Einführung in die Problematik der Sprachein- und -ausgabe

Donnerstag 6. 5. 2004

3. Einführung in Sprachein- und -ausgabe

Richtziel:

Erlernen möglicher Verwendung von Markup-Programmiersprachen für Modellierung der natürlichsprachlichen Mensch-Computer Interaktion.

Schwerpunkte der Unterricht:

Grundlagen der Computerlinguistik

Modellierung und Simulation sprachlicher Prozesse auf dem Computer

Grundlagen der Mensch-Computer Interaktion

Natürlichsprachliche Kommunikation, Discourse und Dialog

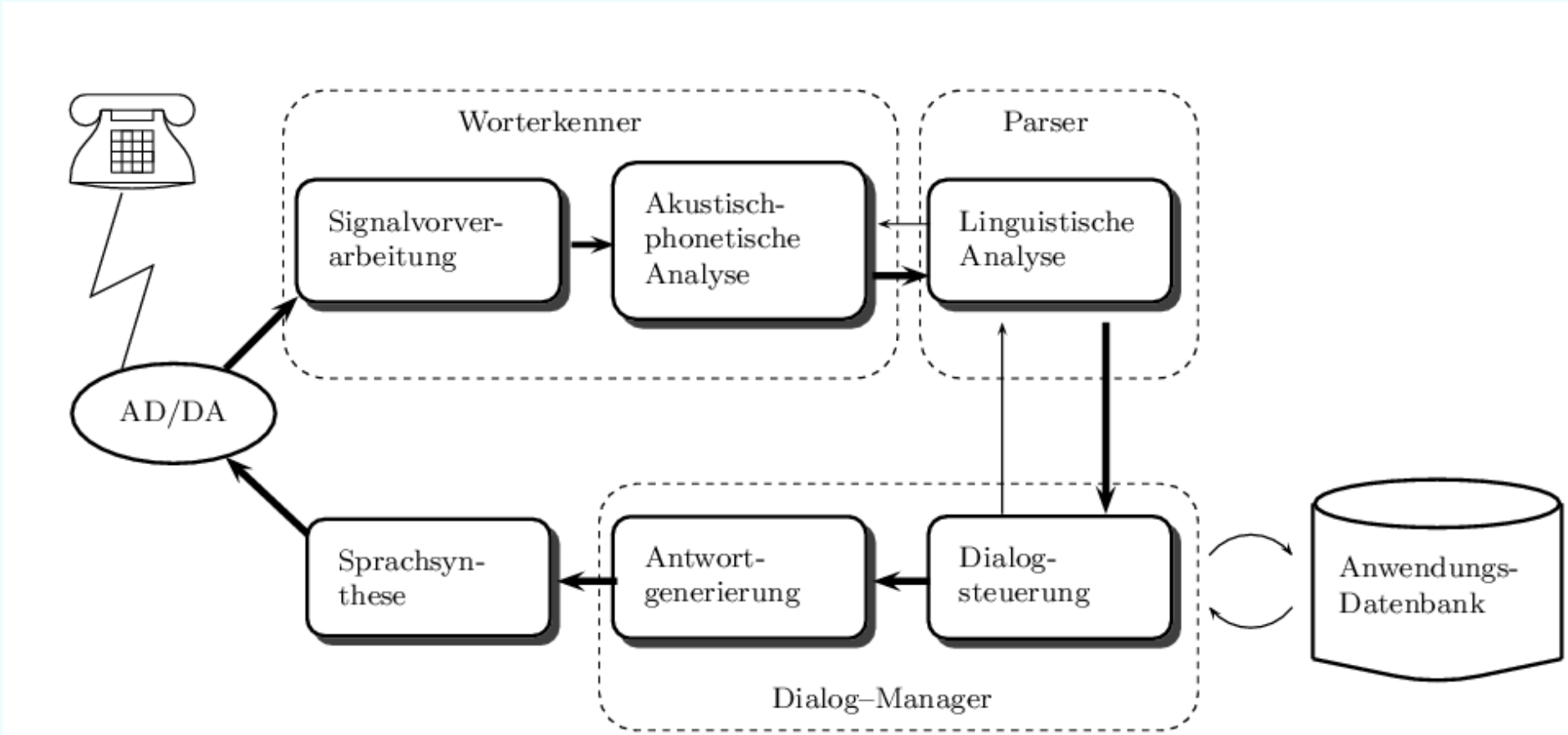
Funktionale Eigenschaften und Blockstruktur des Dialogsystems

Einführung in die Problematik der Sprachein- und –ausgabe

Prinzipien der Dialogsteuerung, symbolische Darstellungen der Dialogführung

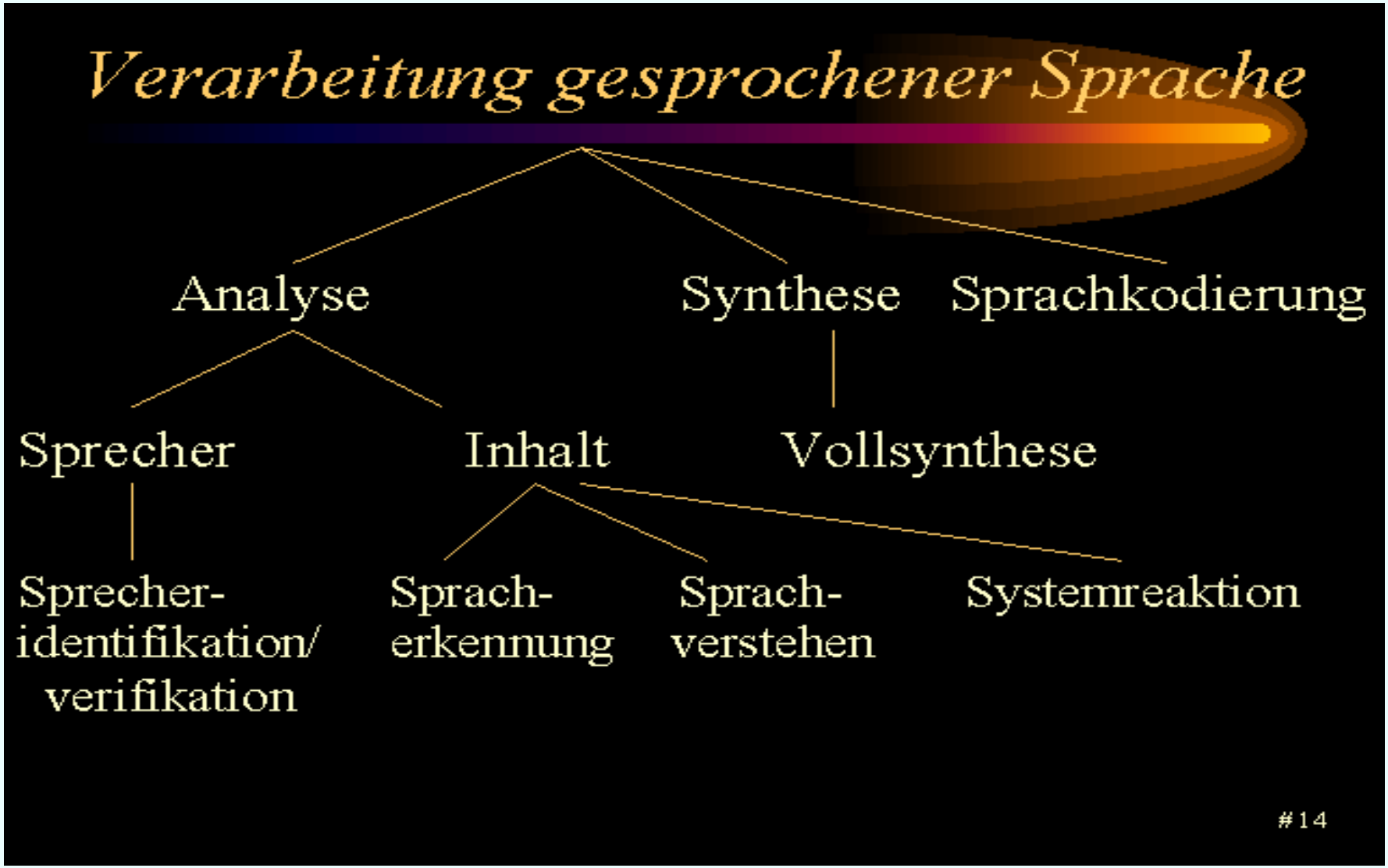
Rolle der Dialogmodelle und Dialogmodellierung

Prinzipielle Struktur des Auskunftsdialogsystems

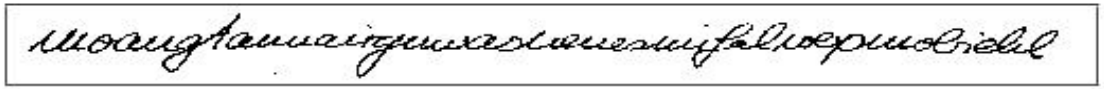
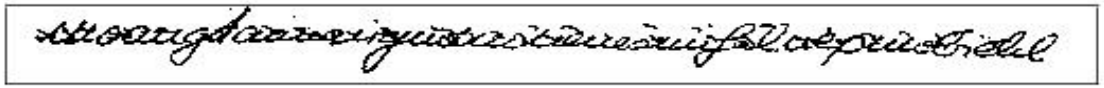
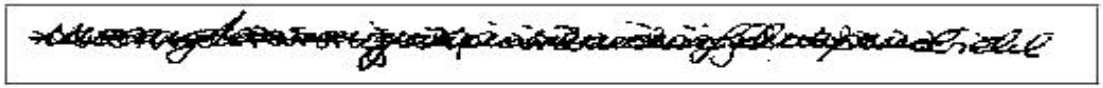


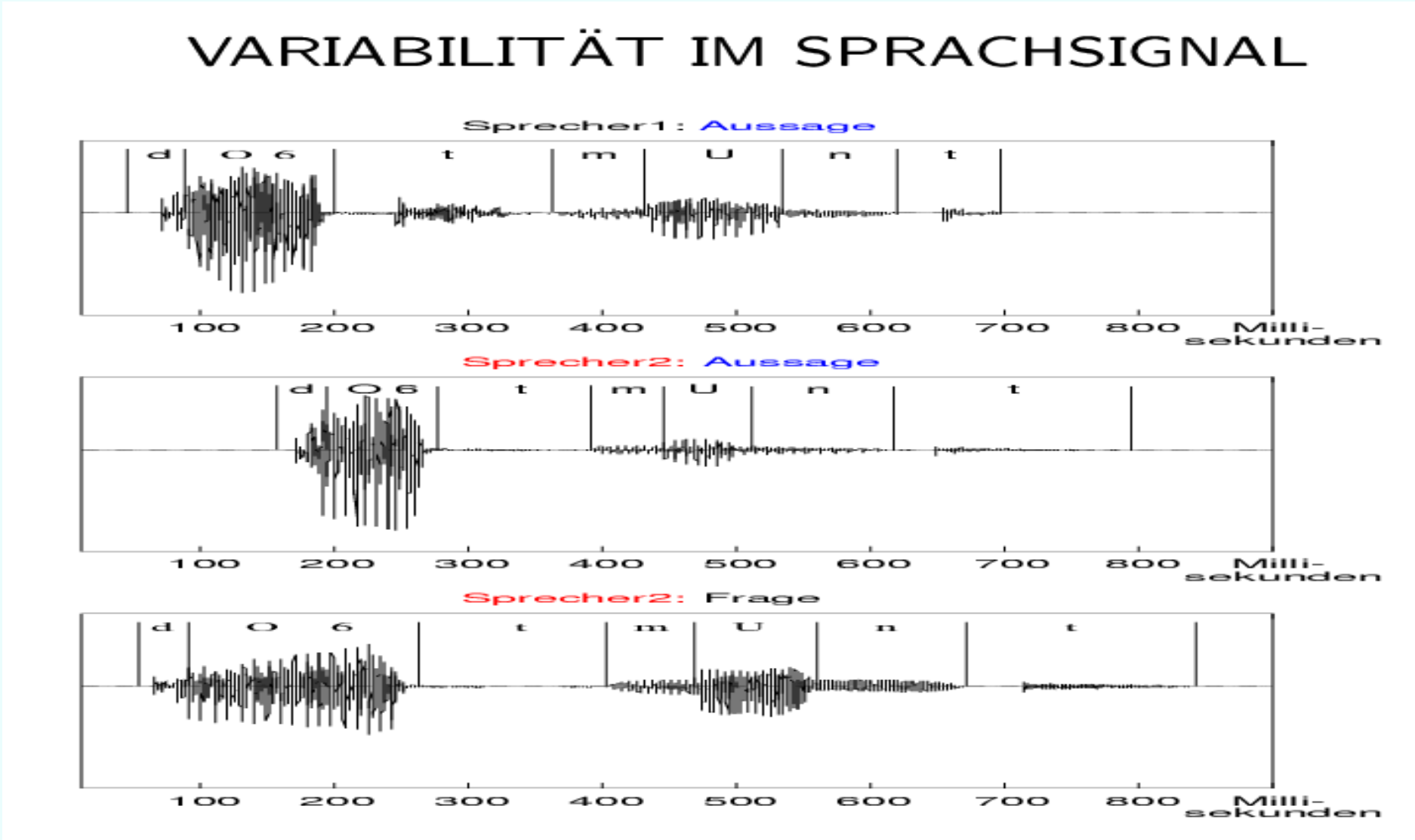
SYSTEMKOMPONENTEN

Modul	Aufgabe	Ergebnis
AD/DA–Wandler	Kodierung	Abtastwerte
Worterkenner	Signalverarbeitung, akust.–phonet. Analyse	Merkmale erkannte Wortkette bzw. Wortgraph
Parser	linguistische Analyse	semantische Struktur
Dialogmanager	Dialogkontrolle, Datenbankabfrage, Antwortgenerierung	Antworttext
Datenbank	effizienter Zugriff auf Datenbestand	Datensätze
Sprachsynthese	sprachliche Ausgabe	Sprachsignal

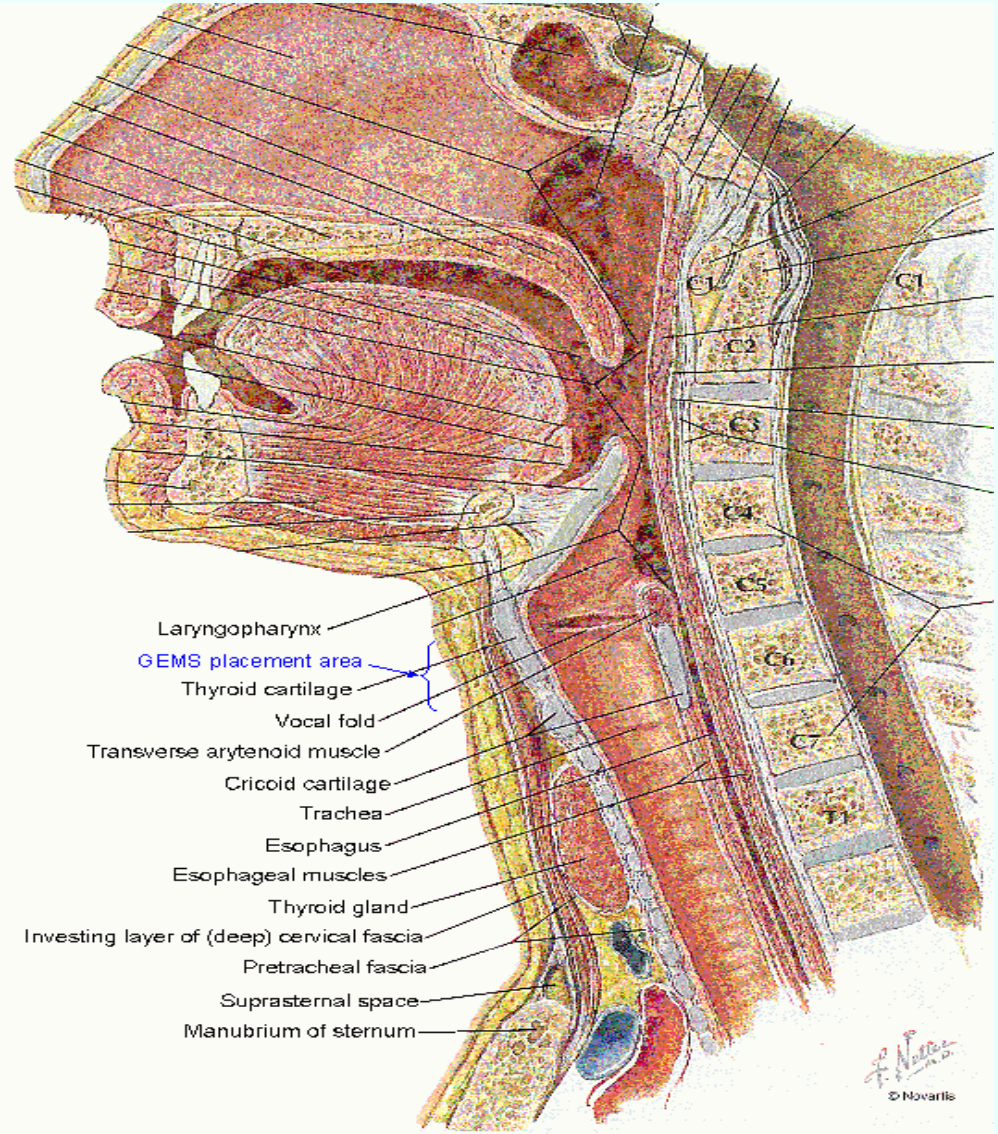

Verarbeitung gesprochener Sprache
Analyse
Synthese
Sprachkodierung
Sprecher
Inhalt
Vollsynthese
Sprecher-identifikation/ verifikation
Sprach-erkennung
Sprach-verstehen
Systemreaktion
#14

WARUM IST SPRACHERKENNUNG SCHWIERIG?

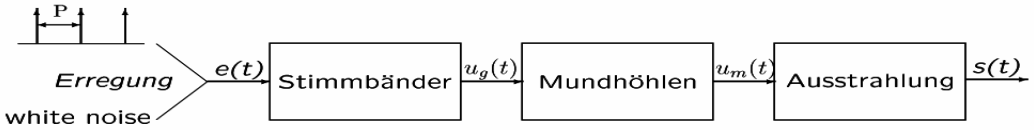
Guten Morgen, Herr Hauptkomissar Thanner. Gibt es irgendetwas Neues im Fall "VERBMOBIL"?	der Text in "Schönschrift"
Morgen, Thanner. Irgendwas Neues im Fall "VERBMOBIL"?	spontan gesprochene Sprache
morgen thanner irgendwas neues im fall verbmobil	Großschreibung? Satzzeichen?
morgenthannerirgendwasneuesimfallverbmobil	kontinuierliche Sprache
moangtannairgnwasneuesimfalwerpmobiehl	Aussprachevarianten
	artikulatorische Verschleifung
	Störungen & Verzerrungen des akustischen Kanals
	Überlagerung durch Fremd-stimmen: Cocktailparty-Effekt


VARIABILITÄT IM SPRACHSIGNAL
Sprecher1: Aussage
d O 6 t m U n t
100 200 300 400 500 600 700 800 Milli-sekunden
Sprecher2: Aussage
d O 6 t m U n t
100 200 300 400 500 600 700 800 Milli-sekunden
Sprecher2: Frage
d O 6 t m U n t
100 200 300 400 500 600 700 800 Milli-sekunden

3. Einführung in Sprachein- und -ausgabe



VOKAL-TRAKT-MODELL



Formale Beschreibung mit der Z–Transformation:

$$S(z) = G(z)\ M(z)\ A(z)\ E(z)$$

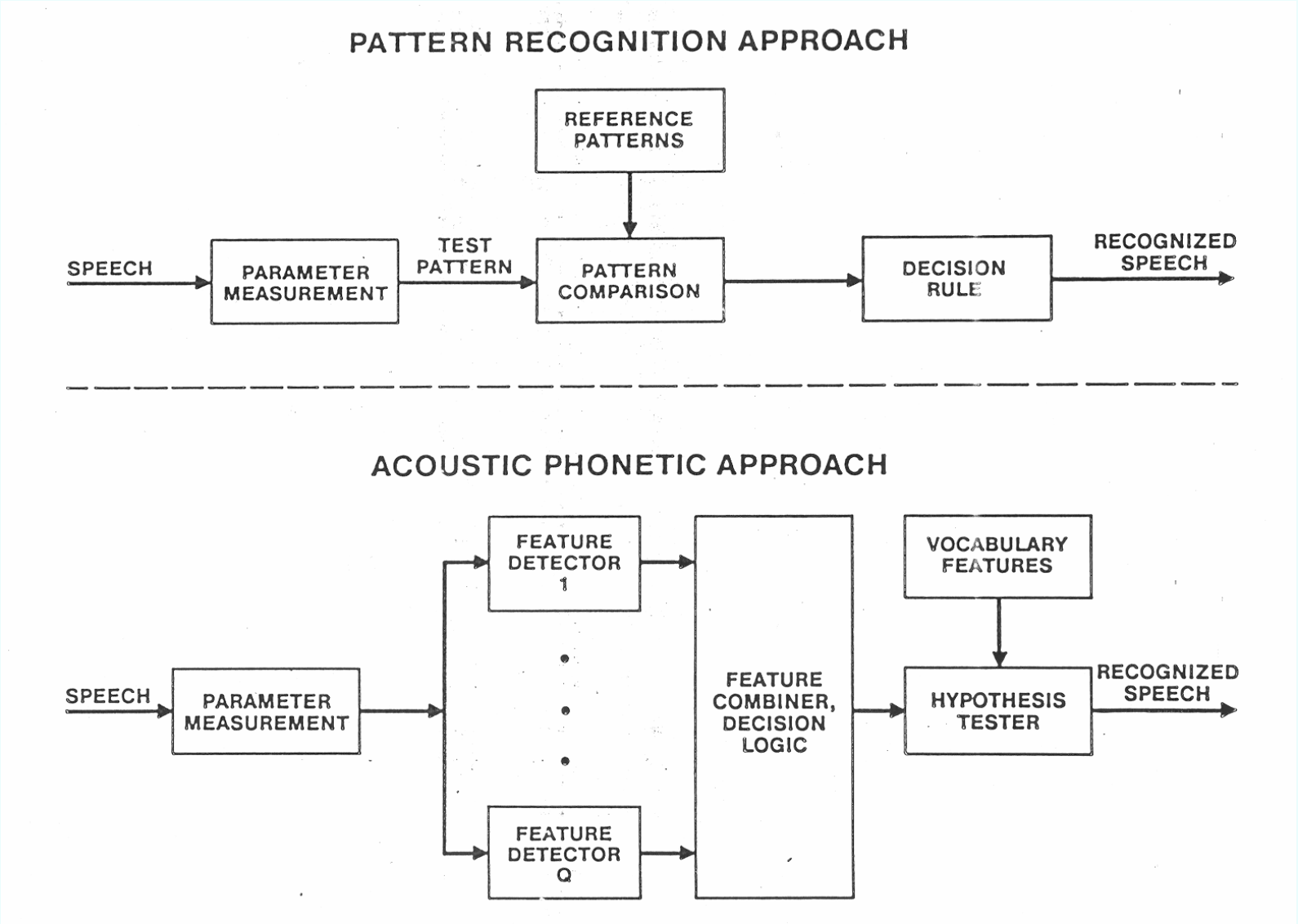
$$E(z) = \sigma \sum_{i=0}^{\infty} (z^{-b})^n = \frac{\sigma}{1 - z^{-b}}$$

$$G(z) = \frac{1}{(1 - e^{-cT} z^{-1})^2}$$

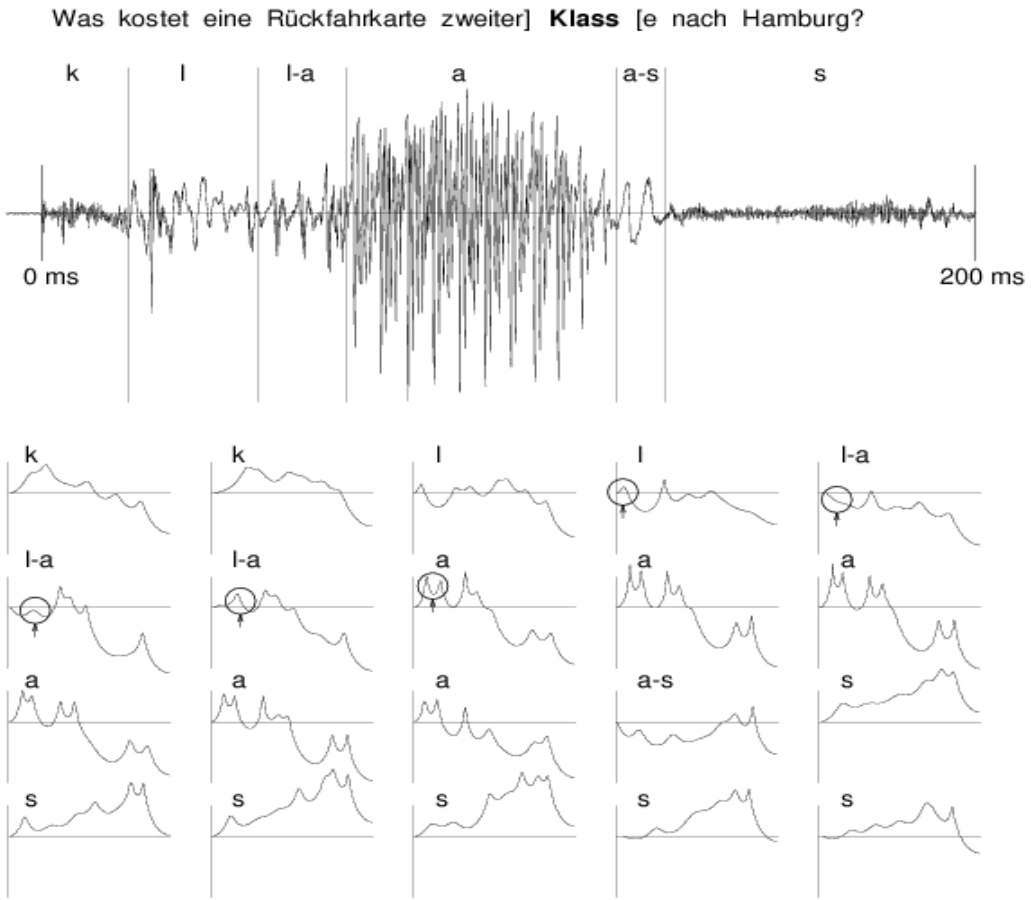
$$M(z) = \frac{1}{\prod_{i=1}^{N} [1 - 2e^{-\alpha_i T} cos\,(\beta_i T)\, z^{-1} + e^{-2\alpha_i T} z^{-2}]}$$

$$A(z) = 1 - z^{-1}$$

wo P . . . Grundperiode der Stimme,
T . . . Abtastperiode,
b . . . Grundperiode : Abtastperiode Verhältnis ($P = bT$),
α_i, β_i, c . . . Koeffizienten des Vokaltrakts und
σ . . . Verstärkungskoeffizient der Erregungsquelle
sind.


PATTERN RECOGNITION APPROACH
REFERENCE PATTERNS
SPEECH
PARAMETER MEASUREMENT
TEST PATTERN
PATTERN COMPARISON
DECISION RULE
RECOGNIZED SPEECH
ACOUSTIC PHONETIC APPROACH
SPEECH
PARAMETER MEASUREMENT
FEATURE DETECTOR 1
FEATURE DETECTOR Q
FEATURE COMBINER, DECISION LOGIC
VOCABULARY FEATURES
HYPOTHESIS TESTER
RECOGNIZED SPEECH

MERKMALBERECHNUNG



Digitalisierung

16 kHz, 16 bit

Zerlegung in Datenfenster

10 msec Zeittakt

Parametrisierung

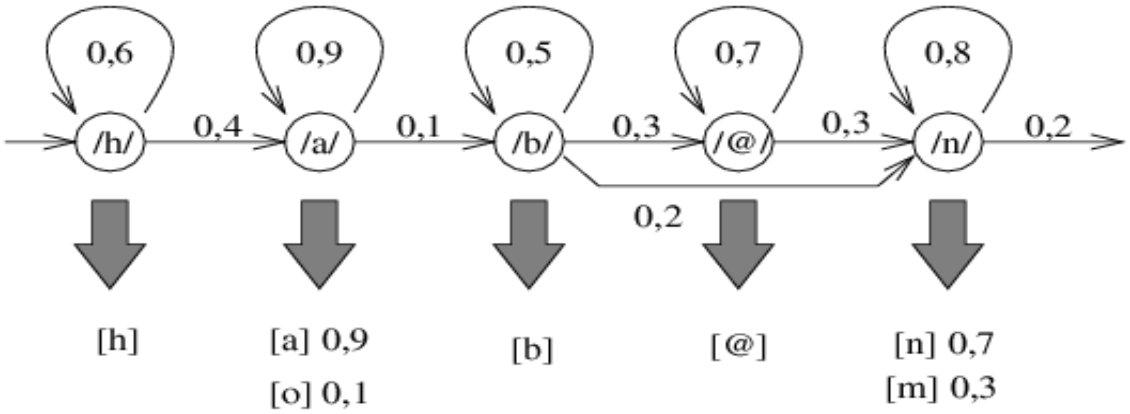
Umrechnung des Sprachsignals in Größen, die dem menschlichen Ohr nachempfunden sind
⟶ Reduktion der Datenmenge

Quantisierung

Zuordnung von Datenfenstern zu Symbolen ⟶ weitere Reduktion der Datenmenge

WORTARTIKULATION

Aussprache des Wortes „*haben*“

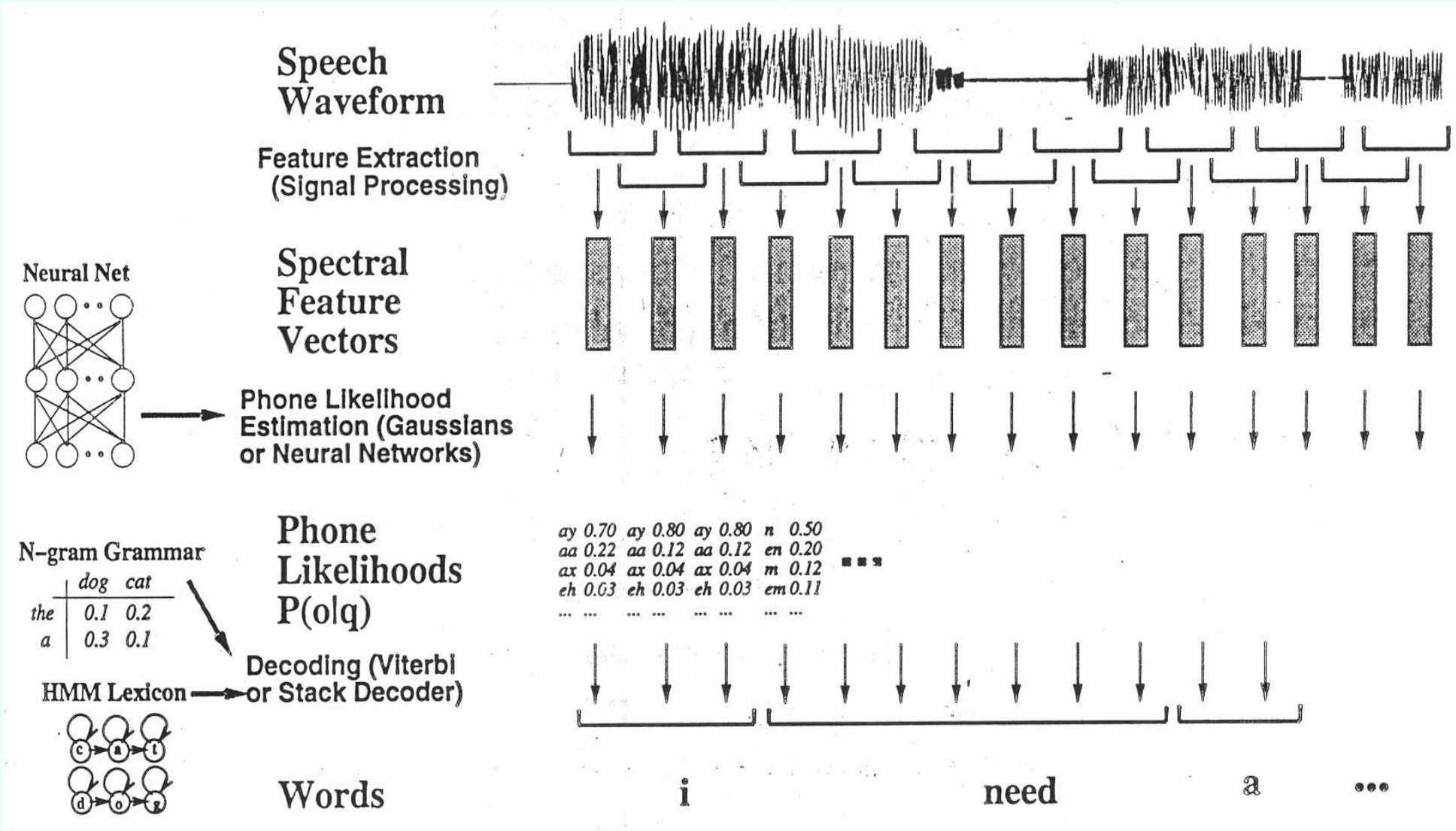


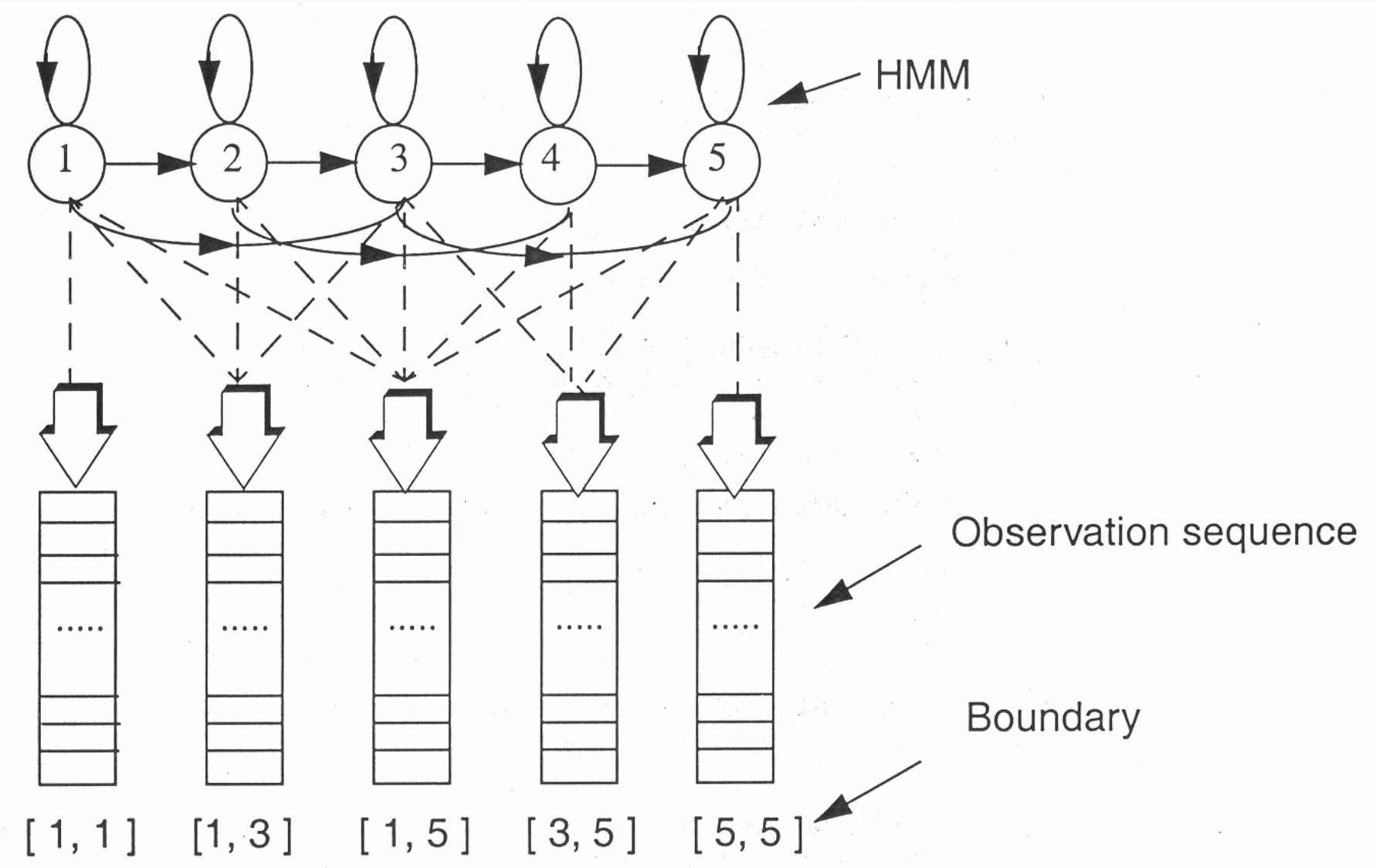
Statistisches Modell menschlicher Wortproduktion:

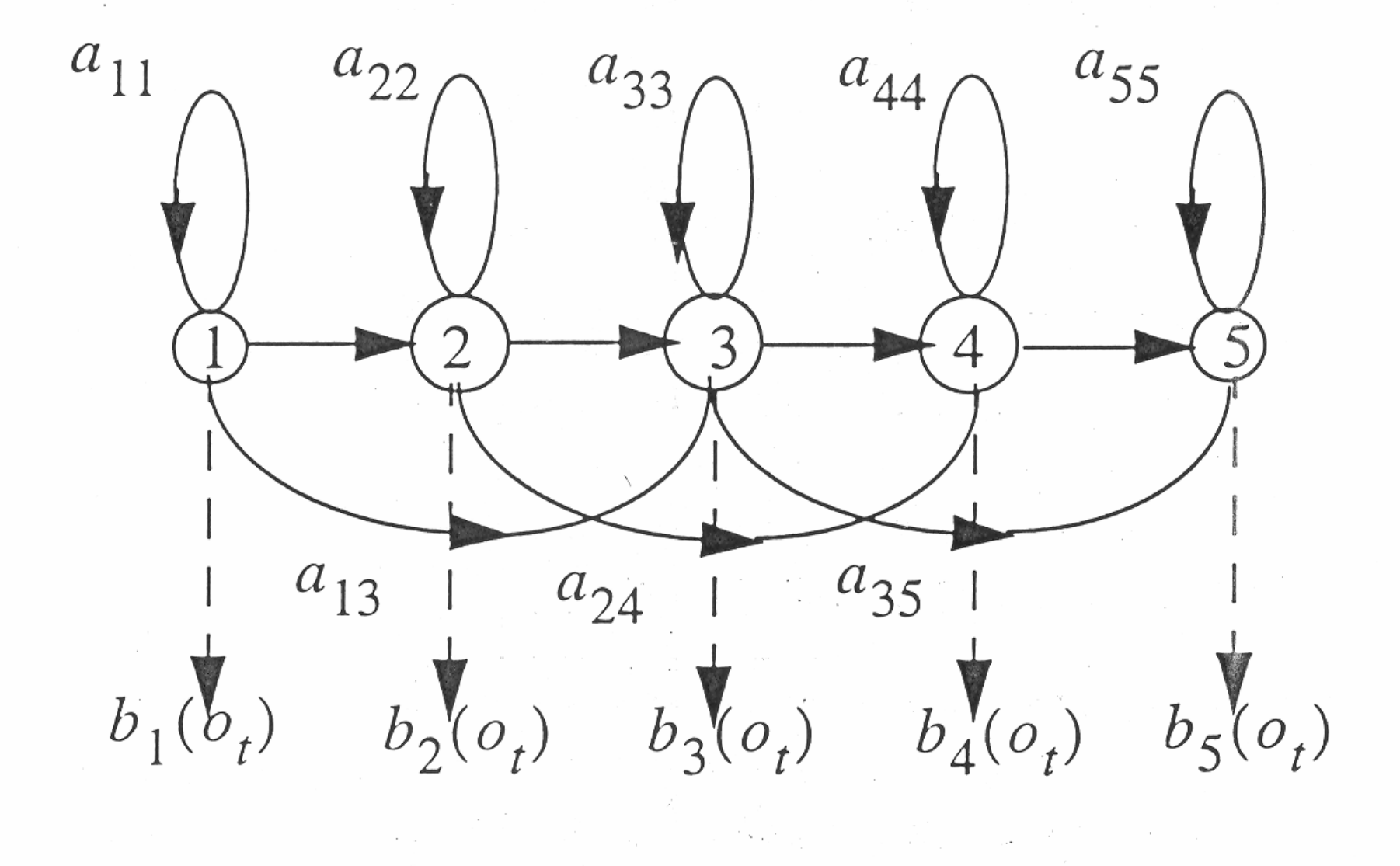
Auslassung, Einfügung — *zeitliche* Verzerrungen

Vertauschung — *spektrale* Verzerrungen

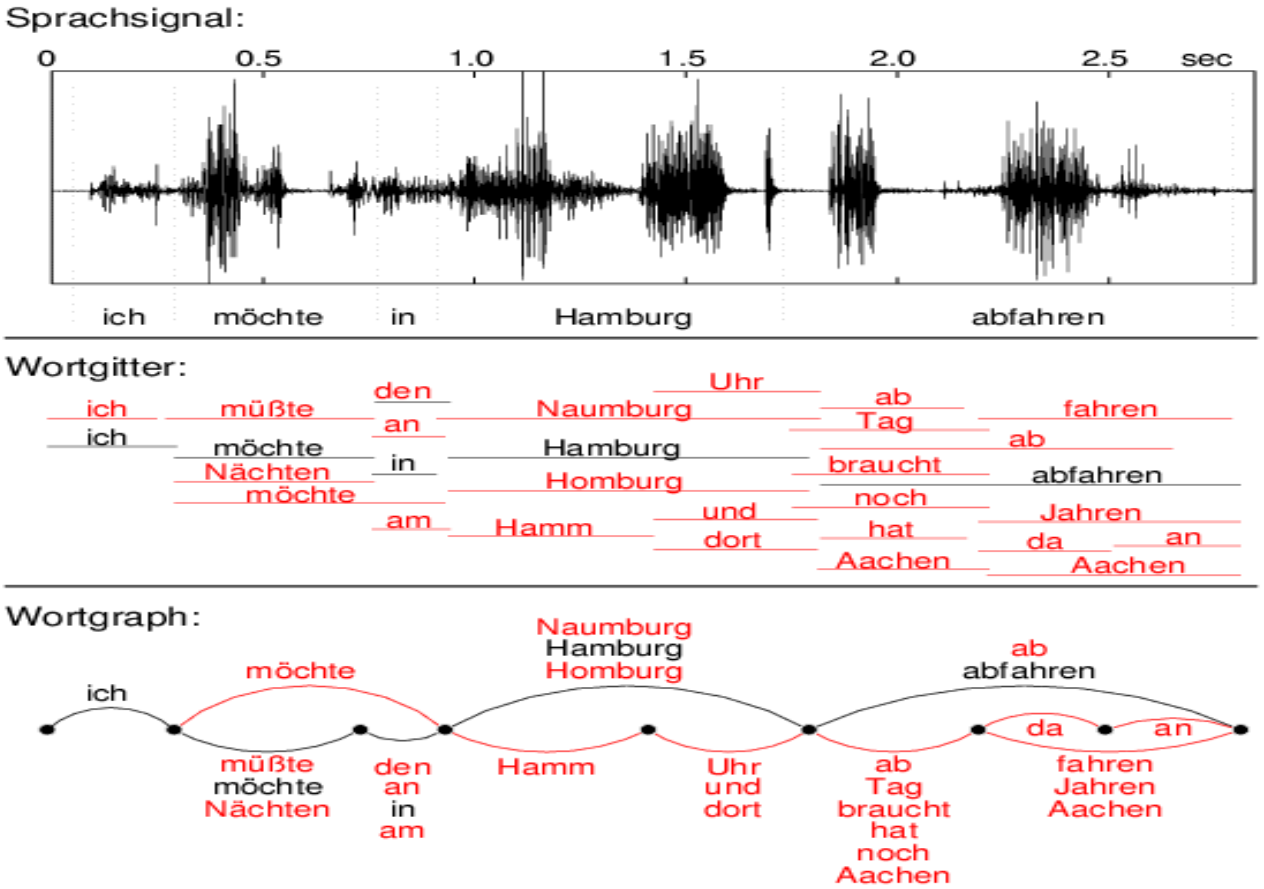
3. Einführung in Sprachein- und -ausgabe




HMM
1
2
3
4
5
Observation sequence
Boundary
[1, 1]
[1, 3]
[1, 5]
[3, 5]
[5, 5]


a_{11}
a_{22}
a_{33}
a_{44}
a_{55}
1
2
3
4
5
a_{13}
a_{24}
a_{35}
$b_1(o_t)$
$b_2(o_t)$
$b_3(o_t)$
$b_4(o_t)$
$b_5(o_t)$

WORTERKENNUNG: VOM SIGNAL ZUM WORTGRAPHEN



Prosodie und prosodische Merkmale

Prosodische Merkmale:

- Verlauf der Grundfrequenz der Sprache (F_0) – Intonation
- Intensität und Gesamtenergie des Sprachsignals
- Dauer (Länge) von Spracheinheiten
- Sprechgeschwindigkeit (Tempo)
- Rhythmus
- Länge von Pausen vor und nach einzelnen Wörtern
- Satz- und Wortakzent
- u. v. a.

Sprachausgabe – Sprachsynthese

- regelbasierte Systeme: hier werden Regeln aufgestellt, die die Verhältnisse in der gesprochenen Sprache möglichst genau wiedergeben sollen
- datenbasierte Systeme: diese Vorgehensweise erzeugt synthetische Sprache aus Einheiten der natürlichen Sprache
- Kopiensynthese: die natürliche Sprache wird einfach reproduziert

TTS – Text to Speech Systems (Text–Sprachsynthese)

Die Übersetzung von Text in Sprache, kurz als Text-Sprachsynthese bezeichnet, ist die Fähigkeit, elektronisch dargestellte Texte in verständlicher, menschlicher Sprache in der Form von Ton auszugeben.
Erste derartige kommerzielle Systeme bestanden einfach darin, Sequenzen von zuvor aufgezeichneten Sätzen, Wörtern, Silben oder Morphemen aneinanderzuhängen und auszugeben. Das Resultat glich gewöhnlich nicht gesprochener Sprache, sondern bestand im wesentlichen in der ''intelligenten'' Anordnung von Sprachteilen. Dies stört bei einem geringen Vokabularumfang von wenigen Wörtern und bei beschränkter Applikationsbreite kaum. Will man jedoch Sprachaufgaben auf eine breitere Basis stellen, so ist eine direkte Ausgabe beliebigen Textes in Sprachform zweckmäßig. Heute sind derartige Sprachsynthesesysteme verfügbar, die direkt von einem Eingabetext ausgehend einen hörbaren sprachlichen Ausdruck produzieren. Der Unterschied zwischen derartigen Systemen hinsichtlich ihrer Leistungsfähigkeit ist groß. Die besten derartigen Systeme sind in der Lage, beliebigen Text in entsprechende Sprachausgabe umzusetzen und dabei unter Umständen noch auf den Akzent des fiktiven Sprechers bzw. auf die semantisch korrekte Interpretation von Abkürzungen Rücksicht zu nehmen.

3. Einführung in Sprachein- und -ausgabe

Als Komponente weist Sprachsynthese einen Tongenerator, dessen Funktion analog zu der des menschlichen Vokaltraktes ist, auf. Zur Simulation benötigen wir ein Modul, dessen Eingabe der Text oder linguistische Information in einer anderen Form ist, welche gesprochen werden soll und deren Output einen Tongenerator treibt. In moderner Technologie sind sowohl Tongenerator als auch die Übersetzung von Text im Input für diesen Tongenerator programm-getriebene Komponenten. Hierbei haben wir zwei Fälle zu unterscheiden:

- **1) die Simulation des Vokaltraktes,**
- **2) die Simulation der Wellenform, die durch entsprechende Verformung des Vokaltraktes und das Hindurchströmen von Atemluft entsteht.**

Die übliche Parameterisierung in diesem Arbeitsbereich wird in der Form der Messung von Resonanzen und Antiresonanzen vorgenommen. In diesem Modell wird der Sprachproduktionsprozeß durch zwei Wellenformen simuliert:

- **a) eine Hauptwelle, welche die Basissprachwelle darstellt und**
- **b) einen Filter, welcher die oralen Einflüsse repräsentiert.**

Die wichtigsten akustischen Parameter der Sprachsynthese sind die Grundfrequenz der Hauptwelle und die Frequenzen der ersten Reihe engerer Resonanzwellen.

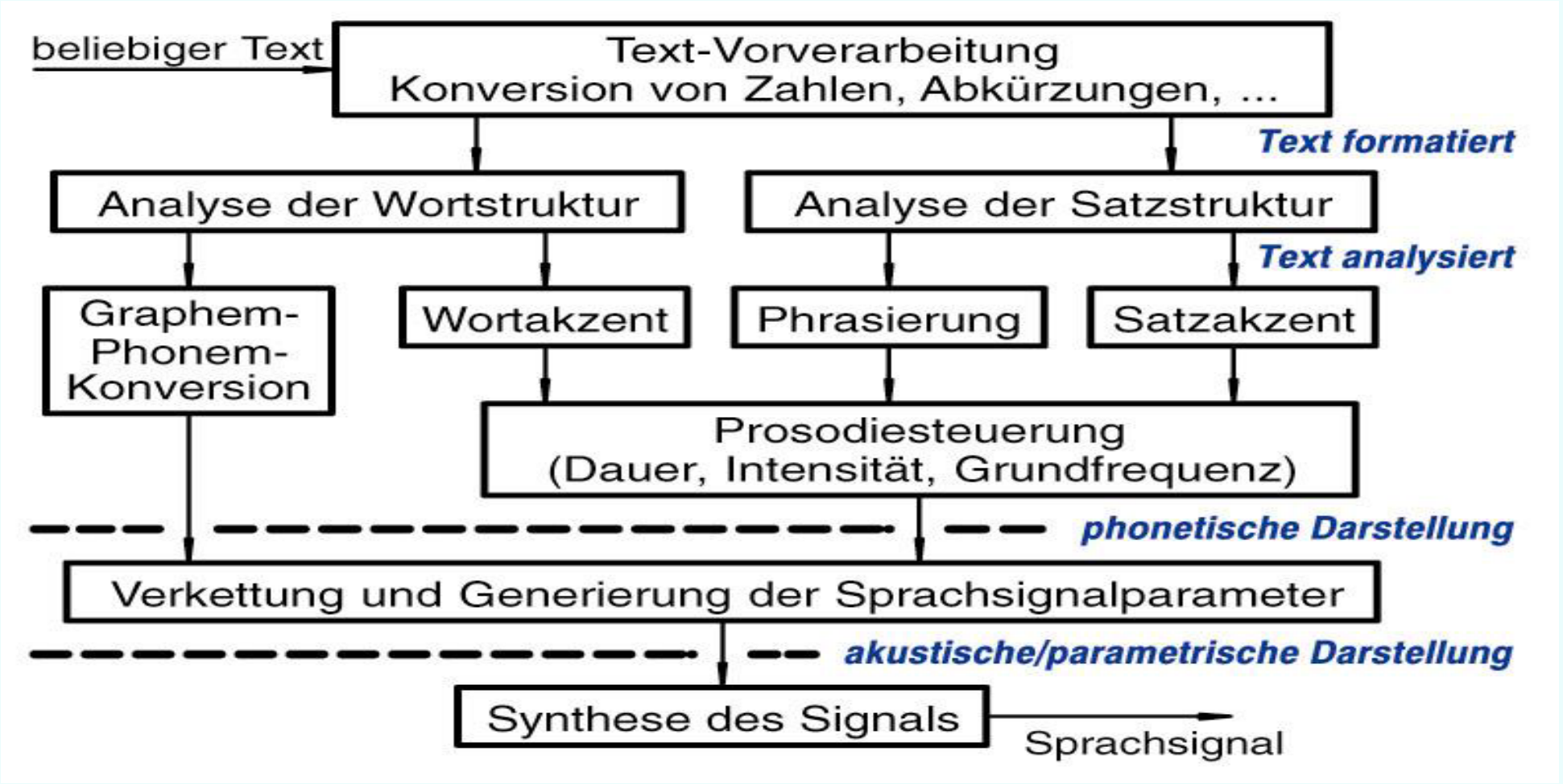
Sprachsynthese als Teil virtueller Realität

Wenn man über "Sprachsynthese" spricht, erinnert das oft an den Klang einer kalten metallischen Stimme, deren monotone, mechanische Intonation sich ständig wiederholt, ohne irgendeinen Zusammenhang zu dem, was gesagt wird. Andererseits denkt man, wenn von "virtueller Realität" die Rede ist, an erstaunlich realistische (leider oft gewalttätige) Filmszenen. Nun hat die Sprachsynthese im Laufe der letzten zehn Jahre jedoch eine beachtliche Entwicklung durchlaufen, und seitdem ist sie Teil der virtuellen Realität.

Sprachsynthese klingt heute viel natürlicher als je, viel realistischer. Und damit ergeben sich völlig neue Anwendungen für Geistes- und Sprachwissenschaften, um die verschiedenen Geheimnisse der menschlichen Existenz und der menschlichen Sprache zu entdecken.


Schritte der Sprachsynthese (TTS)
• Strukturanalyse
– Erkennung von Paragraphen, Sätzen, ...
– „.“ in U.S.A. nicht als Satzende interpretieren
• Text-Vorverarbeitung
– Behandlung von Abkürzungen, Akronymen, Datums- und Zeit-, Zahlen- und -angaben, u.v.a
– Lesart von 8/5, 1998, tedwards@cat..com
→zu sprechender Text
• Text-nach-Phonem-Konvertierung
• Prosodie-Analyse
• Sprachschall-Erzeugung
#314

Blockdiagramm eines TTS-Systems



Verfahren der Sprachsynthese

1. Formant- oder Regelsynthese:

Eine "einfache" Wellenform wird durch entsprechende Filterung in Sprachsignale umgewandelt. Dieses Verfahren benötigt sehr viele Regeln, welche Laute in welchen Kontexten wie realisiert werden. Da alle Parameter des Systems durch Regeln zugänglich sind, lassen sich z.B. Intonation und Lautdauer leicht steuern. Dadurch kann man sehr variationsreiche Sprache generieren. Der größte Nachteil dieser Systeme ist die mangelnde Natürlichkeit der Stimme. Englische Sprachausgaben, die mit Formantsynthese arbeiten, sind: DecTalk und Eloquent.

2. Konkatenationssynthese:

Fast alle derzeit verwendeten Sprachausgabesysteme arbeiten mit diesem Verfahren. Dabei werden sprachliche Äußerungen aufgenommen, es werden Teile daraus ausgeschnitten und wieder zu neuen Äußerungen zusammengesetzt. Die Größe dieser Teile reicht von ganzen Wörtern und Phrasen (z.B. Ansage der Flüge im Flughafen Frankfurt) bis zu Einheiten, die kleiner als Laute sind (z.B. Mikrosegmente). Mit diesen kleinen Einheiten kann man jeden beliebigen Text (einer Sprache) vorlesen lassen. Ein grundsätzliches Problem bei diesem Verfahren ist, dass die aufgenommenen Sprachbausteine sich nicht so leicht in Dauer und Tonhöhe verändern lassen. Die technischen Verfahren, die dies ermöglichen, gehen immer mit einer Qualitätseinbuße bei der Sprachqualität einher und/oder führen zu einer unnatürlicheren Stimme. Der Vorteil dieser Herangehensweise ist die menschliche Qualität der Stimme. Man kann die Menschen, die dahinterstecken, wiedererkennen.

- **Konkatenative Verfahren**
  - Diphonsynthese
  - Halbsilbensynthese
  - Großinventare

3. Artikulatorische Synthese (artikulatorische Modelle)

Dieses Verfahren ist sehr rechenintensiv und wird nur zu Forschungszwecken verwendet. Die Bewegungen der "Sprechorgane" beim Sprechen werden modellhaft nachgebildet und basierend auf der Positionierung der "Sprechorgane" werden die Resonanzeigenschaften im Rachen-, Mund- und Nasenraum berechnet.

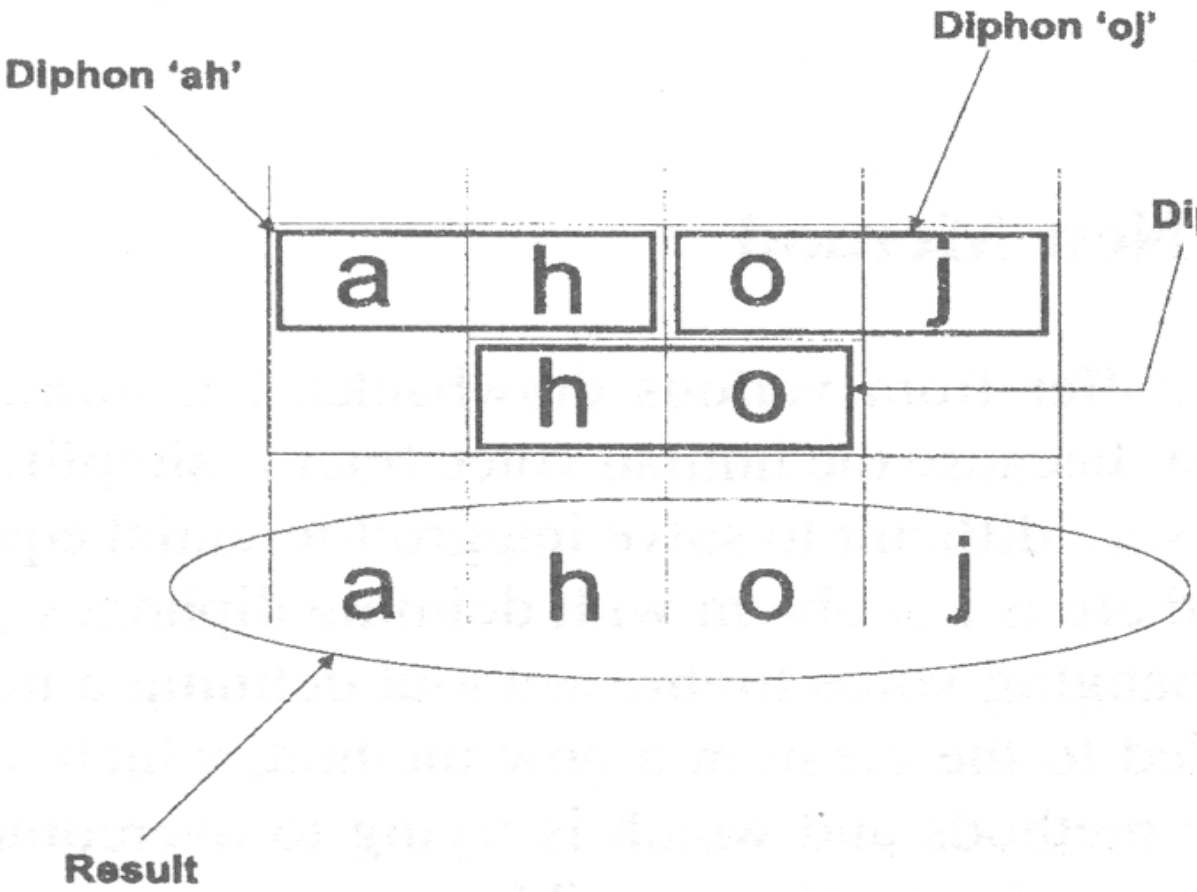


Fig. 1. Synthesis of word "ahoj".

Formantsynthese	Diphonsynthese	Mikrosegmentsynthese
Vorteile • geringer Speicherplatz-bedarf • einfache Veränderung akustischer Parameter	Vorteile • geringer Speicherplatz-bedarf • einfache Veränderung akustischer Parameter	Vorteile • Wiedererkennbare Stimme • einfache Stimmgenerierungs-regeln • geringer Speicherplatzbedarf • Prosodiesteuerung im Zeitbereich • wenige Mikrosegmente • schneller neue Stimmen

Nachteile	Nachteile	Nachteile
• Synthetischer Klang • aufwendige Regelsätze	• hoher Speicherplatzbedarf • aufwendige Resyntheseverfahren zur Prosodiemodellierung	• noch nicht vollendet

http://pcweb.ikp.uni-bonn.de/~kst/sprachsynthese.htm

http://www.ias.et.tu-dresden.de/kom/lehre/

Die Grenzen der bestehenden Sprachsynthese

Gleichzeitig ist die Zeit sicher noch nicht reif für weitere potenzielle Anwendungen von Sprachsynthese:

- **Synthetische Stimmen fehlt immer noch die Ausdruckskraft. Sie können menschliche Emotionen wie Freude, Ärger oder Trauer nicht ausdrücken. Das heißt, dass unsere künstlichen Stimmen noch nicht über die volle, ausgebildete "Stimmenpalette" menschlicher Sprecher verfügen.**
- **Die Anzahl synthetischer Stimmen selbst ist noch sehr beschränkt. Im besten Fall haben wir für eine bestimmte Sprache mehrere erwachsene männliche und weibliche Stimmen zur Verfügung. Aber wo bleiben die Kinderstimmen, wo diejenigen der Jugendlichen, der älteren Menschen? Mit der heute meistverwendeten konkatenativen Technologie bleibt die Schaffung neuer Stimmen mit einem enormen Aufwand verbunden, sogar für personell und finanziell gut dotierte Forschungsteams.**
- **Bisher existiert noch kaum eine Synthese von nicht standardsprachlichen Varietäten. Es gibt keine Dialektsynthese, es gibt keine Synthese informeller Sprache und noch keine Simulation sozialer Varietäten.**