mit begrenzter Kanalkapazität organisiert, wie es die Maskierungsexperimente (s.5.3) nahelegen.
Andererseits ist durch das pipe-lining eine parallele Verarbeitung auf verschiedenen Ebenen möglich, die durch die Rückkopplungszweige sich gegenseitig beeinflussen können.

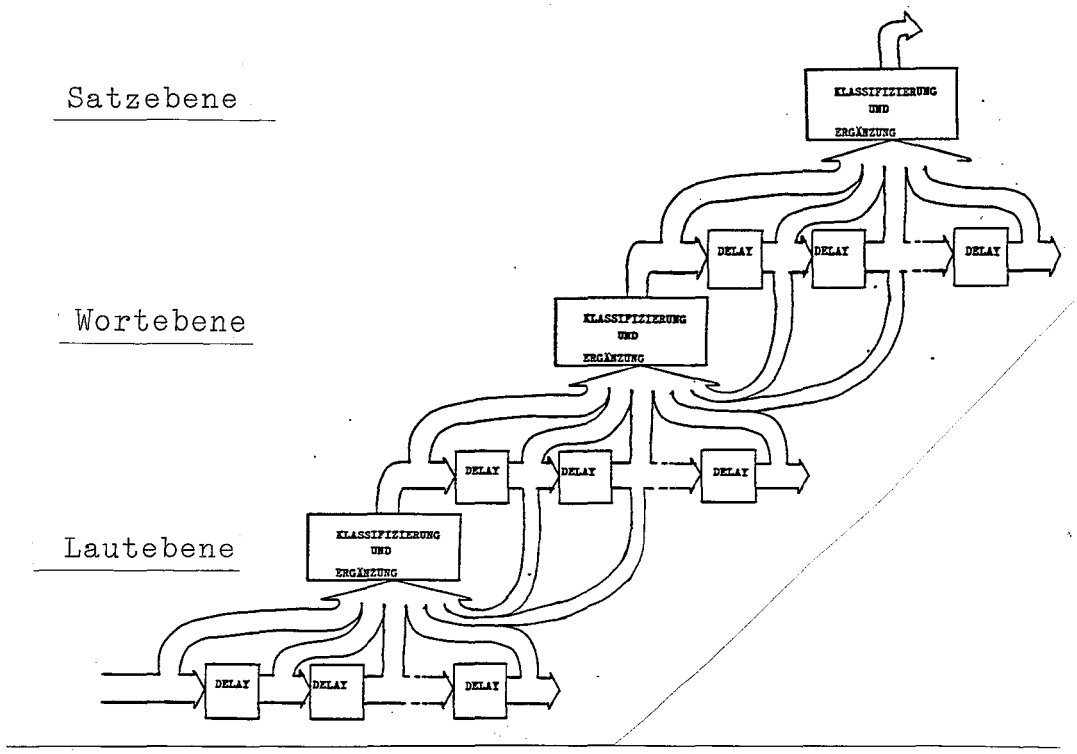


Abb.6.1a Ein Stufenmodell der Sprachverarbeitung

Die Stärken des Modells liegen in seiner Symmetrie, die prinzipiell ähnliche Operationen auf verschiedenen Stufen der Verarbeitung vorsieht und damit grundsätzliche Proportionen menschlichen Spracherkennens zu beschreiben versucht. Damit steht das Modell in Einklang mit neuroanatomischen Untersuchungen, die beim Gehirn eine ziemliche uniforme, sich wiederholende Struktur festgestellt haben, wie sie nachfolgend beschrieben werden soll.
Die Schwächen des Modells liegen in den verschiedenen, gemachten Annahmen. Zeigen die Experimente zur Kategorialen Wahrnehmung noch ziemlich deutlich, daß ein Klassifizierungsmechanismus auf Lautebene existieren muß, so ist die Notwendigkeit einer solchen Konstruktion auf höherer Ebene nicht mehr evident. Auch die Annahme, daß es reine Verzögerungselemente im neurologischen Apparat geben soll, ist ziemlich willkürlich. Zwar gilt es als gesichert, daß bei der Spracherkennung auf den verschiedenen Stufen mehrere Kurzzeitgedächtnisse mit verschiedenen Zeitkonstanten existieren, aber die Repräsentation eines Kurzzeitgedächtnisses durch mehrere "delay"-Elemente ist rein willkürlich. Auch die Aufteilung der Stufen in Laut-, Wort- und Satzebene ist willkürlich. Wahrscheinlicher ist es, daß sehr viele solcher repetativen Strukturen existieren, bei denen die Zuordnung zu einzelnen Funktionsstufen kontinuierlich geschieht.
Die große Frage beim Übersichtsbild 6.1a ist nun:
Wie können die funktionellen Einheiten "Klassifizieren und Ergänzen" realisiert sein?
Betrachten wir dazu die Strukturen des menschlichen Gehirns.

6.2 Ein Modell für "Klassifizieren und Ergänzen"

Die Anforderungen, die an den Funktionsblock "Klassifizieren und Ergänzen" gestellt werden, sind ziemlich hoch. Zum einen sollen die Sprachmuster, die von der Cochlea übermittelt werden (z.B. zeitabhängige Intensitäten der Frequenzbereiche) in "Echtzeit" analysiert und wiedererkannt werden. Dazu wird eine Klassifizierung (Mustererkennung) durchgeführt. Ist das zu klassifizierende Muster auch noch unvollständig, so soll es zum anderen auch noch "richtig" ergänzt werden. Beides möglichst in einem Verarbeitungsgang.
Diese Aufgabe der Klassifizierung und Ergänzung ist auch bei modernen Computern nicht einfach zu bewerkstelligen. Trotz der 100.000-fachen, höheren Schaltgeschwindigkeit der modernen Gatter gegenüber den langsamen Neuronen können heutige Computer im Unterschied zu den neuronalen Netzwerken diese Aufgabe noch nicht in "Echtzeit", also synchron zum Sprechen, durchführen. Das Geheimnis der wesentlich fehleranfälligeren, langsamen Neuronen liegt dabei in der parallelen, fehlertoleranten Organisation. Die Gehirntheoretiker waren deshalb die ersten, die bereits seit über 40 Jahren über die parallelen Organisationsmöglichkeiten vieler, kleiner Einheiten nachdachten. Deshalb bilden Gehirnmodelle eine interessante Alternative zu den konventionellen von-Neumann Computern. Betrachten wir dies näher.

Die Feinstruktur des Großhirns

Bei der anatomischen Betrachtung des Großhirns fällt auf, daß das Gehirn aus einer in Falten gelegten Rinde (**grey matter**) und einem weißlichen Inneren (**white matter**) besteht. Durch spezielle Anfärbemethoden läßt sich erschließen, daß die Großhirnrinde in ca. 1mm breite, kleine Gebiete (Säulen) aufgeteilt ist. Die lokalen Wechselwirkungen der Neuronen jeder Säule sind auf die Säule beschränkt. Zusätzlich gibt es noch weitreichende Wechselwirkungen von den Neuronen einer Säule zu den Neuronen in anderen Säulen, die am "Fuße" der Säulen entlanglaufen. Diese weitreichenden Verzweigungen ("Kabelbaum") bilden das beschriebene "white matter". In Abb.6.2a ist dies gezeigt.

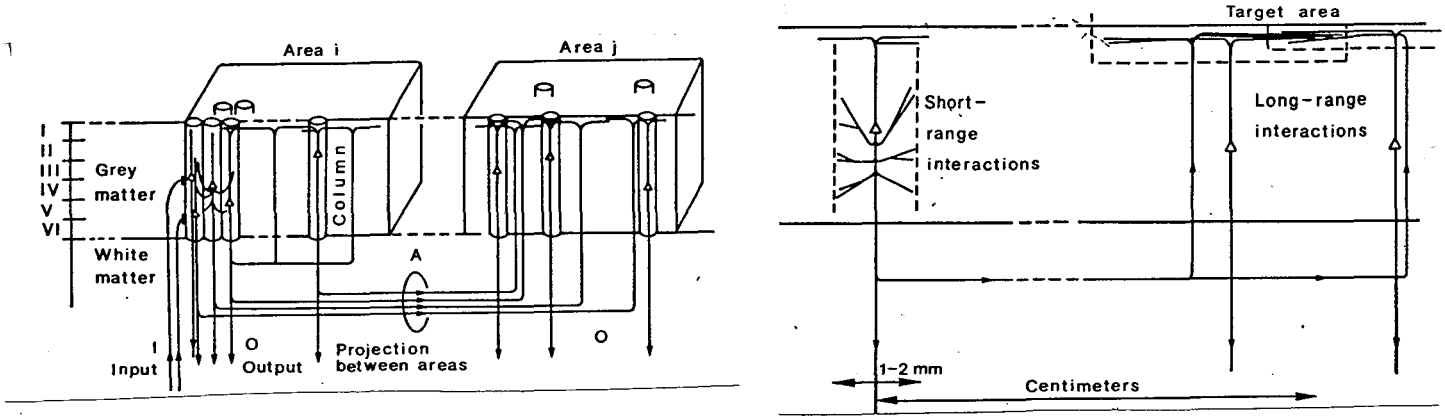


Abb.6.2a Feinstruktur des Großhirns

Die Stärke der Wechselwirkungen der Neuronen i und j innerhalb einer Säule lassen sich durch Koeffizienten M_{ij} modellieren. Die Kopplungskoeffizienten einer Säule oder eines Areals lassen sich somit durch eine Matrix **M** charakterisieren. Ein gesamtes Verarbeitungsgebiet wie das des auditiven Zentrums besteht im Ganzen also aus einer Menge von Säulen oder, modellmäßig

gesprochen, aus einer Menge der sie charakterisierenden Matrizen. Abb. 6.2b zeigt ein solches Schema, das noch durch ein Kontrollzentrum (ARAS) angereichert ist.

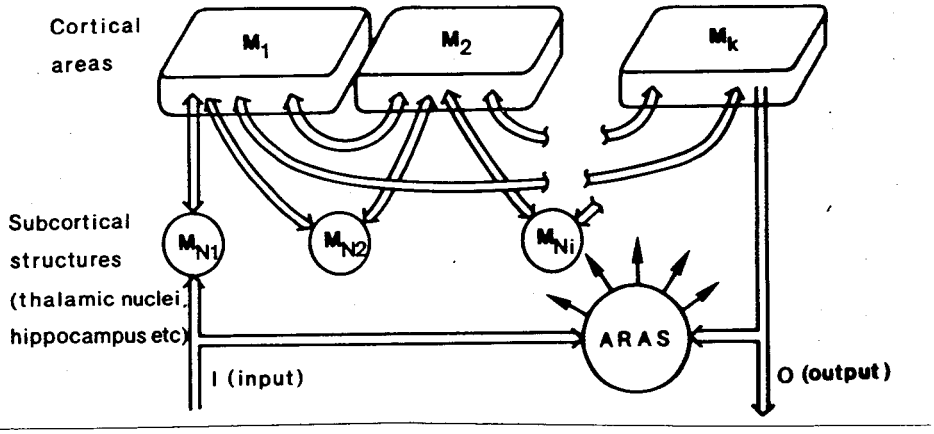


Abb.6.2b Verbindungsschema der Verarbeitungseinheiten

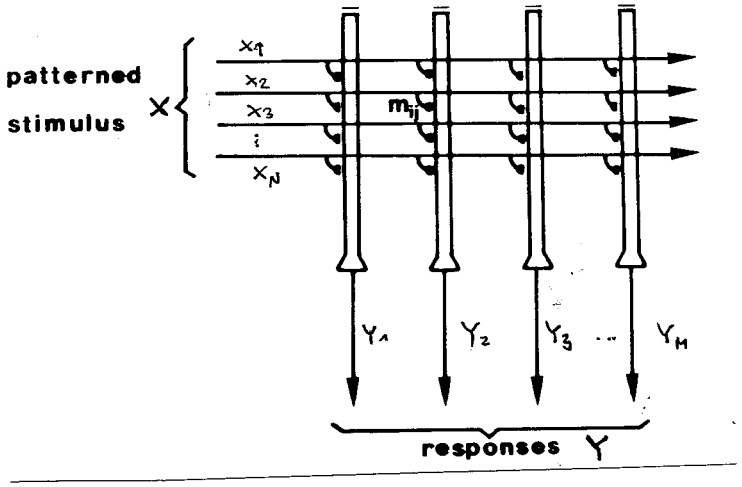
Zweifelsohne lassen sich die Säulen und damit die Matrizen mit den Verarbeitungseinheiten aus Abb.6.1a identifizieren. Wie aber funktionieren diese Verarbeitungseinheiten?
Betrachten wir dazu dieses Matrixmodell genauer.

6.3 Das Lineare Kreuzkorrellations-Matrixmodell

Das lineare, hololologische Matrix-Modell, besser auch Korrelations-Matrix Modell genannt, beruht auf dem linearen Zusammenhang zwischen der Aktivität aller Input-Einheiten $(x_{1...}x_n)=\mathbf{x}$ und dem Output $(y_{1...}y_m)=\mathbf{y}$

$$\mathbf{y} = \mathbf{M}\,\mathbf{x}$$

mit der Matrix der Kopplungswerte $\mathbf{M} = (M_{i,j})$.
Abbildung 6.3a zeigt den schematischen Aufbau mit Modellneuronen.



Assoziative Speicherung

Wie werden nun die Werte von **M** ermittelt?
Folgen wir analog der Hebb'schen Regel, bei gleichzeitigem Auftreten von post- und präsynaptischer Erregung die Synapsenkopplung zu verstärken, so ist

$$\Delta M_{ij} = a\, y_i x_j \qquad (6.3)$$

Angenommen, es existieren p Reize $x^1,\ldots,x^p$ mit den Antworten $y^1,\ldots,y^p$, so ist nach p-maliger Änderung die Matrix **M**

$$M_{ij} = \sum_{k=1}^{p} a_k y_i^k x_j^k$$

Beim Auslesen dieses Assoziativen Speichers durch einen Input-Reiz x^r ist der Output

$$y_i = \sum_j M_{ij} x_j^r = \sum_j \sum_k a_k y_i^k x_j^k x_j^r \qquad (6.3a)$$

$$= a_r y_i^r x^r x^{rT} + \sum_{k\#r} a_k y_i^k x^k x^{rT} \qquad (6.3b)$$

Antwort + Übersprechen

bzw.

$$\mathbf{y} = \mathbf{M}\, x^r = \sum_k \mathbf{y}^k x^r x^k$$

Bei orthogonalen Inputvektoren (d.h. sehr verschiedenen Reizen) fällt mit $x^k x^{rT} = 0$, $k\#r$ das Übersprechen durch andere Reize weg, so daß bei der Normierung

$$x^k x^{kT} = 1/a_k$$

gilt

$$\mathbf{y} = \mathbf{y}^r, \text{ wie gewünscht.}$$

Das Nichtlineare Matrixmodell

Da die Funktion der Neuronen im Gehirn durch nicht-lineare Vorgänge charakterisiert werden kann, gab es schon früh in den ersten Modellierungsansätzen der Gehirnfunktionen nicht-lineare Elemente in den Neuronennetze Sowohl die Existenz einer Potentialschwelle, bei deren Überschreiten eine Reaktion der Neuronen ausgelöst wird, als auch die Kodierung der Reaktion als eine amplitudenkonstante, frequenzmodulierte Folge von Impulsen bieten dabei einen gewissen Schutz vor Störsignalen. Diese Fehlerunempfindlichkeit hat ihren Pendant in den digitalen Schaltungstechnik, die neben der zweiwertigen Logik bei einem hohen Pegelabstand (z.B.CMOS) auch Schwellwertelemente (Schmitt-Trigger) verwendet. In dem Modell von McCulloch und Pitts werden die Neuronenaktivität im Eingang und Ausgang zusätzlich auf zwei Zustände beschränkt. Kodieren wir diese mit -1 und +1, so ist in diesem Modell die Ausgabe zusätzlich zur gewichteten Summe der Eingabewerte auch von einem Schwellwert ϑ abhängig:

$$y = \mathrm{sign}\Big(\sum_j m_j x_j - \vartheta\Big)$$

Zweifelsohne ist die Umkodierung der Eingangsvektoren in binäre Form für die Funktion nicht wesentlich, erleichtert aber die Implementierung der Modelle auf sequentiellen, binär arbeitenden von-Neumann Computern ungemein.

Wir wollen uns deshalb im Folgenden davon ausgehen, daß alle Mustervektoren binär mit 0 und 1 kodiert sind, also große, posive sowie negative Komponentenwerte einem "Sättigungsverhalten" unterworfen sind und nur die Werte 'aktiv' und 'nicht-aktiv' unterschieden werden.
Damit ist

$$y_i' := \min(\max(y_i,0),1) \qquad (6.3c)$$

Das binäre Matrixmodell mit Schwellwert

Was bedeutet nun die Einführung einer Schwelle für unser Lineares Matrixmodell?
Die Gleichung (6.3a) ändert sich mit der Funktion

$$\vartheta(z) = \begin{cases} 1 & z \geq \vartheta \\ 0 & z < \vartheta \end{cases}$$

zu

$$y_i = \vartheta_i(z_i) \qquad (6.4a)$$

mit

$$z_i = \sum_k y_i^k \mathbf{x}\mathbf{x}^{kT} \qquad (6.4b)$$

Wie groß ist dabei die Schwelle ϑ_i?

Angenommen, wir geben eine leicht veränderte Version x des gespeicherten Vektors x^r ein. Gleichung 6.4b läßt sich dann schreiben als

$$z_i^r = y_i^r \mathbf{x}\mathbf{x}^{rT} + \sum_{k \neq r} y_i^k \mathbf{x}\mathbf{x}^{kT}$$

$$= y_i^r g_1 + g_2 \qquad (6.4c)$$

analog zu Gleichung 6.3b.
Um als Antwort y^r zu erhalten muß also gelten

$$z_i^r \begin{cases} \geq \vartheta_i & y_i^r = 1 \\ < \vartheta_i & y_i r = 0 \end{cases}$$

Die Schwelle ϑ ist also von den Indices i und r abhängig und erfüllt die Bedingung

$$g_2 < \vartheta_{ir} \leq g_1 + g_2 \qquad (6.4d)$$

Damit ist nun eine Möglichkeit gegeben, störendes Übersprechen herauszufiltern. Allerdings ist die Abhängigkeit der Schwelle von den Indices i und r sehr ungünstig. Suchen wir ein Schwellwert, der symmetrisch für alle r Muster und alle Komponenten i gilt, so ist die Relation 6.4d sicher auch für $x = x^r$ gültig. Normieren wir außerdem die Vektoren x^k, k=1..p, auf einen konstanten Wert a

$$x^k x^{kT} = a$$

so ist g_2=0, g_1=a und es gilt

$$\vartheta \leq a$$

Ist ϑ = 0, so resultiert das lineare Matrixmodell.
Lassen wir dagegen Verfälschungen unseres ursprünglichen Vektors zu, ohne daß deshalb andere Ausgabe- Komponenten überlagert werden sollen wie in Gleichung 6.3d, so ist dies bei einem Schwellwert größer als Null der Fall, wobei die Obergrenze der Schwelle durch den Wert a gegeben ist. Die optimale Schwelle ist also

$$\vartheta_{opt} = a$$

wobei nur für diejenigen Eingangsvektoren x'^r richtig auf x^r entschieden wird, für die die Relation

$$\max_i \sum_{k \neq r} y_i{}^k x'^r x^{kT} < a$$

erfüllt ist.

Die Annahme konstanter Aktivität

Die bisherigen Betrachtungen für einen Schwellwert, der musterabhängig ist, fußten auf der Annahme |x|=a, also konstanter Aktivität des Input-Musters.
Wann trifft diese Annahme zu?
Betrachten wir beispielsweise die visuelle Informationsverarbeitung beim Menschen. Ausgehend von der Retina beginnt bereits im Auge eine Informationsverarbeitung. Die Erzeugung der "visuellen Felder" geschieht dabei durch einen lokalen Mechanismus, bei dem jede Zelle ihre lokalen Nachbarn beeinflußt und beinflußt wird. Sind im Beinflussungsgebiet a Neuronen vorhanden, so erhält ebenfalls jedes Neuron a Aktivitätssignale, so daß unsere Annahme |x|=a bei diesem Beispiel plausibel erscheint. Andererseits scheint die Lokalität und damit auch die Topologie nicht nur für die Empfindungssensoren(Druck,Schmerz), sondern auch für die höheren Stufen der visuellen Verarbeitung weiterzubestehen, wie beispielsweise Marr und Poggio mit ihrem Modell des Stereo-Sehens zeigen konnten.

Der Hamming-Abstand zweier Vektoren v und w

Für die weiteren Untersuchungen ist es sinnvoll, das Vektorprodukt xx^{kT} etwas anders zu schreiben.
Betrachten wir dazu die beiden Vektoren **v** und **w** gleicher Dimension, deren Komponenten die Werte 0 oder 1 annehmen können.
Die Zahl der Stellen, an denen **v** und **w** in den gleichen Komponenten "1" enthalten, ist mit dem Produkt $\mathbf{vw}^T$ gegeben; die Zahl der Komponenten, die eine "1" enthalten und mit der Komponente gleichen Indexes des anderen Vektors **nicht** übereinstimmen, ist somit für **v**

$$|\mathbf{v}| - \mathbf{vw}$$

und analog auch bei **w**. Damit ist die Gesamtzahl aller nicht-übereinstimmenden Stellen, die "1" bei einem der beiden Vektoren enthalten (die Komponenten mit jeweils Null sind gleich) und somit der **Hamming-Abstand** $d(\mathbf{v},\mathbf{w})$

$$d(\mathbf{v},\mathbf{w}) = |\mathbf{v}| - \mathbf{vw}^T + |\mathbf{w}| - \mathbf{vw}^T = |\mathbf{v}|+|\mathbf{w}| -2\mathbf{vw}^T \qquad (6.4e)$$

und umgeformt

$$\mathbf{vw}^T = 1/2(|\mathbf{v}|+|\mathbf{w}| -d(\mathbf{v},\mathbf{w})) \qquad (6.4f)$$

Die Klassifizierungs-Operation

Seien die Basisvektoren $y^1,..,y^p$ linear unabhängig.
Es gilt

$$y^i y^j = 0$$

d.h. es gibt keine Komponente l, für die Basisvektoren y^i und y^j existieren mit

$$y_l^i = y_l^j = 1 \quad \text{bei } i \neq j$$

Damit ist Gleichung (6.4b)

$$z_i = 1/2\ (|x|+|x^k|-d(x,x^k)) \quad \text{für alle } y_i^k = 1$$

Mit $|x^k| = a$ ist somit

$$z_i = 1/2\ (|x| + a -d(x,x^k)) \qquad (6.4g)$$

für alle x^k mit $y_i^k = 1$.

Sei der Minimalabstand zweier Inputvektoren mit d notiert.

Es sei $|x|=a$, so ist mit 6.4g

$$z_i = a - d(x,x^k)/2 \quad \text{für alle } x^k \text{ mit } y_i^k = 1.$$

Behauptung:
Sei $|x|=a$ und $\vartheta = a-d/4$, so gilt folgendes bei 6.4b: Gibt es ein x^r mit $d(x,x^r)<d/2$, so wird x auf y^r abgebildet.
Gibt es kein solches x^r, so wird x auf den Nullvektor abgebildet.

Beweis:
Sei $d(x,x^r) < d/2$. Dann ist mit der Dreiecksungleichung

$$d(x^r,x) + d(x,x^k) \geq d(x^r,x^k)$$

für jedes k # r

$$d(x,x^k) \geq d(x^r,x^k) - d(x^r,x) > d - d/2 = d/2$$

Damit gilt

$$z_i = \begin{cases} a - d(x,x^k)/2 < a - d/4 = \vartheta & \text{bei } k \# r \\ a - d(x,x^r)/2 > a - d/4 = \vartheta & k=r \end{cases} \quad (6.4h)$$

und somit

$$\vartheta(z_i) = \begin{cases} 0 & k \# r, \text{ wobei } y_i^r = 0 \\ 1 & k=r, \text{ wobei } y_i^r = 1 \end{cases}$$

Da die Ausgangsvektorten **y** linear unabhängig sind, gibt es also für jede Komponente y_i von der gesamten Summe in (6.4b) nur maximal einen Summanden, der ungleich null ist. Ist er von einem Vektor k#r, so ist der Summand kleiner als die Schwelle und die entsprechende Komponente ist Null. Ist der Summand dagegen vom Vektor r, so ist der Summand größer als die Schwelle wenn $y_i^r = 1$ und es resultiert der erste Teil der Behauptung. Existiert kein Eingabevektor $\mathbf{x}^r$ mit der genannten Eigenschaft, so ist auch kein Summand größer als die Schwelle uns es resultiert der Nullvektor in der Ausgabe, wodurch auch der zweite Teil der Behauptung bewiesen ist.

Dieses Ergebnis entspricht der Feststellung in der Kodierungstheorie, daß ein Codewort bei Blockcodes genau dann korrekt erkannt wird, wenn der Abstand vom korrekten Codewort kleiner ist als der halbe minimale Abstand zweier Codeworte.
Zur Demonstration sei ein Beispiel gerechnet.

Beispiel 6.4:

Sei
$$\mathbf{x}_2^1 = (111\ 000\ 000) = \mathbf{y}_2^1$$
$$\mathbf{x}_3^2 = (000\ 111\ 000) = \mathbf{y}_3^2$$
$$\mathbf{x}^3 = (000\ 000\ 111) = \mathbf{y}^3$$

Dann ist

$$\mathbf{M} = \begin{matrix} 1&1&1&0&0&0&0&0&0 \\ 1&1&1&0&0&0&0&0&0 \\ 1&1&1&0&0&0&0&0&0 \\ 0&0&0&1&1&1&0&0&0 \\ 0&0&0&1&1&1&0&0&0 \\ 0&0&0&1&1&1&0&0&0 \\ 0&0&0&0&0&0&1&1&1 \\ 0&0&0&0&0&0&1&1&1 \\ 0&0&0&0&0&0&1&1&1 \end{matrix}$$

$p=3,\ a=3,\ d(\mathbf{x}^i,\mathbf{x}^j)=6,\ \vartheta=3-6/4=1.5$

Sei $\mathbf{x}$ = (110 100 000), so ist im linearen Fall $\mathbf{y}$ = (222 111 000) als Überlagerung von $\mathbf{y}^1$ und $\mathbf{y}^2$. Wird darauf die Schwellwertoperation mit $\vartheta = 1.5$ angewendet, so ist

$\mathbf{y}$ = (111 000 000) = $\mathbf{y}^1$.

Die nichtlineare Operation ist damit deutlich fehlertoleranter als das lineare Modell.

Damit zeigt das Matrix-Modell mit Schwellwert eine interessante Eigenschaft: Es führt eine Klassifizierung der Mustervektoren durch, die angelegt werden. Ist das Input-Muster nicht zu stark abweichend vom gelernten Prototypen, so wird auf die entsprechende Repräsentation des Prototypen entschieden; die Information über die Abweichung geht dabei verloren. Damit ist es möglich, durch stufenweises Hintereinanderschalten der Klassifizierungsoperation eine Informationsreduzierung auf die gewünschte, kontextabhängige Information zu erreichen.
Wie funktioniert aber nun die Ergänzungsoperation?

Die Ergänzung lückenhafter Daten

Betrachten wir nun den Fall, daß $\mathbf{x}$ eine Version eines $\mathbf{x}^r$ ist, bei der einige '1' fehlen. Diese lückenhaften Daten lassen sich beispielsweise bei Ergänzungsoperationen der Autoassoziativen Matrizen vorteilhaft einsetzen. Zweifelsohne hat dieser, von a verschiedene Erwartungswert Konsequenzen für die optimale Schwelle.
Sei $\mathbf{x}$ eine Version eines $\mathbf{x}^r$, bei dem verschiedene, aber nicht alle der binären Komponenten null gesetzt sind, so gilt mit $\mathbf{x}\mathbf{x}^{rT} =: a'$ die Relation

$$0 \leq a' \leq a$$

und es ist in Gleichung 6.4c der Wert $g_1 = a'$. Sind nun die Vektoren $\mathbf{x}^k$, $k=1..p$, linear unabhängig und normiert, so gilt für jeden Vektor $\mathbf{x}^r$, $r \# k$,

$$\mathbf{x}^r\mathbf{x}^{kT} = \sum_i x_i^r x_i^k = 0$$

Bei binären Komponenten ist jeder Summand gleich Null, so daß auch für $\mathbf{x}$ als Teilvektor von $\mathbf{x}^r$ gilt

$$\mathbf{x}\mathbf{x}^{kT} = 0$$

Sei t^- die maximale Zahl von Komponenten von $\mathbf{x}^r$, die von 1 auf 0 gesetzt wurden. Es ist

$$t^- = a - \min(a')$$

Folgen wir wieder der Argumentation für die Bestimmung der optimalen Schwelle für 6.4d, so ist bei lückenhaften Daten die optimale Schwelle

$$\vartheta_{opt} = a - t^-$$

Auch bei lückenhaften Daten wird somit eine Klassifizierung auf den richtigen, gewünschten Prototypen durchgeführt; ist der Inputvektor mit dem Outputvektor identisch, (gleiche Kodierung),

so wird der lückenhafte Input bei der Klassifizierungsoperation zum vollständigen Prototypen ergänzt.

Literatur

Folgende Lehrbücher wurden benutzt:

F.Fallside, W.Woods
Computer speech processing
Prentice-Hall 1985

K.Fellbaum
Sprachverarbeitung und Sprachübertragung
Springer Verlag 1984

Rabiner, Schafer
Digital Processing of Speech Signals
Prentice-Hall 1978

R.DeMori, C.Suen
New Systems and Architektures for
Automatic Speech Recognition and Synthesis
Springer Verlag 1984

und weitere Bücher und Zeitschriften, die aus Platzmangel nicht weiter aufgeführt werden.