Ein- und Ausgabe von Sprache

Markus Kommenda
TU Wien

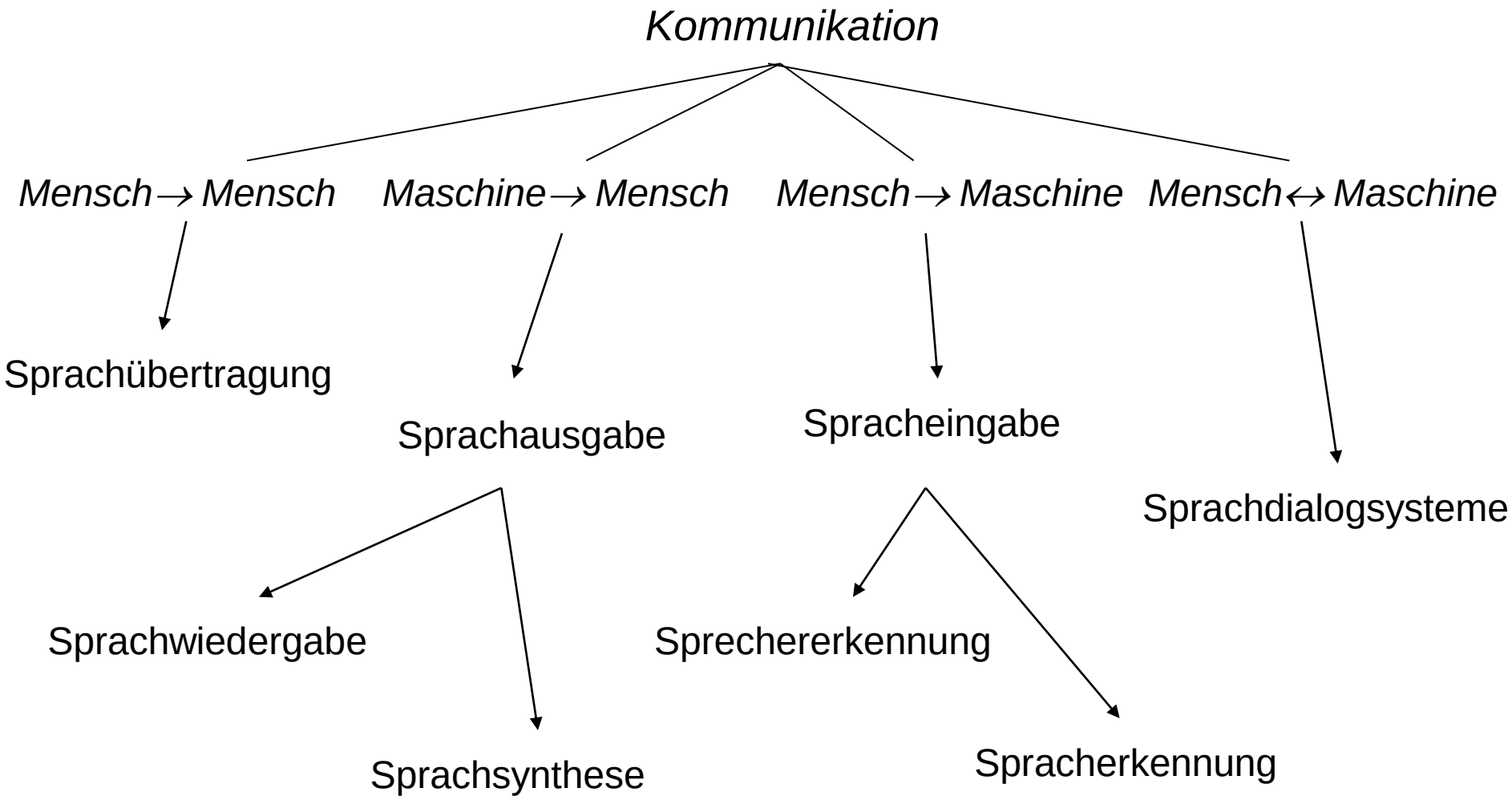
2 Std. VO, LVA-Nr. 185.317
Wintersemester 2010/2011

Inhalt

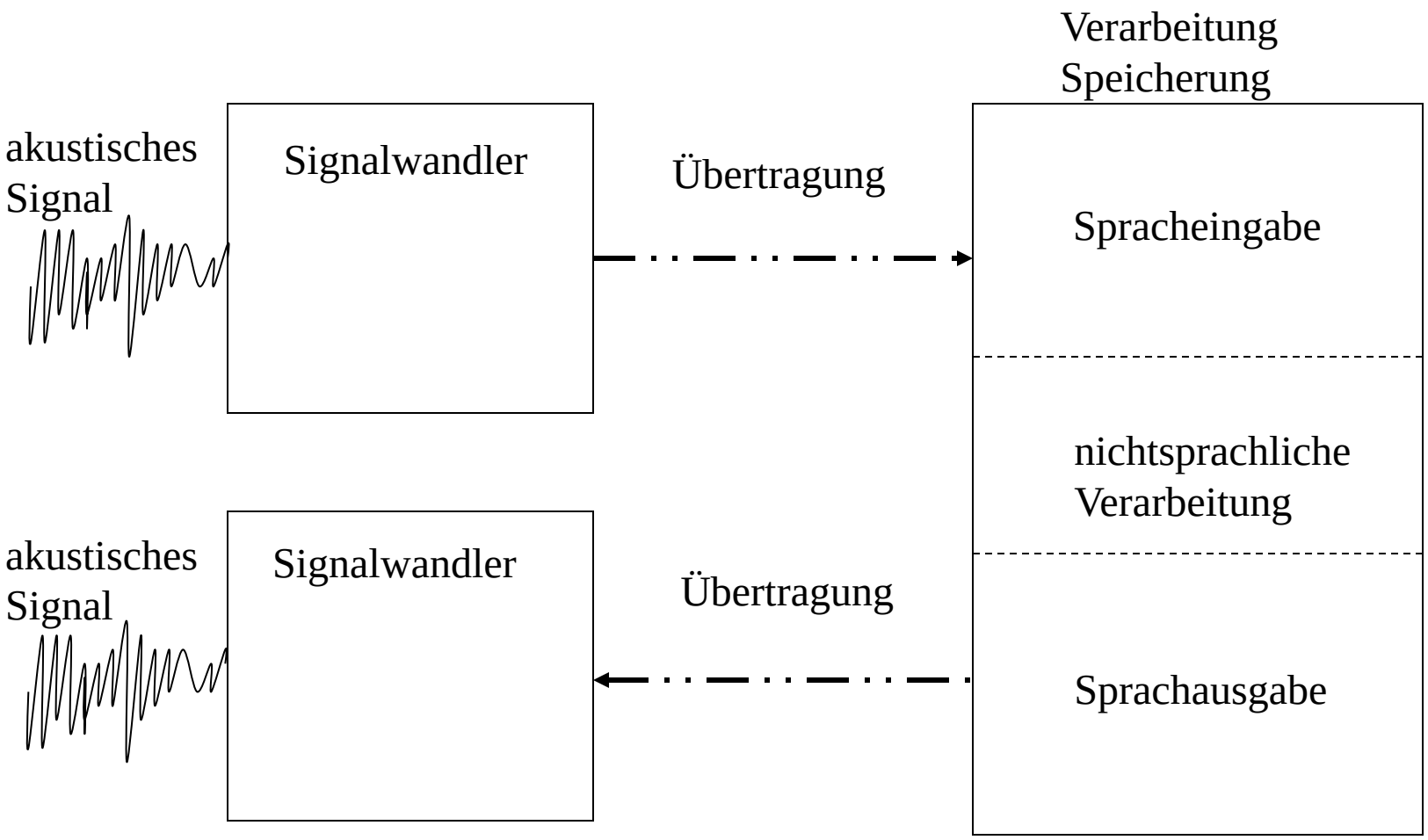
1. Einführung
2. Sprachwissenschaftliche Grundbegriffe
3. Grundlagen der Sprachsignalverarbeitung
4. Sprachausgabe
5. Spracheingabe
6. Sprachdialogsysteme

Inhalt

1. Einführung
Aufgabenstellung der digitalen Sprachverarbeitung

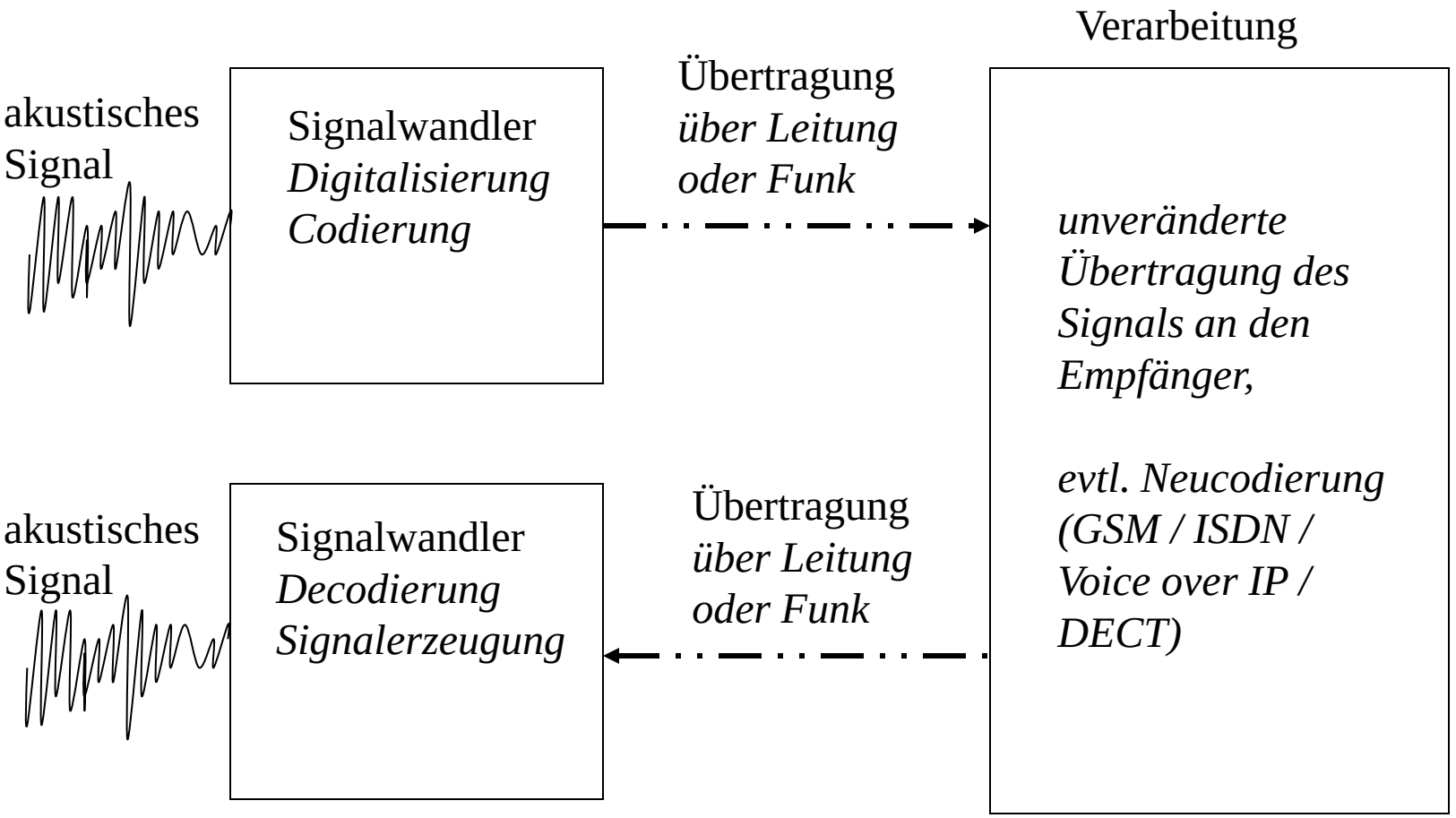


1. Einführung
Aufgabenstellung der digitalen Sprachverarbeitung



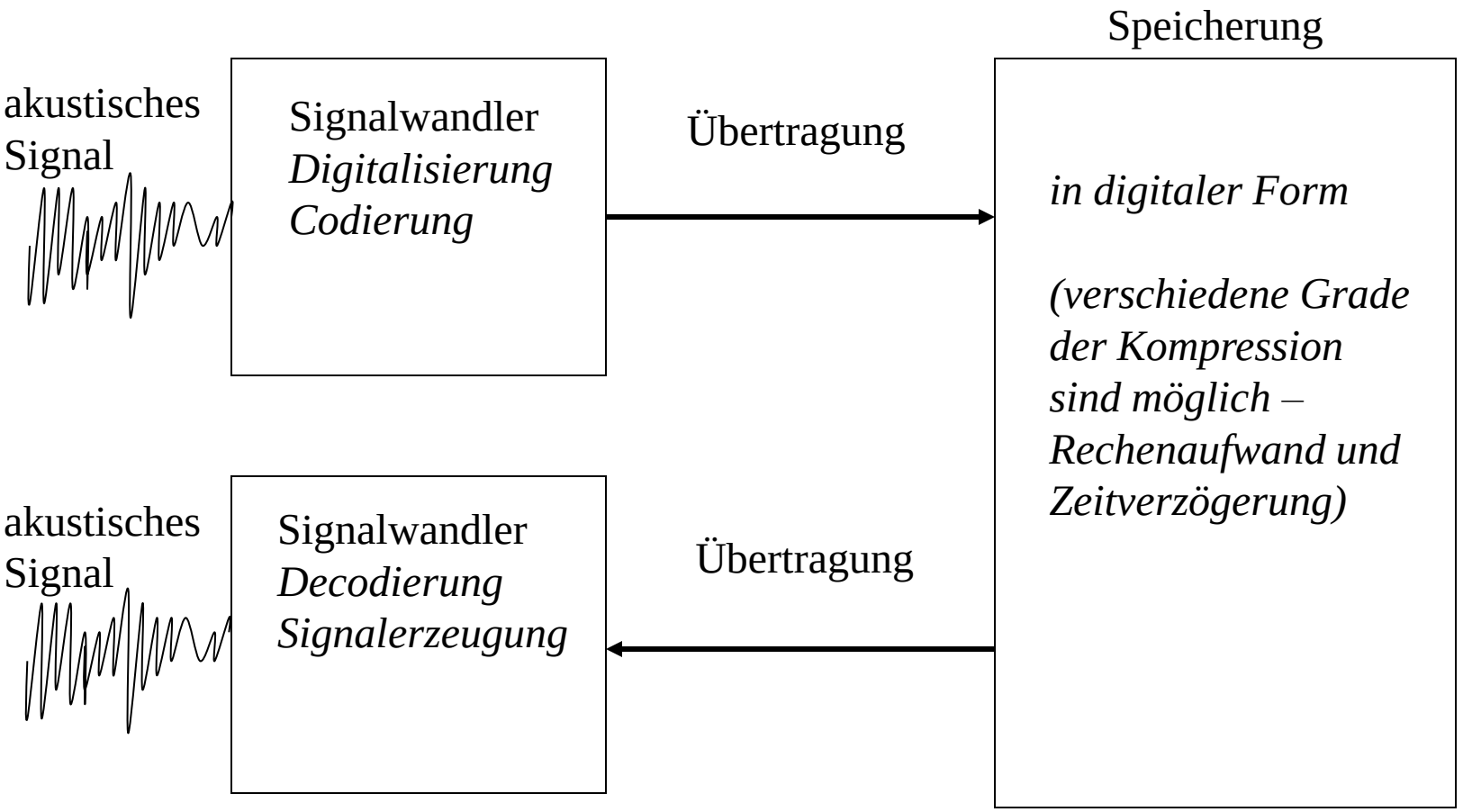
1. Einführung
Anwendungsbeispiel: Telephonie

- Kommunikation von Mensch zu Mensch (im Allgemeinen)



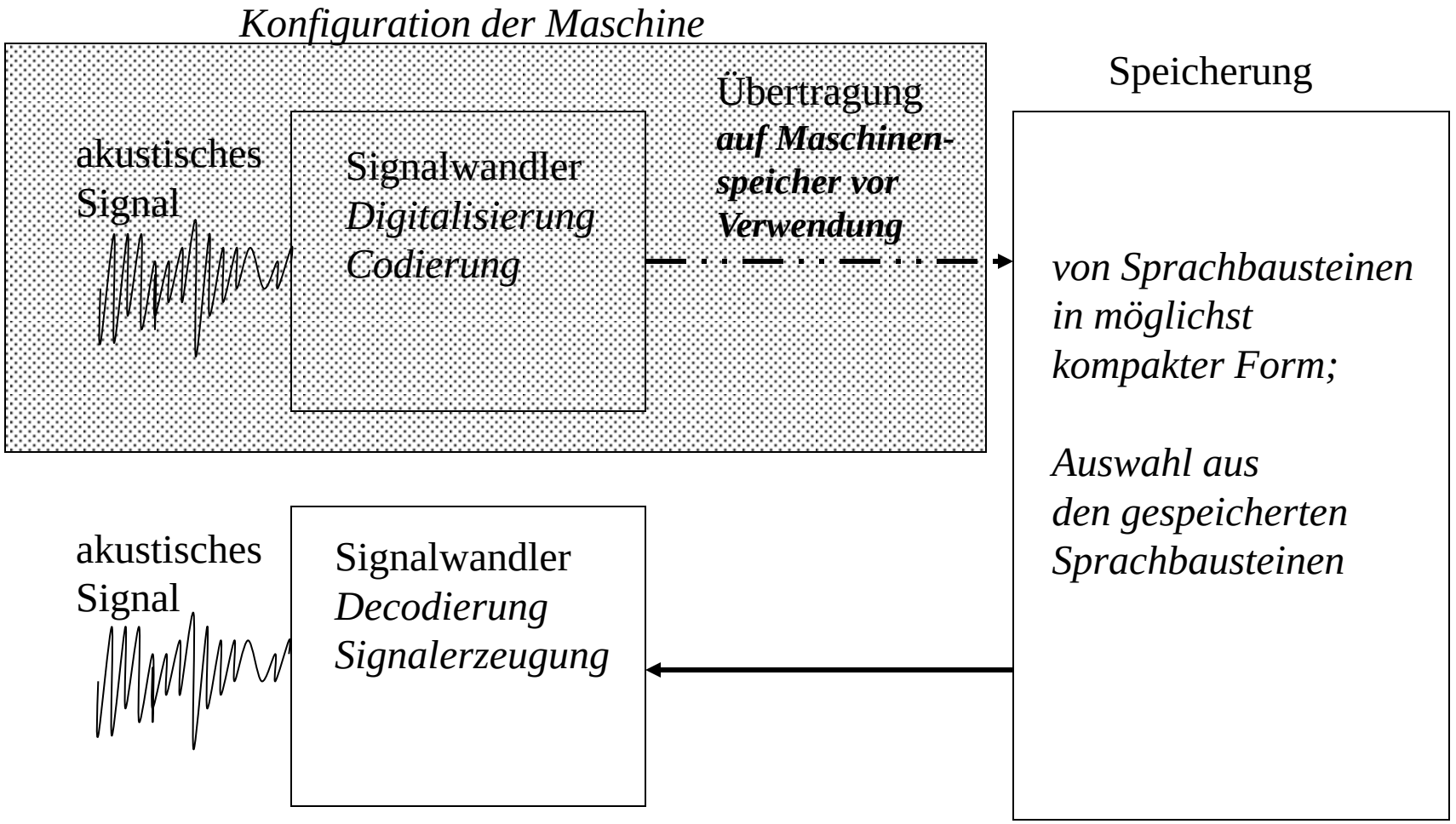
1. Einführung
Anwendungsbeispiel: digitales Diktiergerät / Anrufbeantworter

- Kommunikation von Mensch zu Maschine zu Mensch



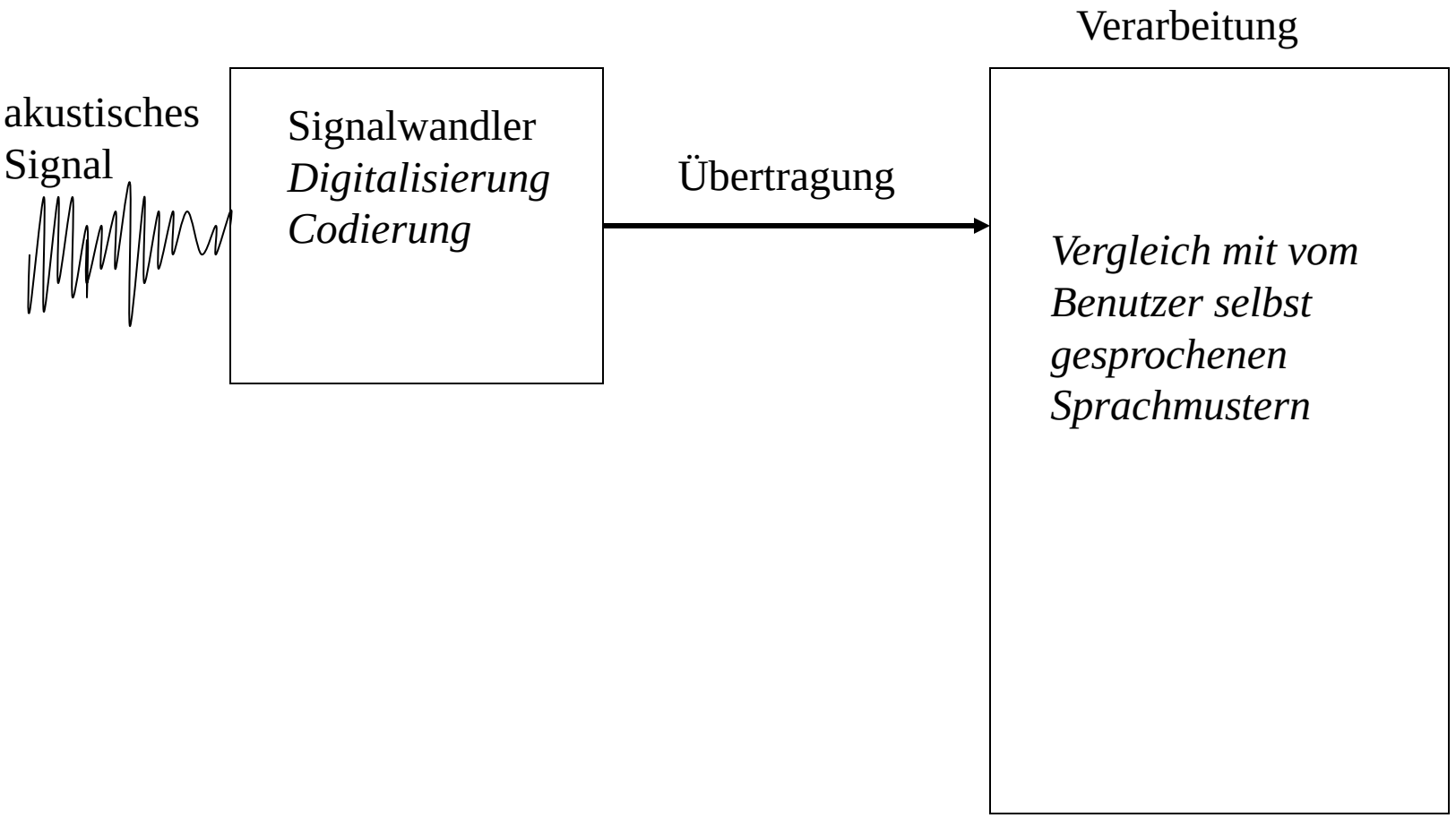
1. Einführung
Anwendungsbeispiel: Sprechende Geräte

- Kommunikation von Maschine (Uhr, Navigationssystem) zu Mensch



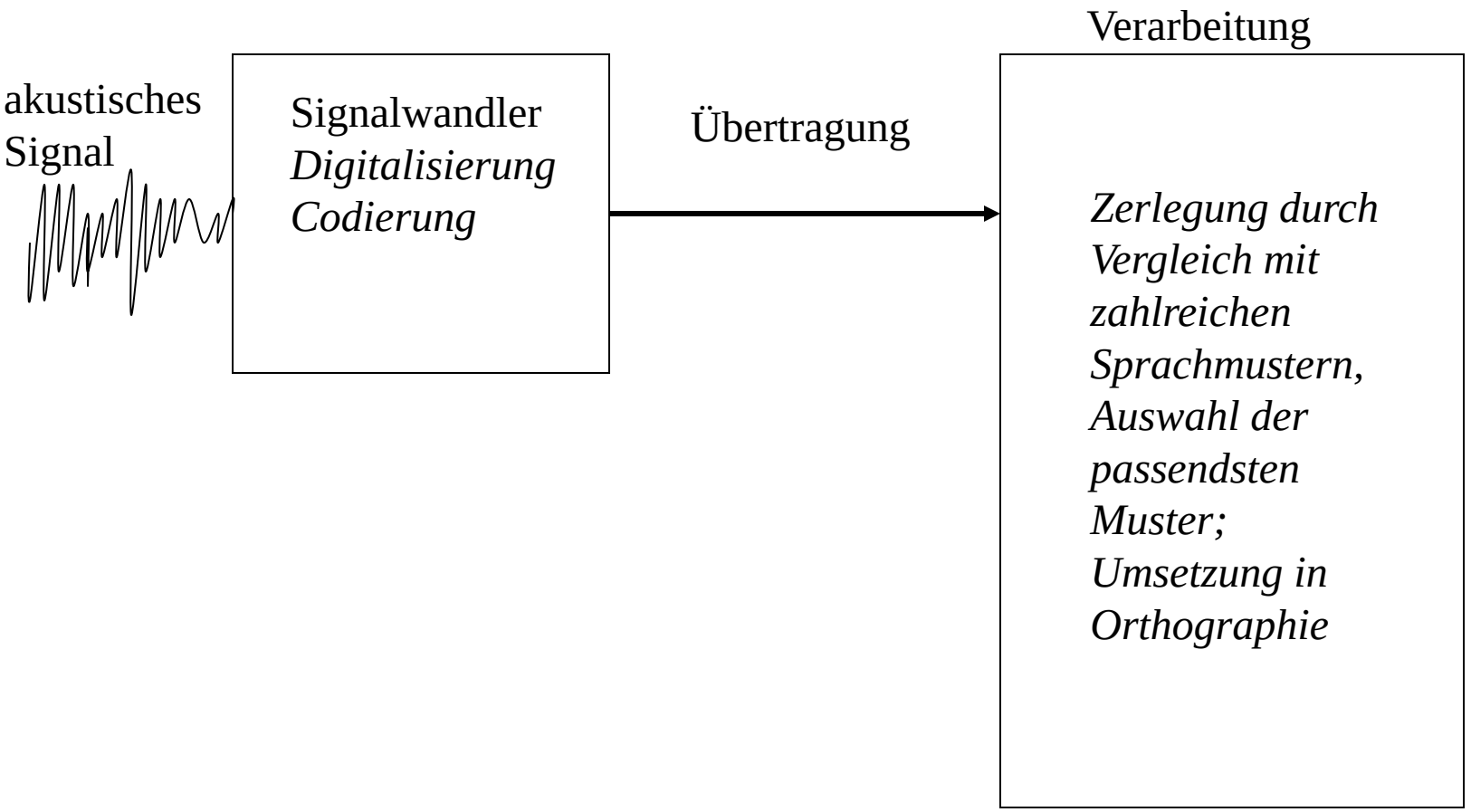
1. Einführung
Anwendungsbeispiel: Sprachwahl im Handy

- Kommunikation von Mensch zu Maschine



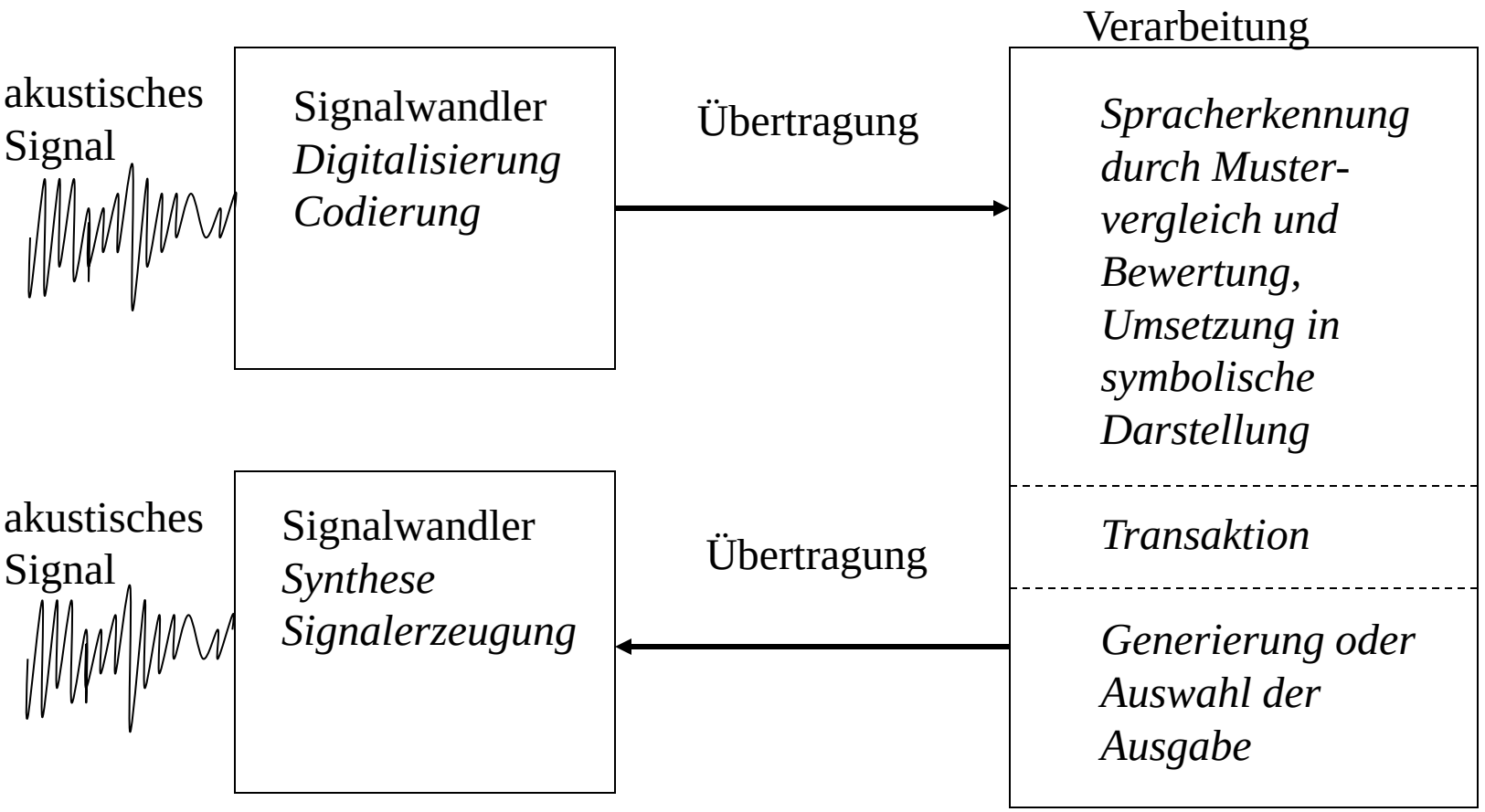
1. Einführung
Anwendungsbeispiel: Diktiersystem

- Kommunikation von Mensch zu Maschine
- Natürliche Eingabe, Befreiung der Hände



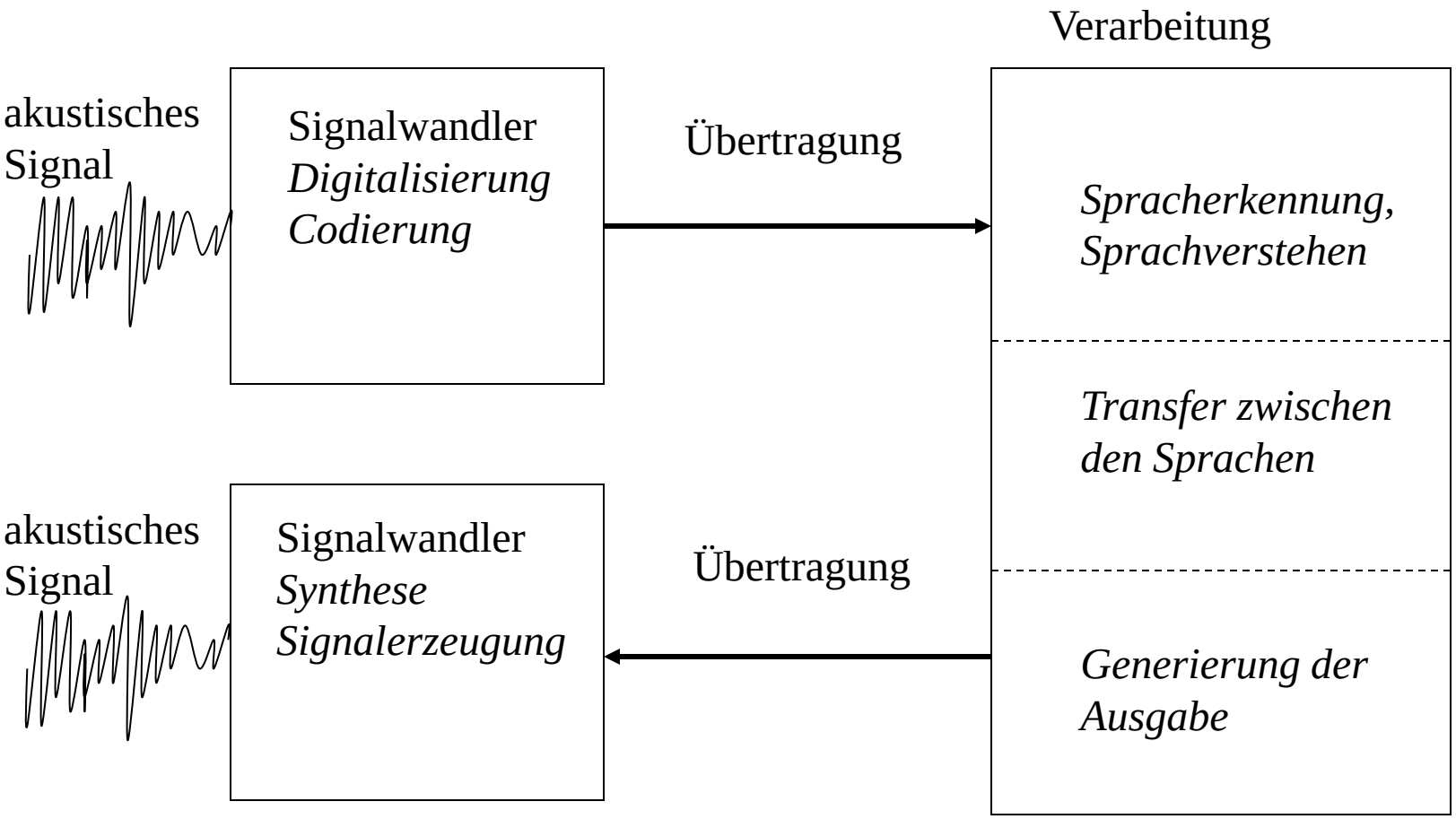
1. Einführung
Anwendungsbeispiel: Dialogsystem

- Kommunikation von Mensch zu Maschine zu Mensch
- Kostensenkung, Erhöhung der Verfügbarkeit von Telefondiensten



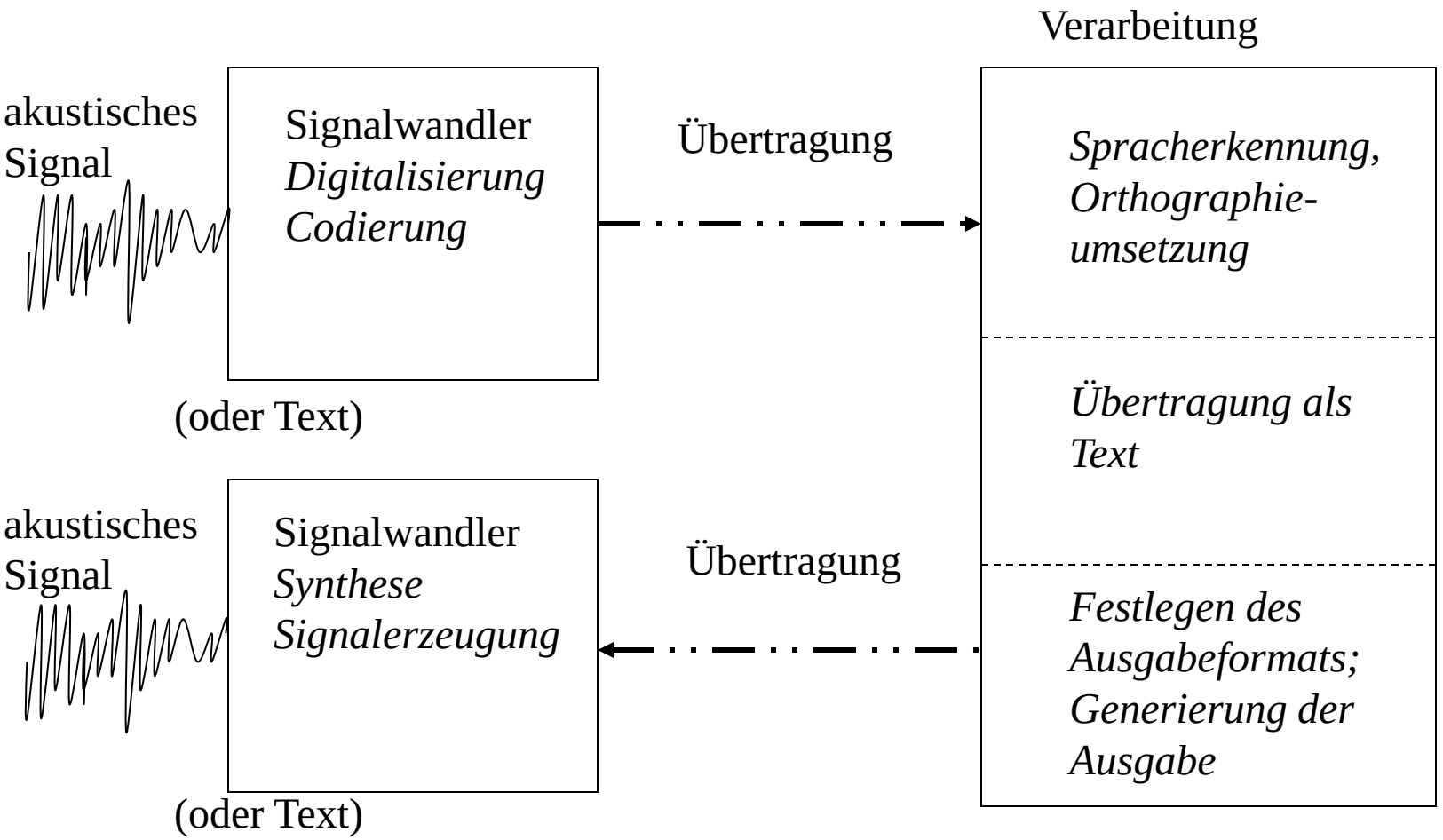
1. Einführung
Anwendungsbeispiel: Sprachübersetzung

- Kommunikation von Mensch zu Mensch
- Überwindung von Sprachbarrieren



1. Einführung
Anwendungsbeispiel: Unified Instant Messaging

- Kommunikation von Mensch zu Mensch, in der Infrastruktur als Text
- Beliebige Wahl Text/Gesprochene Sprache an beiden Endstellen



Inhalt

2. Sprachwissenschaftliche Grundbegriffe

- Sprachproduktion beim Menschen
- Menschliche Sprechwerkzeuge
- Artikulation verschiedener Laute
- Phonetische Beschreibung u. phonologische Klassifikation
- Geschriebene Sprache
- Wortbau
- Silben
- Satzbau
- Prosodie

Sprachproduktion beim Menschen

- Atmung (Respiration)

  Atmungsorgane:

  Bauch- und Brustmuskulatur, Zwerchfell
  Lunge
  Luftröhre

- Stimmgebung (Phonation)

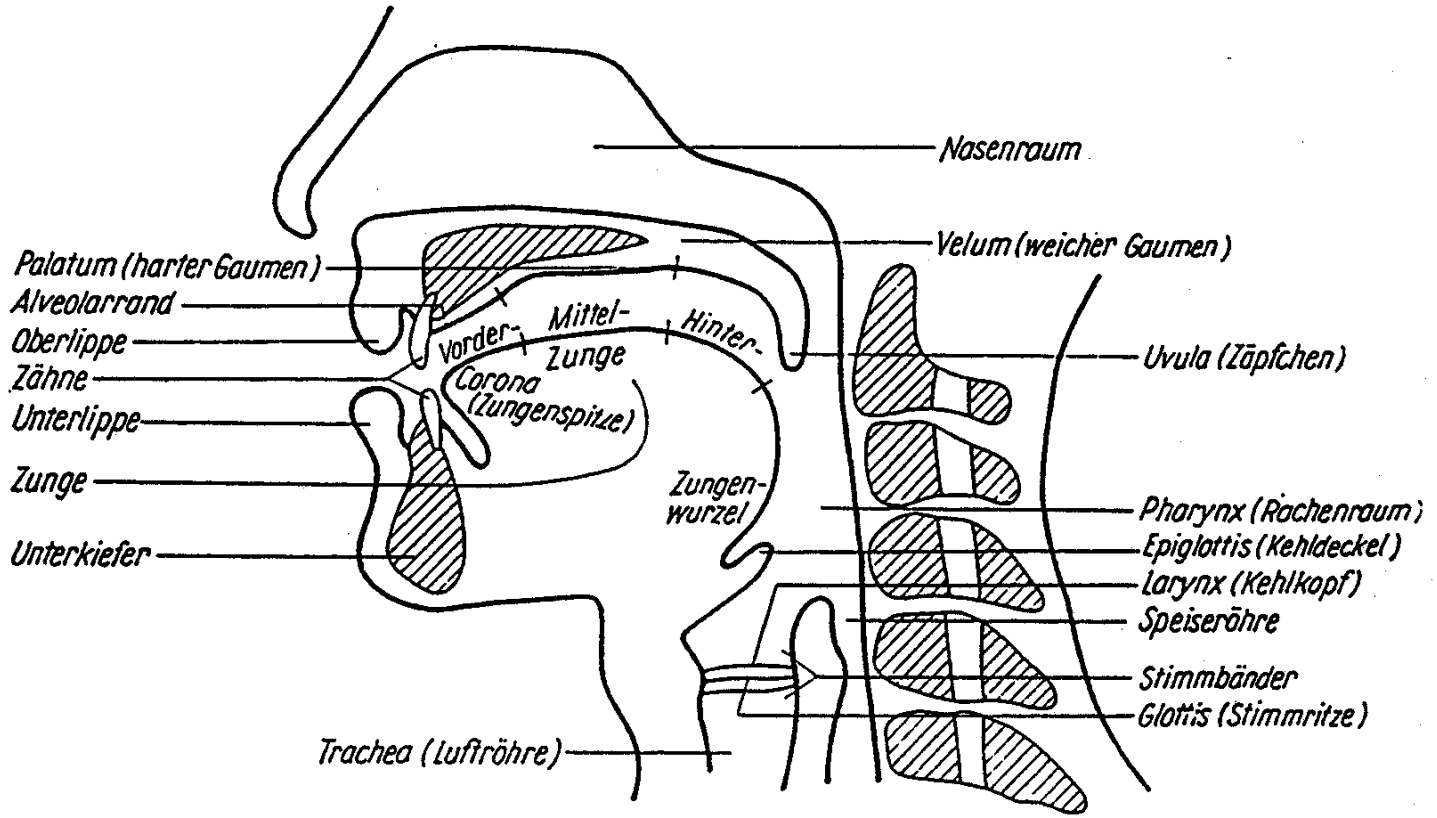
  Kehlkopf mit Stimmlippen („Stimmbänder“)

- Lautbildung (Artikulation)

  „Ansatzrohr“: Rachen-, Nasen- und Mundraum
  Artikulationsorgane (Sprechwerkzeuge)

  aktive: Zunge, Lippen, Gaumensegel, Zäpfchen
  passive: Zähne, Zahndamm, Gaumen

2. Sprachwissenschaftliche Grundbegriffe
Menschliche Sprechwerkzeuge



Artikulation verschiedener Laute

a) Vokale
   - Stimmbänder schwingen
   - Mundraum relativ weit geöffnet, Luftstrom kann ungehindert zirkulieren

b) Konsonanten
   - stimmhaft oder stimmlos
   - Luftstrom im Mundraum behindert oder (zeitweise) abgesperrt

Vokale im Deutschen (1)

Phonetische Umschrift im phonetischen Alphabet der IPA
(International Phonetic Association)

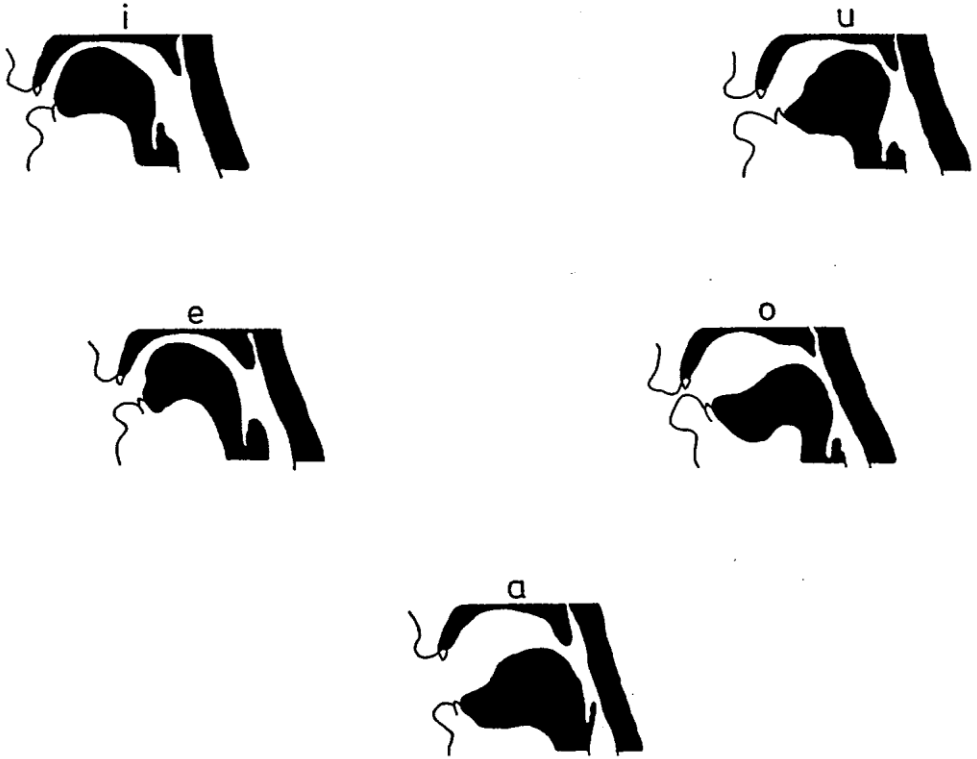
Länge wird i.a. durch Punkte hinter dem Umschriftzeichen ausgedrückt.

[a]	kurzer, vorderer (heller) a-Laut: h**a**tte, M**a**nn
[ɑː]	langer, hinterer (dunkler) a-Laut: **A**bend, t**a**t
[ɛ]	kurzer, offener e-Laut: **E**bbe, B**e**tt
[ɛː]	langer, offener e-Laut: **ä**sen, K**ä**se
[eː]	langer, geschlossener e-Laut: **e**ben, l**e**sen, S**ee**
[ə]	kurzer, unbetonter (schwacher) e-Laut: hatt**e**
[ɪ]	kurzer, offener i-Laut: **i**mmer, b**i**tte
[iː]	langer, geschlossener i-Laut: **I**da, M**ie**te, n**ie**
[ɔ]	kurzer, offener o-Laut: **o**ffen, T**o**chter

Vokale im Deutschen (2)

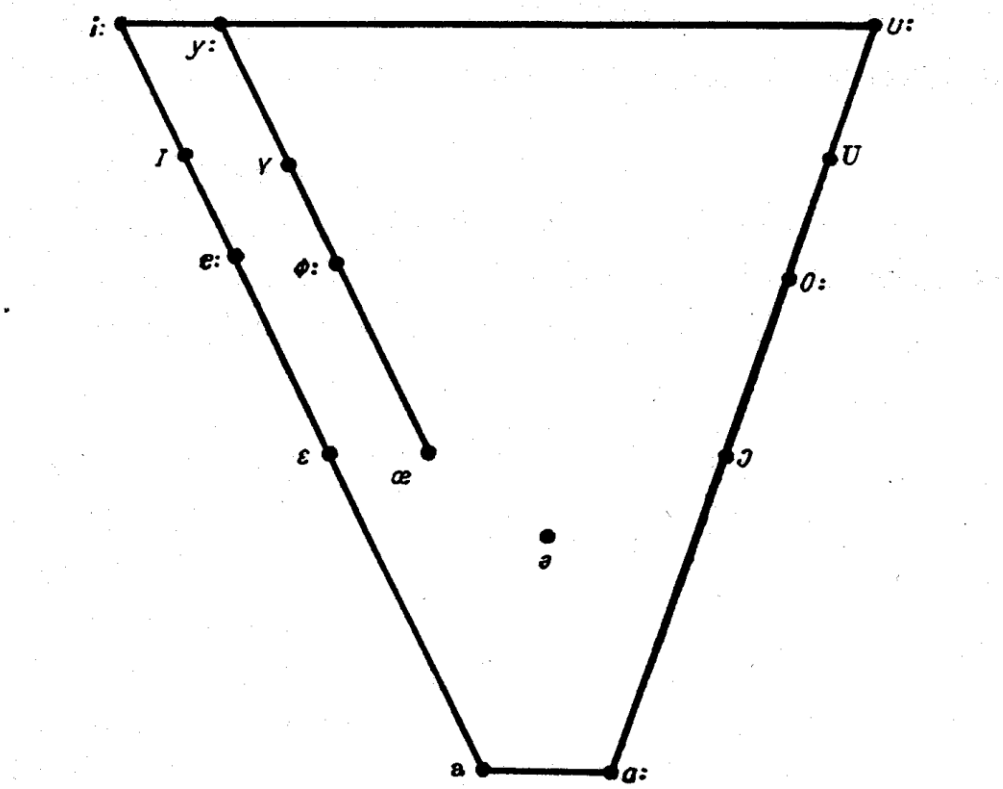
[oː]	langer, geschlossener o-Laut: **O**fen, S**o**hn
[U]	kurzer, offener u-Laut: **U**lm, M**u**tter
[uː]	langer, geschlossener u-Laut: **U**hr, M**u**t, d**u**
[œ]	kurzer, offener o-Umlaut: H**ö**lle, G**ö**tter
[øː]	langer, geschlossener o-Umlaut: H**ö**hle, G**oe**the
[ʏ]	kurzer, offener u-Umlaut: **ü**ppig, H**ü**tte
[yː]	langer, geschlossener u-Umlaut: **Ü**bel, H**ü**te
[ai]	Diphthong: **ei**ne, n**ei**n, Bäcker**ei**
[au]	Diphthong: **au**s, L**au**t
[ɔy]	Diphthong: **eu**ch, Fr**äu**lein, n**eu**

2. Sprachwissenschaftliche Grundbegriffe
Vokale: Zungenstellung, Resonanzraum



Quelle: K. Fellbaum, Sprachverarbeitung und Sprachübertragung, Springer Verlag Berlin. Heidelberg, New York, 1984

2. Sprachwissenschaftliche Grundbegriffe
Vokalviereck



Konsonanten im Deutschen (1)

[b]	stimmhafter, bilabialer Verschlusslaut: **b**itte, E**bb**e
[p]	stimmloser, bilabialer Verschlusslaut : **P**eter, Ma**pp**e, kna**pp**
[m]	stimmhafter, bilabialer Nasal: **M**utter, A**mm**e, La**mm**
[d]	stimmhafter, alveolarer Verschlusslaut: **D**ieb, I**d**a
[t]	stimmloser, alveolarer Verschlusslaut: **T**ag, Wa**tt**e, Wel**t**
[n]	stimmhafter, alveolarer Nasal: **N**ame, Ka**nn**e, Ma**nn**
[g]	stimmhafter, velarer Verschlusslaut: **G**arten, le**g**en
[k]	stimmloser, velarer Verschlusslaut: **K**arl, we**ck**en, Sa**ck**
[ŋ]	stimmhafter, velarer Nasal: A**ng**el, e**ng**
[v]	stimmhafter, labio-dentaler Reibelaut: **W**iege, e**w**ig
[f]	stimmloser, labio-dentaler Reibelaut: **F**inger, **V**ater, A**ff**e, Ha**ff**

2. Sprachwissenschaftliche Grundbegriffe
Konsonanten im Deutschen (2)

[z]	stimmhafter, alveolarer Reibelaut: **s**ingen, lang**s**am
[s]	stimmloser, alveolarer Reibelaut: wa**s**, e**ss**en, Ha**ss**
[ʃ]	stimmloser, präpalataler Reibelaut: **Sch**üler, E**sch**e, ra**sch**
[j]	stimmhafter, palataler Reibelaut : **j**a, A**j**ax
[ç]	stimmloser, palataler Reibelaut : **Ch**ina, i**ch**
[x]	stimmloser, velarer Verschlusslaut: su**ch**en, a**ch**
[l]	stimmhafter, alveolarer Lateralengelaut: **L**aut, a**ll**e, Nu**ll**
[r]	stimmh., alveolarer intermittierender Laut (Zungenspitzen-r): **r**ot, Eh**r**e, He**rr**
[ʀ]	stimmhafter, uvularer intermittierender Laut (Zäpfchen-r): **r**ot, Eh**r**e, He**rr**
[h]	stimmloser glottaler Reibelaut: **h**alt, A**h**orn
[ʔ]	glottaler Verschlusslaut: er-achten, ver-eisen

2. Sprachwissenschaftliche Grundbegriffe
SAMPA: Ein maschinenlesbares phonetisches Alphabet

Symbol	Wort	Transkription
S	waschen	"vaS=n
Z	Genie	Ze"ni:
C	sicher	"zIC6
x	Buch	bu:x
N	Ding	dIN
I	Sitz	zIts
E	Gesetz	g@"zEts
O	Trotz	trOts
U	Schutz	SUts
Y	hübsch	hYpS
9	plötzlich	"pl9tslIC
@	bitte	"bIt@
y:	süß	zy:s
2:	blöd	bl2:t
aI	Eis	aIs
aU	Haus	haUs
OY	Kreuz	krOYts

IPA-Symbole, die in ASCII nicht vorkommen, werden durch ASCII-Symbole ersetzt.

Konsonanten: Unterscheidungskriterien

- Stimmhaftigkeit: stimmhaft – stimmlos
  z.T. gekoppelt mit Stärke (fortis – lenis)

- Artikulationsart
  - Enge, Reibung (Engelaute, Reibelaute, Frikative)
  - Verschluss, Sprengung (Verschlusslaute, Plosive)
  - Nasenöffnung (Nasenlaute, Nasale)

- Artikulationsort
  - Lippen (labial)
  - Zähne (dental)
  - Zahndamm (alveolar)
  - Harter Gaumen (palatal)
  - weicher Gaumen (velar)
  - Zäpfchen (uvular)
  - Stimmritze (glottal)

Konsonanten: Unterscheidungskriterien

Artikulationsort

Artikulationsart		bi-labial	labio-dental	dental	alveo-lar	palatal	velar	uvular	glottal
Ver-schluss-laute	sth.	b			d		g		ʔ
	stl.	p			t		k		
Reibe-laute	sth.		v	z		j			
	stl.		f	s	ʃ	ç	x		h
Nasale		m			n		ŋ		
Laterale					l				
Inter-mittierende					r			R	

Phonetik und Phonologie

Zwei grundsätzlich verschiedene Betrachtungsweisen

Phonetik

untersucht und beschreibt Laute, genauer Phone,
unter dem Gesichtspunkt ihrer

- physiologischen (artikulatorischen) und
- physikalischen (akustischen) Eigenschaften

Phonologie

- abstrahiert von der konkreten (artikulatorischen, akustischen) Realisierung der Laute
- erforscht sie hinsichtlich ihrer Funktion im sprachlichen System
- klassifiziert Phone zu Phonemen
  - Kriterium: bedeutungsunterscheidende („distinktive“) Funktion

2. Sprachwissenschaftliche Grundbegriffe
Phon und Phonem

- Unterschiedliche Realisierungen des R-Lauts
  (Zungenspitzen-R, Zäpfchen-R, Reibe-R, vokalisches R, ...):
  im Deutschen kein Bedeutungsunterschied
  ➔ verschiedene **Phone** (üblicherweise in [ ] gesetzt)

- Unterschiedliche Aussprache von E:
  - langes, geschlossenes [eː] in *Beet*
  - kurzes, offenes [ɛ] in *Bett*
  unterschieden (bei gleicher lautlicher Umgebung)
  zwischen Wörtern verschiedener Bedeutung
  ➔ verschiedene **Phoneme** (üblicherweise in / / gesetzt)

2. Sprachwissenschaftliche Grundbegriffe
Was bedeutet "Sprache als System"

- Ferdinand de Saussure (frz. Gründerfigur der modernen Sprachwissenschaft und des Strukturalismus in den Geisteswissenschaften):

  Sprache ist ein System von Unterscheidungen
  ... auf einer Reihe von hierarchischen Repräsentationsstufen

- Noam Chomsky (amerikan. bedeutendster Linguist der zweiten Hälfte des 20. Jhdts.):

  Sprache ist ein hochspezialisiertes, komputationales System des menschlichen Geistes *(mind)*
  ... Ausgangspunkt für ein besseres Verständnis des Prozesscharakters sprachlicher Leistungen

Graph und Graphem

Buchstaben in geschriebener Sprache:

- Graph:
  konkrete Erscheinungsform eines Buchstaben,
  z.B.: a, *a*, **a**, `a`

- Graphem:
  Klasse graphischer Varianten (üblicherweise in < > gesetzt)

Morph und Morphem

- Morph

  konkrete lautliche oder schriftliche Erscheinungsform eines Morphems
  z.B. [bilt] (*Bild*) und [bild] (wie in *Bilder*)

- Morphem

  kleinste bedeutungstragende Einheit der Sprache;
  je nach Verwendungsart werden unterschieden:

  - freie Morpheme (Stämme): z.B. *lieb*, *Fest*
  - gebundene Morpheme: Präfixe, z.B. *ver-*, *prä-*
    Suffixe, z.B. *-lich*, *-keit*
    u.a.

Silbenstruktur

Silben bestehen aus bis zu drei Teilen:

1. Anlaut (onset): ein oder mehrere Konsonanten
2. Inlaut (nucleus): ein Vokal oder Diphthong
3. Auslaut (coda): ein oder mehrere Konsonanten

nur Inlaut: *Ei*, *oh*, *eh*, *au* (englisch: *a*, *I*; französisch: *Eau*, *on*)

Anlaut + Inlaut: *bei*, *frei*, *Schrei*, *Spreu*

Inlaut + Auslaut: *Eis*, *eilt*, *eilst*

Anlaut + Inlaut + Auslaut: *rot*, *Trost*, *sprangst*

Der Inlaut muss in jeder Silbe vorhanden sein.

Silbengrenzen stimmen nicht mit Morphgrenzen überein.

Morphologie

a) Gestalt, Aufbau eines Wortes aus Morphen/Morphemen
b) Teilgebiet der Sprachwissenschaft, das den Aufbau von Wörtern aus Morphemen untersucht und beschreibt

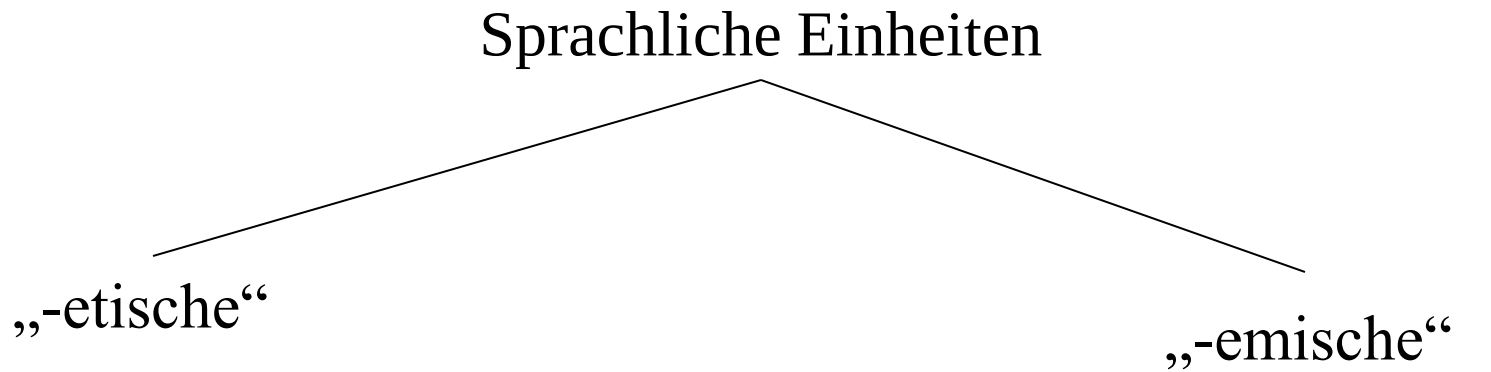
Syntax

a) Aufbau von Wortgruppen, insbesondere von Sätzen
b) Teilgebiet der Sprachwissenschaft, das den Aufbau von Sätzen und Wortgruppen untersucht und beschreibt

Semantik

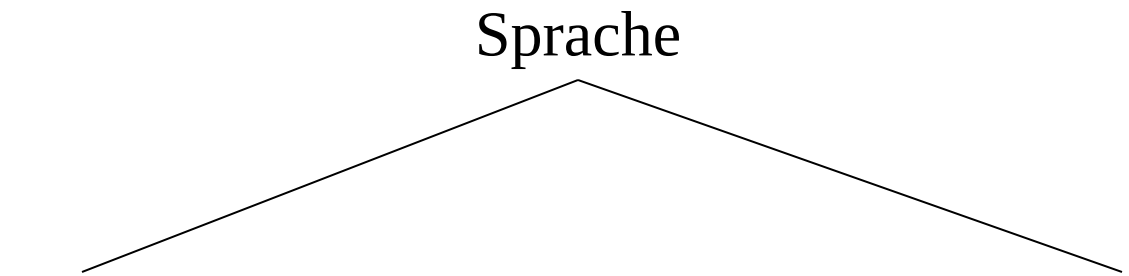
a) Bedeutung eines Morphems, Wortes oder einer Wortgruppe
b) Teilgebiet der Sprachwissenschaft, die Bedeutung von Morphemen und die Zusammensetzung der Bedeutung größerer Einheiten (Wortgruppen, Sätze), untersucht und beschreibt.

2. Sprachwissenschaftliche Grundbegriffe
Abstrakte Klassen und konkrete Realisierungen



„-etische“	„-emische“
Phon, Graph, Morph ...	Phonem, Graphem, Morphem ...
konkret	abstrakt
Varianten	Klassen
Realisierung (in einer Äußerung)	Funktion (im Sprachsystem)

2. Sprachwissenschaftliche Grundbegriffe
Langue / Parole



und deren Verarbeitung auf

Signalebene

Symbolebene

Zergliederung (Analyse) von Sätzen und Wortgruppen in kleinere Einheiten.

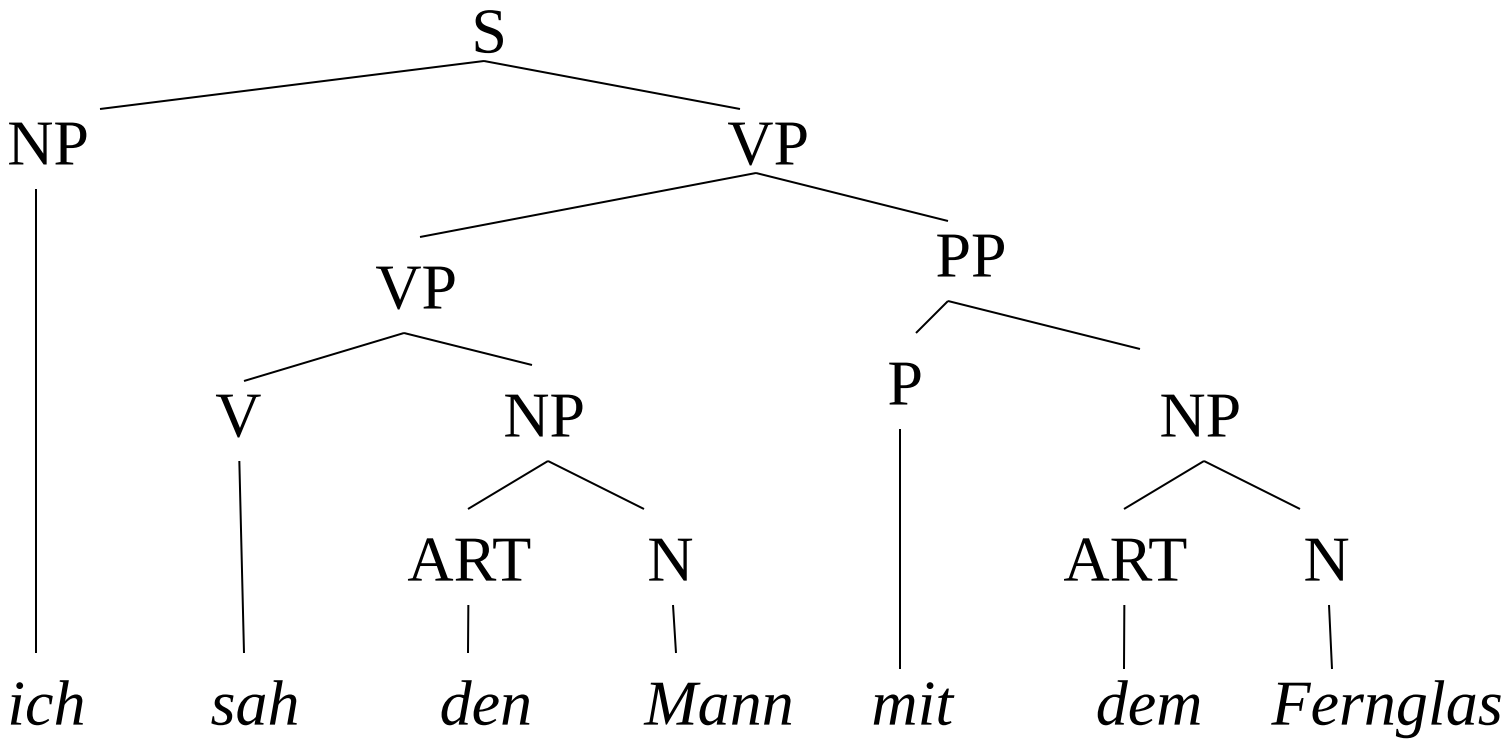
(*Ich* (*sah* ((*den Mann*) (*mit* (*dem Fernglas*))))))

(*Ich* ((*sah* (*den Mann*)) (*mit* (*dem Fernglas*))))

Syntaktische Ambiguität: es gibt mehrere Zergliederungen des Satzes

Syntax: Baumstrukturen

$$(Ich\ ((\ sah\ (den\ Mann)_{NP}\ )_{VP}\ (\ mit\ (\ dem\ Fernglas\ )_{NP}\ )_{PP}\ )_{VP}\ )_{S}$$



S = Satz, NP = Nominalphrase, VP = Verbalphrase, PP = Präpositionalphrase, ART = Artikel, N = Nomen, V = Verb, P = Präposition

Syntax: Substitutionstest

Ich sah den Mann mit dem Fernglas

Er sah den Mann mit dem Fernglas
Wer sah den Mann mit dem Fernglas
Der alte kranke Lehrer sah den Mann mit dem Fernglas

Ich kenne den Mann mit dem Fernglas
Ich fürchte und verachte den Mann mit dem Fernglas

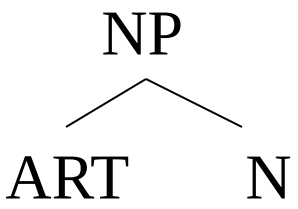
Ich sah einen Mann mit dem Fernglas
Ich sah diesen Mann mit dem Fernglas

Grammatische Kategorien werden durch Substitutionstest bestimmt. Austauschbare Wortgruppen gehören der gleichen Kategorie an.

Syntax: Kontextfreie Grammatiken

Grammatikregeln entsprechen Teilbäumen

NP → ART N



ART → *den*

Kontextfreie Grammatiken sind für die Beschreibung natürlicher Sprachen nicht ganz ausreichend, da Merkmale wie Kasus, Numerus usw. nicht angemessen behandelt werden können.

Syntax: Notwendigkeit von kontextfreien Grammatiken

- Reguläre Grammatiken sind für natürliche Sprachen nicht ausreichend, da rekursive Selbsteinbettung damit nicht beschrieben werden kann.

(Der Hund, der (die Katze, die (die Maus) jagt), ärgert), bellt.

NP → ART N REL-S

REL-S → REL-PRO NP V

NP = Nominalphrase, REL-S = Relativsatz,
ART = Artikel, N = Nomen, V = Verb, REL-PRO = Relativpronomen

Ambiguität

Ambiguität besteht, wenn ein Satz mehrere Bedeutungen hat.

- Lexikalische Ambiguität
  auf der Bank (Sitzgelegenheit oder Geldinstitut)

- Ambiguität der Wortart
  flying planes can be dangerous
  he saw that gasoline can explode

- Strukturelle syntaktische Ambiguität
  Ich sah den Mann mit dem Fernglas auf dem Berg

Prosodie

Lautliche Eigenschaften

- Erstrecken sich über mehrere Phone oder Phoneme
  ➔ „suprasegmental“
  z.B. Betonung einer Silbe: be<u>wun</u>dern [bəvˈʊndɐn]
- relevant für die Bildung größerer Einheiten
  (Wörter, Wortgruppen, Sätze, ...)
  z.B. Wortbetonung
  Gliederung eines Satzes,
  Satzbetonung
  Satzmelodie
  Unterscheidung Aussage – Frage
- zunehmend wichtig in Sprachsynthese und -erkennung

Prosodie: Beispieltext

Die Kurzsichtigkeit der Argumente für eine fortgesetzte Lohndrosselung ist daher besonders befremdlich bei denjenigen, die immer wieder betonen, der technische Fortschritt müsse genutzt und das Strukturwandeltempo beschleunigt werden: Nur so könne die Wirtschaft dem internationalen Wettbewerb gewachsen bleiben.

Prosodie: Phrasierung des Beispieltexts

di: 'kʊrtʂ,zɪçtɪçkaɪ̯t −↑
de:ɐ̯ ʔargu'mɛntə −↑
fy:ɐ̯ ʔaɪ̯nə fɔrtgəzɛtʂtə 'lo:n,drɔsəlʊŋ −↑
ʔɪst dɛshalp bəzɔndɐs bəfrɛmtlɪç baɪ̯ 'de:n,je:nɪgən =
di: ʔɪmɐ vi:dɐ bə'to:nən =↑
de:ɐ̯ tɛçnɪʃə 'fɔrt,ʃrɪt −↑
mʏsə gə'nʊtʂt =↑
ʔʊnt das strʊk'tu:ɐ̯,vandəl,tɛmpo −↑
bə'ʃlɔʏ̯nɪçt ,ve:ɐ̯dən =↓
nu:ɐ̯ zo: kœnə di: 'vɪrtʃaft −↑
de:m ʔɪntɛrnatʂi̯ona:lən 'vɛtbə,vɛrp −↑
gə'vaksən ,blaɪ̯bən =↓

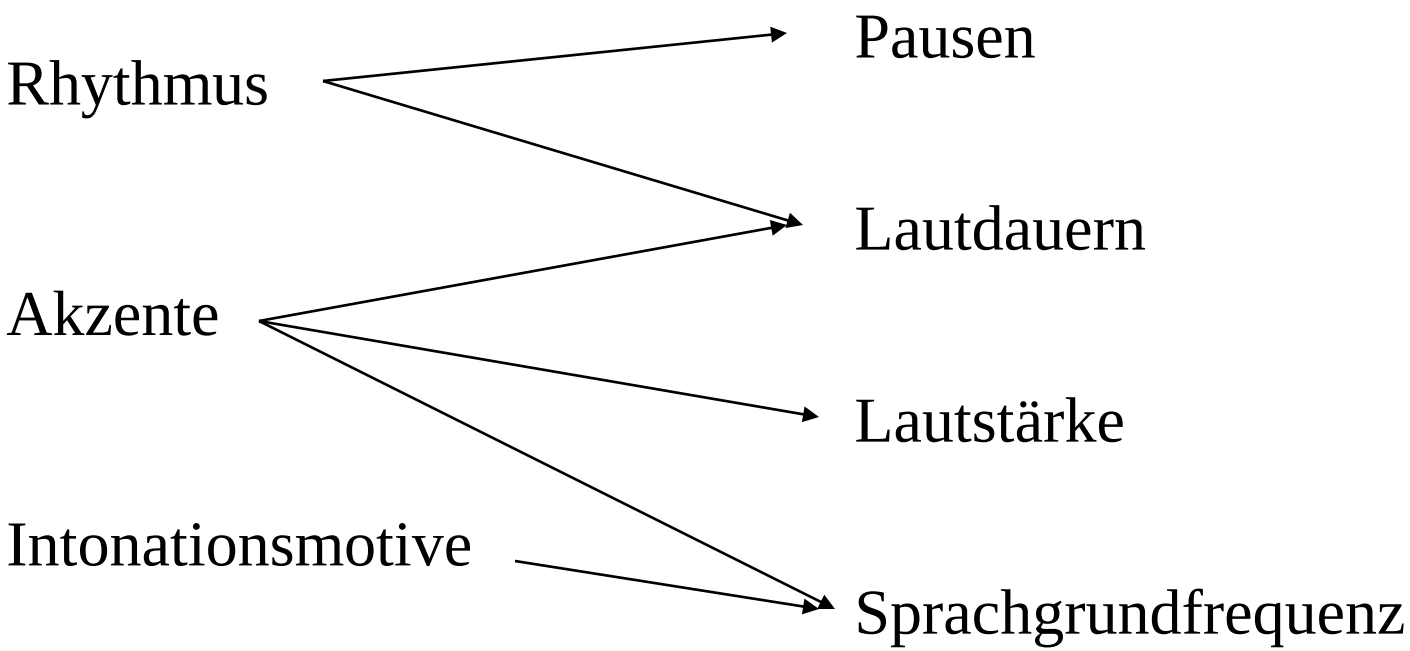
=/- ... Phrasengrenze mit/ohne Sprechpause
↑/↓ ... Satzmelodie steigt / fällt
' ... Hauptbetonung
, ... Nebenbetonung

26

Prosodie und prosodische Parameter

Prosodie

prosodische Parameter



Literatur

DUDEN Aussprachewörterbuch.
Mannheim: Bibliographisches Institut. 1983.

Fleischer, W. et al. (Hrsg.): *Kleine Enzyklopädie Deutsche Sprache.*
Leipzig: Bibliographisches Institut. 1983.

Jurafsky, D. & Martin, J. H.: Speech and Language Processing
Upper Saddle River: Prentice Hall. 2000.

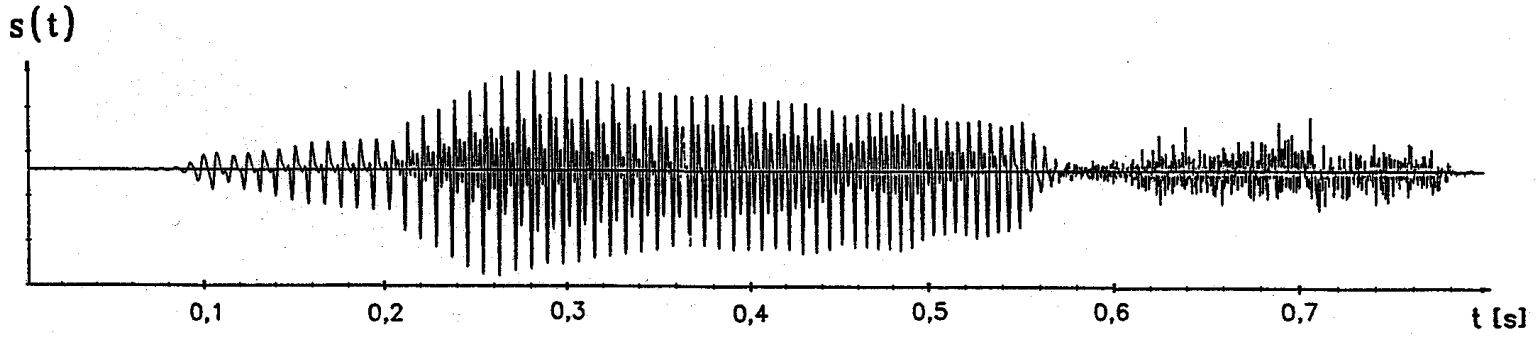
Ladefoged, P.: A Course in Phonetics.
New York: Harcourt Brace & Co. 1993.

Inhalt

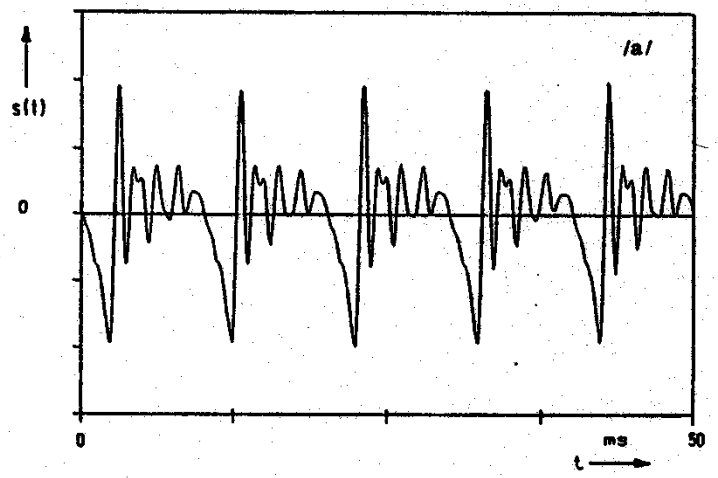
3. Grundlagen der Sprachsignalverarbeitung

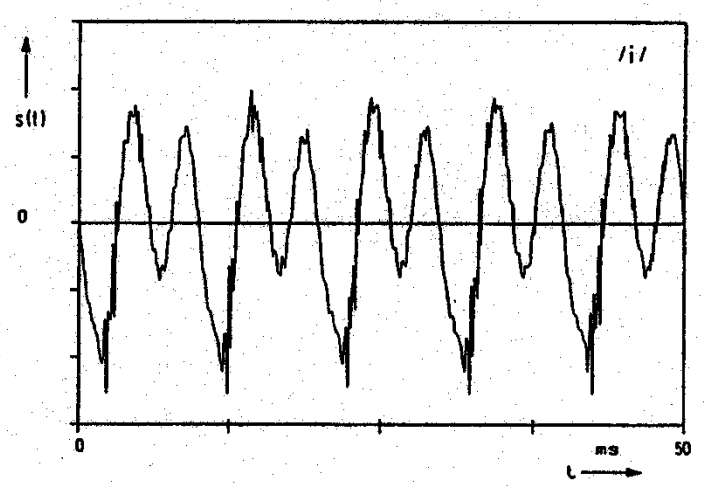
- Das Sprachsignal im Zeitbereich
- Das Sprachsignal im Frequenzbereich
- Spektrum – Sonagramm
- Formantstruktur der Vokale
- Digitalisierung des Sprachsignals
- Quelle-Filter-Modell

Sprachsignal im Zeitbereich:



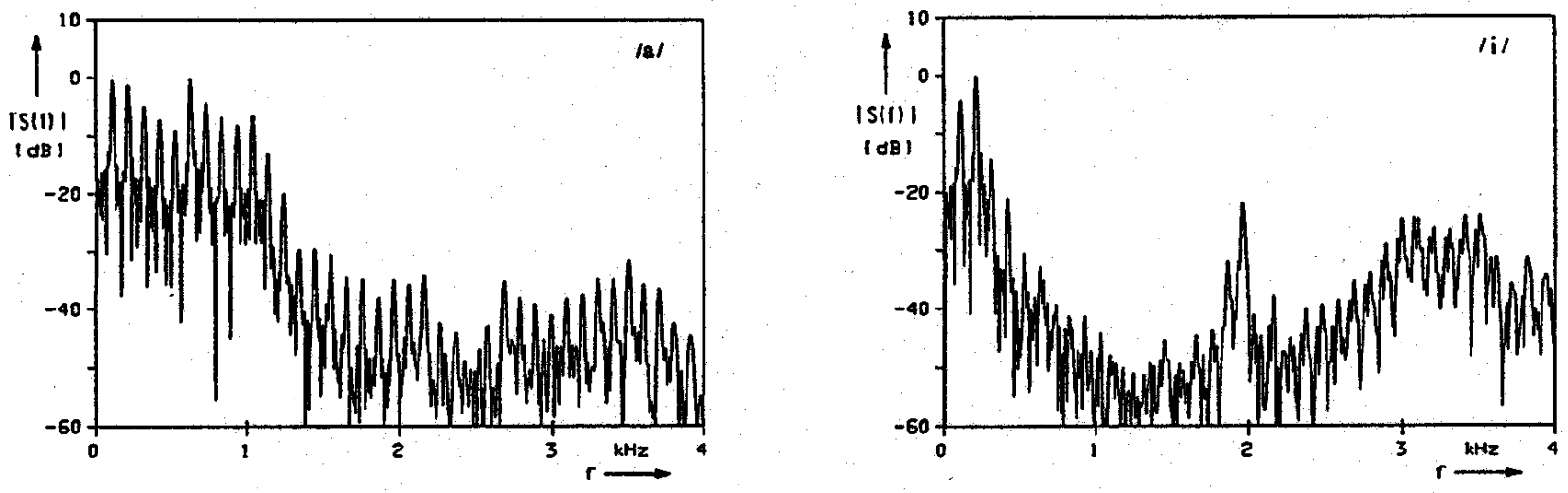
Vokale:





Quelle: K. Fellbaum, Sprachverarbeitung und Sprachübertragung, Springer Verlag Berlin, Heidelberg, New York, 1984

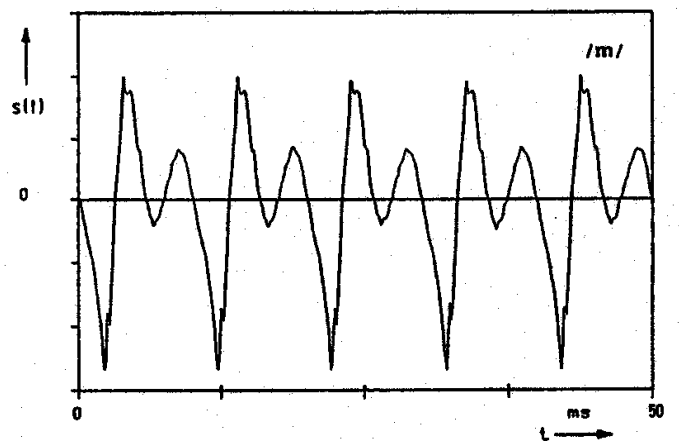
Im Frequenzbereich: Spektrum



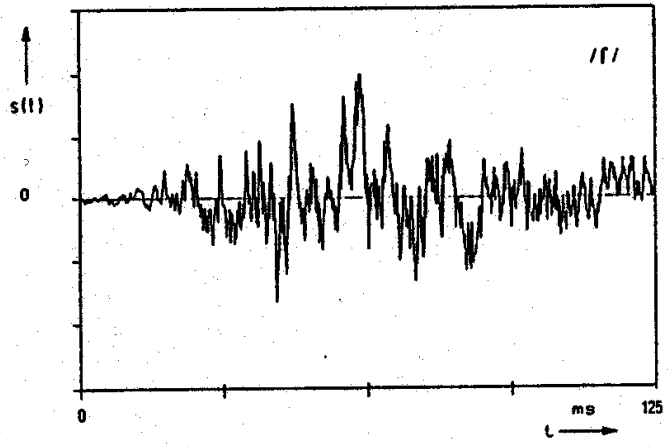
Quelle: K. Fellbaum, Sprachverarbeitung und Sprachübertragung, Springer Verlag Berlin, Heidelberg, New York, 1984

Konsonanten:

Nasal:

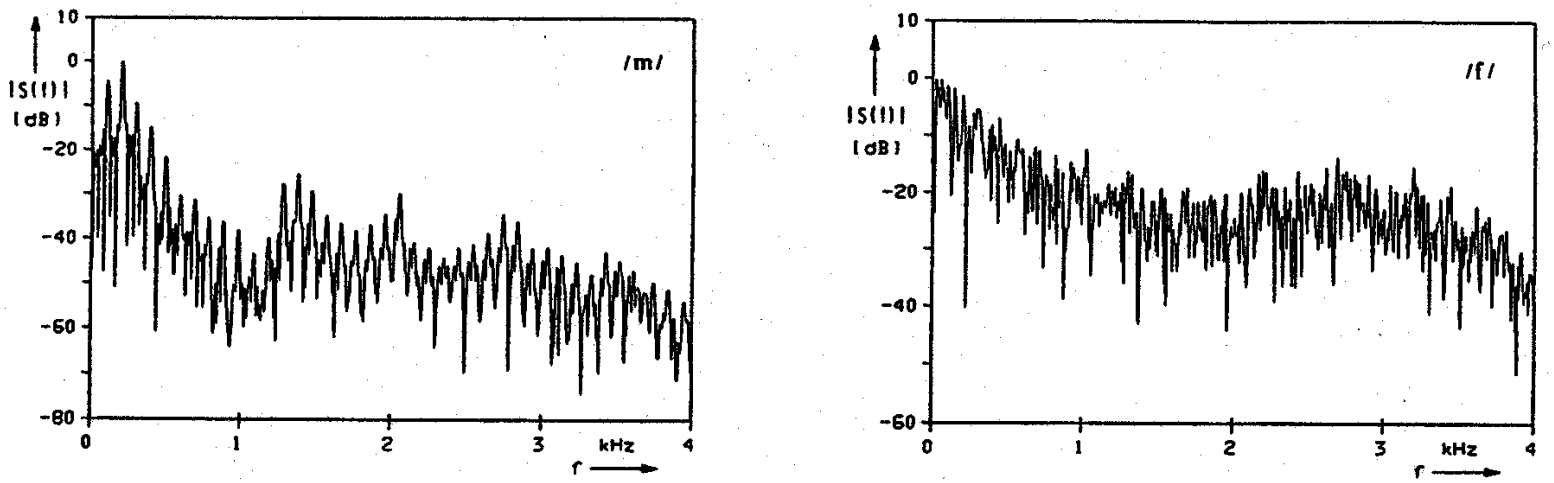


stimmloser Frikativ:



Quelle: K. Fellbaum, Sprachverarbeitung und Sprachübertragung, Springer Verlag Berlin, Heidelberg, New York, 1984

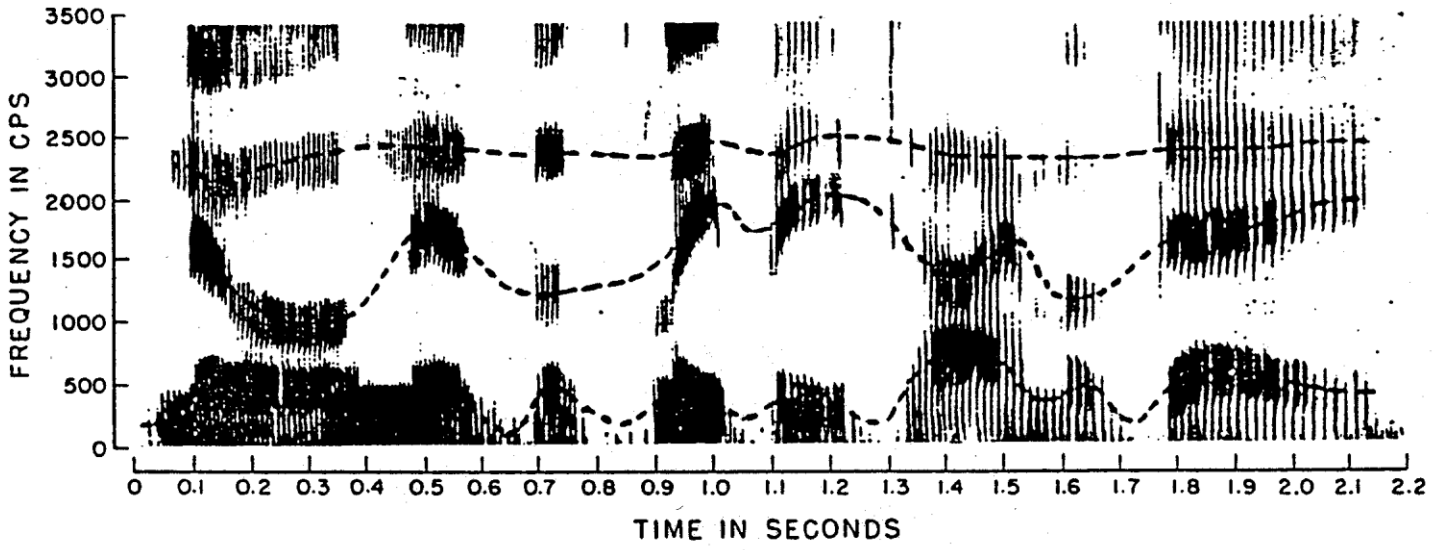
Spektren:



Quelle: K. Fellbaum, Sprachverarbeitung und Sprachübertragung, Springer Verlag Berlin, Heidelberg, New York 1984

Sonagramm:

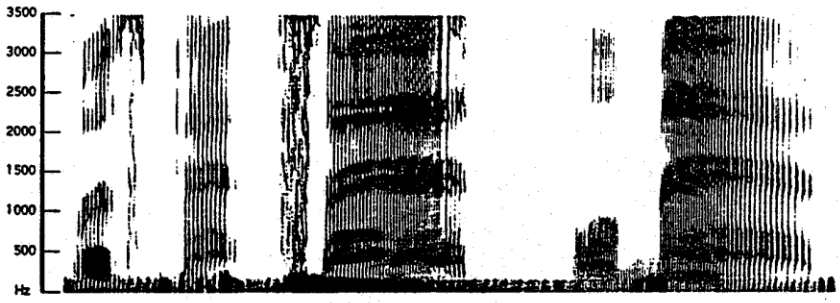
Darstellung über Zeit und Frequenz



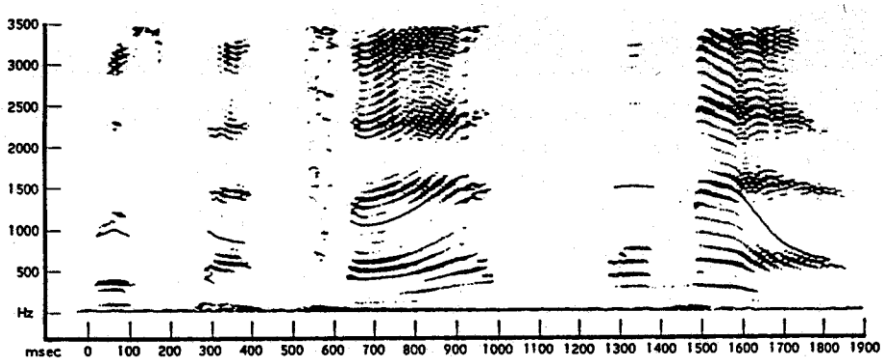
„*Noon is the sleepy time of day.*“

3. Grundlagen der Signalverarbeitung

Breitbandsonagramm:



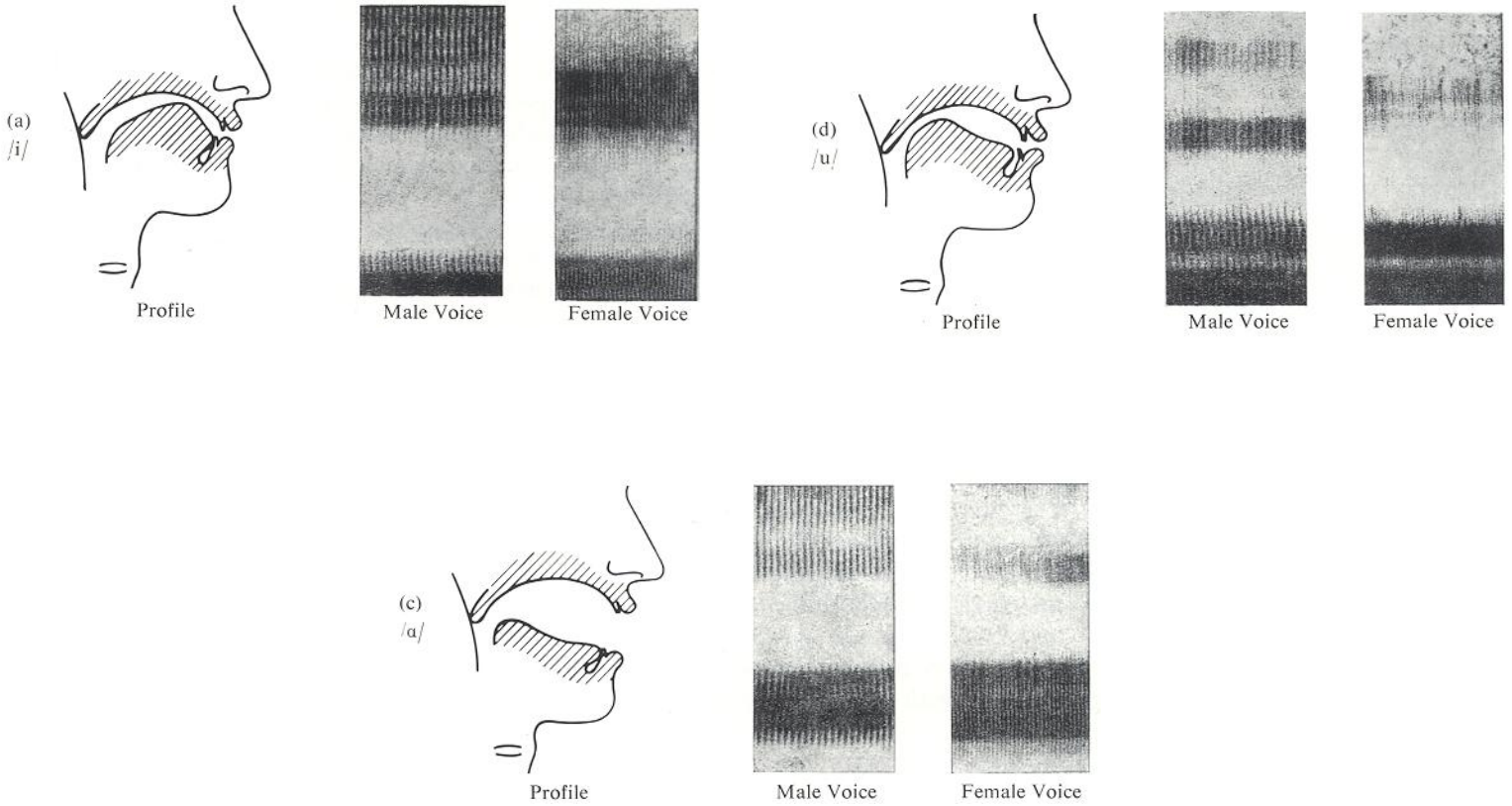
Schmalbandsonagramm:



„Is Pat sad or mad?“

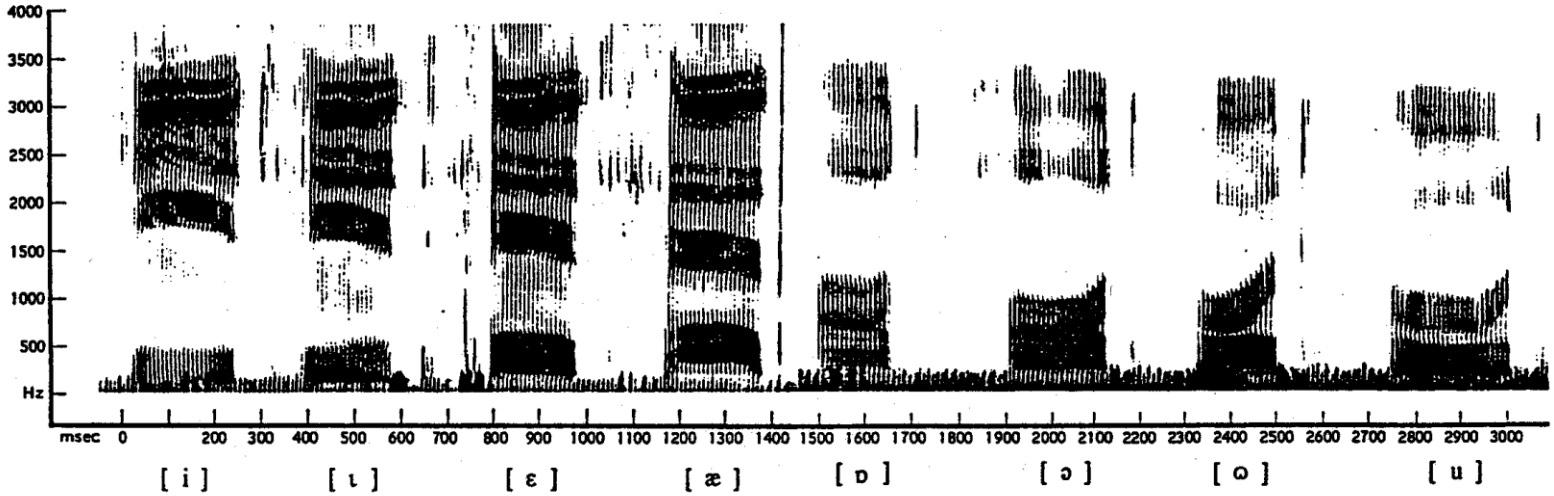
Quelle: P. Ladefoged, A Course In Phonetics (Third Edition), Harcourt Brace & Company, 1993

Formantstruktur der Vokale:



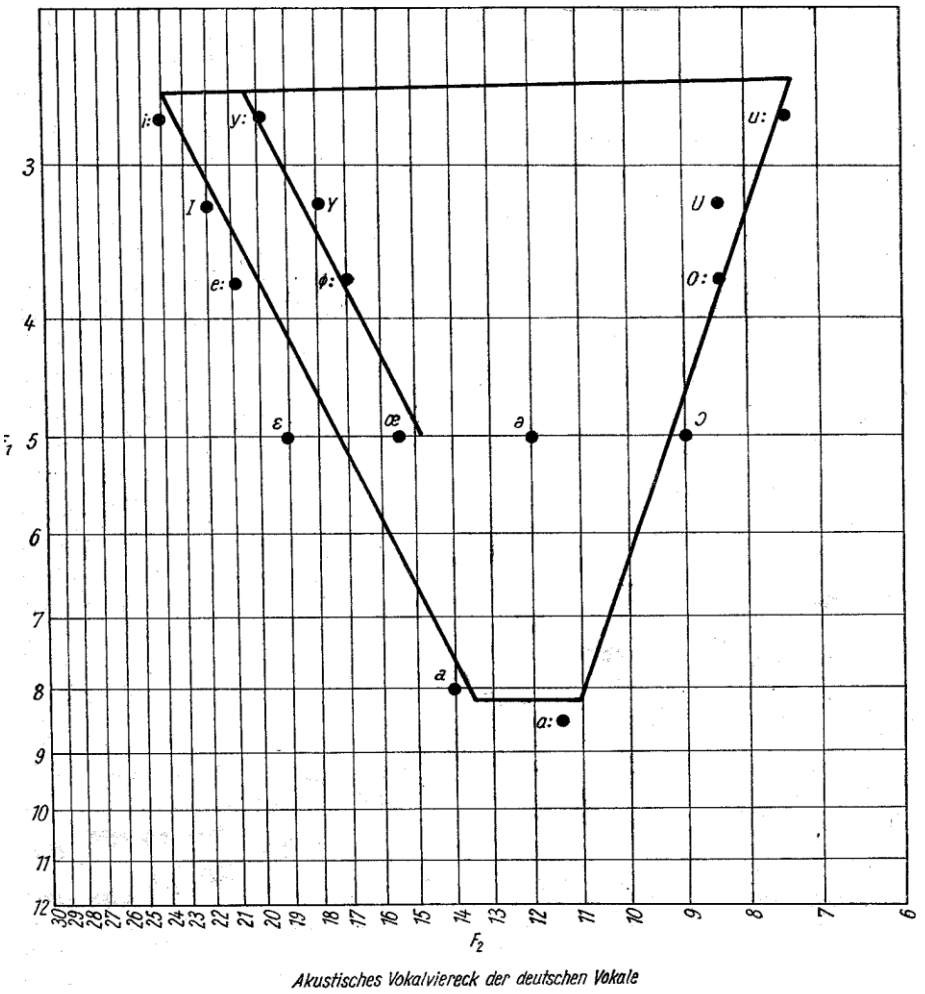
Formanten von 8 englischen Vokalen

(aus *heed*, *hid*, *head*, *had*, *hod*, *hawed*, *hood*, *who´d*):



Quelle: P. Ladefoged, A Course In Phonetics, Harcourt Brace & Company, 1993

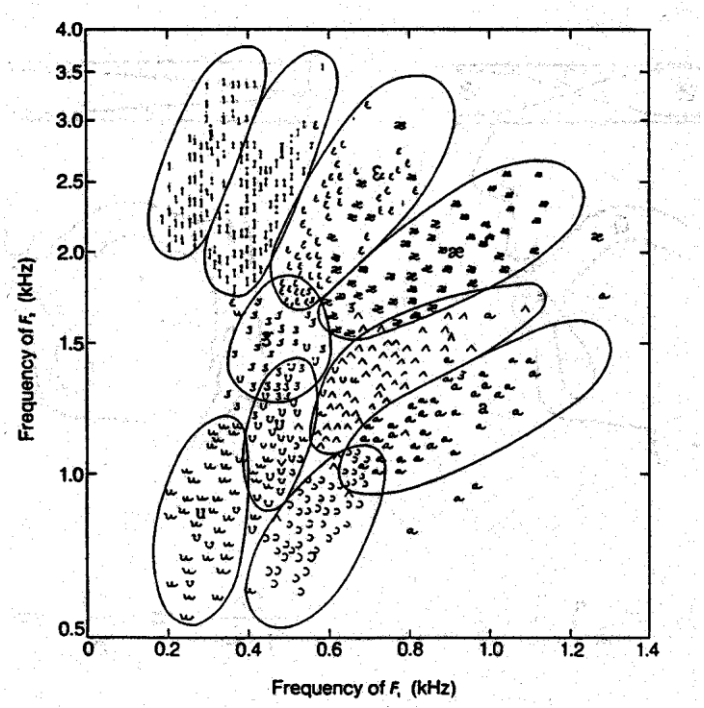
Akustisches Vokalviereck:



Akustisches Vokalviereck der deutschen Vokale

Quelle:
Hans-Heinrich Wängler
Atlas deutscher Sprachlaute
Akademie-Verlag Berlin 1981

Streuung der Formantwerte:

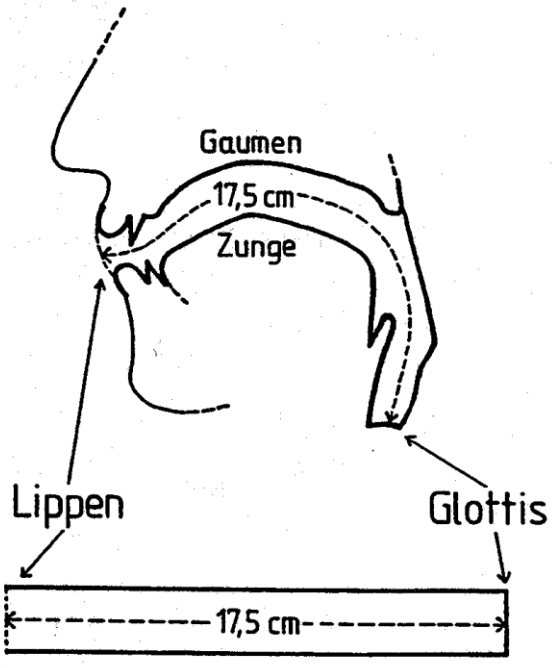


76 Sprecher, 10 englische Vokale

(aus *heed*, *hid*, *head*, *had*, *hod*, *hawed*, *hood*, *who´d*, *hud*, *heard*)

Akustische Theorie:

Vokaltrakt (Mund- und Rachenraum)

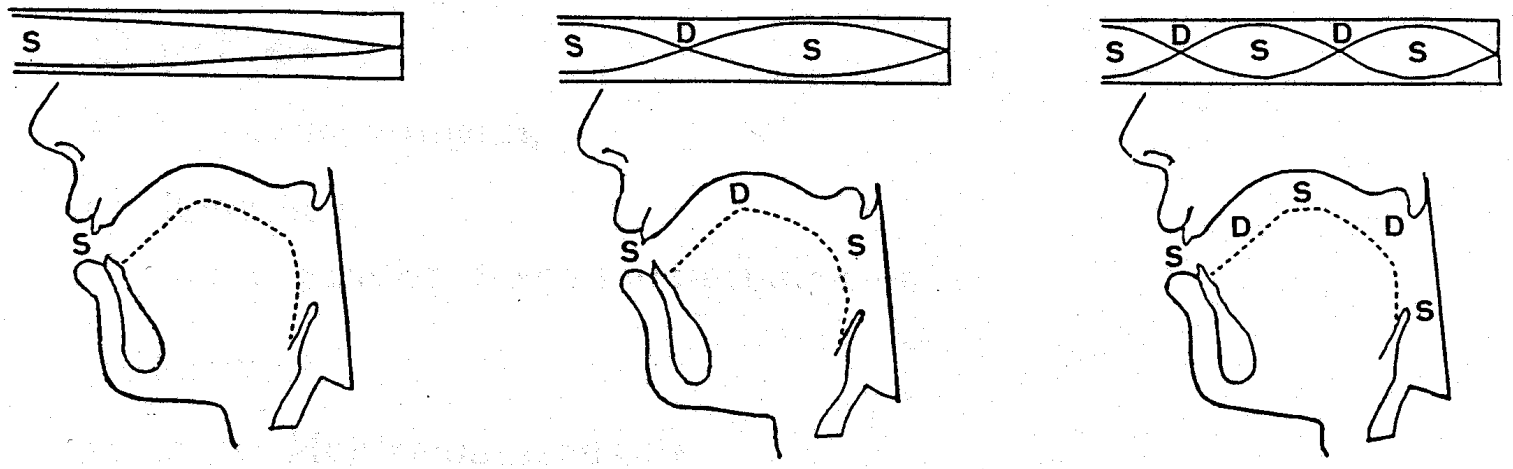


angenähert durch

einseitig geschlossenes

zylindrisches Rohr

Formanten als Resonanzen im Vokaltrakt:



$f_1 = \frac{c}{\lambda} = \frac{c}{41}$

$\sim \frac{350\text{m/s}}{0{,}7\text{m}} = 500\text{Hz}$

$f_2 = 3\, f_1$

~ 1500 Hz

$f_3 = 5\, f_1$

~ 2500 Hz

- Zylindrisches Rohr

  ~ neutrale Zungenposition [ə]

- Querschnittsänderungen beim einseitig geschlossenen Rohr:

  Verengung in der Nähe eines Amplitudenmaximums von

  Schallschnelle (S): f_{Res} ↓

  Druck (D): f_{Res} ↑

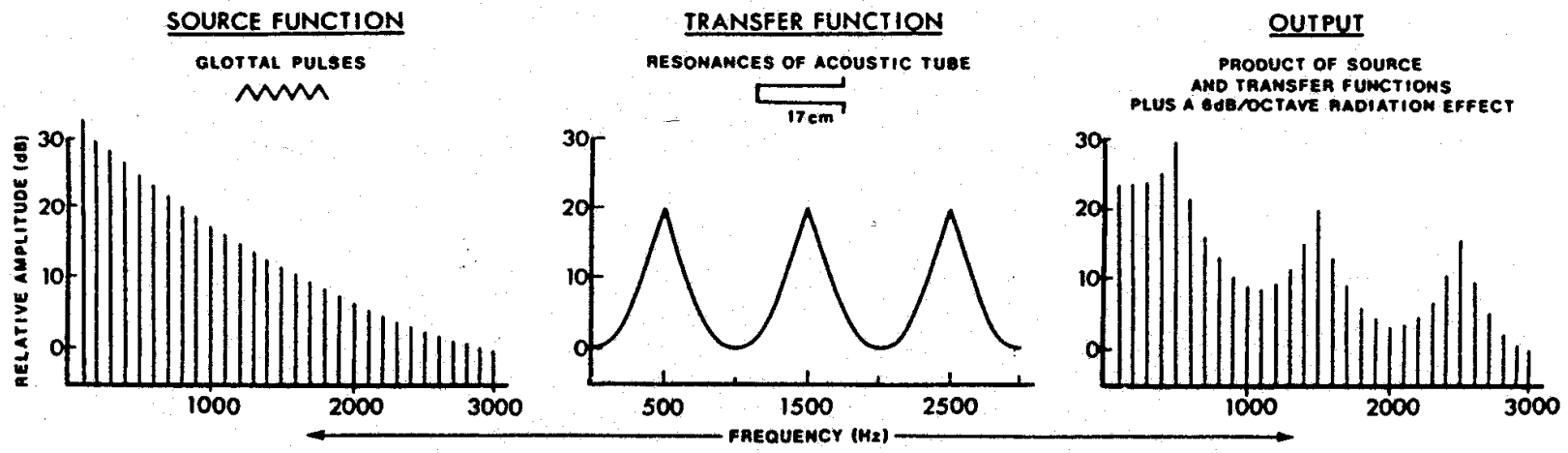
Quelle-Filter-Theorie:

- Quelle:
  - Glottisschwingung
  - Rauschen, hervorgerufen durch Luftturbulenzen
- Filter:

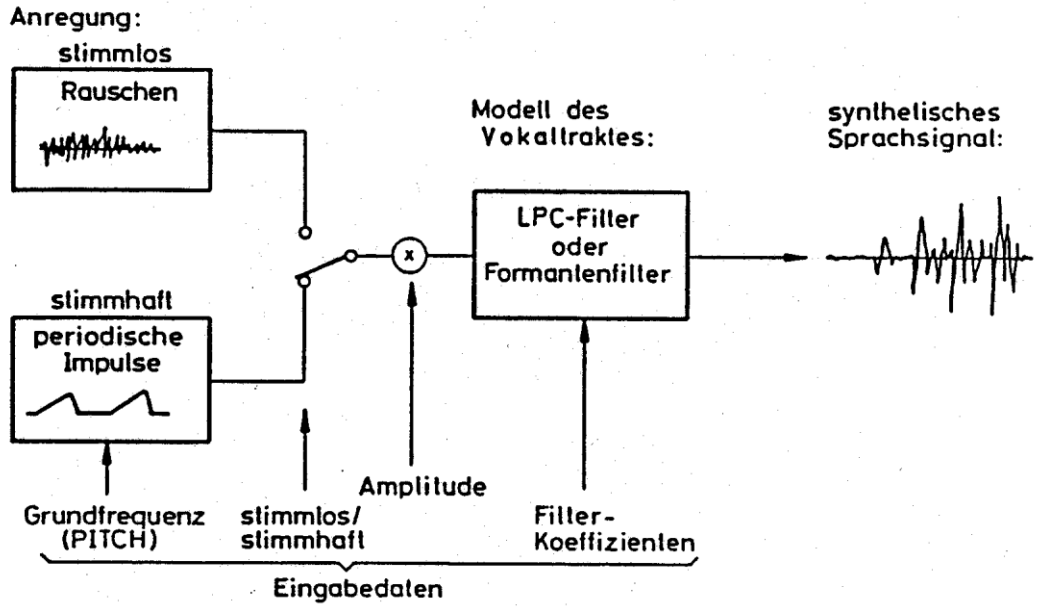
  variabler Hohlraumresonator,

  gebildet vom Mund-, Nasen- und Rachenraum (Ansatzrohr)

im Frequenzbereich:



Quelle-Filter-Modell:



Grundlage zahlreicher Verfahren für

- Sprachwiedergabe, Sprachsynthese
- Spracherkennung

Literatur:

Fellbaum, K.: SPRACHVERARBEITUNG UND SPRACHÜBERTRAGUNG
Berlin, ...: Springer Verlag (1984)

Fellbaum, K.:
http://www.kt.tu-cottbus.de/teleteaching/

Flanagan, J.L.:
SPEECH ANALYSIS, SYNTHESIS AND PERCEPTION.
Berlin, ...: Springer Verlag (1972)

Hasegawa-Johnson, M.:
http://www.ifp.uiuc.edu/~hasegawa/notes/index.html

Hess, W.:
http://www.ikp.uni-bonn.de/dt/lehre/materialien/grundl_ssv/

Neppert, J., Pétursson, M.: ELEMENTE EINER AKUSTISCHEN PHONETIK.
Hamburg: Helmut Buske Verlag (1986)

O´Shaughnessy, D.: SPEECH COMMUNICATION (Second Edition)
Reading (MA), ...: Addison-Wesley Publishing Company (2000)

Wängler, H-H.: ATLAS DEUTSCHER SPRACHLAUTE
Berlin: Akademie-Verlag (1981)

Inhalt

4. Sprachausgabe

Historischer Rückblick und Einführung

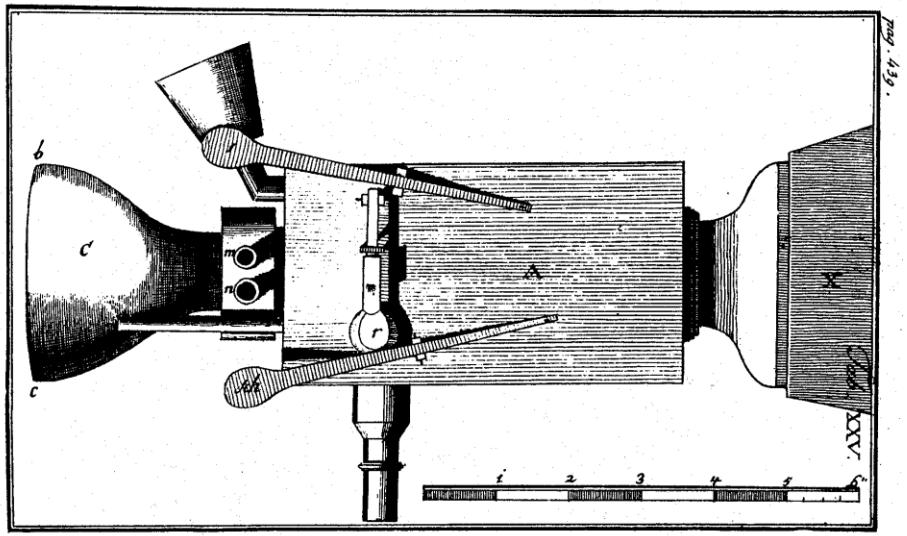
4.1 Sprachwiedergabe (reproduktive Verfahren)

4.2 Sprachsynthese (unbegrenztes Vokabular)

Historisches

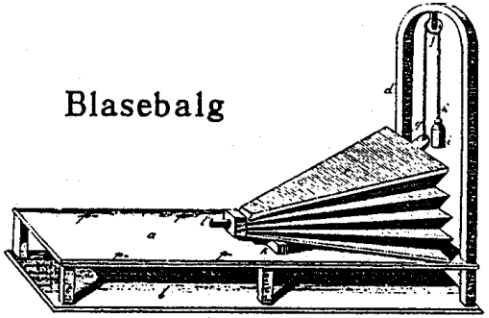
- Künstliche Sprache – ein alter Menschheitstraum
- Wolfgang v. Kempelen (1791):

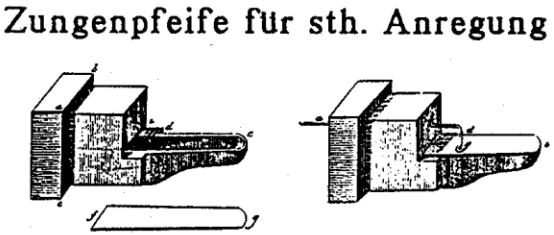
Es ist möglich eine alles sprechende Maschine zu machen.

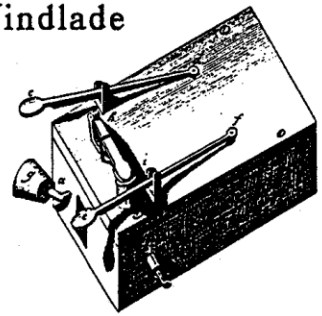


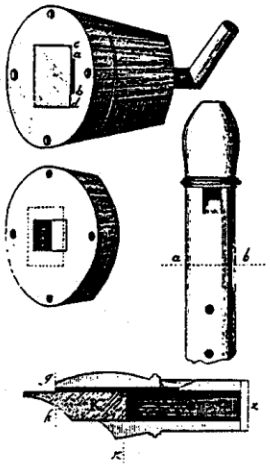
Hörprobe: „Mama“ (3x), „Papa“ (3x)

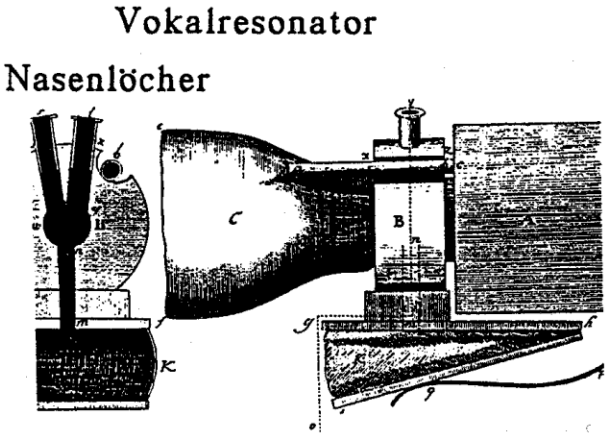
4. Sprachausgabe

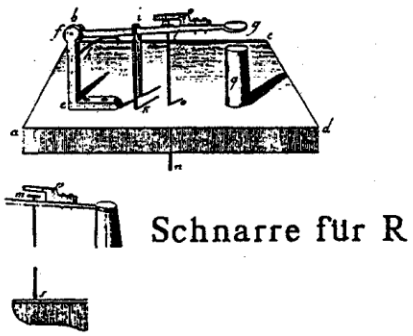










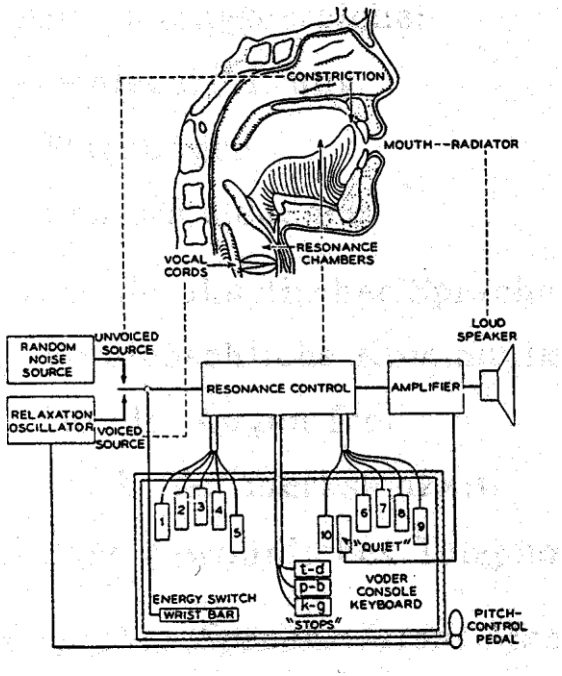


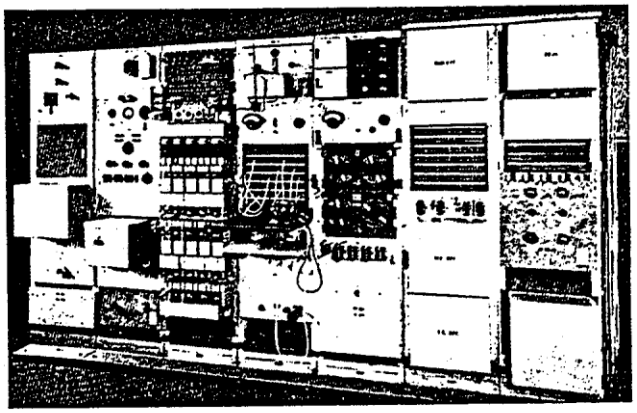
4. Sprachausgabe

bedeutende Fortschritte erst durch elektronische Modelle:

Homer Dudley (1939): Vocoder (Voice Coder),

Voder (Voice Operation Demonstrator)





4. Sprachausgabe

- Durchbruch mit Aufkommen leistungsfähiger Elektronik:
  - dedizierte Chips
  - PC („Computer Speech“)
- Ziel: Verbesserung der Benutzerschnittstelle
- Vorteile akustischer Sprachausgabe:
  - an menschliche Kommunikation angepasst (?)
  - Hände, Augen frei
  - weckt Aufmerksamkeit
  - über gewöhnliches Telefon übertragbar
  - miniaturisierbar

Hörprobe: Speech Output Module (Hewlett-Packard)

4. Sprachausgabe

- Anwendungsbeispiele:
  - Rechenzentrum
  - Büro
  - industrielle Prozesse
  - Lagerverwaltung
  - Fertigungskontrolle
  - Auto, Flugzeugcockpit
  - Ansage- und Auskunftsdienste
    (Zeitansage, Wetterdienst, Fahrplanauskunft, ...)
  - Bestell- und Buchungssysteme
    (Versandhäuser, Verlage, Reisebüros, ...)
  - akustische Ausgabe als Ergänzung zum Bildschirm
    (Multimedia-Präsentationen, Computer-unterstützter Unterricht, ...)
  - Behindertenhilfe (Blindenlesegerät, Sprechhilfen)

Sprachausgabe

	Sprachwiedergabe	**Sprachsynthese**
Funktion	Wiedergabe eines vorher gesprochenen Vokabulars	Zusammensetzung von Lautelementen
Vokabular	begrenzt	unbegrenzt
Sprecher	meist erkennbar	nicht erkennbar, oft verschiedene Stimmen wählbar
Prosodie	im Vokabular enthalten	muss künstlich hinzugefügt werden

	Sprachwiedergabe	**Sprachsynthese**
Aufwand	niedrig	hoch
Datenrate	1 .... 100 kbit / s	60 .... 300 bit / s
Sprachqualität	gut .... sehr gut	meist unnatürlich, aber verständlich
Entwicklungsstand	zahlreiche Bausteine auf dem Markt	verfügbar, aber weiterhin Gegenstand der Forschung

4. Sprachausgabe

Hörproben:

- *LPC-Baustein*

  Bitrate ~ 2 kbit / s

  (3 Sätze, 2x)
- *Teilbandcodierung*

  CNET, Lannion (Frankreich)

  Bandbreite 0 .... 7 kHz, Bitrate 64 kbit / s

  (4 Sätze, 4 Sprecher, männliche und weibliche Stimme)
- *„Karlchen“ (Fahrplanauskunft)*

  SPRAUS-VS

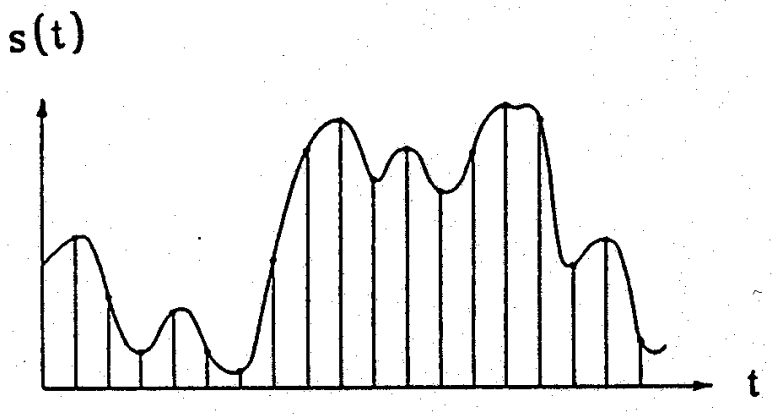
  Daimler-Benz-Forschungsinstitut, Ulm
- *SAMT*

  FTZ Darmstadt

Verarbeitung des Sprachsignals

a) im Zeitbereich:

Signalformcodierung

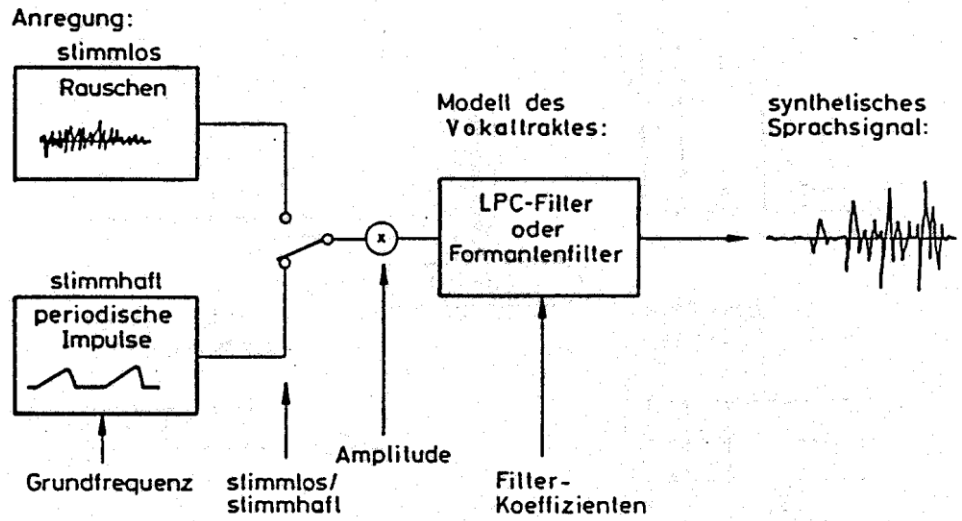


nur für einfache Aufgaben,

insbesondere Sprachwiedergabe, Übertragung

b) parametrische Verfahren:

Quelle-Filter-Modell



für Sprachwiedergabe, Sprachsynthese

und andere komplexere Aufgaben, z.B.: Spracherkennung

c) hybride Verfahren:

Kombination von Signalformcodierung und parametrischen Verfahren

In der Praxis gebräuchlich für Sprachübertragung (GSM, VoIP,...), Sprachwiedergabe und Sprachsynthese

Sprachwiedergabe:

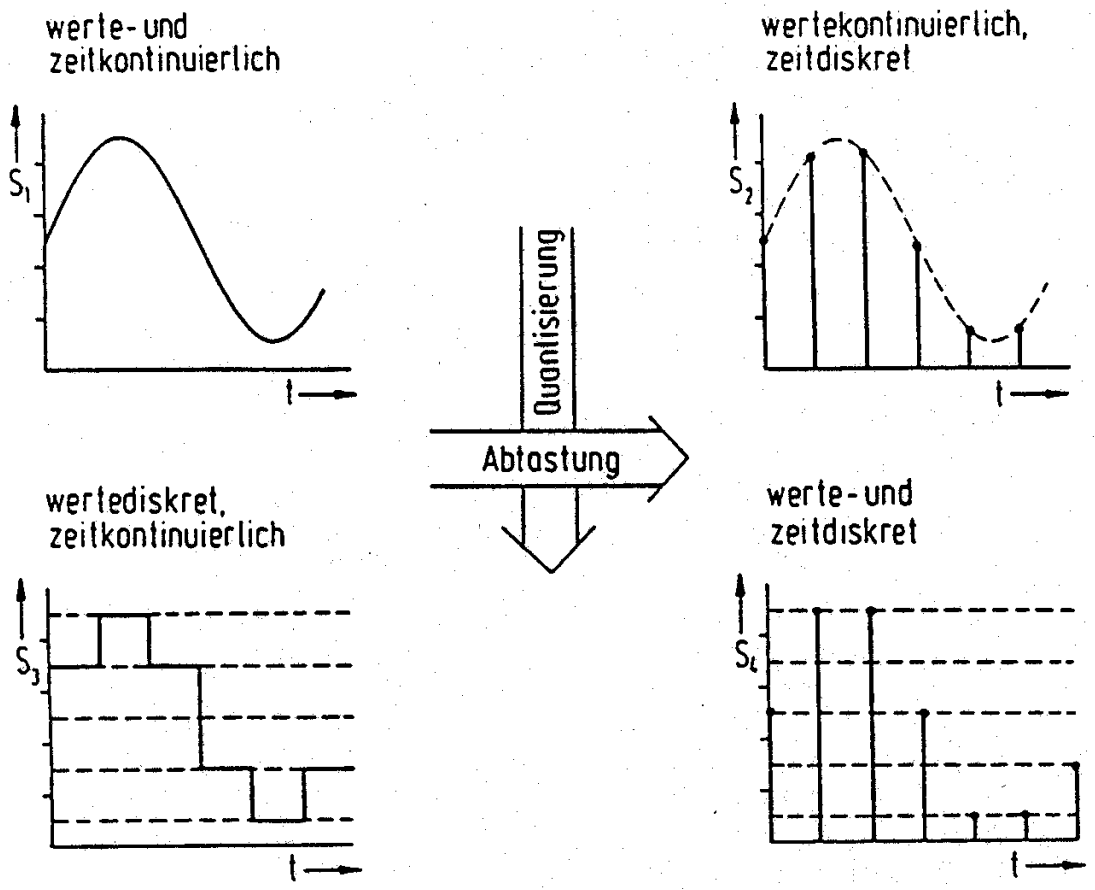
a) Signalformcodierung

Grundprinzip:

- analoges Sprachsignal
  (1) abtasten
  (2) quantisieren
- quantisierte Abtastwerte als Zahlenwert darstellen (digitalisieren), spez. im Dualsystem, d. h. binär
  (3) codieren
- Binärcodes können einfach als Folge von Spannungsimpulsen übertragen werden:

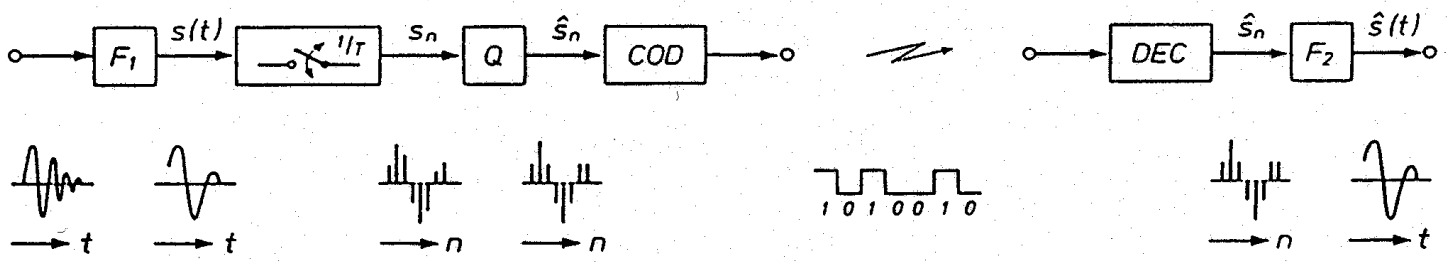
Pulscodemodulation (PCM)

Abtasten und Quantisieren:



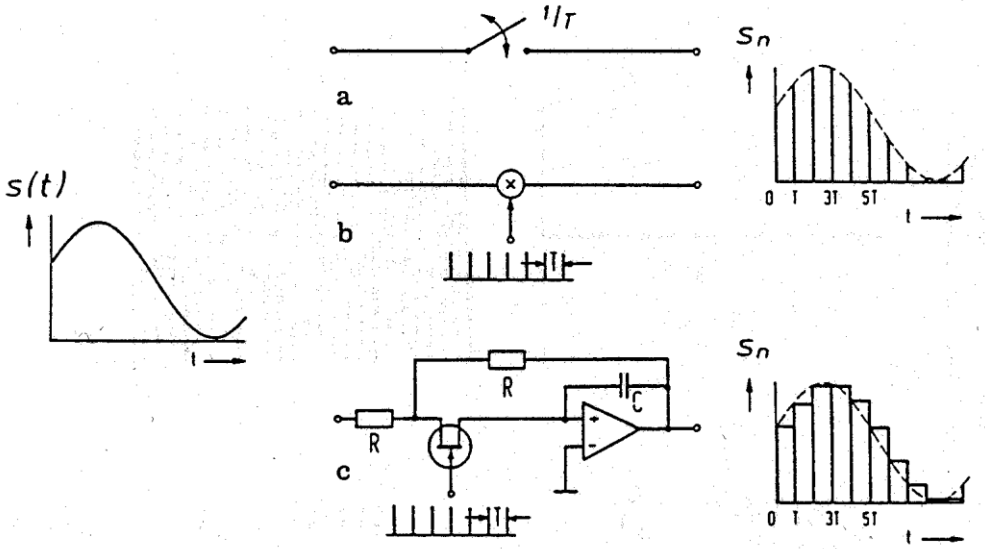
Blockdiagramm eines PCM-Systems:

Sender Empfänger



4.1 Sprachwiedergabe

Abtasten: $s(t) \rightarrow s_n = s(nT)$



a) symbolische Darstellung
b) Abtastung als Pulsamplitudenmodulation
c) Realisierung als Abtast-Halteglied

Abtasttheorem (Claude Shannon, 1948):

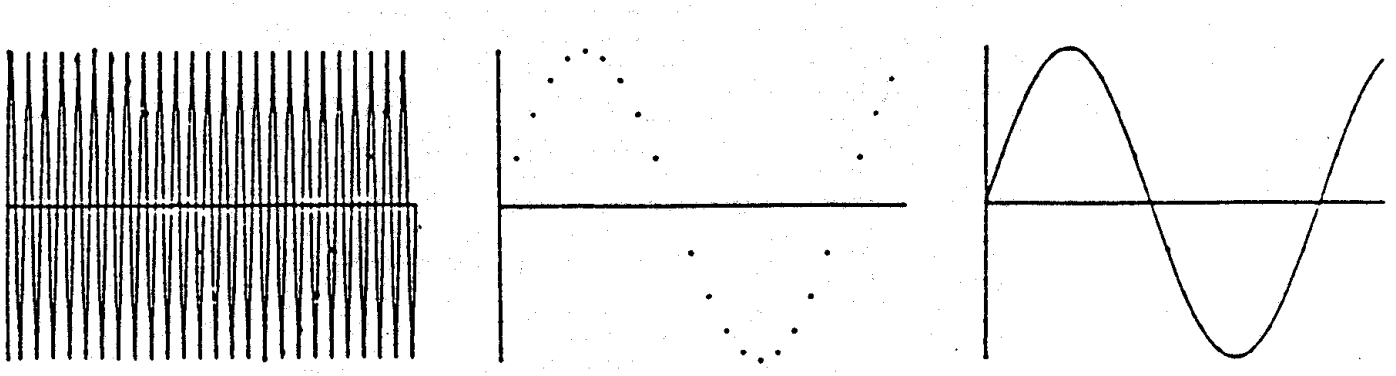
- Jedes **bandbegrenzte Signal** (Grenzfrequenz f_g) ist **vollständig durch** die Folge seiner **Abtastwerte** (Abtastfrequenz f_a) **bestimmt**, sofern:

$$f_a \geq 2 f_g$$

- für Sprachsignal in Telefonqualität (fg = 3,4 kHz) standardisiert:

$$f_a = 8 \text{ kHz}$$

Unterabtastung führt zu Aliasing:



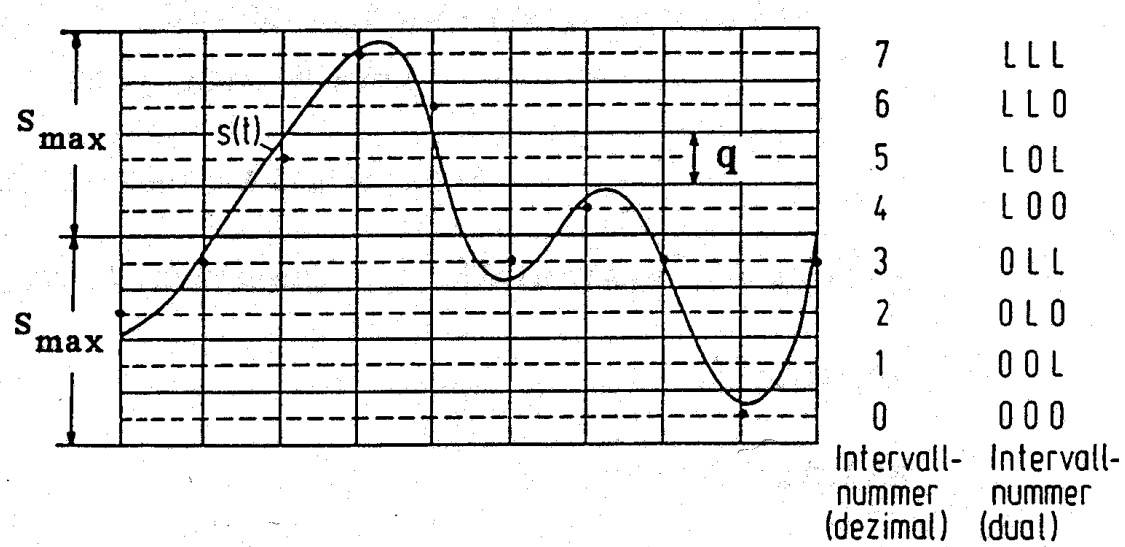
z.B.: $f_0 = 21$ Hz $f_a = 20$ Hz $f_0 - f_a = 1$ Hz

4.1 Sprachwiedergabe

Hörprobe:

- *Wetterbericht*
- *Wiedergabe nach Abtastung mit*
  - *500 Hz*
  - *1,2 kHz*
  - *3,0 kHz*
  - *8,0 kHz*

Quantisieren: $s_n \rightarrow \hat{s}_n$



Aussteuerbereich: $-s_{max} \leq s(t) \leq +s_{max}$

bei Binärcodierung mit B Bits:

2^B Abtastintervalle, Intervallbreite $q = \frac{2s_{max}}{2^B}$

Quantisierungsfehler:

$$\varepsilon_n = \hat{s}_n - s_n , \quad -q/2 \leq \varepsilon_n \leq +q/2$$

- vereinfachende Annahmen:
  - ε_n statistisch unabhängig (untereinander und vom Signal)
  - ε_n gleichverteilt über Quantisierungsintervall

  → Quantisierungsrauschen

- Rechnung liefert Signal-Rauschleistungsverhältnis (SNR):

$$\text{SNR [dB]} = \text{SNR}_0 + 6\ \text{B}$$

SNR_0: bestimmt durch Signaleigenschaften, ~ -5 dB für Sprache

6 B: d. h.: Verbesserung 6 dB / bit

Hörprobe:

- *Wetterbericht (Bandbreite 300 Hz ... 3,4 kHz)*
  nach Abtastung mit 8 kHz
  und Quantisierung mit

  - *3 bit pro Abtastwert → Datenrate 24 kbit / s*
  - *4 bit 32 kbit / s*
  - *5 bit 40 kbit / s*
  - *6 bit 48 kbit / s*
  - *7 bit 56 kbit / s*

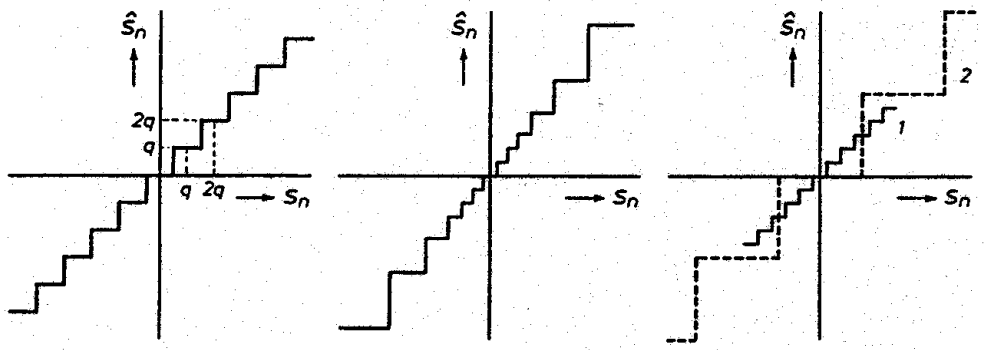
- berechnetes SNR (bzw. Qualität der Hörproben) setzen optimale Aussteuerung des Quantisierers voraus!
- SNR verschlechtert sich in leisen Passagen
- Sprachsignal weist aber große Dynamik auf
- für Telefonqualität wäre Auflösung von 12 bit erforderlich

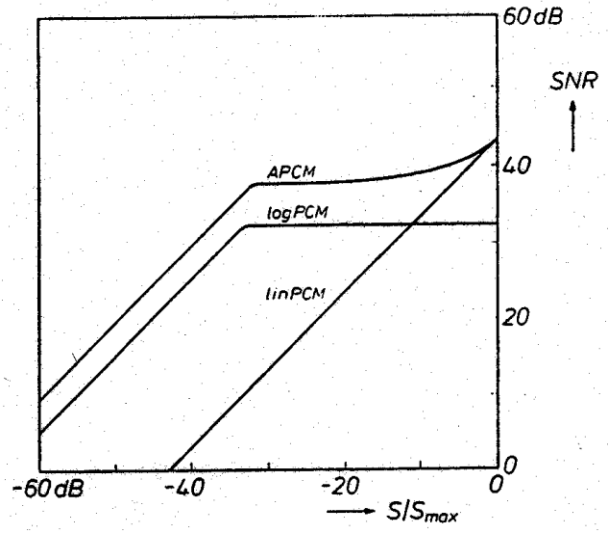
Abhilfe:

Quantisierungsrauschen $\leftrightarrow$ Stufenhöhe

an Lautstärke anpassen: Kompandierung

gleichförmige - ungleichförmige Quantisierung





linear	logarithmisch	adaptiv	SNR
↓	↓	↓	in Abhängigkeit von
linPCM	logPCM	APCM	der Aussteuerung

Weitere Möglichkeiten

zur Reduktion der Redundanz bei PCM:

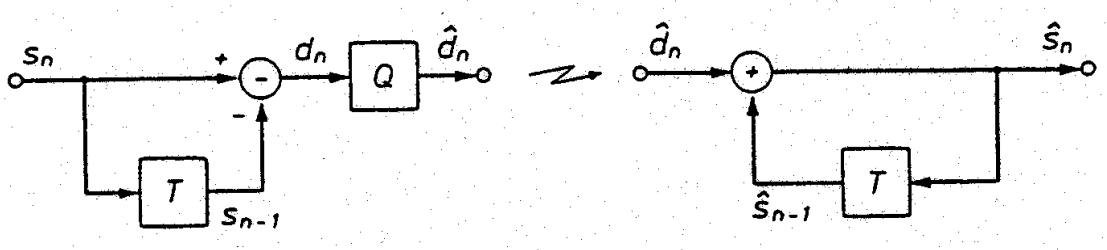
- Ausnützen der statistischen Eigenschaften des Sprachsignals:
  statistische Abhängigkeit (Korrelation)
  aufeinanderfolgender Abtastwerte

- Grundidee:
  nicht Absolutwerte, sondern Differenz
  aufeinanderfolgender Abtastwerte codieren:

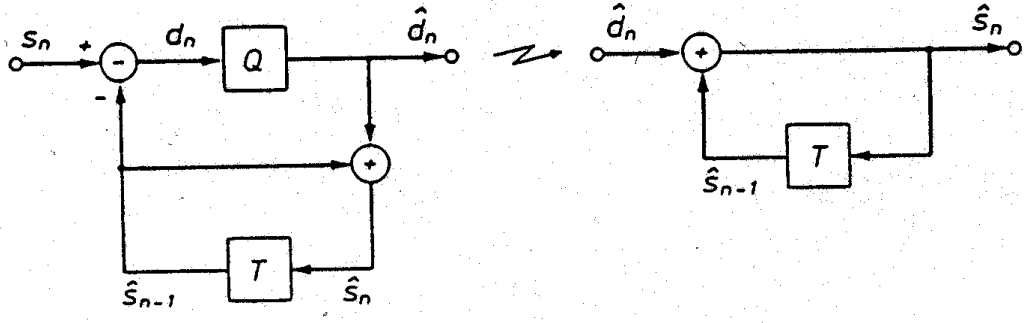
$$\hat{s}_n = \hat{s}_{n-1} + \hat{d}_n$$

$\rightarrow$ Differenz-Pulscodemodulation (DPCM)

Prinzipschaltung eines DPCM-Systems:



besser:

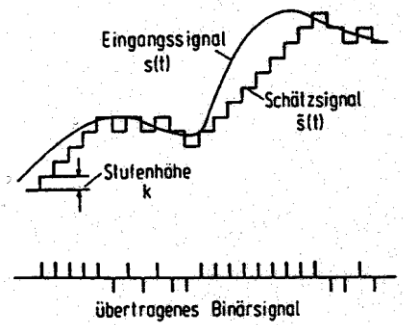


Sender enthält Abbild des Empfängers, Quantisierer in geschlossener Schleife

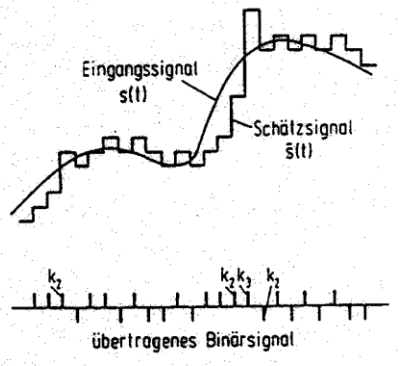
→ Quantisierungsfehler werden nicht akkumuliert

4.1 Sprachwiedergabe

- für Differenzwerte $\hat{d}_n$ werden weniger Bits benötigt,
  im Extremfall nur 1 Bit: $\hat{d}_n = \pm \triangle$
  $\rightarrow$ Deltamodulation (DM)
- besonders zweckmäßig mit adaptiver Quantisierung:
  $\rightarrow$ adaptive Differenz-Pulscodemodulation (ADPCM)
  adaptive Deltamodulation (ADM)



DM: Quantisierer übersteuert



Anpassung der Stufenhöhe: ADM

4.1 Sprachwiedergabe

- Grundprinzip der DPCM:

$$\hat{s}_n = \hat{s}_{n-1} + \hat{d}_n$$

kann auch folgendermaßen interpretiert werden:

$\hat{s}_{n-1}$ liefert bereits einen **Schätzwert** $\tilde{s}$ für $\hat{s}_n$,

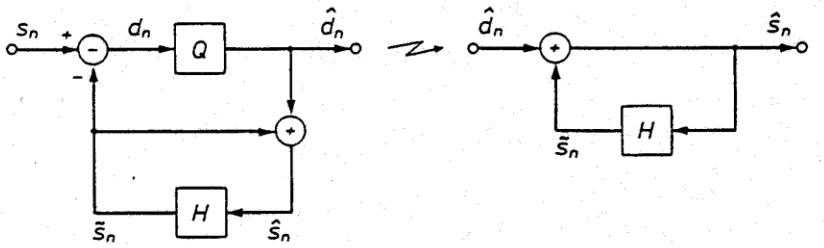
nur der „Schätzfehler“ d_n muss übertragen
(bzw. bei Sprachwiedergabe: gespeichert) werden

- Verallgemeinerung dieses Prinzips:
nicht nur den letzten Abtastwert zur Schätzung heranziehen,
sondern **Linearkombination** von p vorangegangenen Abtastwerten:

$$\hat{s}_n = \tilde{s} + \hat{d}_n = \sum_{i}^{p} a_i \hat{s}_{n-i} + \hat{d}_n$$

→ lineare Prädiktion

Verallgemeinertes Prinzip der DPCM:



- Verzögerungselement T der ursprünglichen Prinzipschaltung ersetzt durch

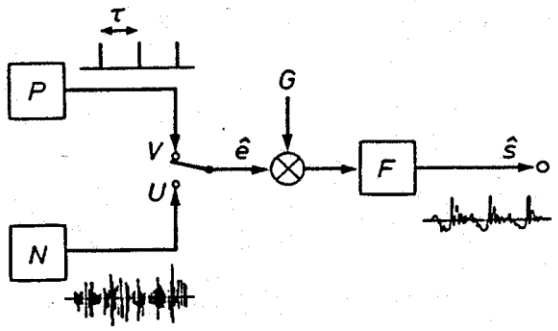
  allgemeines lineares Filter H: Prädiktor
- mit adaptiver Quantisierung: ADPCM
- zusätzlich kann das Prädiktorfilter H
  adaptiv bzw. zeitvariant gemacht werden,
  vgl. später: Linear Predictive Coding (LPC)

b) parametrische Verfahren (Vocoder)

Prinzip:

- Spracherzeugungsmodell (Quelle-Filter-Modell)

  z.B.:



- Beschreibung (Codierung, Übertragung, Speicherung, Rekonstruktion) des Sprachsignals mit Hilfe der Modellparameter
- Setzt Extraktion der zeitlichen Verläufe dieser Parameter aus dem Sprachsignal voraus: Sprachanalyse

Modellparameter - Sprachanalyse

Quelle: ← EXA (Excitation analysis)

- Art der Anregung:
  stimmhaft / stimmlos (/ gemischt) V / U
- Sprachgrundfrequenz (für sth. Segmente)
  bzw. Sprachgrundperiode τ

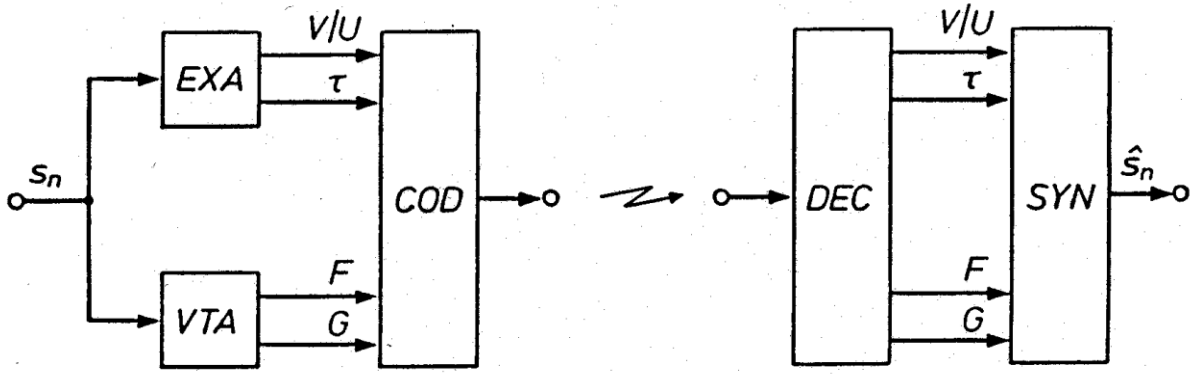
Filter: ← VTA (Vocal tract analysis)

- Verstärkungsfaktor G
- Filtercharakteristik
  Approximation der Einhüllenden
  des Sprachsignalspektrums F

Vocoder

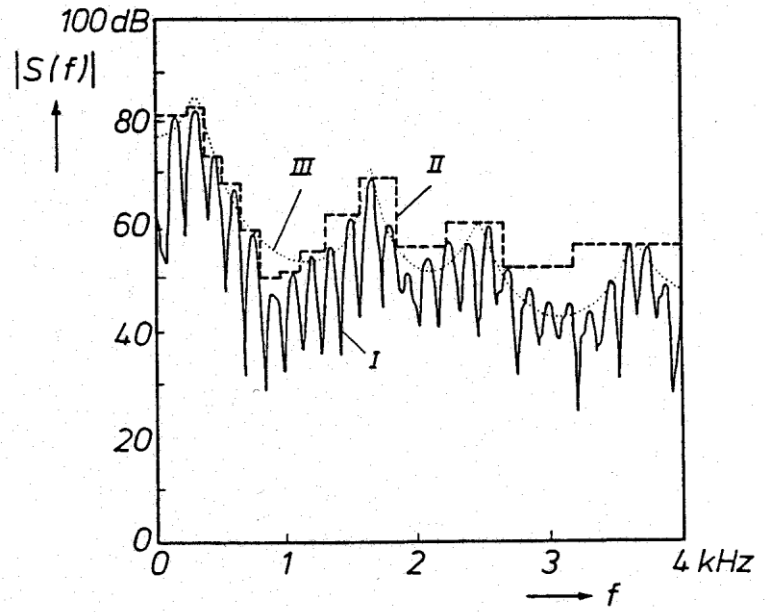
Sender Empfänger

Sprachanalyse Codierung Decodierung (Sprachsynthese)
(Signalanalyse) Signalsynthese



Vocoder: verschiedene Typen

Unterscheiden sich nach der Art der Charakterisierung des Vokaltrakts: verschiedene Approximationen der Einhüllenden des Sprachsignalspektrums (I):

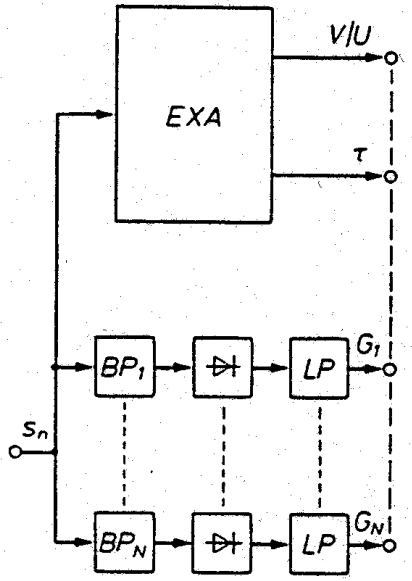


- stückweise konstant (II)
  → Kanalvocoder

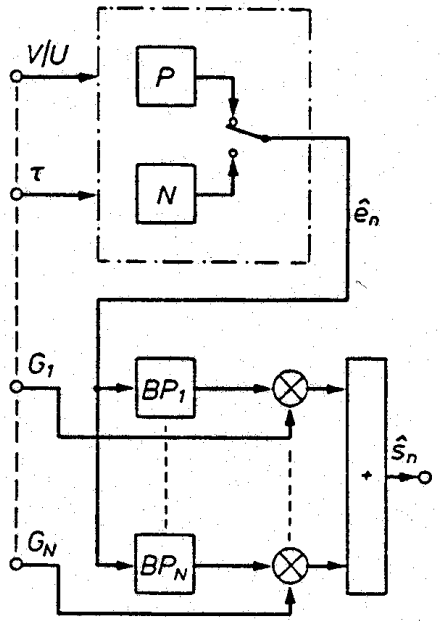
- durch Allpol-Modell (III)
  → LPC-Vocoder

Kanalvocoder:

Sender

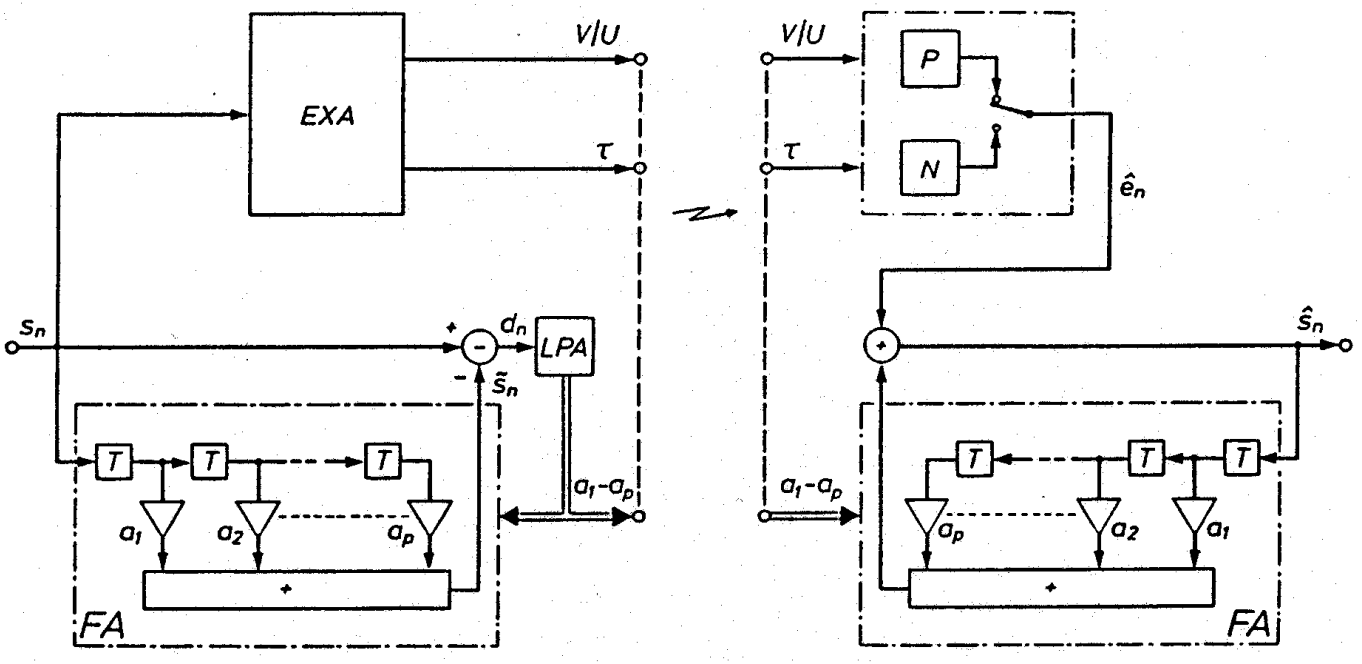


Empfänger

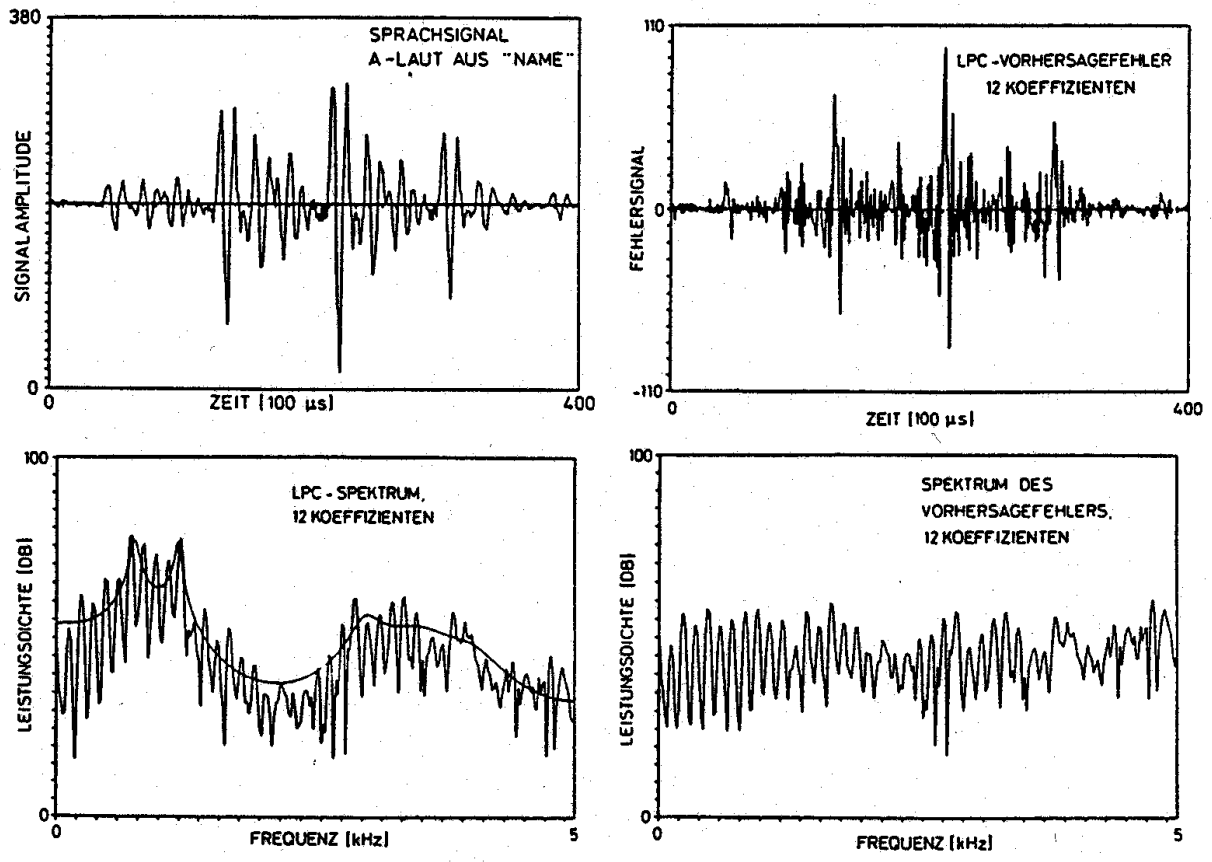


LPC-Vocoder:

Sender Empfänger



LPC-Analyse:


SPRACHSIGNAL
A-LAUT AUS "NAME"
SIGNALAMPLITUDE
380
0
0
ZEIT [100 µs]
400
LPC-VORHERSAGEFEHLER
12 KOEFFIZIENTEN
FEHLERSIGNAL
110
0
-110
0
ZEIT [100 µs]
400
LPC-SPEKTRUM,
12 KOEFFIZIENTEN
LEISTUNGSDICHTE [DB]
100
0
0
FREQUENZ [kHz]
5
SPEKTRUM DES
VORHERSAGEFEHLERS,
12 KOEFFIZIENTEN
LEISTUNGSDICHTE [DB]
100
0
0
FREQUENZ [kHz]
5

Hörproben:

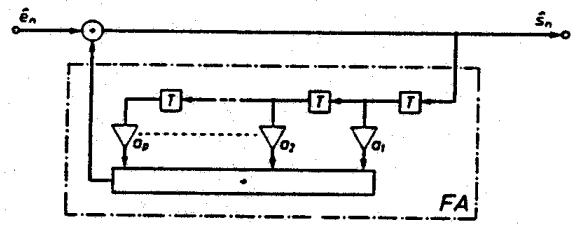
- *Kanalvocoder (Sennheiser)*
  Demonstration verschiedener Experimente und Einsatzmöglichkeiten im künstlerischen Bereich

- *LPC (4 kHz Bandbreite, 12 Prädiktorkoeffizienten)*
  - *Restsignal*
  - *Original*
  - *Restsignal*

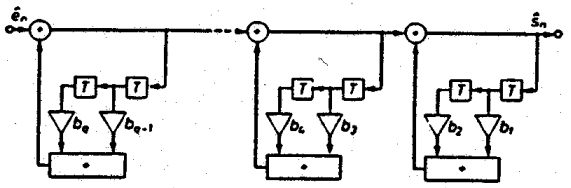
Signalsynthese bei LPC:

Rekonstruktionsfilter kann unterschiedlich realisiert werden, z.B.:

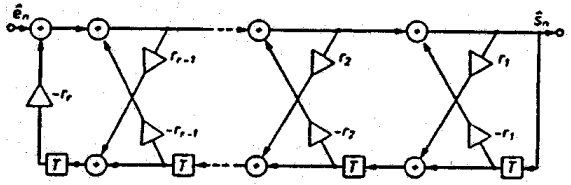
- direkte Filterstruktur
  (Analysefilter in
  Rückkopplungsschleife)



- Kaskadenschaltung von
  Bandpässen 2. Ordnung
  → Formantenfilter



- Kreuzgliedstruktur
  (Lattice-Filter)



Vorteile parametrischer Verfahren:

- niedrige Datenrate: 0,5 ... 5 kbit / s
- stützen sich auf ein Spracherzeugungsmodell und auf
- Parameter, die (mehr oder weniger gut)
  interpretiert und manipuliert werden können
  ↓
  gezielt gesteuert, interpoliert, ...
  ↙ ↘
  z.B.: Sprachgrundfrequenz z.B.: Formantübergänge an Lautgrenzen
  notwendig für Sprachsynthese
- Parameterextraktion ↔ Datenreduktion
  liefert bedeutungsrelevante Information
  → Spracherkennung

Nachteile parametrischer Verfahren:

- algorithmischer Aufwand höher als bei Signalformcodierung,

  insbesondere für Parameterextraktion

- erreichbare Sprachqualität niedriger

  (auch bei Erhöhung der Datenrate)

→ c) hybride Verfahren:

Kombination von Signalformcodierung und parametrischen Verfahren

für mittlere Bitraten (2,4 - 16 kbit /s),

z.B.: LPC + Übertragung des Restsignals

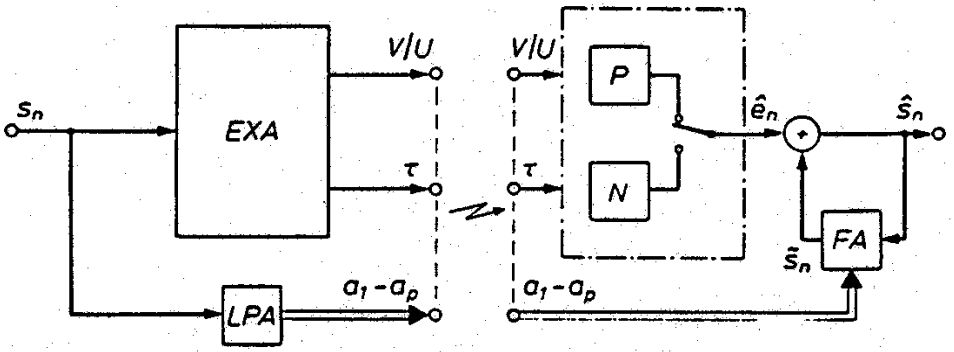
In der Praxis gebräuchlich für Sprachwiedergabe und Sprachsynthese

Zahlreiche standardisierte Varianten für Sprachübertragung (GSM, VoIP,...),

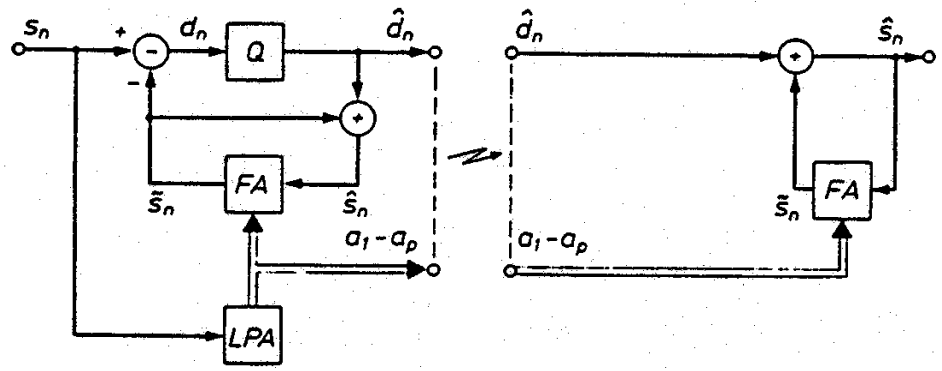
4.1 Sprachwiedergabe

Ansatz für hybride Verfahren:

- LPC:



- DPCM mit adaptiver Prädiktion:



Sprachsynthese:

- Ausgabe beliebiger Meldungen

  „unbegrenztes Vokabular“

  (unrestricted vocabulary speech synthesis)

- liegen i.a. zunächst als geschriebener Text vor

  „Sprachsynthese aus Text“, „Vorleseautomat“

  (text-to-speech synthesis)

→ 2 Teilaufgaben:

a) Signalgenerierung

„Synthese“ im doppelten Sinn:

- Rekonstruktion des Signals
  aus Parametern eines Spracherzeugungsmodells
  (Gegenstück zur Signalanalyse)
- Zusammensetzung lautlicher Segmente
  Hauptschwierigkeit: Übergänge, Koartikulation

b) Textumsetzung

Bestimmung von

- Aussprache und
- Prosodie

a) Signalgenerierung

Signalrekonstruktion

- wie bei Sprachwiedergabe mit parametrischem Verfahren
- Parameter i.a. zunächst aus natürlicher Sprache extrahiert
  (aus geeigneten Musterwörtern)

aber:

- gespeichert für kleinere lautliche Einheiten,
- zusätzlich Modifikationen beim Zusammensetzen,
  Steuerung prosodischer Parameter (Grundfrequenz, Dauer, ...)

Zusammensetzung lautlicher Segmente

- Grundidee:
  aus begrenztem Inventar von Segmenten
  lassen sich beliebige Wörter aufbauen

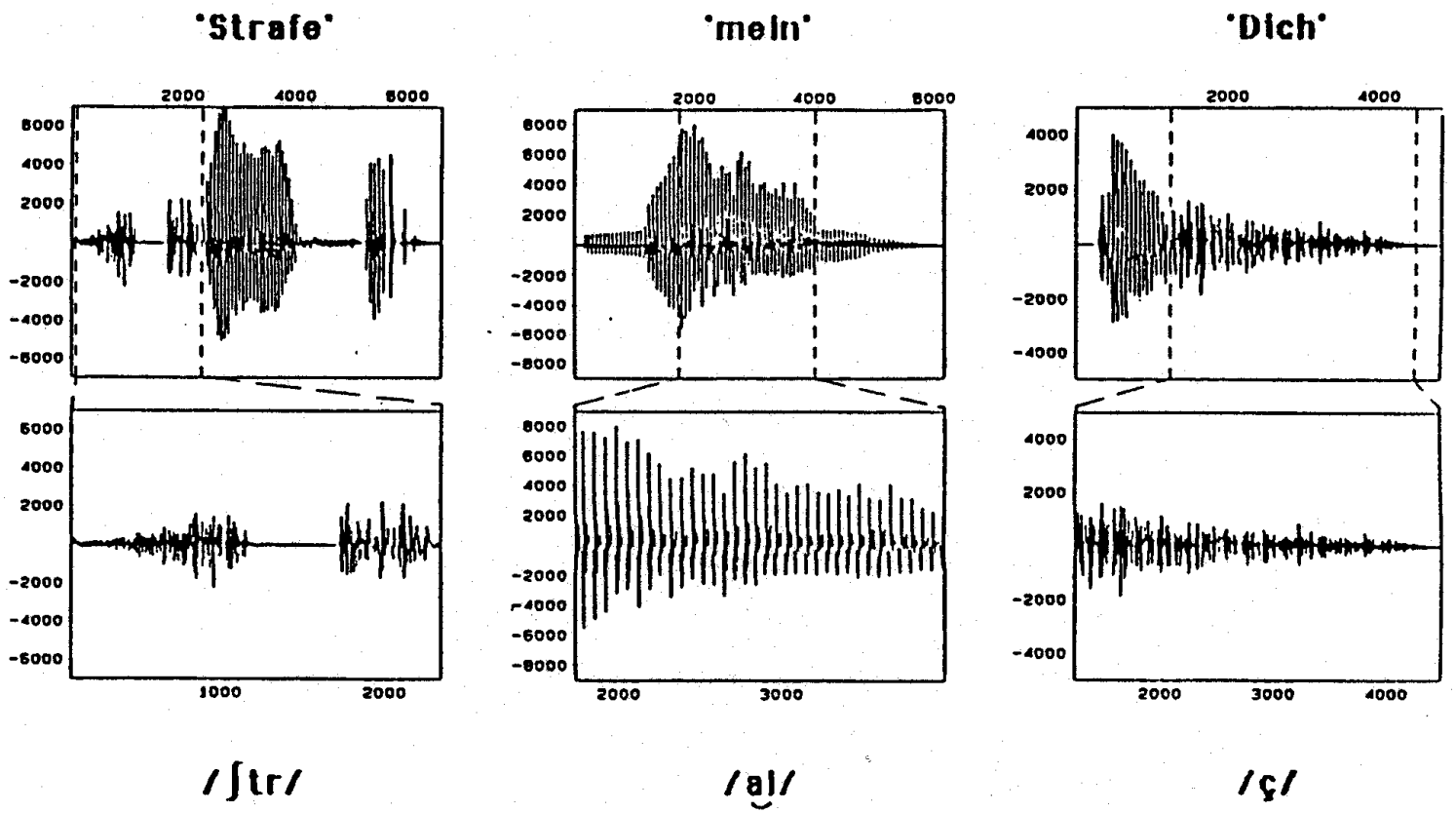
- ca. 40-50 Phoneme genügen, um alle unterschiedlichen
  Wörter einer Sprache zu beschreiben

  → „Phonemsynthese“, besser: Einzellautsynthese

- auch andere (größere) Einheiten begrenzter Anzahl sind möglich,
  z.B.: Doppellaute, Konsonanten- oder Vokalfolgen, Halbsilben, ...

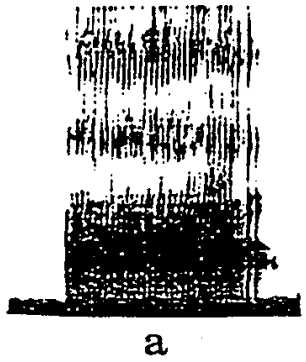
4.2 Sprachsynthese

Zusammensetzung lautlicher Segmente zu einem neuen Wort:

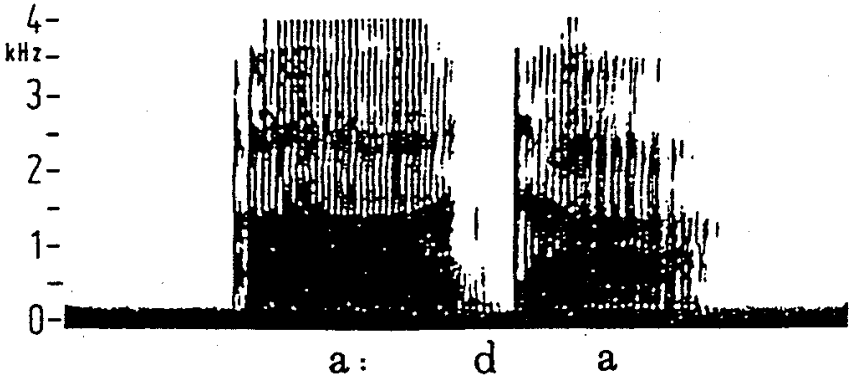


Problem: Übergänge, Koartikulation

z.B.: isoliert gesprochener Vokal



im Kontext:

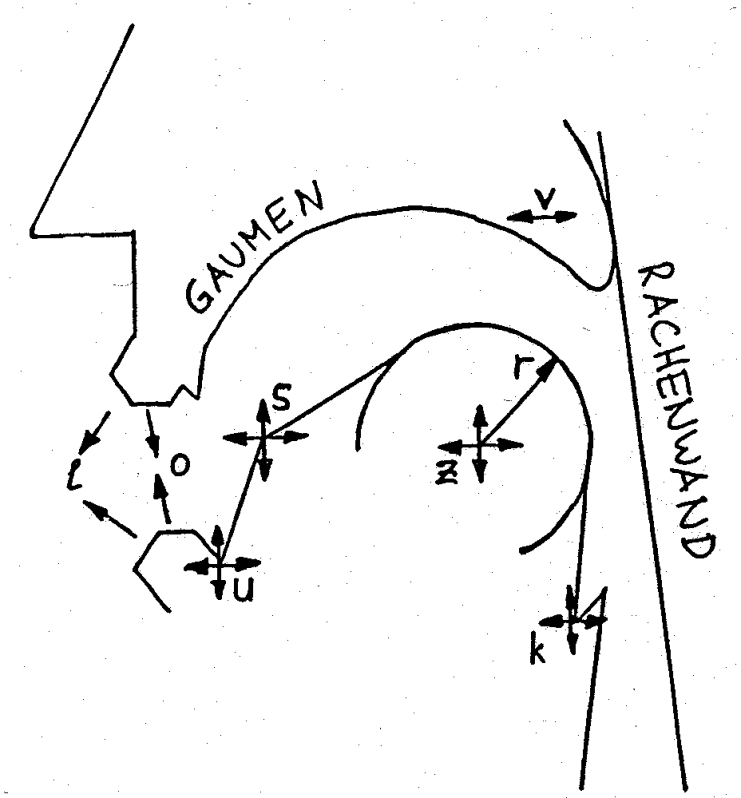


Abhilfe:

- Berücksichtigung lautlicher Varianten (kontextabh.)
  → Allophonsynthese (~ 100 Einheiten)
- Speichern der Lautübergänge
  → Diphonsynthese (~1500 Einheiten)
  z.B.:
- Verwendung größerer Segmente, z.B.:
  → Halbsilbensynthese (~2000 Einheiten)
- algorithmische Beschreibung koartikulatorischer Effekte,
  z.B.: Glättung von Parameterverläufen, bes. an Segmentgrenzen,
  setzt geeignete Parameter voraus (z.B. Formanten)
  → regelgesteuerte Synthese
- Einbeziehung eine höheren Modellebene:
  Simulation der Sprechbewegungen
  → artikulatorische Synthese

Artikulatorisches Modell zur Sprachsynthese

k Kehlkopfdeckel
l Lippenvorstülpung
o Lippenöffnung
r Zungenradius
s Zungenspitze
u Unterkiefer
v Velum
z Zungenkörper



4.2 Sprachsynthese

Aktuelle Synthesesysteme: Das „Unit Selection“-Prinzip

- Große Datenbanken (z.B. 300MB) von Sub-Phon-Einheiten als Material für konkatenative Synthese
- Jedes Sub-Phon wird in vielen Varianten gespeichert, mit unterschiedlichen prosodischen und Signaleigenschaften
- Zur Laufzeit ermitteln effiziente Suchverfahren mit Hilfe einer global zu optimierenden Kostenfunktion jene Einheiten, die für eine Zieläußerung am besten zusammengefügt werden können
  - Kriterien: Nähe zu den Zielparametern,
    glatte Übergänge mit verfügbaren Nachbareinheiten
- Die Originalsignale im Einheiteninventar müssen so nicht oder nur wenig verändert werden, daher natürlicherer Klang
- Variable Size Unit Selection: Enthält auch größere Einheiten, etwa häufige Funktionswörter

b) Textumsetzung (Transkription)

Eingabe: Text in orthographischer Form

⇓

Ausgabe: Lautschrift + zus. Marker für Prosodie

Aussprache, Wortakzent

Akzent, Rhythmus, Melodie auf Satzebene

Zusätzliche Aufgaben:

- Textvorverarbeitung (Abkürzungen, Zahlen, Sonderzeichen, ...)
- Benutzerdialog, Interface zum Synthetisator, ...

Bestimmung von Aussprache und Wortakzent

„Transkription im engeren Sinn“,

„Graphem-Phonem-Umsetzung“

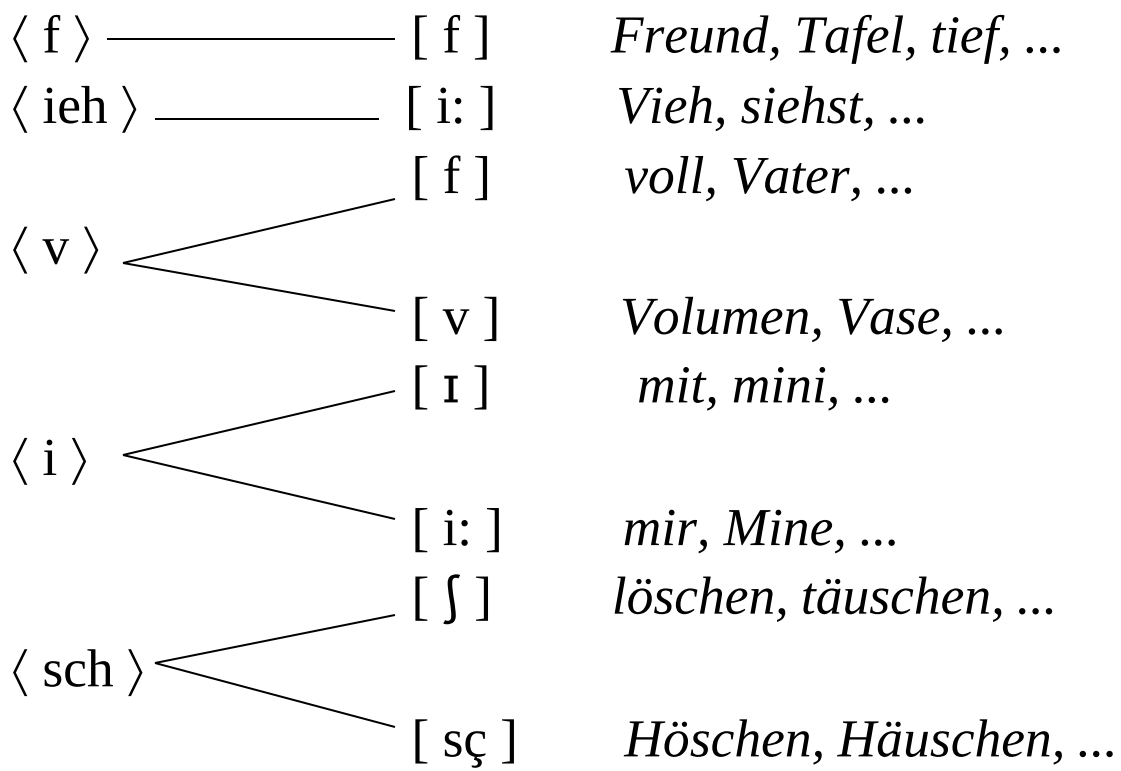
- Schrift (Orthographie) spiegelt gesprochene Sprache wider,
  aber z.B. auch Herkunft eines Wortes

$\Rightarrow$ Zuordnung Buchstaben $\rightarrow$ Laute

i.a. nicht eindeutig

4.2 Sprachsynthese

z.B. (dt.):



in anderen Sprachen z.T. noch schlimmer,

z.B. engl.: ⟨ ea ⟩ in *head*, *heat*, *hear*, *heard*, *heart*, ...

Prinzipiell mögliche Verfahren:

- Lexikon

  z.B. (engl.): head [hed]
  heat [hi:t]
  hear [hiə]

  aber: kein Lexikon ist komplett !

- Regelkatalog

  z.B. (dt.): ⟨ a ⟩ vor Doppelkonsonant → [a] (Wall)
  ⟨ a ⟩ vor zwei Konsonanten → [a] (Wald)
  ⟨ a ⟩ vor einzelnem Kons. → [a:] (Wal)

  aber: keine Regel ohne Ausnahme !

Beispieltext:

Das Büro der Zukunft:

Kein Papiertiger

Mehr als die Hälfte aller unselbständig Erwerbstätigen ist in Büros beschäftigt. In der Fertigung wird seit Jahren wirkungsvoll rationalisiert. Der Mitarbeiter-Anteil in den Büros wächst. Ebenso wächst die Informationslawine. Es gilt dem steigenden Informationsfluß und der anfallenden Routinearbeit mit Hilfe des Computers Herr zu werden. Bislang ist aber nur jeder 20. Arbeitsplatz EDV-unterstützt.

Besonderheiten im Deutschen:

Kenntnis der Wortstruktur ist Voraussetzung für

- Anwendung der Ausspracheregeln

  z.B.:

kalt	*malt*
besteingefahren	*besteigen*

- Bestimmung des Wortakzentmusters

  z.B.:

geben	*gebeten*

- Wortartbestimmung $\rightarrow$ Satzanalyse $\rightarrow$ Prosodiesteuerung

4.2 Sprachsynthese

Verschiedene Ansätze in deutschsprachigen Systemen:

- Umsetzung nach Regeln

  +

  kleine Lexika

  für häufige Ausnahmen

  und Funktionswörter

- Abtrennung von Präfixen, *(ge-, ver-, ...)*

  Derivationssuffixen *(-ung, -bar, ...)*

  und Flexionsendungen *(-t, -en, ...)*

  mit Affixlisten

- Zerlegung von Komposita *(Sandstrand)*

  mit Listen von Konsonantenverbindungen und deren Häufigkeiten: „Klusteranalyse“

Literatur:

Fellbaum, K.:
http://www.kt.tu-cottbus.de/teleteaching/

Hasegawa-Johnson, M.:
http://www.ifp.uiuc.edu/~hasegawa/notes/index.html

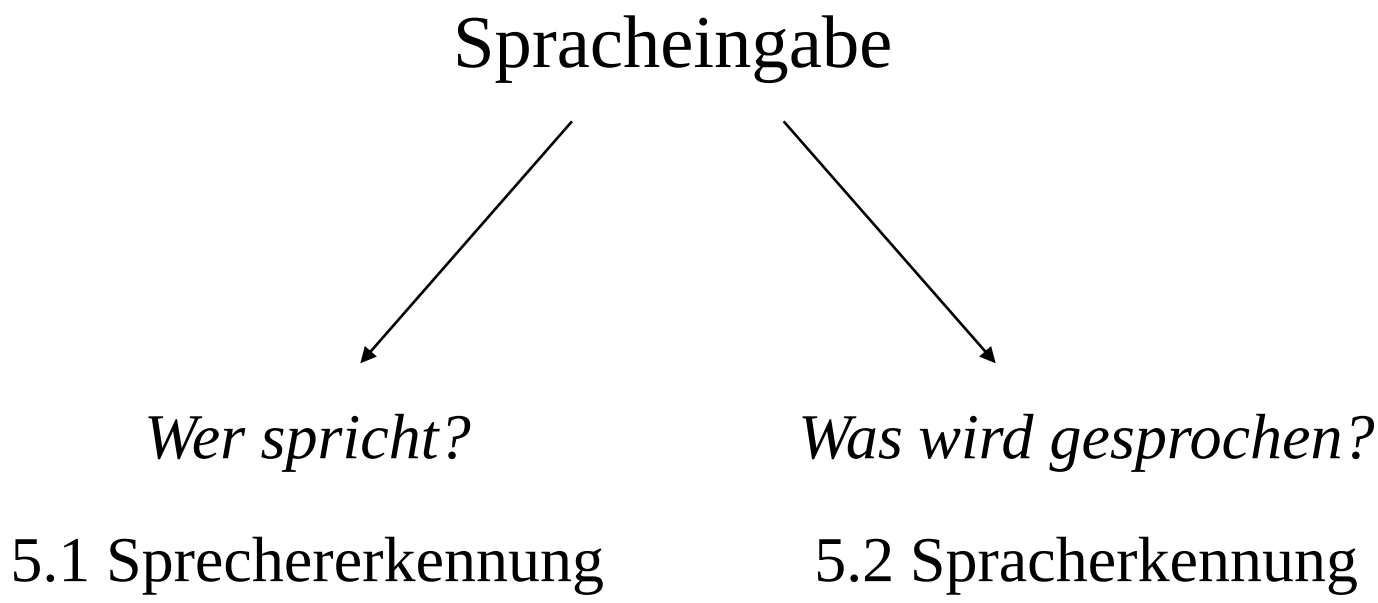
O´Shaughnessy, D.: SPEECH COMMUNICATION (Second Edition)
Reading (MA), ...: Addison-Wesley Publishing Company (2000)

Sluyter, R.J.: DIGITALIZATION OF SPEECH
Philips Technical Review, Vol.41, Nr. 7/8 (1983/84), S. 201-223

Huang, X. et al: Spoken Language Processing, Theory, Algorithms, System Development. Prentice Hall (2001)

Hunt, A. J. & A. W. Black: „Unit Selection in a Concatenative Speech Synthesis System using a Large Speech Database“, Proc. ICASSP-96, Atlanta (1996).
http://www.cs.cmu.edu/~awb/papers/icassp96.ps

Keller, E. et al. (Hg.): IMPROVEMENTS IN SPEECH SYNTHESIS.
Cost 258: The Naturalness of Synthetic Speech.
New York,...: John Wiley (2002)


Spracheingabe
Wer spricht?
Was wird gesprochen?
5.1 Sprechererkennung
5.2 Spracherkennung

Drei Verarbeitungsschritte:

1. Signalvorverarbeitung
    - Digitalisierung und Analyse des Sprachsignals
    - Merkmalsgewinnung
2. Mustervergleich
3. Klassifizierung

5.1 Sprechererkennung

a) Sprecheridentifikation

v.a. in der Kriminalistik, z.B. bei erpresserischen Anrufen

- Sprecher unbekannt (mehrere Kandidaten)
- nicht kooperativ

b) Sprecherverifikation

z.B.: Zutrittskontrolle, telefonische Konto-Auskunft

- Mustervergleich mit nur einem Kandidaten,
  ausgewählt durch persönliches Kennzeichen:
  „Identitätsziel“ (z.B. Name, Codewort)
- kooperativer Sprecher

Gegenüberstellung:	**Identifikation**	**Verifikation**
Typische Anwendung	telefonische Erpressung	telefon. Kontoauskunft
Identitätsziel	unbekannt	bekannt
Kooperationsbereitschaft	nicht vorhanden	vorhanden
Sprachmaterial	beliebiger Text	vorgegebener Text (Codesatz)
Auswertung	textunabhängige automatische Verarbeitung	automatisch (Vergleich von zwei Texten gleichen Inhalts)
Entscheidung	Text stammt mit Wahrscheinlichkeit p_n von Sprecher n	beide Texte stammen vom gleichen Sprecher oder nicht
Erkennungsergebnis abh. von der Sprecheranzahl	in hohem Maße	ja
Stand der Technik	praktisch einsetzbar	einsetzbar (in Verbindung mit anderen Verfahren)

Signalvorverarbeitung

- Ziele: - Datenreduktion
  - Extraktion Sprecher-relevanter Eigenschaften
  - Elimination Text-relevanter Eigenschaften

  in Lernphase und Arbeitsphase
- meist spektrale Analyse

  z.B. mit Filterbank (vgl. Kanalvocoder),

  $\rightarrow$ Matrix R von Parametern

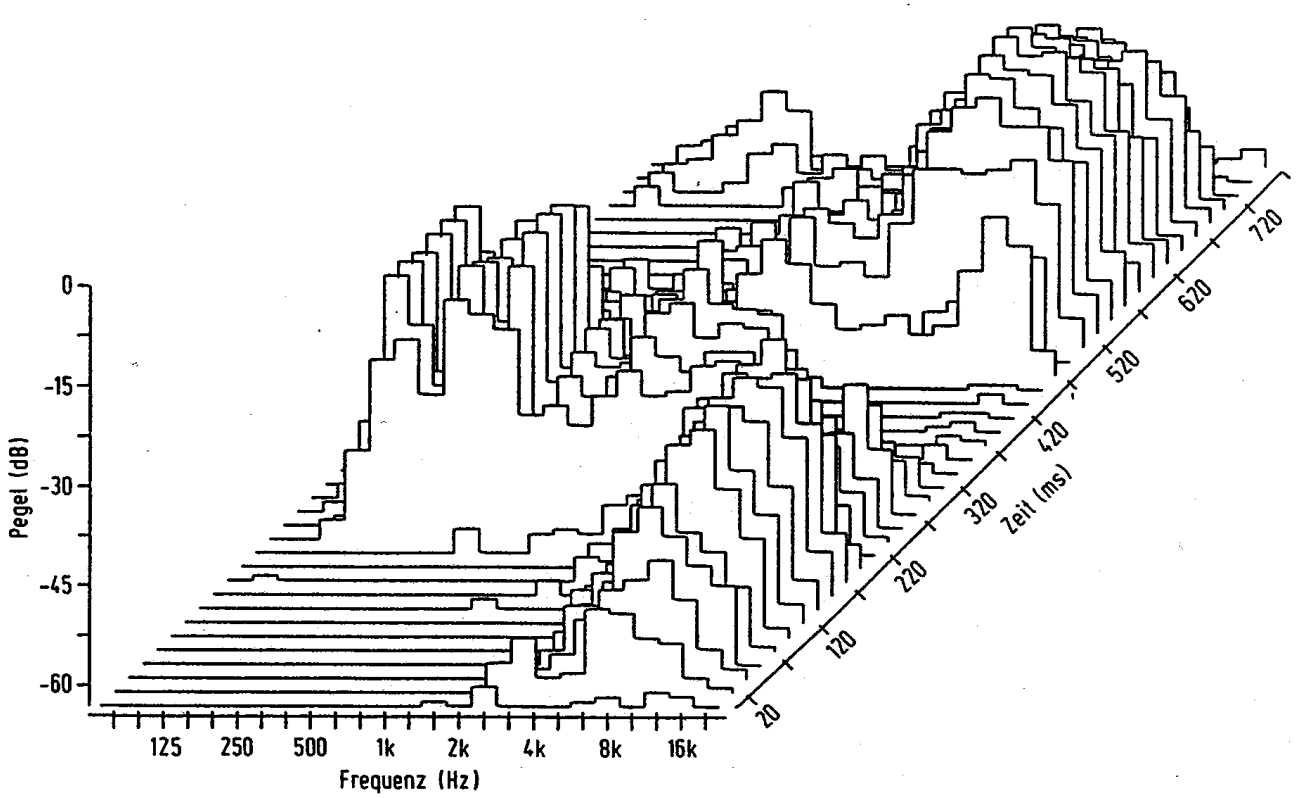
  r_{nm}: Ausgang des Kanals m zum Zeitpunkt n T
- weitere Verarbeitung: Merkmalsgewinnung

  z.B. zeitliche Mittelung, evtl. Auswahl spezieller Segmente (z.B.[n], [m])

  $\rightarrow$ Merkmalsvektor X

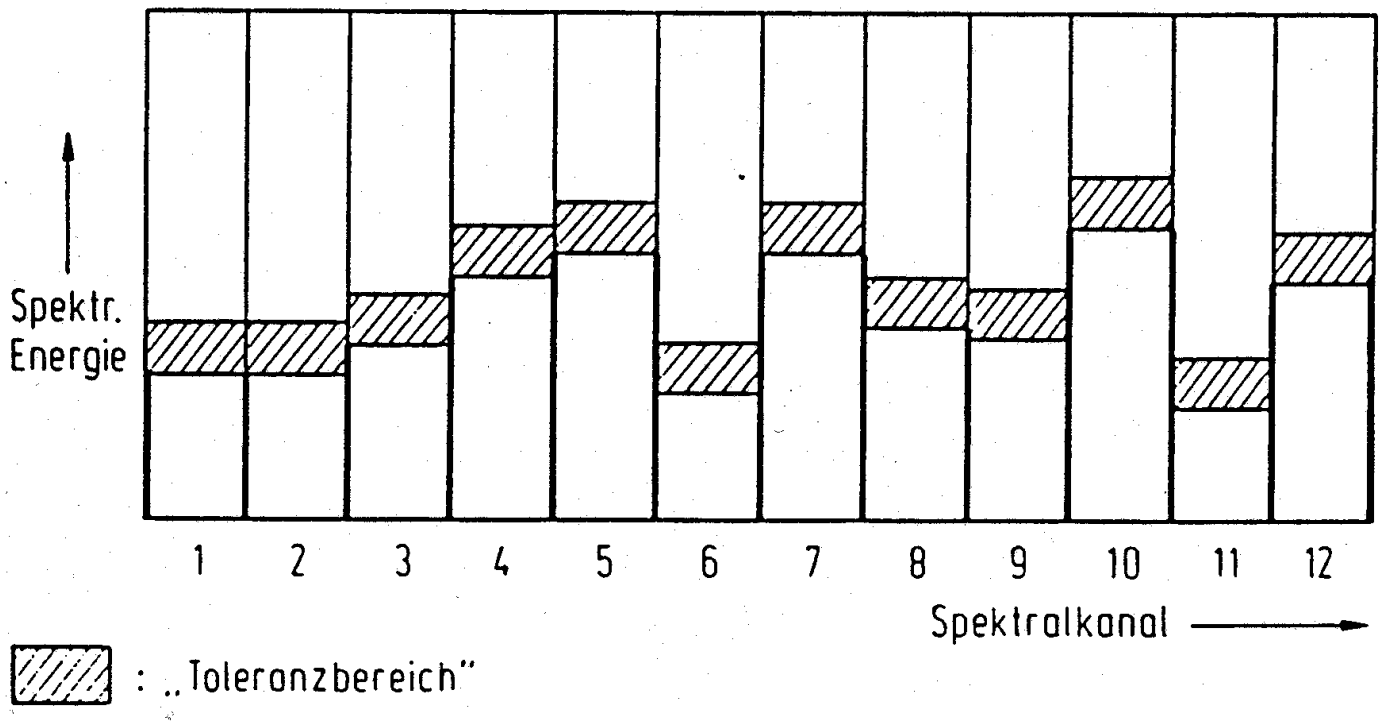
Parametrische Beschreibung: Matrix R

für das Musterwort *sechs*, in dreidimensionaler Darstellung



zeitliche Mittelung

→ Referenz für Merkmalsvektor X:



Mustervergleich

aktueller Merkmalsvektor ↔ Referenzvektor(en)

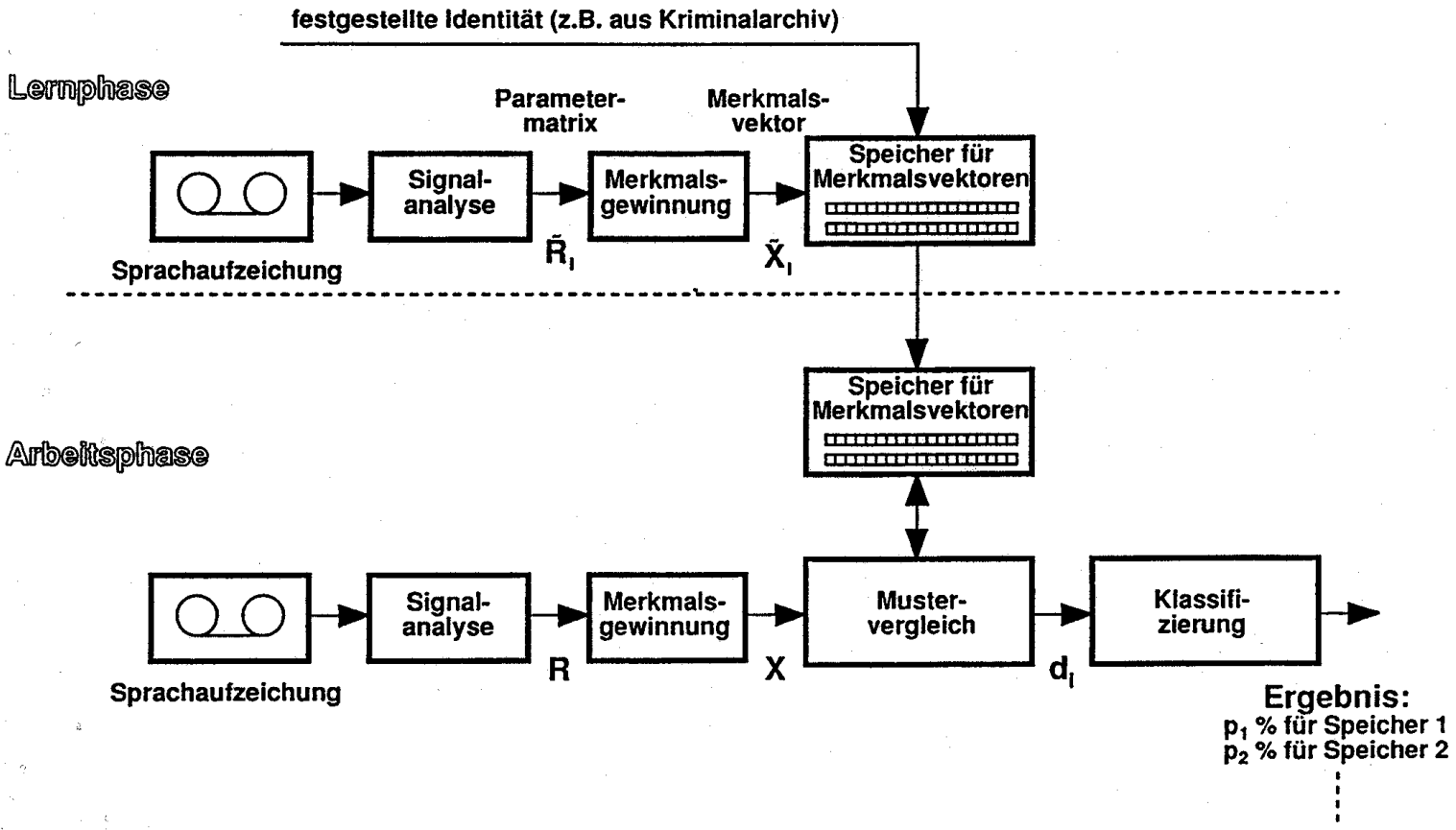
- Vergleichsmaß i.a. euklidischer Abstand

Klassifizierung

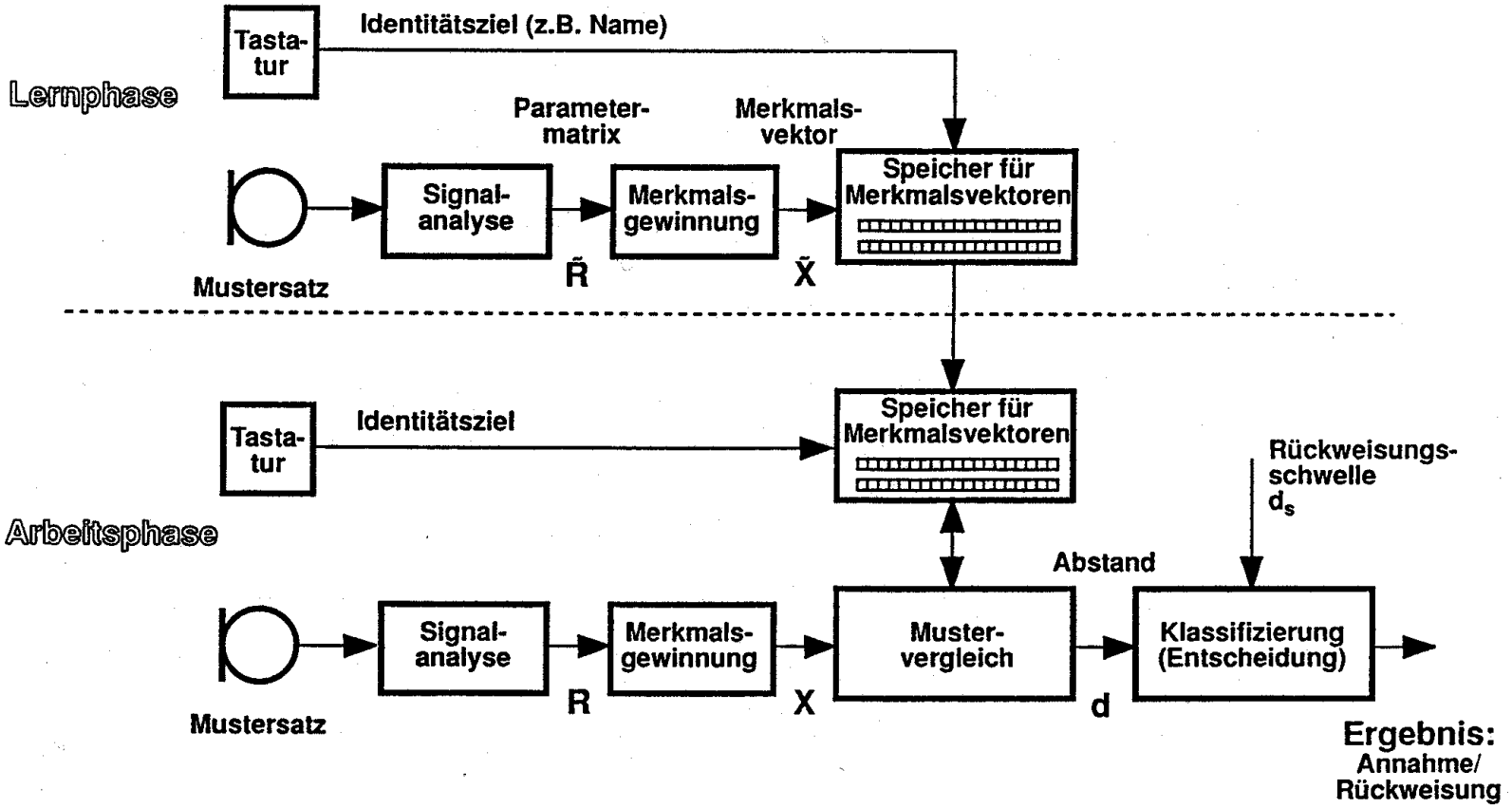
- bei Verifizierung:
  Rückweisungsschwelle für Abstand

- bei Identifizierung:
  Suche des Kandidaten mit minimalem Abstand,
  übrige Abstände geben Aufschluss über Zuverlässigkeit

Identifikationssystem:



Verifikationssystem:



Beurteilung von Systemen zur Sprecherverfikation:

1. Falsche Akzeptanz (False Acceptance Ratio)

   gibt an, wie oft ein nicht berechtigter Sprecher für den berechtigten Sprecher gehalten wird

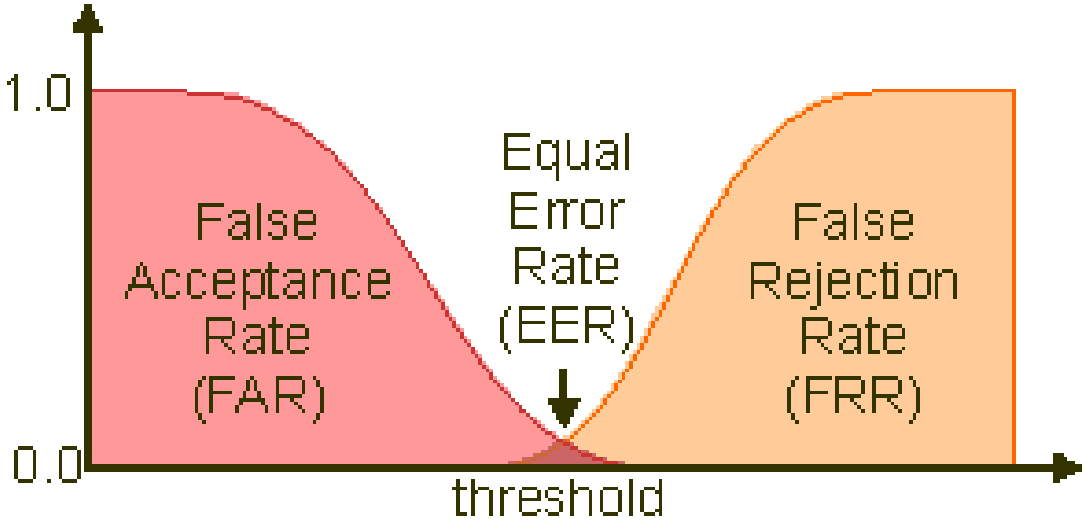
2. Falsche Zurückweisung (False Rejection Ratio)

   gibt an, wie oft einem berechtigten Sprecher der Zugriff verweigert wird, weil er nicht als berechtigter Sprecher erkannt wird

Ein Maß kann auf Kosten des anderen erhöht werden.

Equal Error Rate:

jene Rückweisungsschwelle, bei der falsche Akzeptanz und falsche Zurückweisung gleich häufig sind



5.2 Spracherkennung

Unterschiedliche Zielsetzungen möglich:

Gesprochene Sprache soll

- interpretiert werden und einen Prozess auslösen:

  akustische Steuerung oder Befehlseingabe, Sprachwahl im Handy

- erkannt und (z.B. als Schrifttext) ausgegeben werden:

  Datenerfassung, z.B. akustische Eingabe von Zahlen- oder Wortlisten Diktiersysteme

- verstanden und beantwortet werden:

  akustischer Dialog, z.B. Auskunftssystem, Flugreservierung

5.2 Spracherkennung

	Einzelwort-erkennung	Diktiersystem	Dialogsystem
Sprecher-abhängigkeit	sprecherabhängig (muss für jeden Sprecher neu trainiert werden)	sprecherabhängig bzw. sprecheradaptiv	sprecherunabhängig
Größe des Vokabulars	~ 200 Wörter	60.000 Wörter und mehr, die immer aktiv sind	einige 1000 Wörter, von denen immer nur eine Teilmenge aktiv ist
Art der Eingabe	isoliert gesprochen	unbeschränkt, auch komplexe Sätze sind möglich	nur bestimmte Muster werden bei jedem Dialogschritt erkannt

Eigenschaften von Spracherkennungssystemen

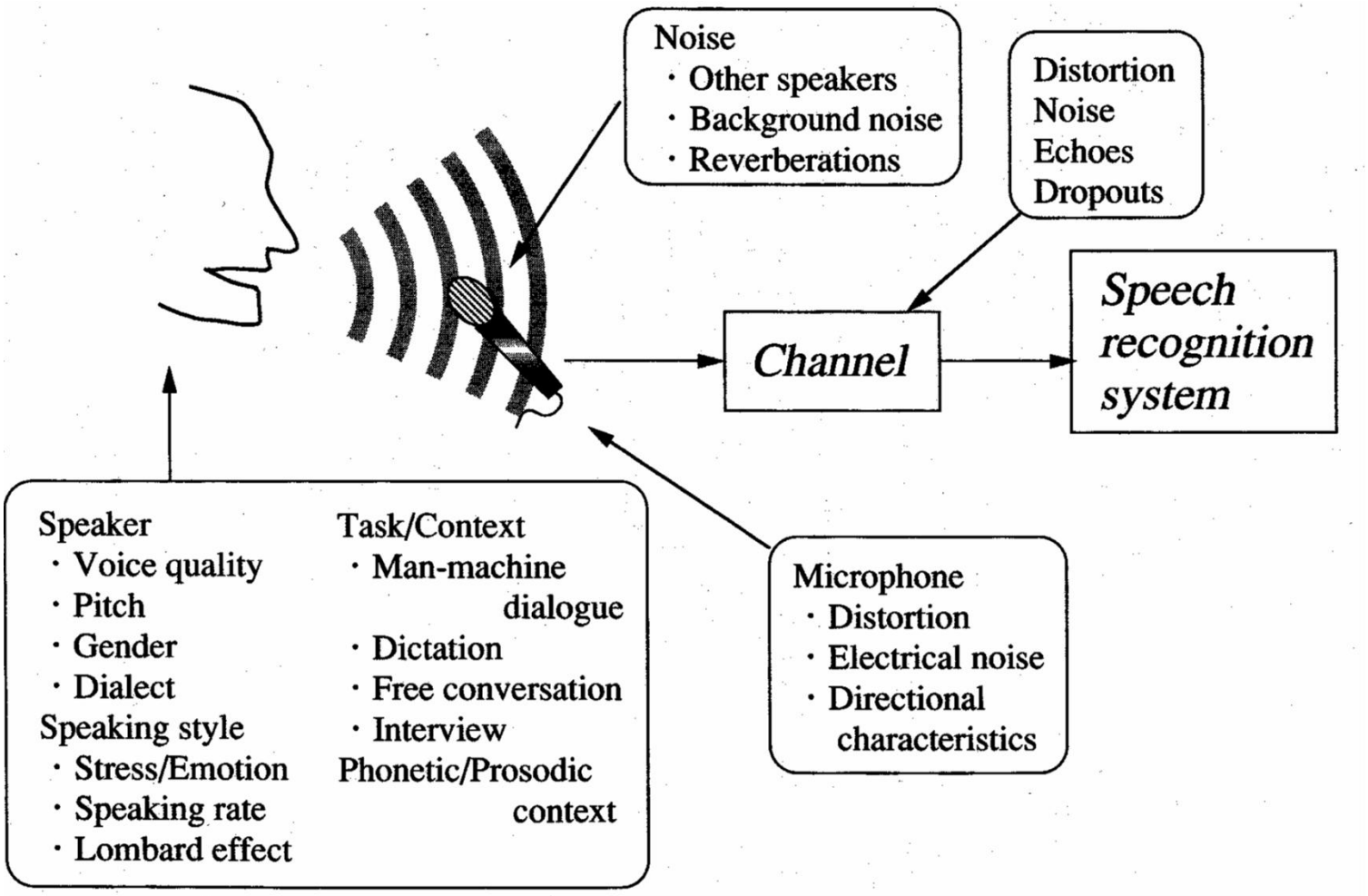
- Sprecherabhängig / sprecherunabhängig / sprecheradaptiv
- Größe des Vokabulars
- Einzelworterkennung / kontinuierliche Sprache / Spontansprache
- Echtzeit / größerer Zeitbedarf / Zeitverzögerung
- Qualität der Eingabe (Qualitätsmikrofon / Telefon / Handy)
- Inkrementelle Ausgabe von Zwischenergebnissen
- Bedarf an Rechenleistung, Festspeicher und Arbeitsspeicher
- Zuverlässigkeit (Fehlerrate)

Vorteile	*Anwendungsbeispiele*
- natürliche Kommunikationsform	*Programmbedienung, Spiele*
- Hände und Augen entlastet	*Paketverteilung, Qualitätskontrolle, Auto, Flugzeugcockpit*
- Bediener kann sich frei bewegen	*Lagerverwaltung*
- Fernbedienung über Telefon möglich	*Bestell-, Buchungssysteme, Fernabfrage (Messwerte, ...)*
- Hilfe für Behinderte	*Gerätebedienung, sprachgesteuerter Rollstuhl*
- Ersatz für Tasten und mechanische Schalter, Miniaturisierung	*Dateneingabe in staubiger oder feuchter Umgebung, Teilnehmerwahl (Telefonzelle, Handy)*
- schnell, direkt	*Notschalter, Diktierschreibmaschine*

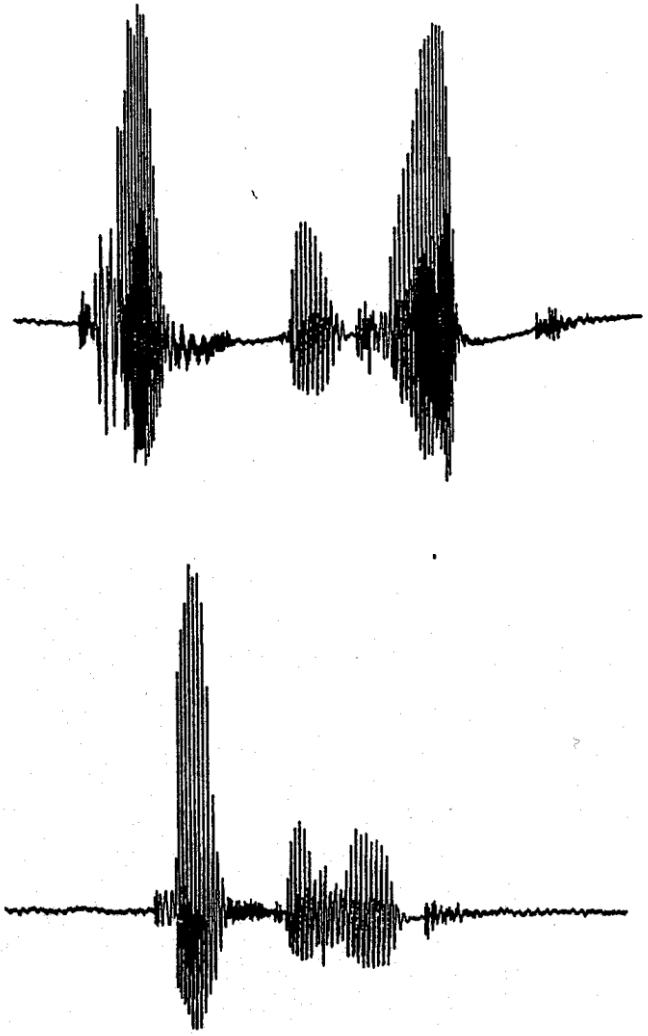
Warum ist Spracherkennung so schwierig?

- ein Wort (Satz, ...) $\leftrightarrow$ viele verschiedene akustische Realisierungen
  Intra-Sprecher-Variabilität - Inter-Sprecher-Variabilität
  (Alter, Geschlecht, Dialekt, Sprechsituation, Stimmung, Gesundheit...)
- isoliert gesprochene Wörter - zusammenhängende Sprache
- Vokabulargröße
- Spontansprache
- Umgebungsgeräusche, weitere Sprecher
- gestörte Sprachübertragung

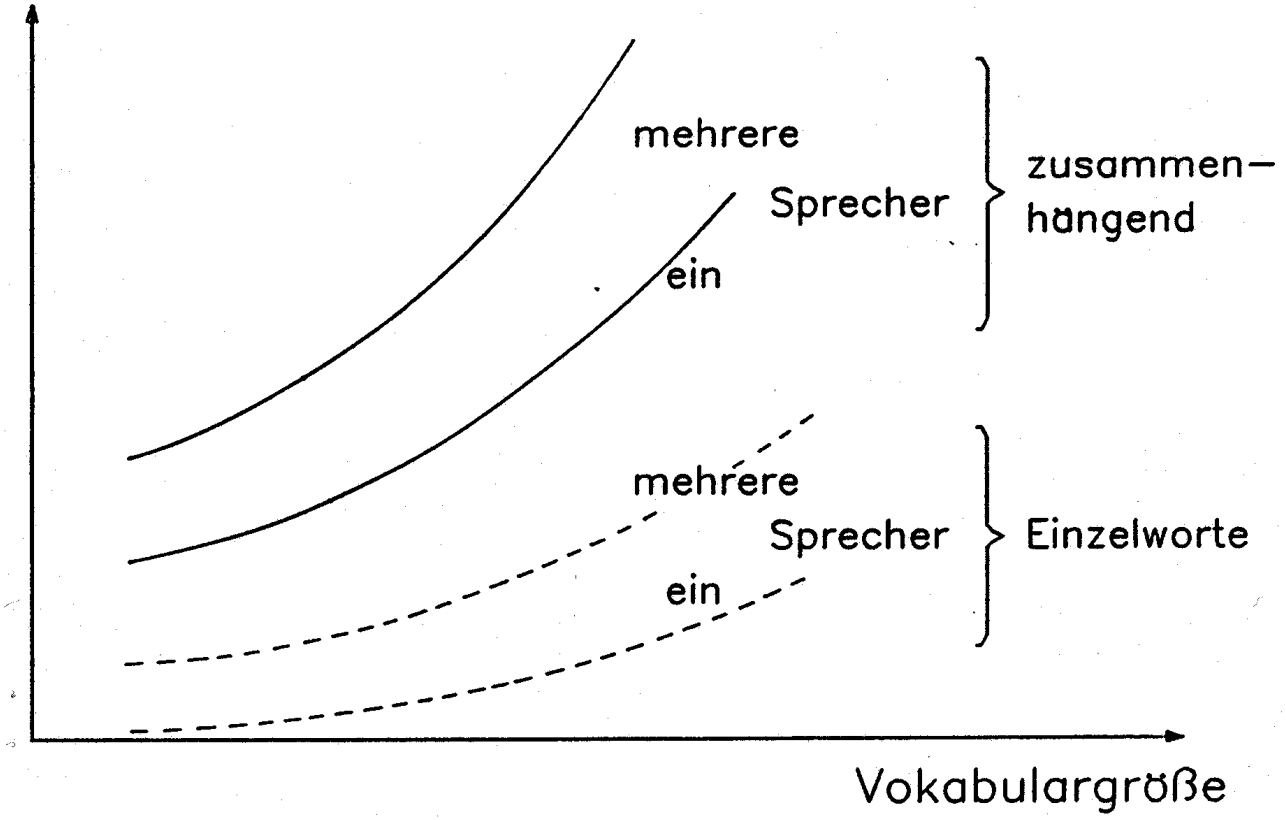
5.2 Spracherkennung



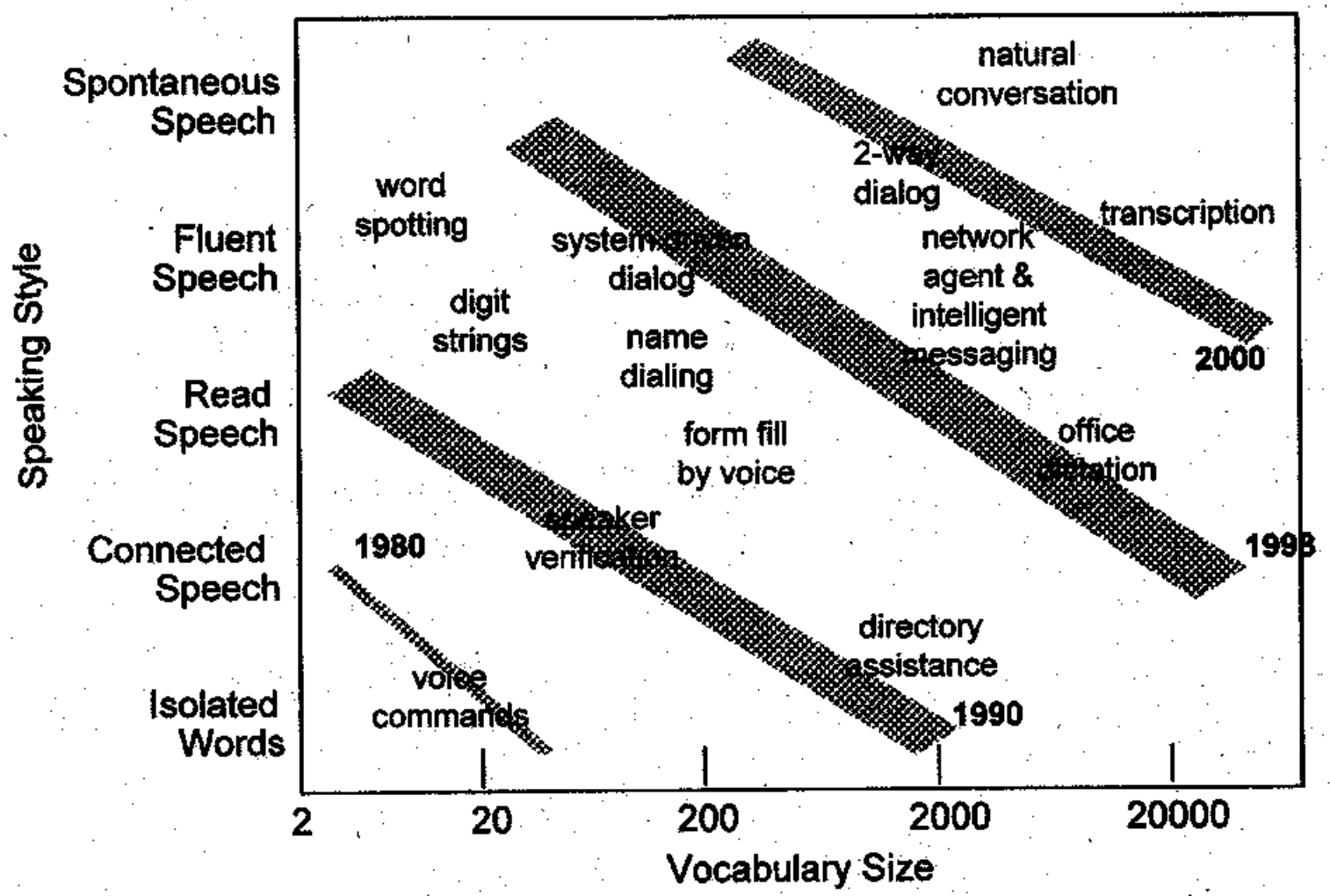
zwei Proben des Worts
„Tastendruck“
vom selben Sprecher:



Komplexität von Spracherkennungssystemen



Fortschritt bei Spracherkennungsanwendungen



Einzelworterkennung

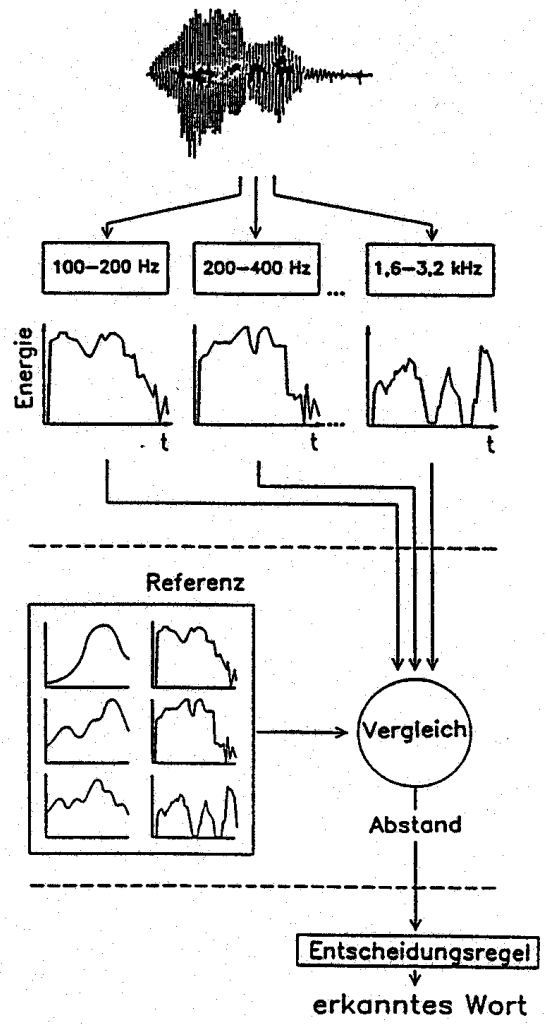
1. Signalvorverarbeitung
    - Analyse des Sprachsignals,
    - Elimination von sprecherspezifischen Eigenschaften,
    - Extraktion laut-typischer Merkmale
2. Mustervergleich
    - mit zeitlicher Anpassung des Testmusters an die Referenz
3. Klassifizierung

5.2 Spracherkennung

Signal-
vorverarbeitung
z.B.:

Mustervergleich

Klassifizierung



Signalvorverarbeitung - Aufgaben und Ziele

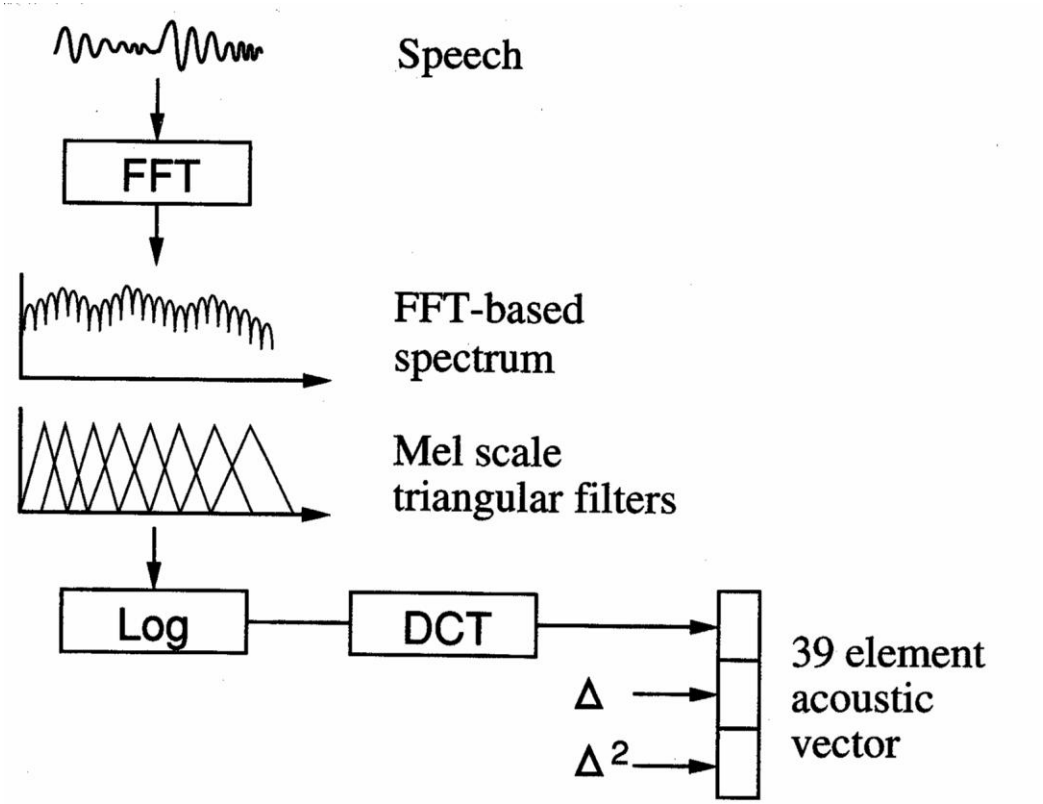
- Datenreduktion
- Parameterextraktion
- Gewinnung zuverlässiger Merkmale
- Normalisierung
- Detektion von Wortanfang und -ende
- Störgeräuschunterdrückung

Analyseparameter: zahlreiche Möglichkeiten, z.B.:

- Energie, Nulldurchgangsrate, Grundfrequenz
- Autokorrelationskoeffizienten
- Prädiktorkoeffizienten
- Kurzzeitspektrum
- Frequenzgang des „Inversen Filters“ (vgl. LPC-Analyse)

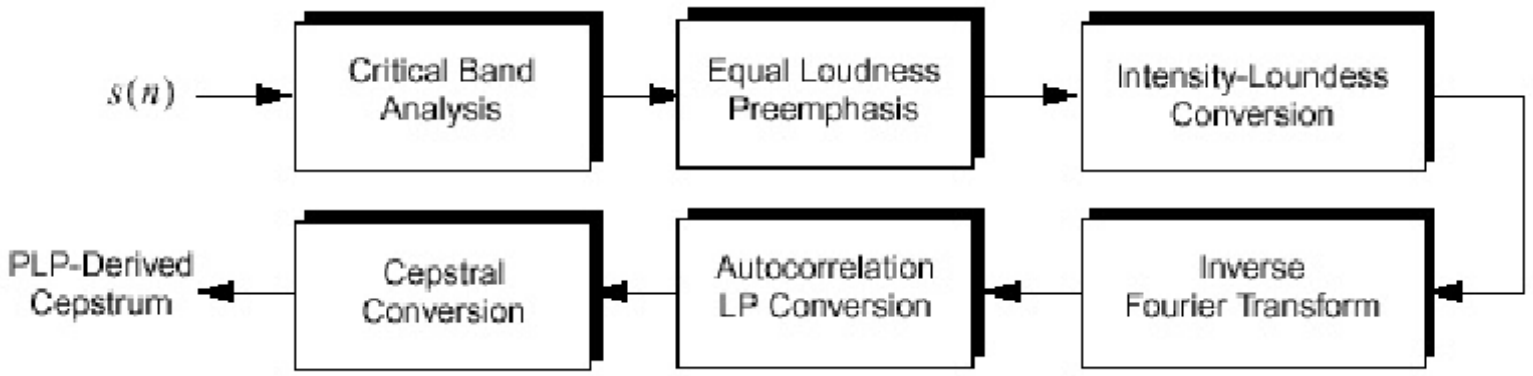
Aktuell gebräuchliche Merkmalssets (I):

Mel Frequency Cepstral Coefficients (MFCC)



Aktuell gebräuchliche Merkmalssets (II):

Perceptually Linear Prediction (PLP)



Mustervergleich

- Abstandsbestimmung:

  über zeitlichen Verläufen der Merkmalsvektoren (Matrizen) für aktuelles Signal und eine Referenz

  z.B. Euklidischer Abstand

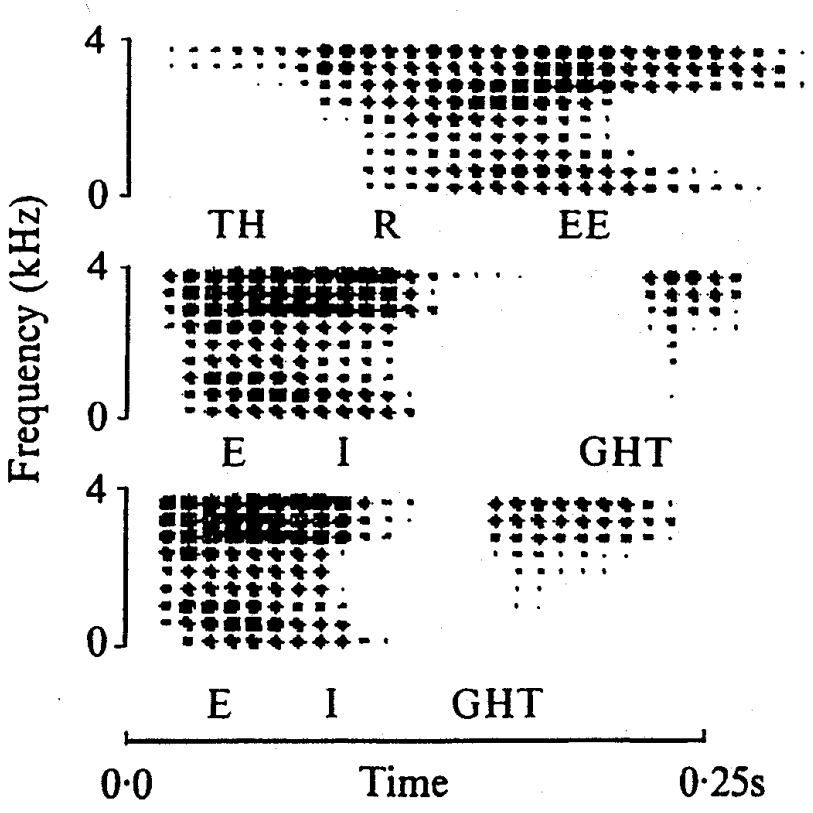
- Hauptproblem: Zeitanpassung
- Lösungsansätze
  - Dynamic Time Warp (DTW)

    (Dynamische Programmierung)
  - Hidden Markov Models (HMMs)
  - Neural Networks

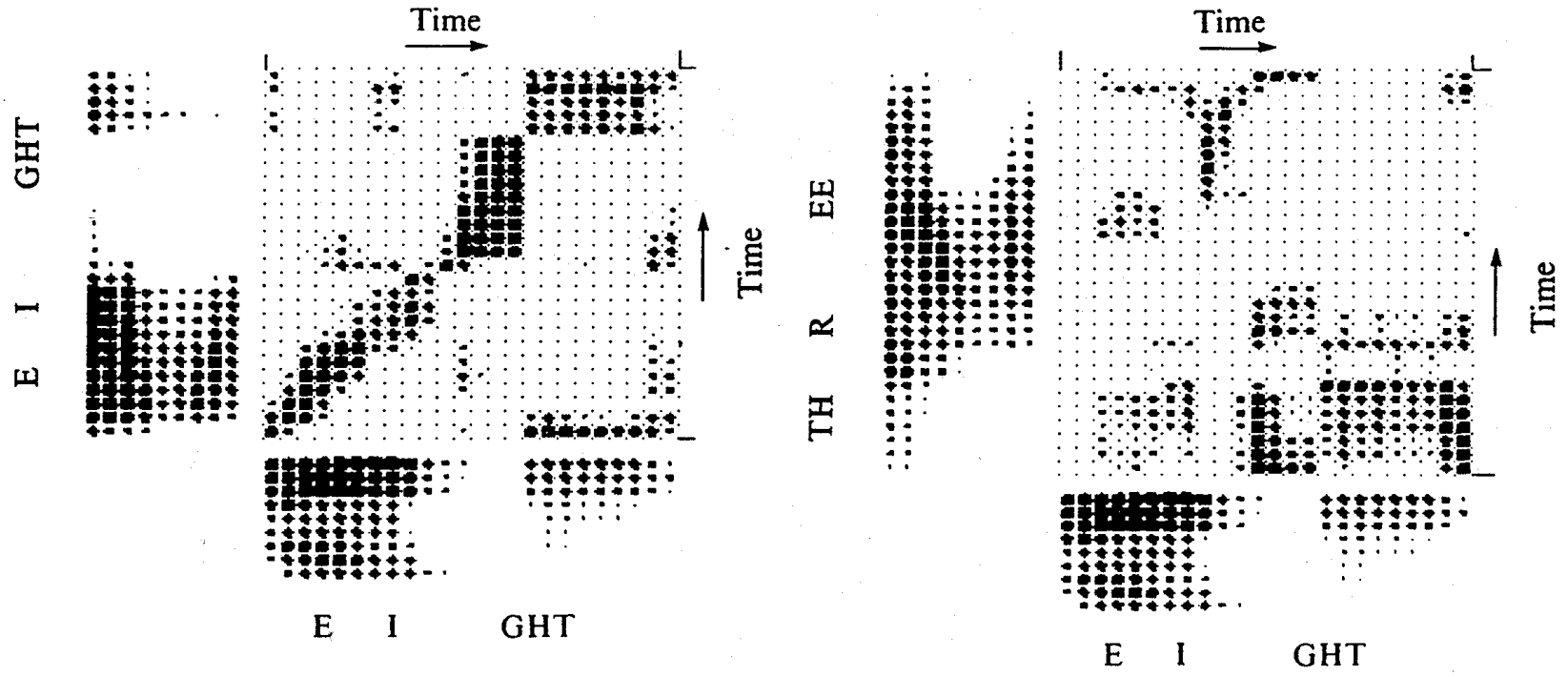
Kurzzeitspektrogramme

(Filterbank mit 9 Kanälen)

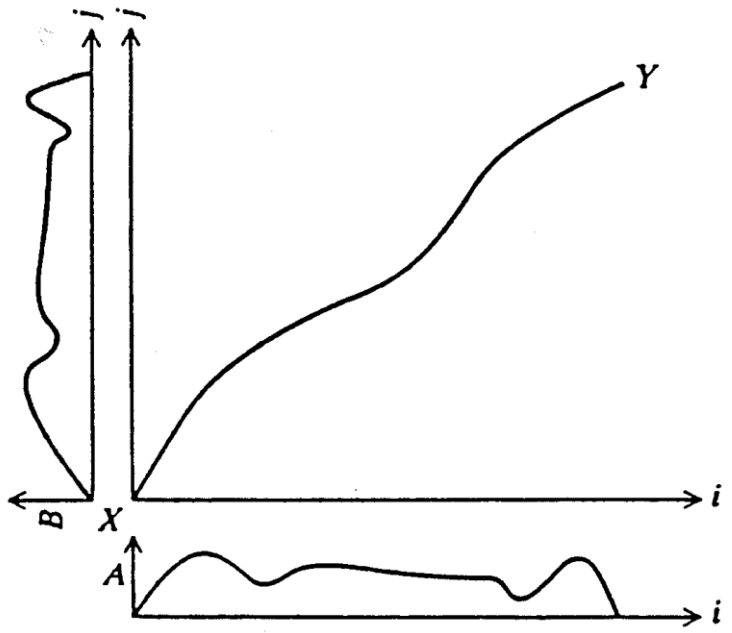
z.B.:



Euklidischer Abstand (Schwärzung ≙ Ähnlichkeit)



Dynamic Time Warp (DTW)



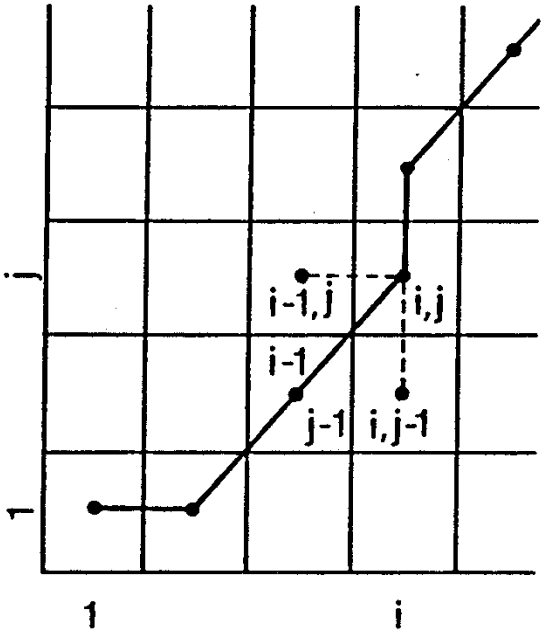
Verzerrungsfunktion weicht mehr oder weniger stark von einer Geraden zwischen den detektierten Anfangs- und Endpunkten ab

Prinzip der dynamischen Programmierung:

- Voraussetzung: diskrete Entscheidungsstufen mit trennbaren Kosten
- Algorithmus: auf jeder Stufe *i*

  bestimme für jeden erlaubten Punkt *(i,j)*

  den besten erlaubten Vorgänger



Beispiel:

7	4	9	1
4	3	3	12
1	2	2	1

Lokale Distanzen

12	8	13	7
5	4	6	17
1	3	5	6

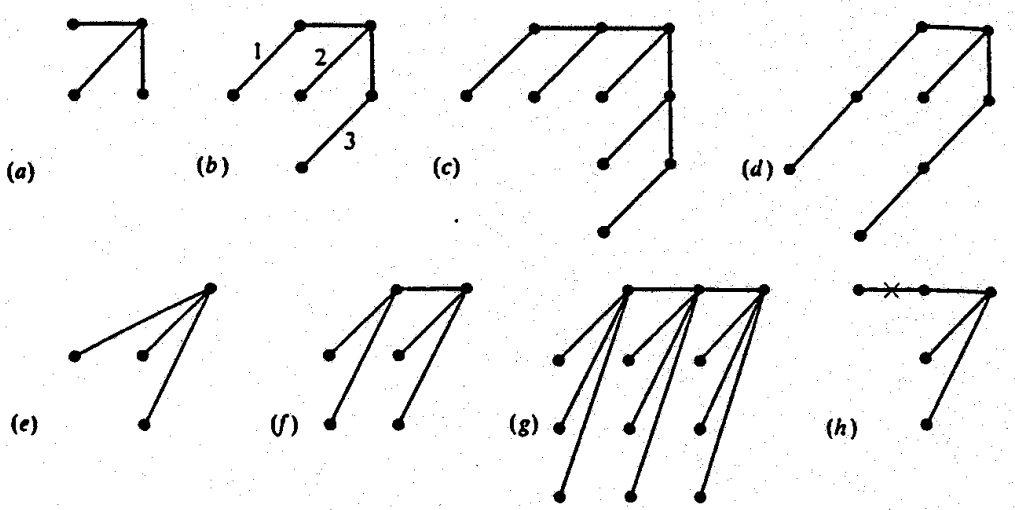
Kumulierte Distanzen

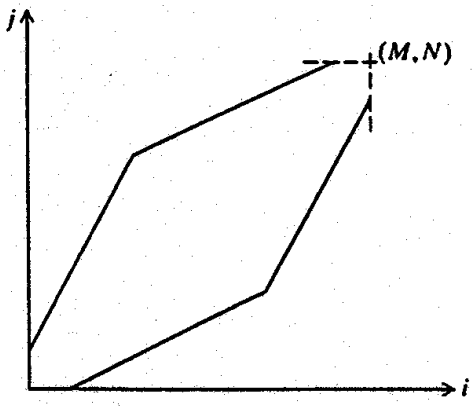
5.2 Spracherkennung

Einschränkungen für den Suchraum:

- Verzerrungsfunktion monoton
- Steigung in bestimmten Grenzen

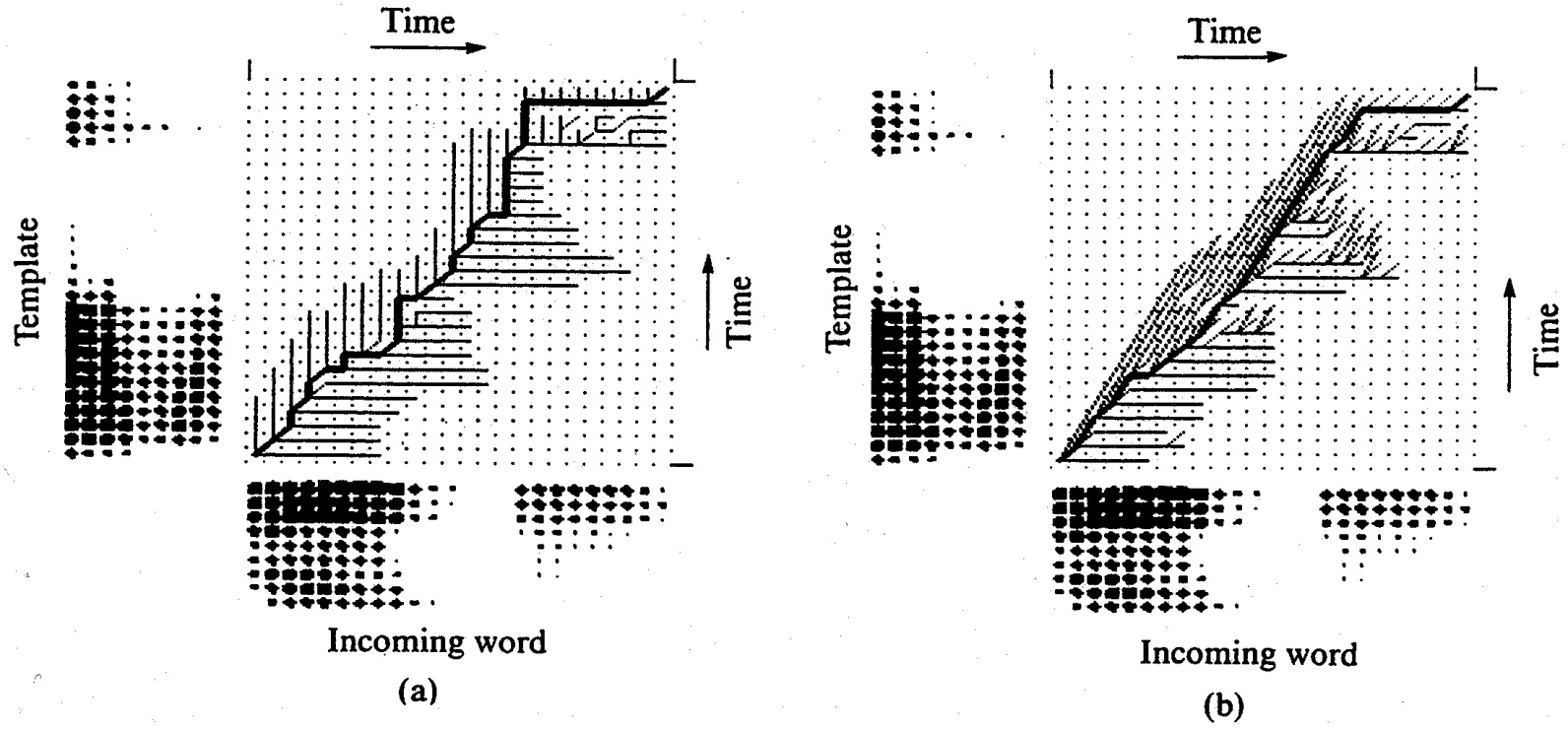
z.B.:





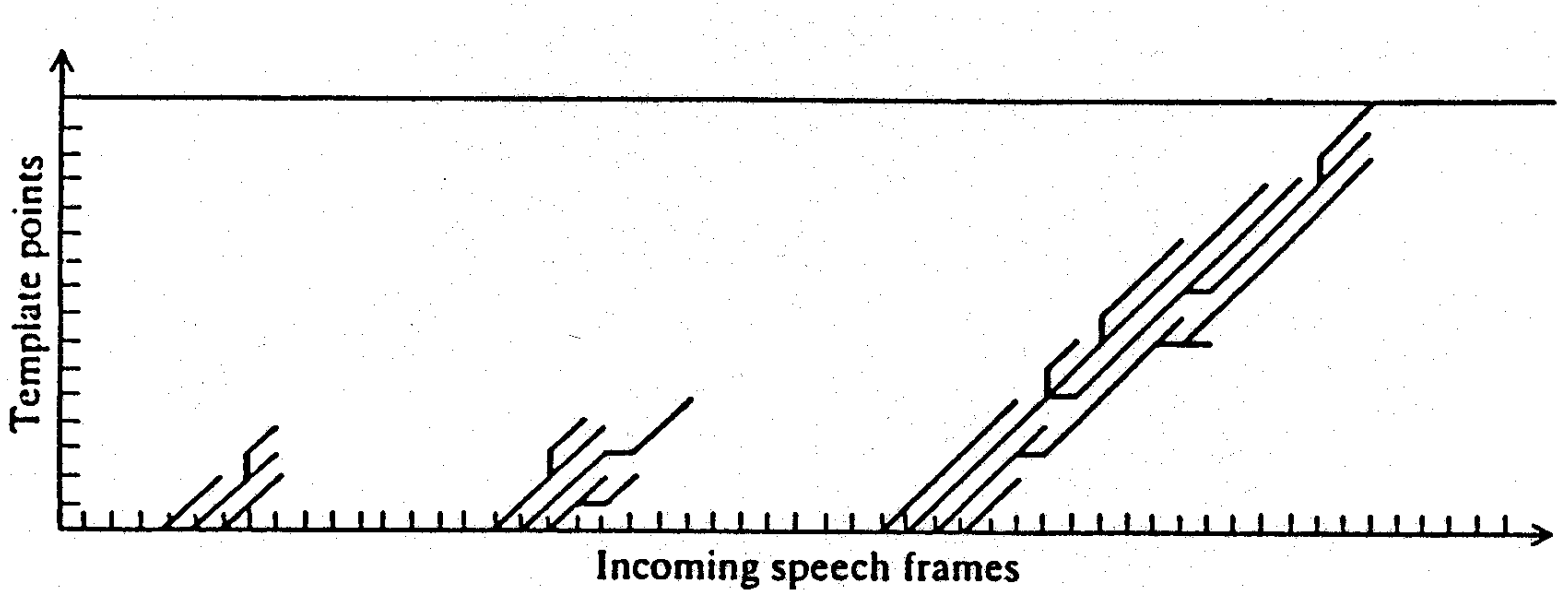
Untersuchte Pfade

z.B.:



Word spotting

- DTW sucht laufend vorgegebenes Wort
- obere Grenze für Kosten



Spracherkennung mit statistischen Verfahren

Grundprinzip:

- Statistische Modelle zur Beschreibung der unterschiedlichen möglichen Realisierungen
- Je ein Modell für jede Erkennungseinheit (z.B. Wort)
- Beschreibt die Wahrscheinlichkeit, dass dieser Einheit eine bestimmte Folge von akustischen Merkmalsvektoren entspricht (Beobachtungsfolge)
- Modellparamter werden in Trainingsphase gewonnen

- **Bayes'sche Regel**
  - $P(a,b) = P(a|b)\,P(b) = P(b|a)\,P(a)$
  - Wahrscheinlichkeit von *a* **und** *b* = Wahrscheinlichkeit von *b* * bedingte Wahrscheinlichkeit von *a* bei gegebenem *b*

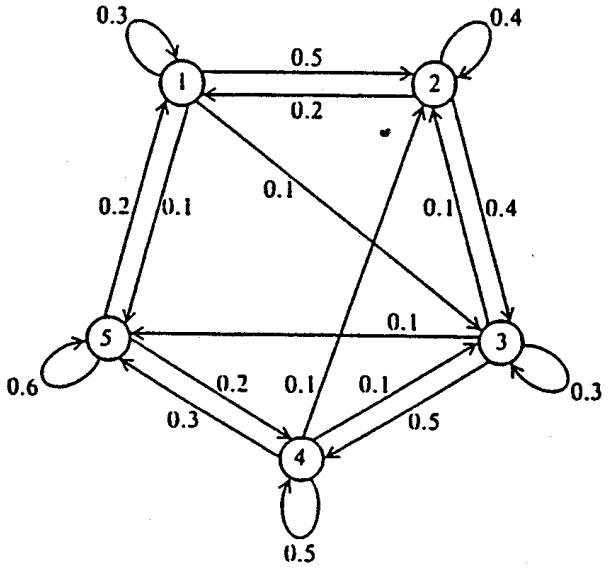
- Das **Spracherkennungsproblem:**
  - Finde die wahrscheinlichste Folge **w** von Wörtern, gegeben eine Folge von akustischen Eingaben (Vektoren) **a**
  - $\text{ArgMax}_{\mathbf{w}}\ P(\mathbf{w}|\mathbf{a}) = \text{ArgMax}_{\mathbf{w}}\ P(\mathbf{a}|\mathbf{w})\,P(\mathbf{w}) / P(\mathbf{a})$
    $= \text{ArgMax}_{\mathbf{w}}\ P(\mathbf{a}|\mathbf{w})\,P(\mathbf{w})$

- **Akustisches Modell:** $P(\mathbf{a}|\mathbf{w})$ / **Sprachmodell:** $P(\mathbf{w})$

Markov-Modelle

Zustände und Übergangswahrscheinlichkeiten

z.B.:



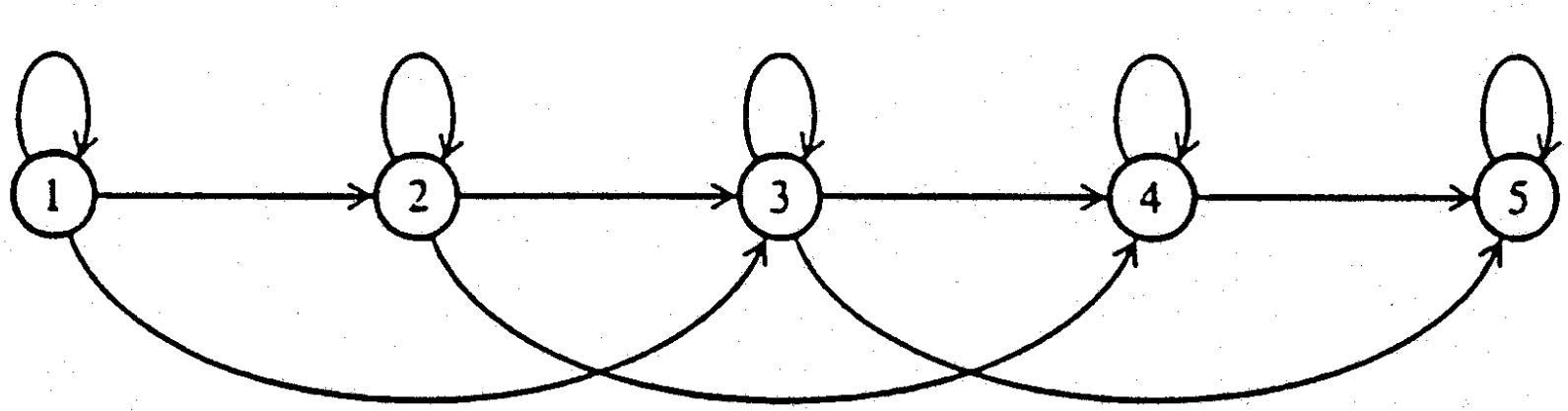
Zustandsdiagramm

$$\mathbf{A} = \begin{bmatrix} 0.3 & 0.5 & 0.1 & 0 & 0.1 \\ 0.2 & 0.4 & 0.4 & 0 & 0 \\ 0 & 0.1 & 0.3 & 0.5 & 0.1 \\ 0 & 0.1 & 0.1 & 0.5 & 0.3 \\ 0.2 & 0 & 0 & 0.2 & 0.6 \end{bmatrix}$$

Matrix der Übergangs-wahrscheinlichkeiten

Einschränkung: links-rechts-Modell

z.B.:



Hidden Markov Model (HMM):
ein ***doppelt*** eingebetteter stochastischer Prozess

1. Innere Zustände des Modells und Zustandsübergänge

   - nicht direkt sichtbar, sondern nur über die Sequenz der beobachteten Ereignisse, z.B. MFCC-Merkmalsvektoren, statistisch zu schließen

2. Emissionswahrscheinlichkeiten in jedem Zustand für die verschiedenen beobachtbaren Ereignisse

   - können diskrete oder kontinuierliche Wahrscheinlichkeitsverteilungen sein

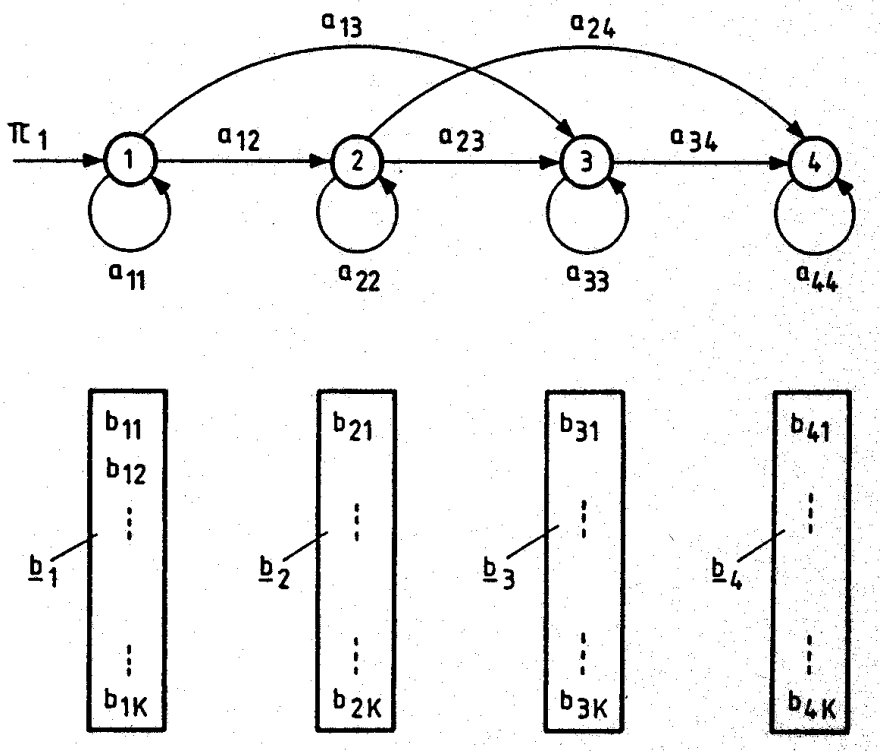
Hidden Markov Model $\lambda = (\pi, A, B)$

mit: π Vektor der Startwahrscheinlichkeiten

A Matrix der Übergangswahrscheinlichkeiten

B Matrix der Emissionswahrscheinlichkeiten

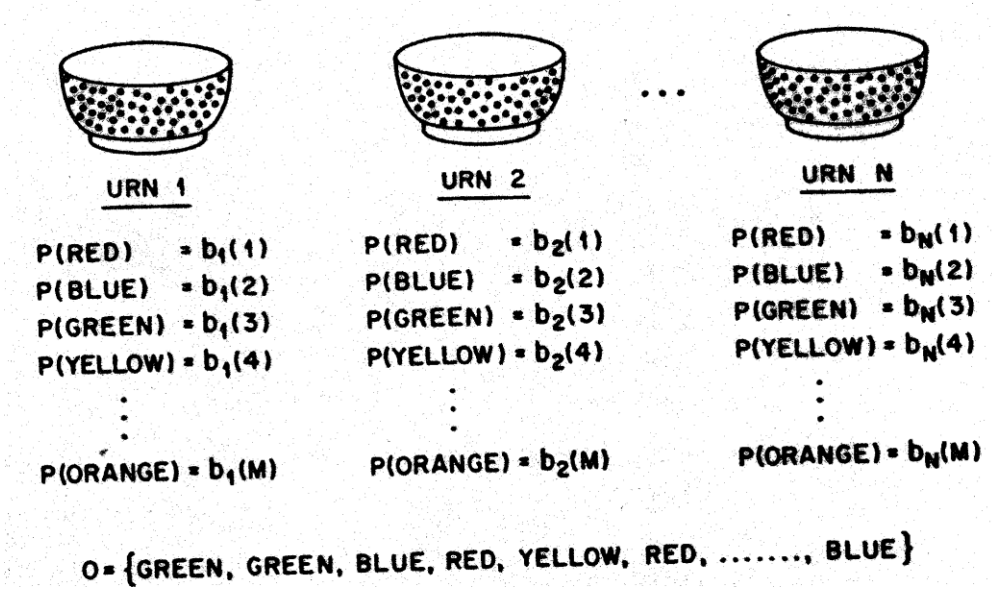
z.B.:



Eine einfache Illustration: Das Ball-Urnen Modell

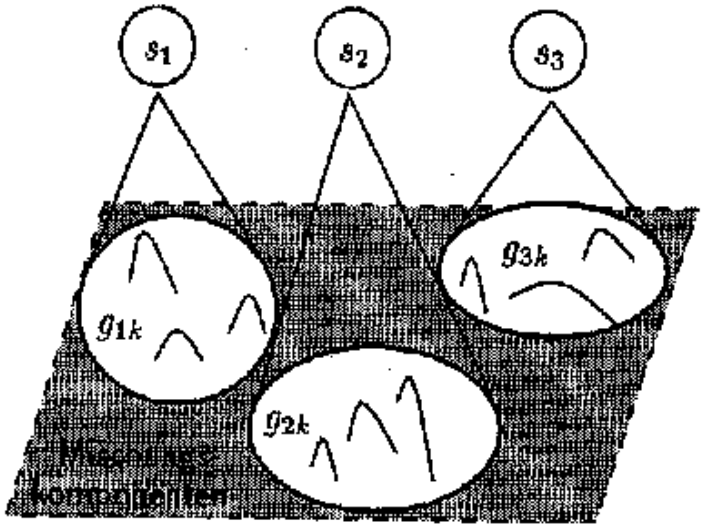
Hinter einem Vorhang werden nacheinander aus einer Reihe von Urnen farbige Bälle gezogen (mit Zurücklegen).

Die Verteilung der Farben in jeder Urne und der Zeitpunkt des Fortschreitens zur nächsten Urne sind dem Beobachter unbekannt, bekannt ist nur die gesamte Beobachtungsserie
(z.B. "grün, grün, blau, rot, gelb, rot, blau").



Anwendung von HMMs für Spracherkennung:

- ein HMM für jede zu erkennende Einheit
  - z.B. Wort, Phon, Phon plus linker und rechter Kontext (Triphon)
- die Emissionswahrscheinlichkeiten b_i beziehen sich auf die Merkmalsvektoren in den Frames
- typischerweise werden für die b_i multivariate Gaußverteilungen angesetzt bzw. Mischverteilungen mehrerer multivariater Gaußverteilungen (um Aussprachevarianten erfassen zu können)

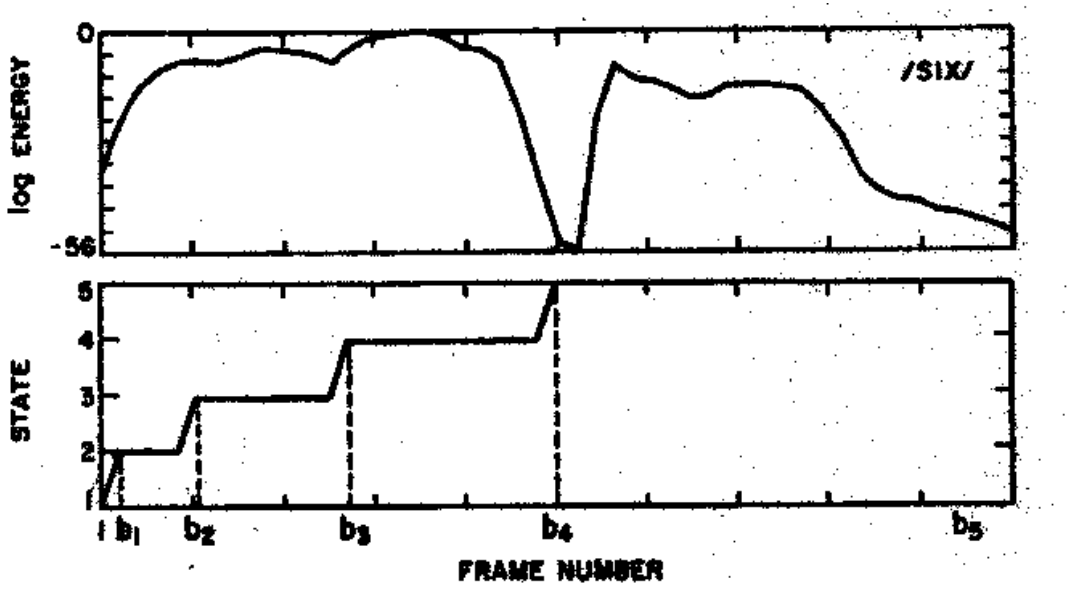


5.2 Spracherkennung

Die einzelnen HMM-Zustände werden durch den Trainingsalgorithmus festgelegt; ihre Bedeutung bleibt unbekannt (hidden)

- "We leave it to the computer to learn what we have failed to understand. The computer might do the job but can it tell us how?" (G. Fant, 'Speech Research in Perspective', in *Speech Communication* vol. 9(3), pp.171-176, 1990)

Illustration: Zustandsverlauf eines Wort-HMMs für ein Vorkommnis des Wortes *„six"*



Die 3 Teilprobleme der HMM-Methode:

Erkennung:

1. Gegeben die Beobachtungsfolge $O = O_1O_2...O_T$ und Modell $\lambda = (\pi, A, B)$; wie berechnen wir effizient $P(O|\lambda)$, die Wahrscheinlichkeit der Beobachtungsfolge gemäß dem Modell?

   $\rightarrow$ *Forward-Backward Prozedur*

Training:

2. Gegeben die Beobachtungsfolge $O = O_1O_2...O_T$ und Modell $\lambda = (\pi, A, B)$; wie wählen wir jene Zustandsfolge $Q = Q_1Q_2...Q_T$, welche die Beobachtungsfolge am besten erklärt?
   $\rightarrow$ *Viterbi-Algorithmus* (maximale Wahrscheinlichkeit von O)
3. Wie berechnen wir die Modellparameter $\lambda = (\pi, A, B)$ so, dass die Wahrscheinlichkeit $P(O|\lambda)$ maximiert wird?
   $\rightarrow$ *Baum-Welch-Methode* (oder: Standard-Gradientenverfahren)

(siehe z.B. L. Rabiner, 'A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition' in *Proceedings of the IEEE*, vol. 77(2), pp.257-285, 1989)

Lösung Problem 1: Forward-Backward Prozedur (Erkennung mit HMMs)

$$P(O \mid \lambda) = \sum_{i_1, i_2, \ldots i_T} \pi_{i_1} b_{i_1}(O_1) a_{i_1 i_2} \ldots a_{i_{T-1} i_{T-2}} b_{i_T}(O_T)$$

Vorwärtsvariable $\alpha_t(i) = P(O_1, O_2, \ldots O_t, i_t = q_i \mid \lambda)$

Lösung durch Rekursion:

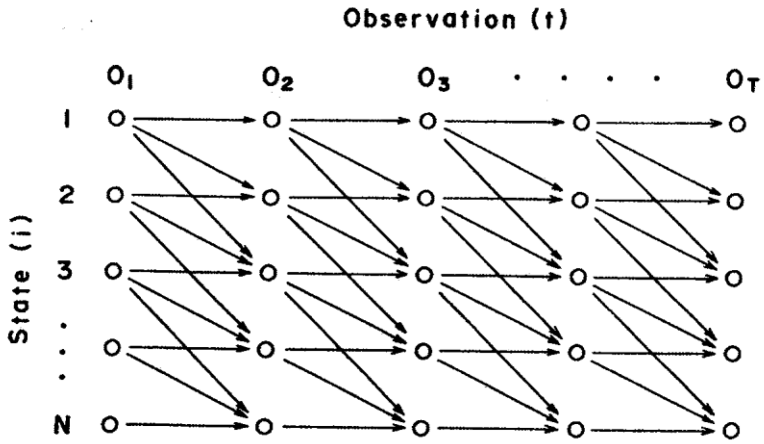
Schritt 1: $\alpha_1(i) = \pi_i b_i(O_1) \qquad (1 \le i \le N)$

Schritt 2: *Für* $t = 1, 2, \ldots, T-1 \qquad (1 \le j \le N)$

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij} \right] b_j(O_{t+1})$$

Schritt 3: *Dann*

$$P(O \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i)$$



Lösung Problem 2: Viterbi-Algorithmus

bestimme P*, die größte Wahrscheinlichkeit, und
F, die Zustandsfolge, die zur größten Wahrscheinlichkeit führt

Lösung durch Rekursion:

Schritt 1: Initialisierung

$$\delta_1(i) = \pi_i b_i(O_1) \qquad (1 \le i \le N)$$
$$\varphi_1(i) = 0$$

Schritt 2: Rekursion

für $2 \le t \le T,\ 1 \le j \le N,$

$$\delta_t(j) = \max_{1 \le i \le N} \left[\delta_{t-1}(i) a_{ij}\right] b_j(O_t)$$
$$\varphi_t(j) = \operatorname*{argmax}_{1 \le i \le N} \left[\delta_{t-1}(i) a_{ij}\right]$$

Schritt 3: Terminierung

$$P^* = \max_{1 \le i \le N} \left[\delta_T(i)\right]$$
$$i_t^* = \operatorname*{argmax}_{1 \le i \le N} \left[\delta_T(i)\right]$$

Schritt 4: Zustandssequenz - Backtracking

für $t = T-1, T-2, ..., 1,$

$$i_t^* = \varphi_{t+1}(i_{t+1}^*)$$

Lösung für Problem 3: Baum-Welch-Schätzung (Erwartungsmaximierung)

Verwende die Vorwärtsvariablen aus Problem 1 und analoge Rückwärtsvariablen, um iterativ zu schätzen:

Startwahrscheinlichkeiten (π):

... durch die geschätzte Wahrscheinlichkeit, zum Zeitpunkt t=1 im Zustand q_i zu sein

Übergangswahrscheinlichkeiten (A):

... durch das Verhältnis der erwarteten Anzahl von Übergängen von Zustand q_i zu q_j

zur erwarteten Anzahl von Übergängen, die von Zustand q_i ausgehen

Emissionswahrscheinlichkeiten (B):

... durch das Verhältnis der Wahrscheinlichkeitsschätzung, in Zustand q_i zu sein und Symbol k zu beobachten,

zur Wahrscheinlichkeitsschätzung, in Zustand q_i zu sein

Praktischer Ablauf des Training von HMMs:

- Jedes HMM eines Spracherkenners wird in mehreren Durchläufen mit möglichst vielen unterschiedlichen Beispielen der zu erkennenden Einheit (z.B. einem bestimmten Phon) trainiert (mit Baum-Welch o.ä.)
- Dadurch wird das HMM so parametrisiert, dass es die Wahrscheinlichkeit dieser Trainingsdaten maximiert (zumindest bis zu einem lokalen Maximum) – *Maximum Likelihood (ML)* Ansatz
- Es gibt auch andere Trainingsverfahren, bei denen direkt die Unterscheidungsschärfe zwischen den verschiedenen HMMs maximiert wird – *Minimum Classification Error (MCE)* Ansatz
- Je detaillierter der Erkenner (mehr HMMs, mehr Zustände, mehr Mischverteilungskomponenten, mehr Merkmale), desto mehr Trainingsdaten werden benötigt

Besondere Stärken von HMMs:

- Reiche mathematische Struktur; sprachwissenschaftliches Wissen kann eingebracht werden
  - bei der Festlegung der Topologie der HMMs (Anzahl Zustände, mögliche Zustandsübergänge)
  - Verbinden (*Tying*) von Parametern zur Reduktion des Trainingsdatenbedarfs (manuell oder automatisch)
- Es existieren effiziente Verfahren zur Anpassung eines gut trainierten Erkenners an neue Sprecher oder Umgebungsbedingungen, mit relativ wenig zusätzlichen Trainingsdaten
  - *Maximum Likelihood Linear Regression (MLLR)*
  - *Maximum A-Posteriori Estimation (MAP)*

- Auf Grund dieser Stärken haben sich HMMs in der Spracherkennung gegenüber alternativen Methoden, insbesondere Neural Networks durchgesetzt

Erweiterungen:

→ Sprecherunabhängigkeit

- Adaption, z.B. Frequenznormalisierung
- Mittelung über mehrere Referenzmuster (DTW)
- Training mit Mustern von mehreren Sprechern (HMM)

→ großes Vokabular

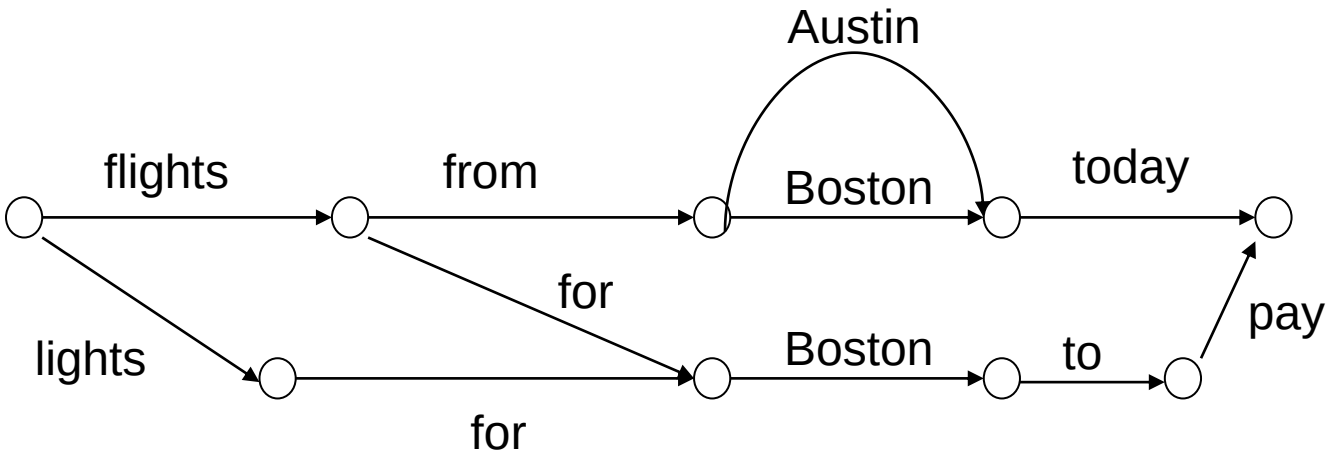
- Übergang auf kleinere Erkennungseinheiten,
  z.B.: HMMs für Einzellaute
  \+ (auf nächsthöherer Ebene) Wortmodelle

→ zusammenhängende Sprache

- „word spotting“
- Erwartungsgesteuerte Analyse
  z.B.: Syntaxgraph, statistisches Modell für Wortfolgen

Sprachmodelle

dienen zur Einschränkung von Worthypothesen des Spracherkenners, die sich aus der HMM-Analyse ergeben



Worthypothesengraph;
in einer bestimmten Anwendung sind nicht alle Pfade gleich wahrscheinlich

Statistische Sprachmodelle modellieren die Wahrscheinlichkeit von Wörtern bzw. Wortfolgen

$P(w_j|w_1,\ldots,w_{j-1}) \sim P(w_j|w_{j-n-1},\ldots,w_{j-2},w_{j-1})$

- Unigramm-Wahrscheinlichkeit $P(w_j)$
- Bigramm-Wahrscheinlichkeit $P(w_j|w_{j-1})$
- Trigramm-Wahrscheinlichkeit $P(w_j|w_{j-2},w_{j-1})$

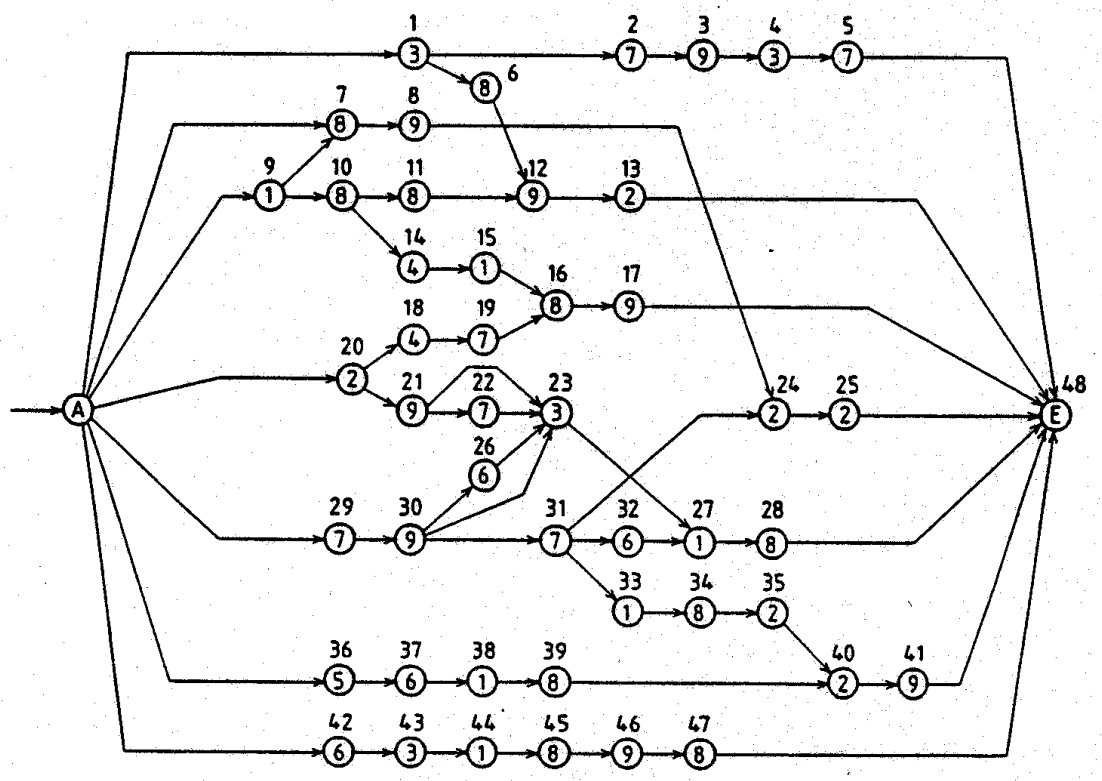
Kombination von Uni-, Bi- und Trigramm-Wahrscheinlichkeiten:

$$\hat{P}(w_3 \mid w_1, w_2) = \lambda_3 \frac{F(w_3 \mid w_1, w_2)}{\sum_k F(w_k \mid w_1, w_2)} + \lambda_2 \frac{F(w_3 \mid w_2)}{\sum_k F(w_k \mid w_2)} + \lambda_1 \frac{F(w_3)}{\sum_k F(w_k)}$$

Syntaxgraph

als Alternative zu n-gramm-Modellen

z.B.:



1. Adjektive
2. Adverbien
3. Artikel
4. Hilfsverben
5. Konjunktionen
6. Präpositionen
7. Pronomina
8. Substantive
9. Vollverben

Trainingsdaten für Spracherkenner

- Große Datenmengen (100 - 500 Stunden) notwendig für sprecherunabhängige Systeme
- Daten müssen repräsentativ für die erwarteten Sprecher sein (Alter, Geschlecht, Dialekt)
- Daten müssen phonetisch transkribiert werden (in der Praxis oft orthographische Transkription in Verbindung mit phonetischem Wörterbuch)
- Sprecherspezifisches Training bei sprecherabhängigen oder sprecheradaptiven Systemen (z.B. Diktiersysteme)

Beurteilung von Spracherkennern

- Wortfehlerrate = Fehler / Wörter
- Arten von Fehlern: Einfügung, Tilgung, Ersetzung

Literatur:

Fellbaum, K.:
http://www.kt.tu-cottbus.de/teleteaching/

Furui, S.: Digital Speech Processing, Synthesis, and Recognition.
Marcel Decker (2. Aufl., 2001)

Jelinek, F.: Statistical Methods for Speech Recognition.
MIT Press (1999)

Jurafsky, D. & Martin, J. H.: Speech and Language Processing
Upper Saddle River: Prentice Hall (2000)

Hasegawa-Johnson, M.:
http://www.ifp.uiuc.edu/~hasegawa/notes/index.html

Huang, X. et al:
Spoken Language Processing, Theory, Algorithms, System Development.
Prentice Hall (2001)

O´Shaughnessy, D.: Speech Communication.
IEEE Press (2. Aufl., 2000)

Rabiner, L., Juang, B.-H.: Fundamentals of Speech Recognition.
Prentice Hall (1993)

Schukat-Talamazzini, E.G.: Automatische Spracherkennung (1995)
http://www.minet.uni-jena.de/www/fakultaet/schukat/asebuch.html

Inhalt

6. Sprachdialogsysteme

- Einführung, Anwendungen
- Aufbau eines Sprachdialogsystems
- Dialogsteuerung
- Dialogbeschreibungssprachen
- Erkennungsgrammatiken

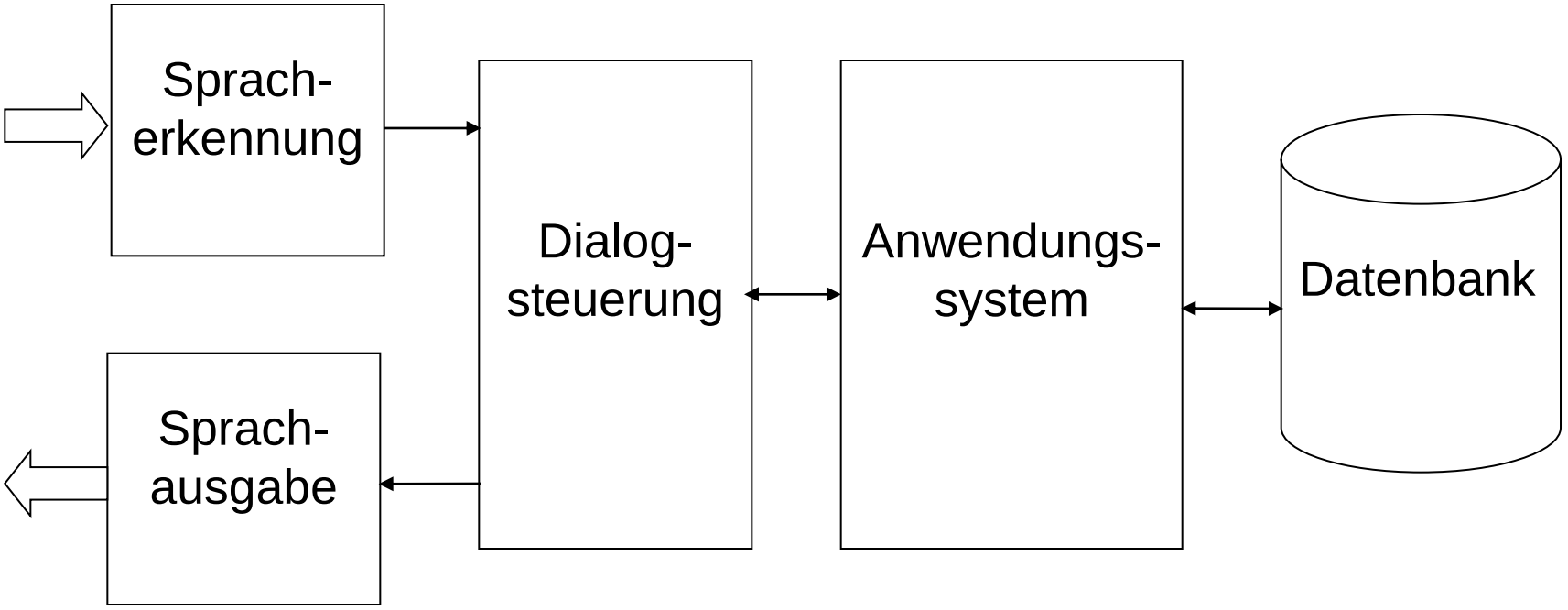
Anwendungen von Sprachdialogsystemen

- Auskunftssysteme (Fahrpläne, Aktienkurse, Postgebühren)
- Transaktionen (Aktienhandel, Flugbuchung)
- "Voice Browsing"
- Informationszugriff (Vorlesen von e-mail, "Voice Portal")

Wirtschaftlicher Nutzen durch

- Kosteneinsparung durch geringeren Personaleinsatz
- verbesserter Kundendiensts durch kürzere Wartezeiten
- Neuartige Anwendungen wie Vorlesen von e-mail

Aufbau eines Dialogsystems



Dialoginitiative

1. Systeminitiative
   - bei Systemen, die nur unregelmäßig benutzt werden
2. Benutzerinitiative
   - erfahrene Benutzer können ohne Aufforderungen des Systems Kommandos eingeben
3. gemischte Initiative
   - beispielsweise für Rückfragen des Benutzers oder Aktivierung einer Hilfefunktion
   - Überbeantwortung von Fragen durch den Benutzer

Barge-In

- "Barge-In" ist die Unterbrechung der Ausgabe eines Dialogsystems durch eine neue Eingabe des Benutzers
- Vorteile:
  - Möglichkeit der Unterbrechung langer Ausgaben (z.B. umfangreiche Fahrplanauskünfte, Vorlesen von e-mail)
  - Zeitersparnis durch schnellere Beantwortung von Fragen
- Probleme:
  - Unterbrechung der Systemausgabe durch Störgeräusche und Störung des Dialogablaufs

Verifikation

- Verifikation ist Bestätigung von Benutzereingaben
- Explizite Verifikation: Eingabe muss ausdrücklich bestätigt werden.
- Implizite Verifikation: Eingabe wird wiederholt und gilt als akzeptiert, wenn der Benutzer nicht widerspricht.

Explizite Verifikation	**Implizite Verifikation**
System (S): Wollen Sie ein Paket oder eine Briefsendung schicken? **Benutzer (B):** ein Paket **S:** Sie wollen also ein Paket schicken? **B:** ja	**S:** Wollen Sie ein Paket oder eine Briefsendung schicken **B:** ein Paket **S:** In welches Land wollen Sie das Paket schicken? **B:** nein, kein Paket sondern einen Brief

Dialogmodellierung

1. Ablaufdiagram
   - klarer Ablauf des Dialogs
   - wenig Flexibilität bei der Dialogführung
   - unübersichtlich bei komplexeren Dialogen
2. Erfragen von Informationseinheiten ("slot-filling")
   - System fragt nach fehlenden Informationen.
   - Überbeantwortung kann gut behandelt werden.
   - flexibler Dialogablauf
3. Planbasierter Ablauf
   - Dialogsystem bekommt ein Ziel vorgegeben und versucht, dieses durch Planung des Dialogs zu erreichen.

Dialogbeschreibungssprache VoiceXML

- Mit VoiceXML können Sprachdialogsysteme spezifiziert werden.
- VoiceXML ist eine XML-Applikation und wird durch eine DTD (Document Type Description) definiert.
- Dialogführung durch "slot-filling" (Form Interpretation Algorithm)
- Verarbeitung ist mit dem Ausfüllen von Formularen in HTML-Seiten vergleichbar.
- VoiceXML ist beim WWW Consortium als Standard eingereicht worden und wird von zahlreichen Firmen unterstützt.

VoiceXML-Beispiel: Programmcode

```
<?xml version="1.0"?>
<vxml version="1.0">
 <form>
  <field name="drink">
   <prompt>Would you like coffee,
      tea, milk, or nothing?</prompt>
   <grammar src="drink.gram"
      type="application/x-jsgf"/>
  </field>
  <block>
   <submit next="http://www.drink.example/drink2.asp"/>
  </block>
 </form>
</vxml>
```

VoiceXML-Beispiel: Dialog

S (System): Would you like coffee, tea, milk, or nothing?

B (Benutzer): Orange juice.

S: I did not understand what you said.

S: Would you like coffee, tea, milk, or nothing?

B: Tea

S: (setzt den Dialog mit dem VoixeXML-Programm drink2.asp fort)

Erkennungsgrammatiken

- Die Anzahl der möglichen Eingaben bei jedem Dialogschritt wird durch eine Erkennungsgrammatik eingeschränkt.
- Einschränkung der Eingaben ist bei sprecherunabhängigen Systemen erforderlich, um die Anzahl der Erkennungs-fehler zu verringern.
- Erkennungsgrammatiken sind normalerweise kontextfreie oder reguläre Grammatiken

Java Speech Grammar Format

- Java Speech Grammar Format (JSGF) ist eines von mehreren ähnlichen Formaten für Erkennungsgrammatiken

`<xyz>` Grammatische Kategorie `xyz`

`*` Wiederholung (0 bis n-mal)

`+` Wiederholung (1 bis n-mal)

`(...)` Gruppierung

`[...]` Gruppierung, optional

`|` Auswahl aus Alternativen

`/n/` Alternative hat Gewichtung `n`

Erkennungsgrammatik in Java Speech Grammar Format

```
#JSGF V1.0 ISO8859-1 en;

grammar com.acme.commands;

<basicCmd> = <startPolite> <command> <endPolite>;

<command> = <action> <object>;

<action> = /10/ open |/2/ close |/1/ delete |/1/ move;

<object> = [the | a] (window | file | menu);

<startPolite> = (please | kindly | could you |
   oh  mighty  computer) *;

<endPolite> = [ please | thanks | thank you ];
```

Literatur:

Balentine, B., Morgan D. P.:
How to build a speech recognition application.
Enterprise Integration Group (1999)

W3C speech activities (VoiceXML etc.):
http://www.w3.org/voice

SALT forum:
http://www.saltforum.org