

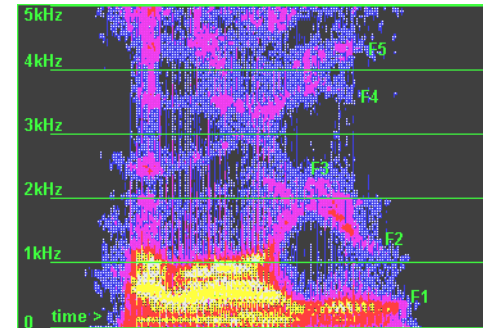
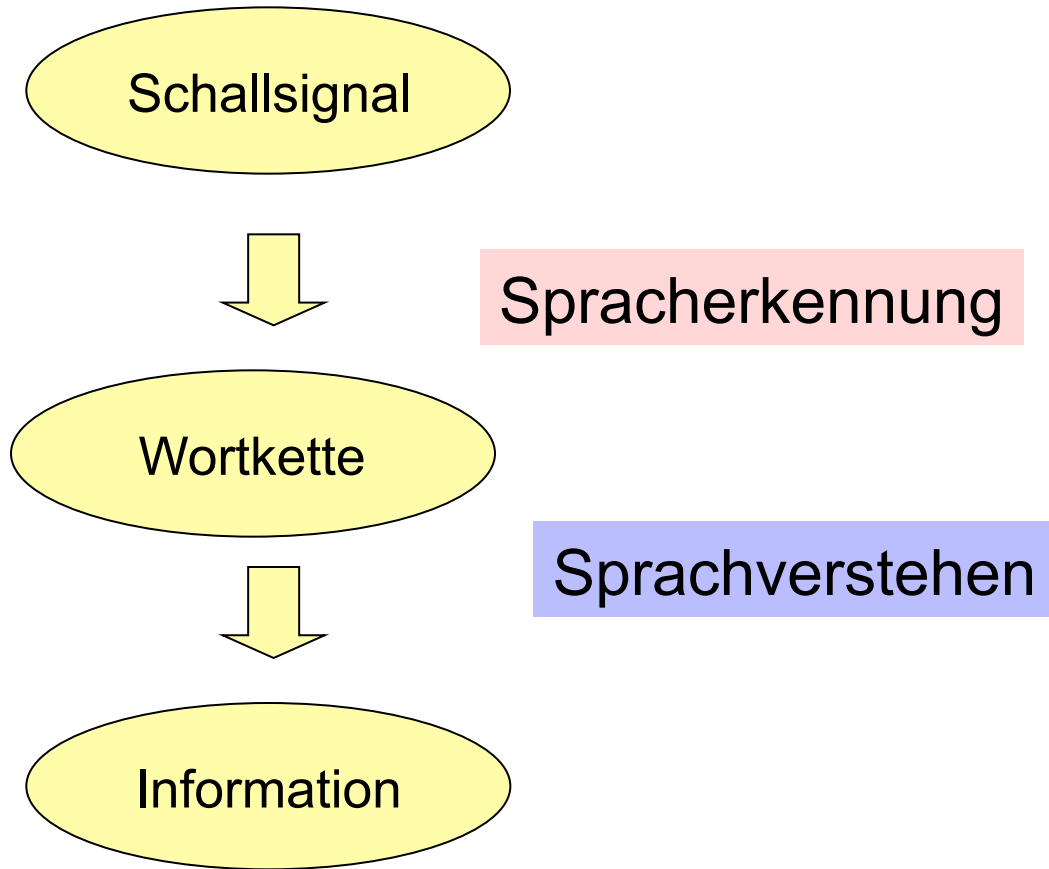
# Einführung in die Computerlinguistik

Verarbeitung gesprochener Sprache

WS 2014/2015

Vera Demberg

# Sprachverarbeitung

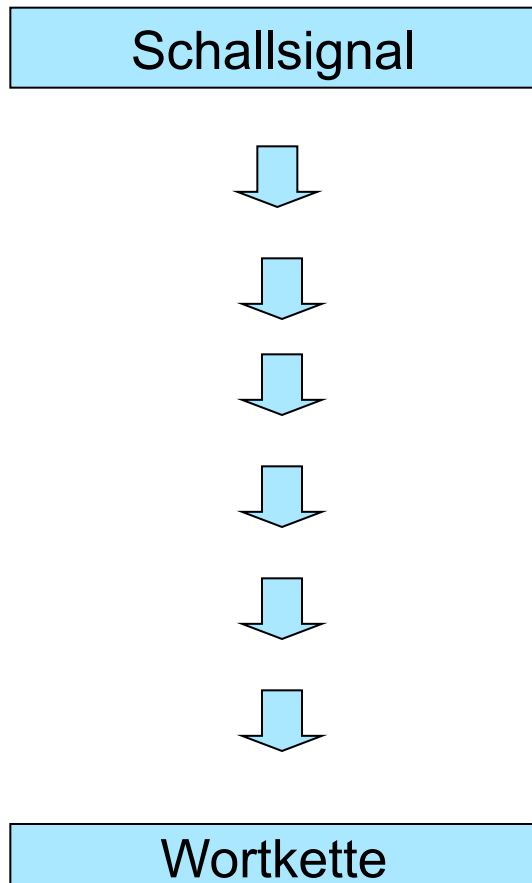


A yellow arrow pointing downwards from the spectrogram to the text below.

Laura schläft

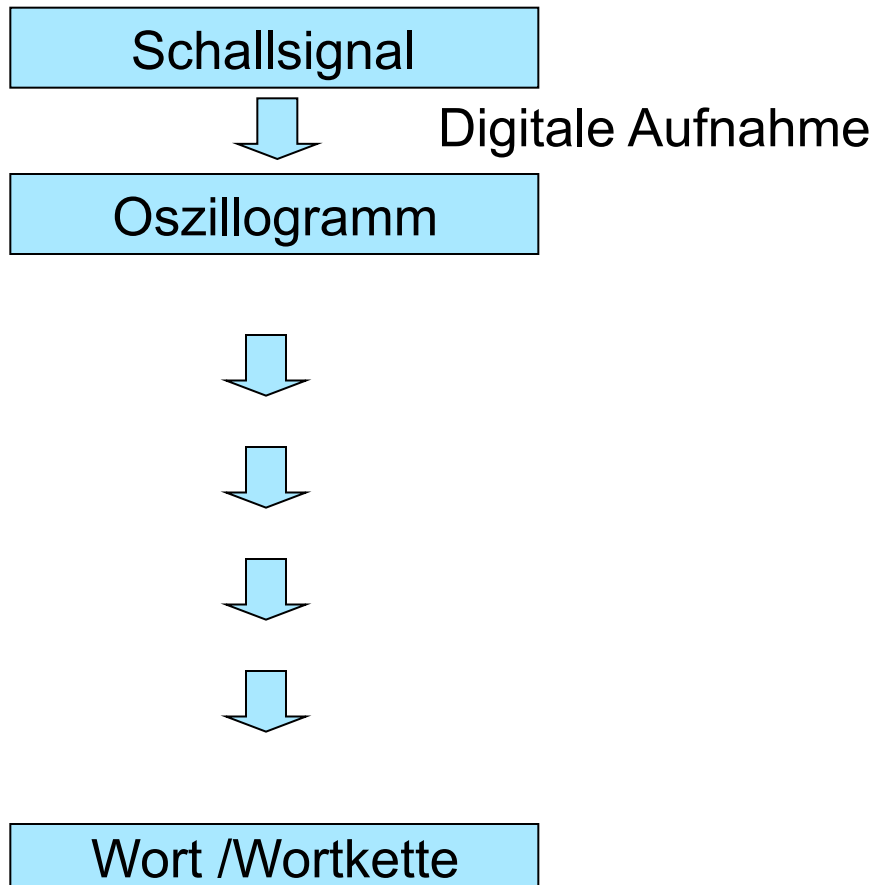


# Spracherkennung

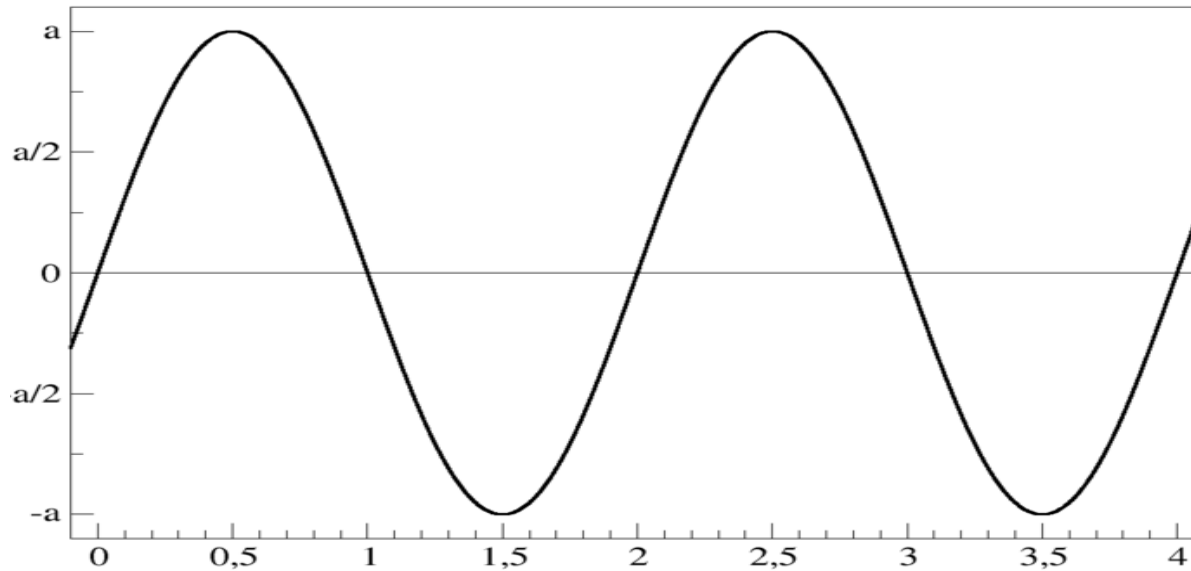


- Die Grundaufgabe der Spracherkennung:
- Gegeben ein kontinuierliches Schallsignal.
- Welche Wortkette wurde vom Sprecher geäußert?

# Spracherkennung

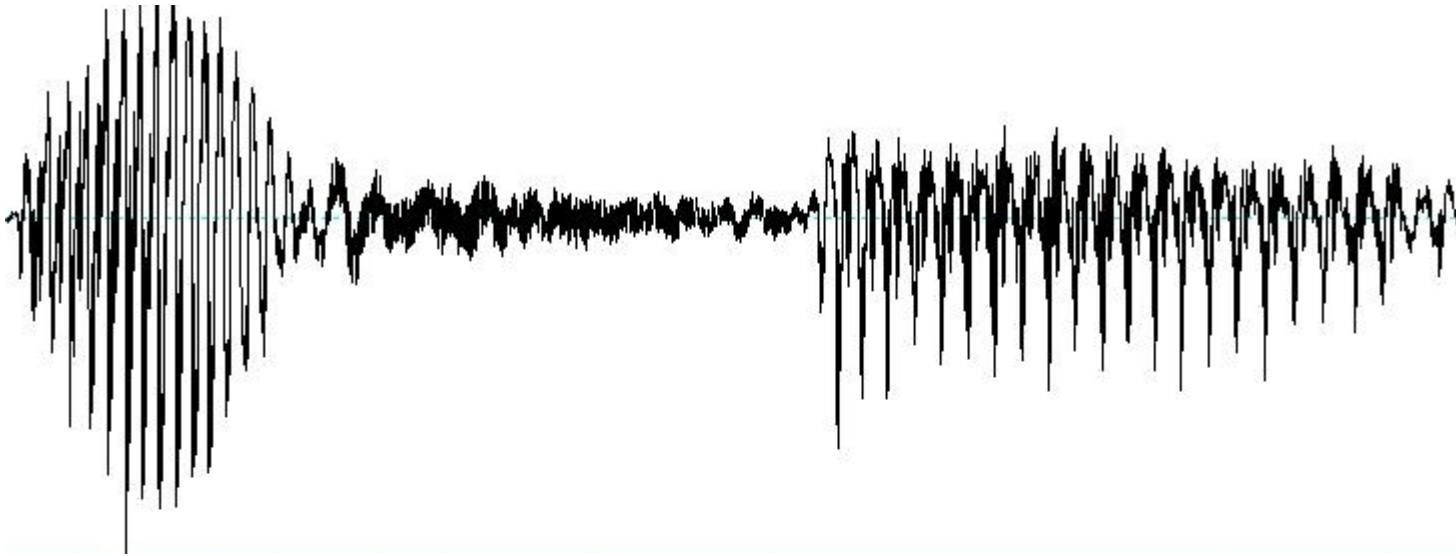


# Reine Schwingung



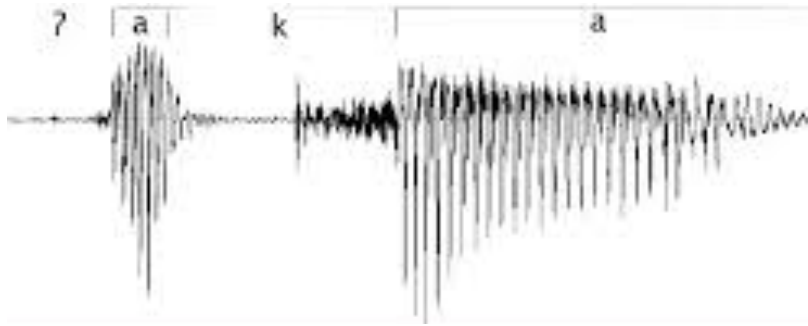
# Ein Oszillogramm

- Das Oszillogramm für „afa“



# Oszillogramme

aka



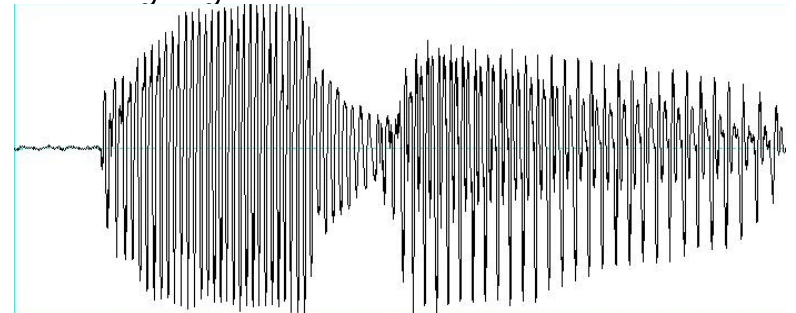
ama



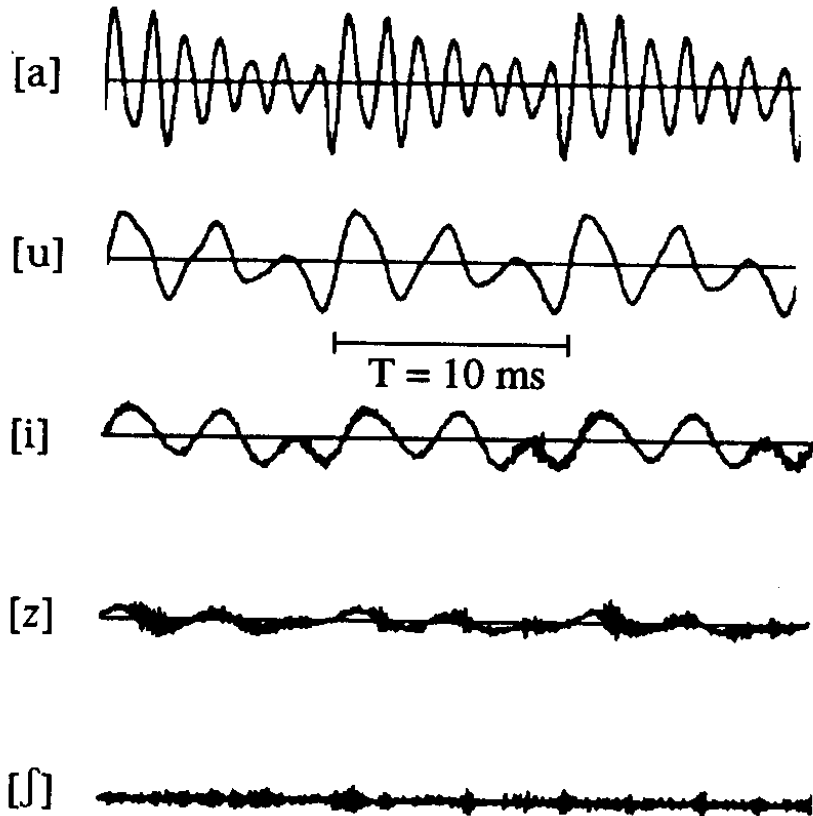
acha



ydy



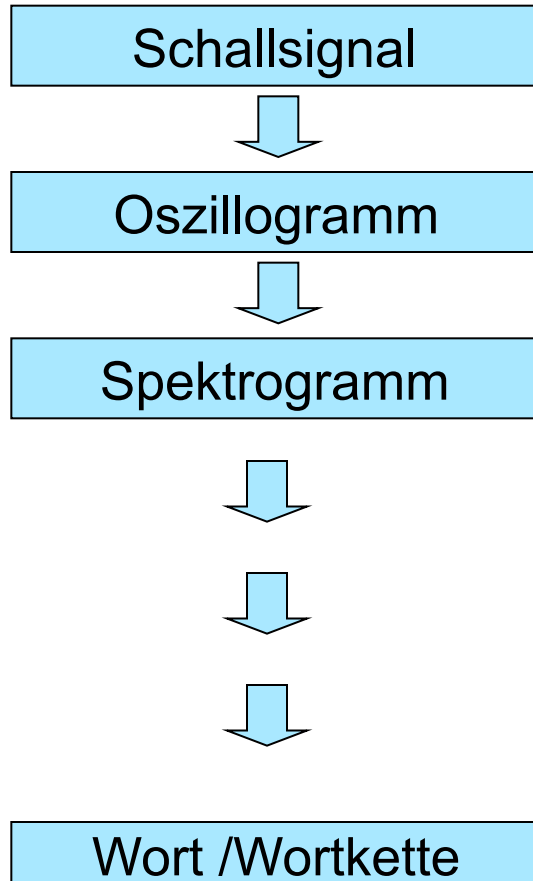
# Einzelne Laute als Oszillogramme



- Laute werden charakterisiert durch Kombination von Schwingungen verschiedener Frequenzen
- Im Oszillogramm **schwer erkennbar** (Überlagerung)
- Deshalb: Überführung in Zeit-Frequenz-Diagramm (**Spektrogramm**) mittels Komponentenanalyse (Fourier-Transformation)



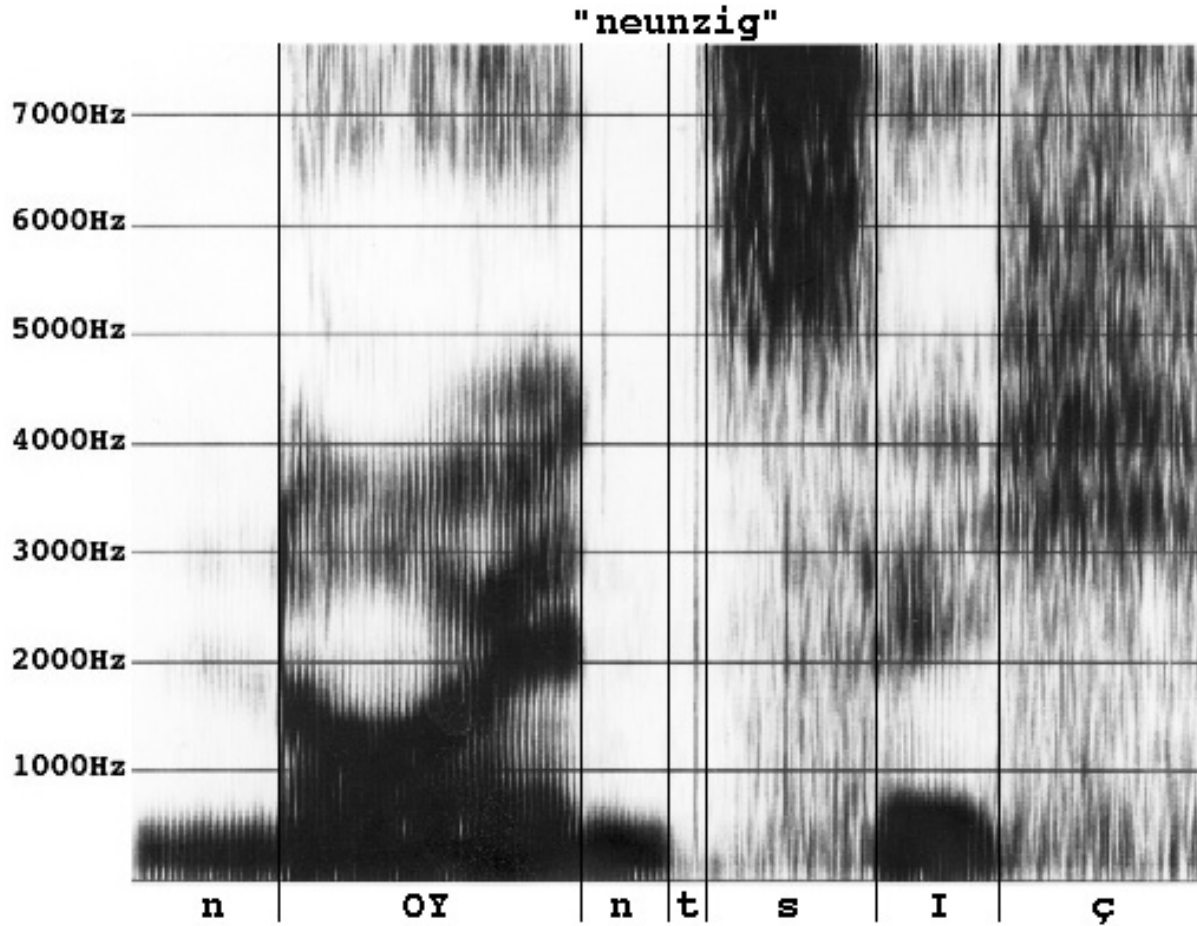
# Spracherkennung: (Vereinfachtes) Schema



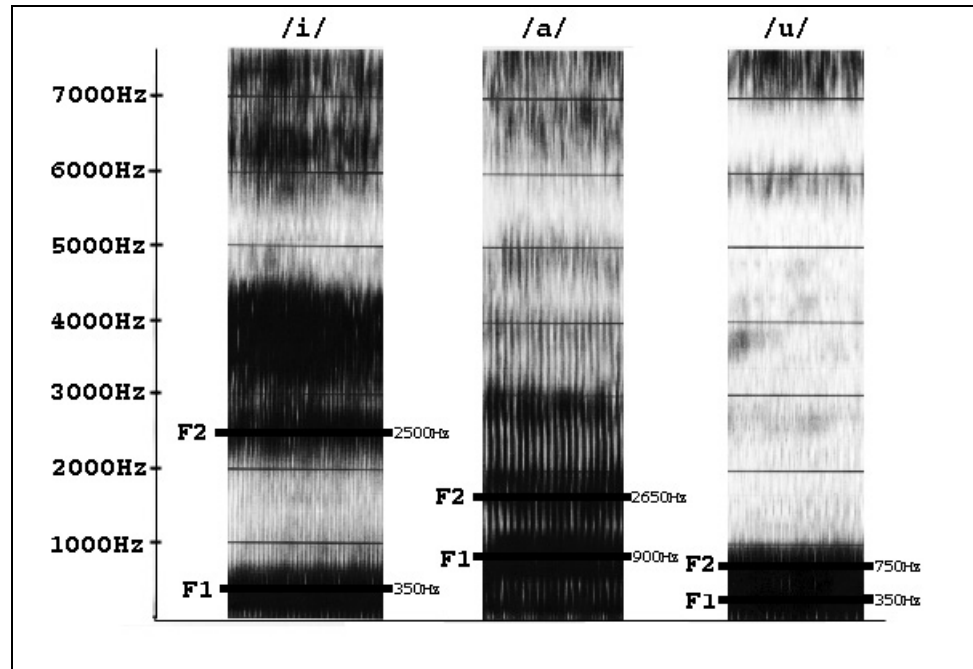
Digitale Aufnahme

Zerlegung in Einzelfrequenzen

# Spektrogramm für eine Aufnahme von „neunzig“

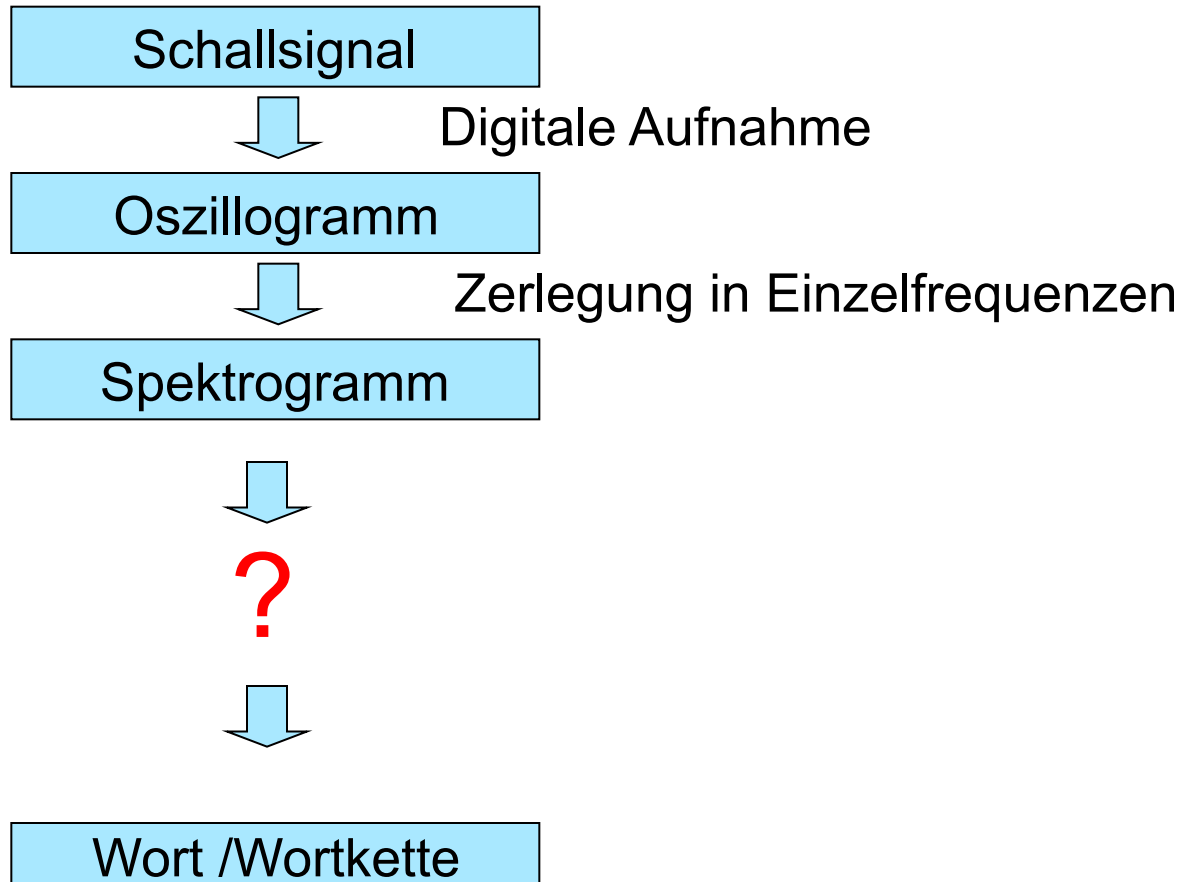


# Spektrogramm für die Vokale i,a,u



- Dunkle Färbung: große Schallenergie in einem bestimmten Frequenzbereich.
- Die **Formanten** (Obertöne) F1 und F2 sind für die charakteristische Vokalqualität verantwortlich.
- Der Verlauf des **Basisformanten** F0 (hier nicht sichtbar) gibt die Intonation der Äußerung wieder.

# Spracherkennung: (Vereinfachtes) Schema



# Spracherkennung: Erster Versuch

- Identifikation von Lautgrenzen im Spektrogramm (Segmentierung)
- Abgleich der Spektrogramm-segmente mit einer Datenbank "idealer" Laute (Identifikation)
- Verknüpfung der identifizierten Laute zu Wörtern und Sätzen.
- **Funktioniert nicht**, wegen der **Varianz des Signals**.

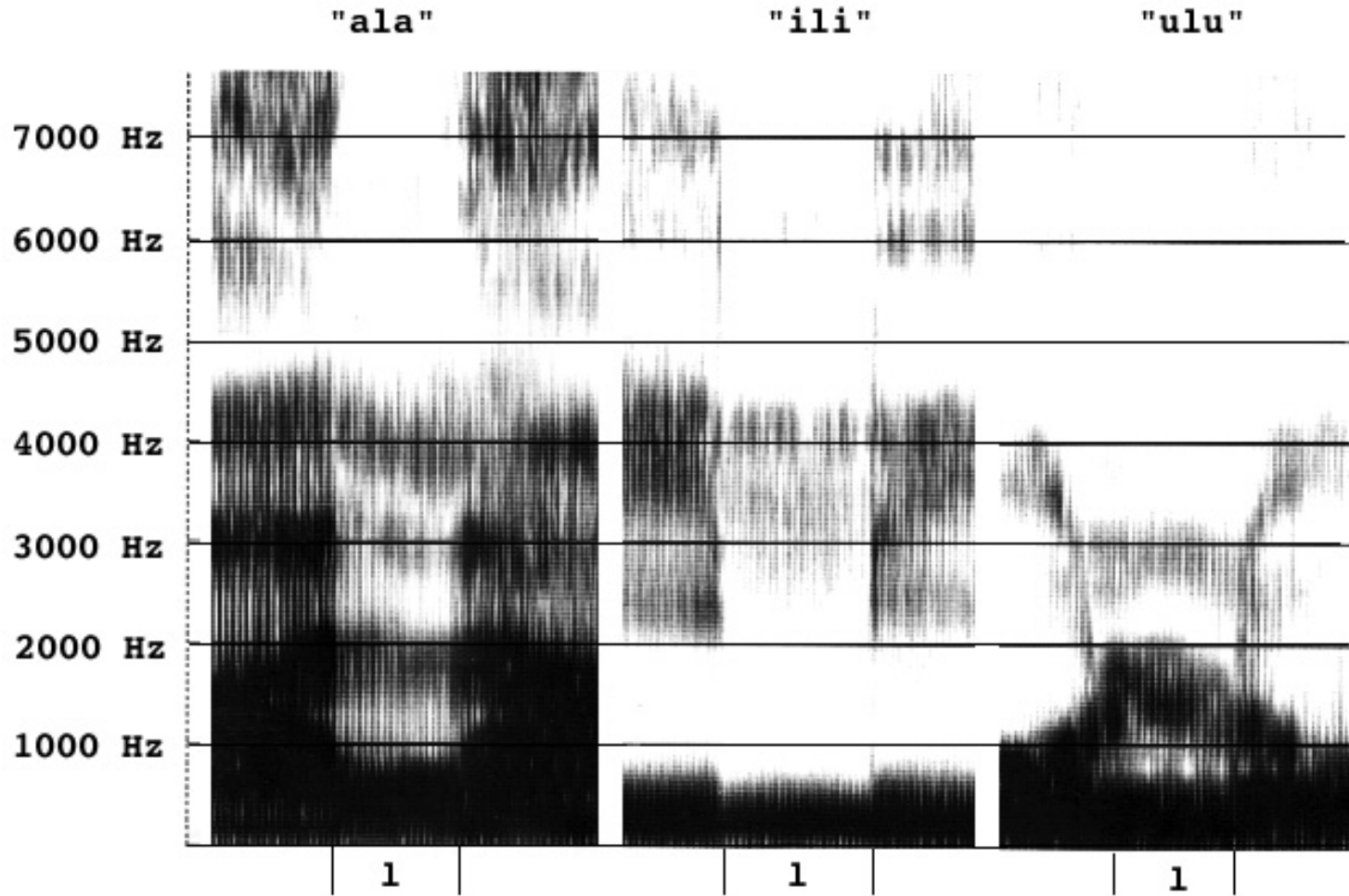
# Problem 1: Varianz des Signals

- Gleicher Laut/ gleiches Wort wird nicht immer gleich ausgesprochen
  - Verschiedene Dialekte
  - Verschiedene Sprecher
  - Unterschiedliche Sprechgeschwindigkeit
  - Physischer und emotionaler Zustand des Sprechers
  - Abhängig von Tonhöhe und Akzent
- Sprachexterne Einflüsse verändern das Signal
  - Raumakustik, Hall, Entfernung
  - Medium: direkte Kommunikation, Telefon, Handy
  - Mikrofonqualität und -charakteristik
  - Hintergrundgeräusche

# Spracherkennung: Zweiter Versuch

- Identifikation von Lautgrenzen im Spektrogramm (Segmentierung)
- Erstellung eines Trainingskorpus mit Lautannotationen (alignierte phonetische Annotation)
- Bestimmung von Merkmalsmustern für die Spektrogrammsegmente
- Training eines statistischen Laut-Klassifikators
- **Funktioniert nicht**, vor allem wegen der **Kontinuität des Signals**.

# Kontextabhängigkeit

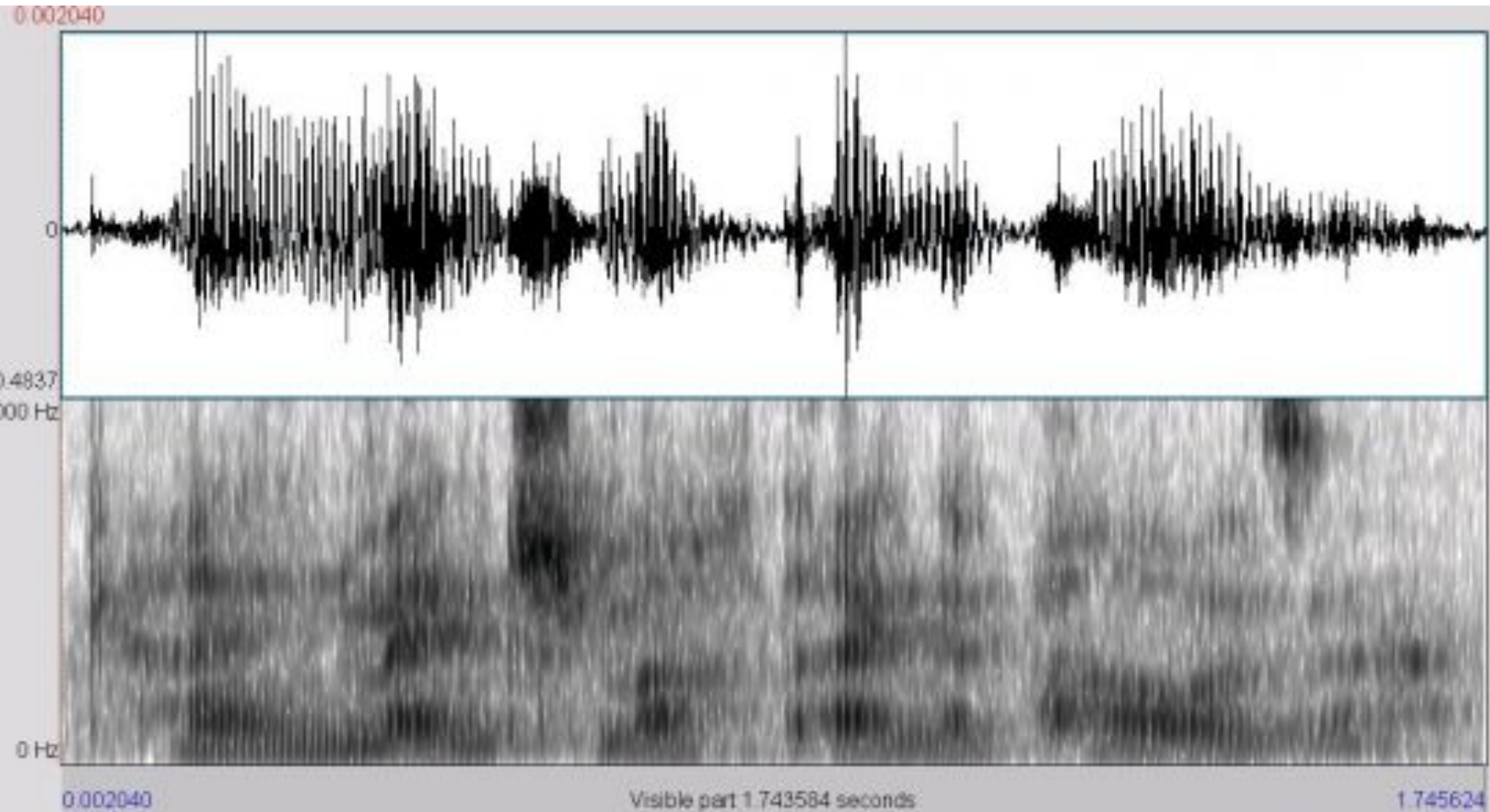




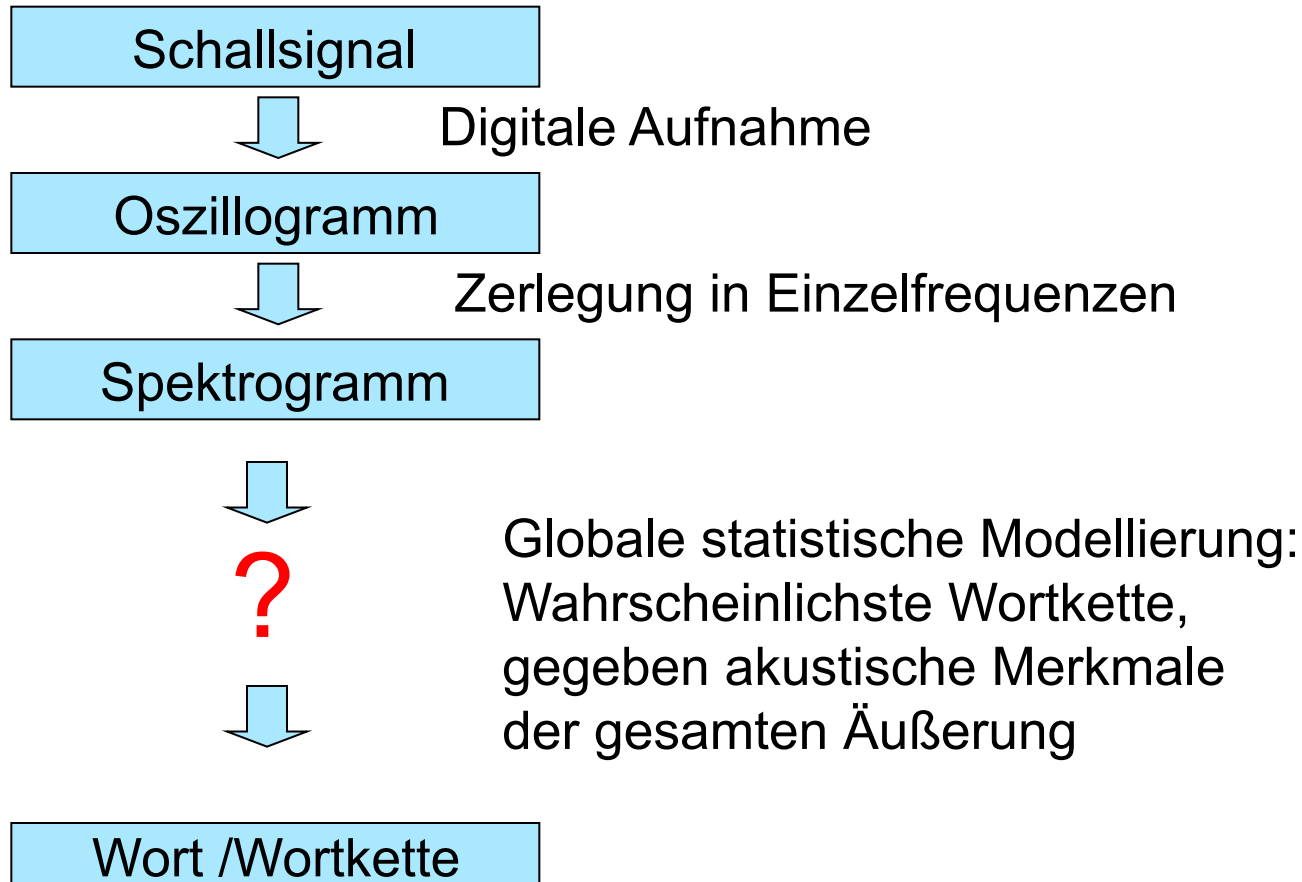
## Problem 2: Kontinuität des Signals

- Die **Laute** eines Wortes lassen sich schwer gegeneinander abgrenzen
  - Wo hört Laut 1 auf, wo fängt Laut 2 an?
  - Dazu kommt das Phänomen der **Koartikulation**: Laute beeinflussen sich gegenseitig.
    - In Lautfolgen wie [am], [um], [an] kann man nicht den Vokal vom Nasal trennen: Vokal hat Nasal-Qualität und umgekehrt.
    - /k/ wird verschieden realisiert in Koffer, Kind, Kabel
- **Wörter** sind nur in der Orthografie sauber getrennt.
  - In der gesprochenen Sprache gibt es zwischen Wörtern meistens keine Pause
  - Pausen kommen in spontaner Sprache auch innerhalb von Wörtern vor

# „Kein Mensch macht eine Pause.“



# Spracherkennung: (Vereinfachtes) Schema



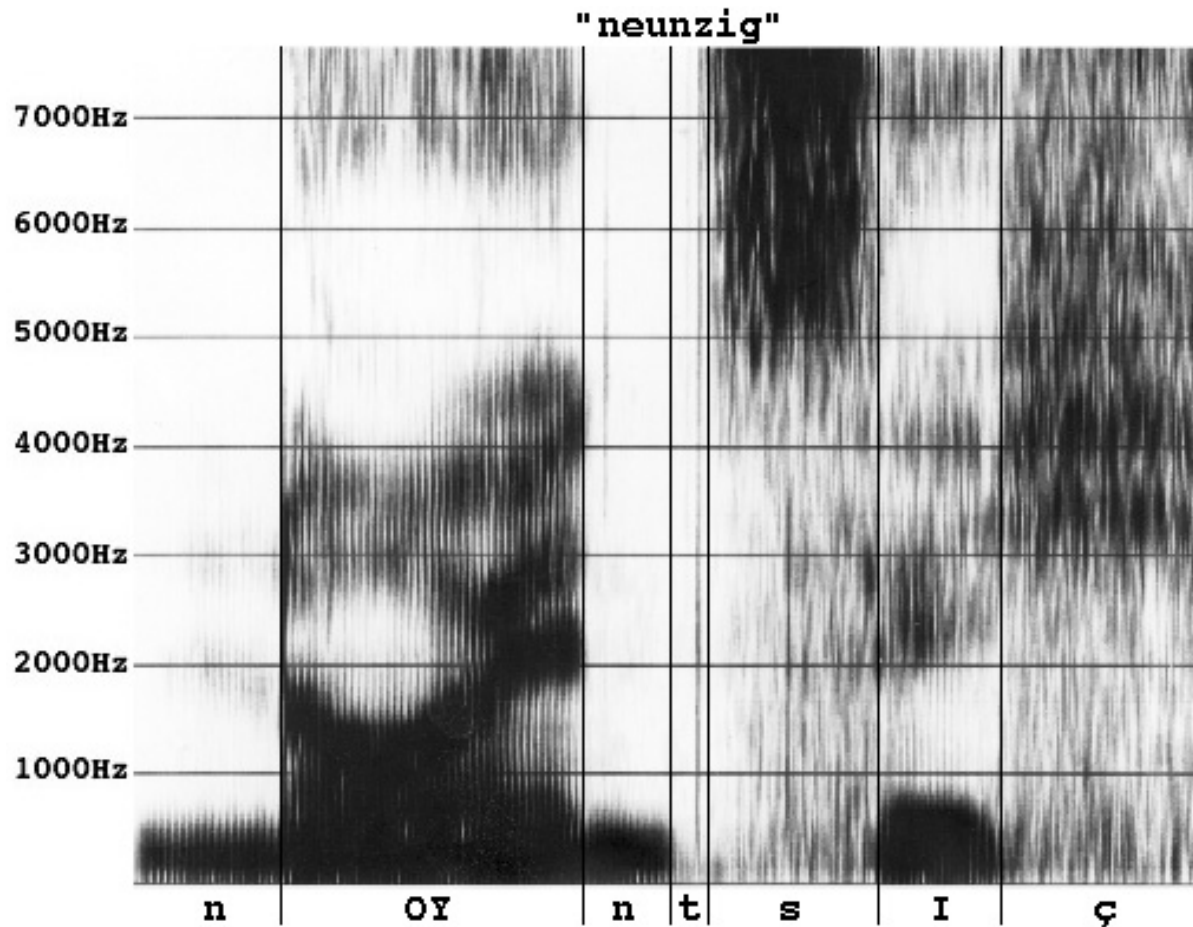
# Statistische Modellierung: Allgemeines Schema

- Manuelle Korpusannotation
- Merkmalspezifikation
- Automatische Merkmalsextraktion
- Training eines statistischen Modells
- Evaluierung

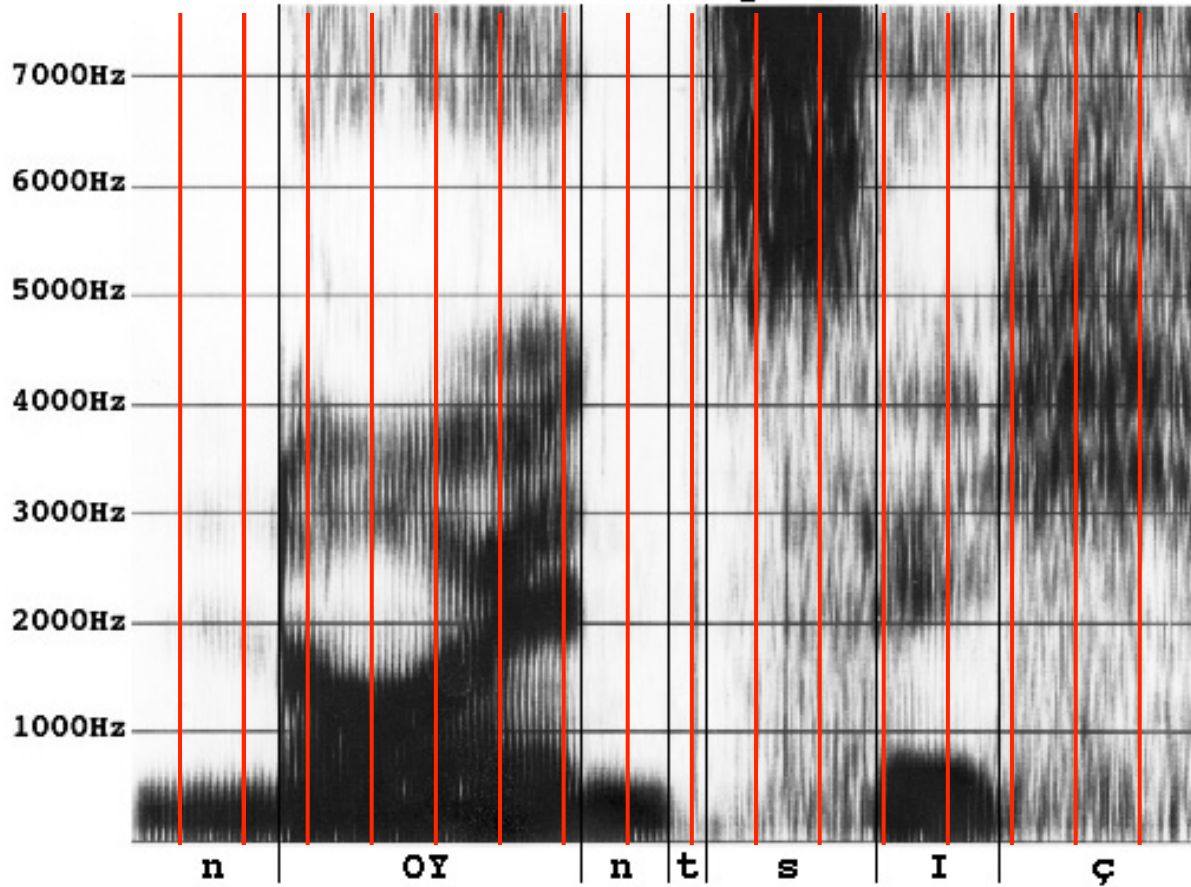
# Merkmalspezifikation

- Was sind die Einheiten, von denen wir ausgehen?
  - Zerlegung des Signals in „Beobachtungen“:  
Zeitfenster von z.B. 30 ms

# Spektrogramm für ein deutsches Wort



# Merkmalspezifikation: Zeitfenster "neunzig"

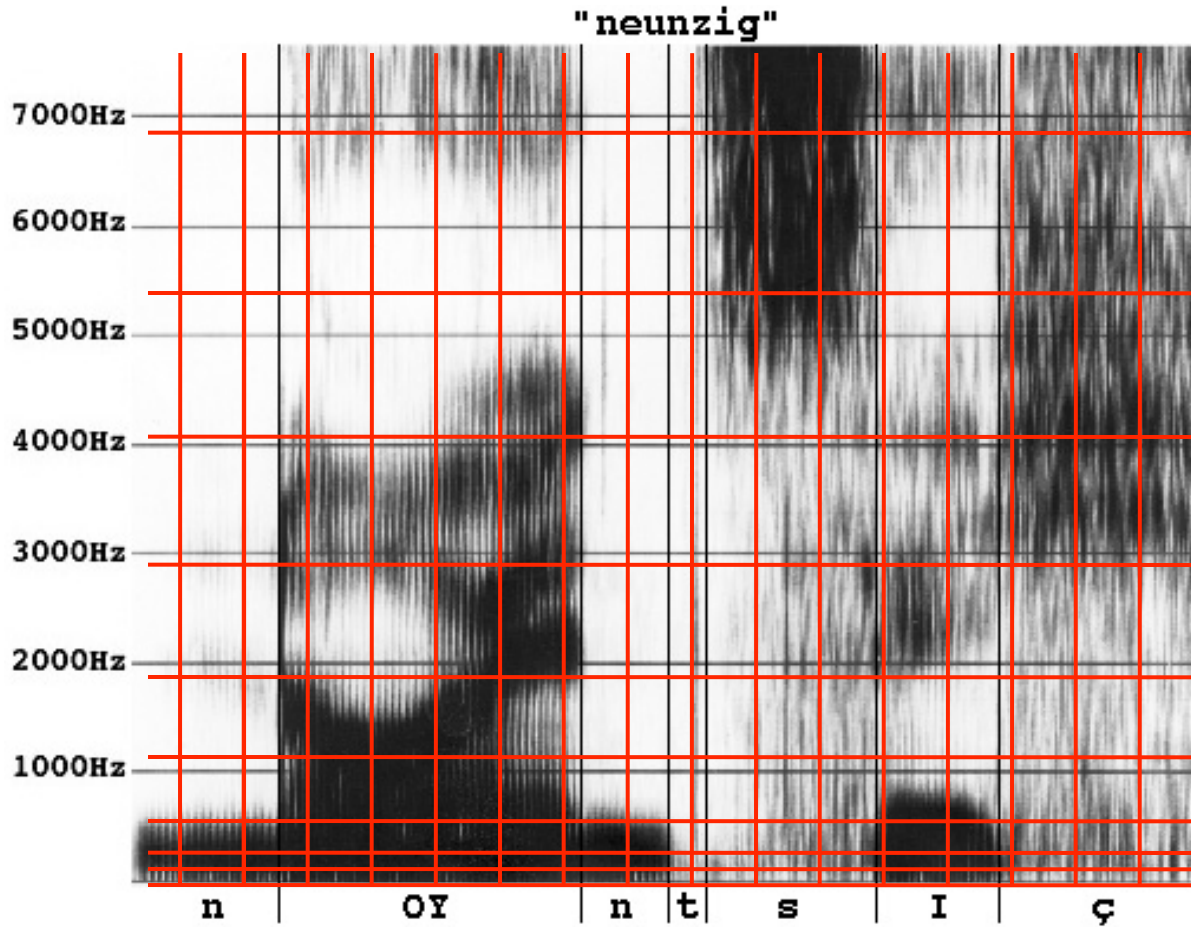


# Merkmalspezifikation/-extraktion

- Zerlegung des Signals in „Beobachtungen“:  
Zeitfenster von z.B. 30 ms
- Zerlegung jeder Beobachtung in Frequenzintervalle  
(z.B. Vierteltonschritte im Standard-12-Ton-System)



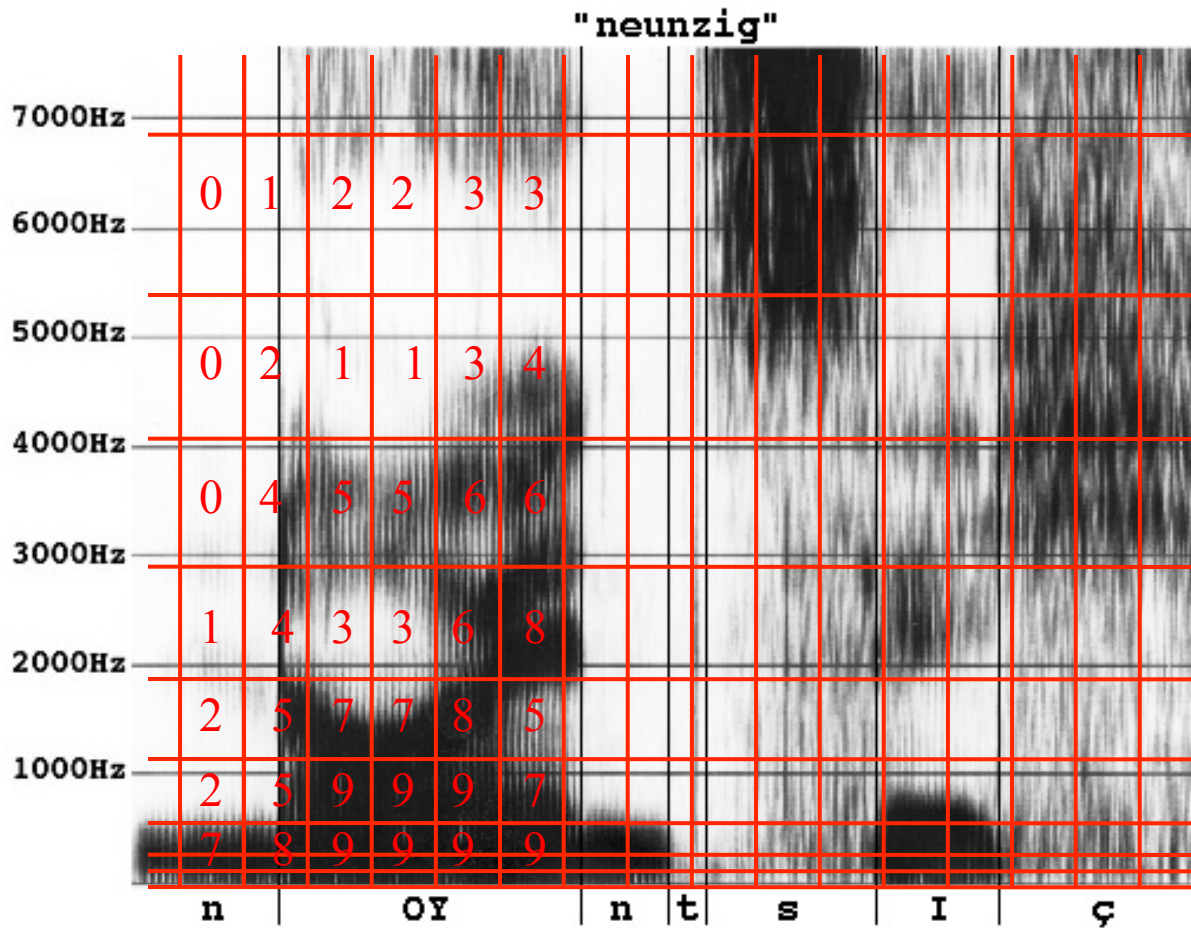
# Spektrogramm für ein deutsches Wort



# Merkmalspezifikation/-extraktion

- Zerlegung des Signals in „Beobachtungen“:  
Zeitfenster von z.B. 30 ms
- Zerlegung jeder Beobachtung in Frequenzintervalle  
(z.B. Vierteltonschritte im Standard-12-Ton-System)
- Bestimmung des Schalldrucks (Schallenergie) in  
jedem Zeit-Frequenz-Fenster
- Resultat: Eine Folge von Einzelbeobachtungen, die  
durch Merkmalsvektoren charakterisiert sind

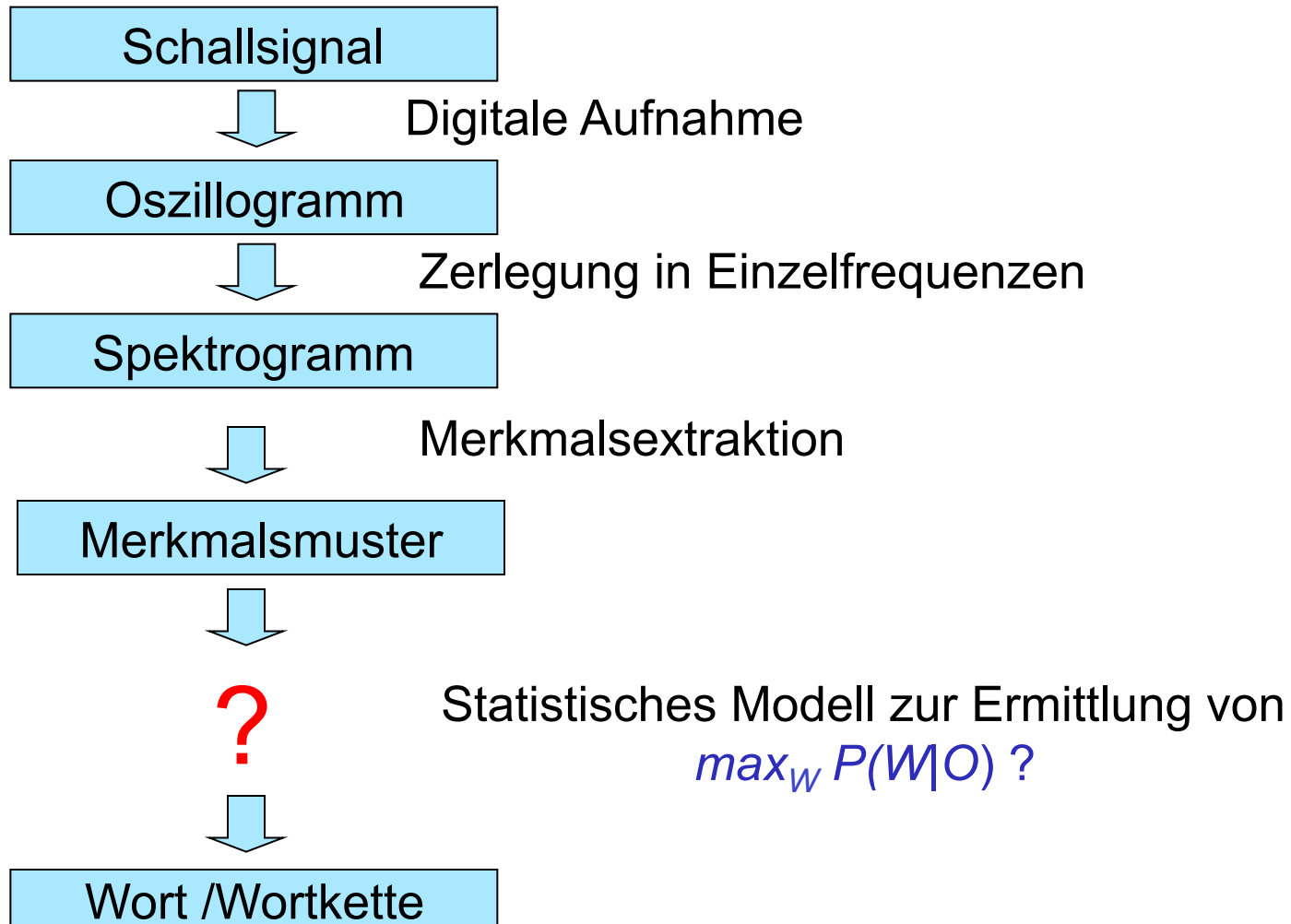
# Spektrogramm für ein deutsches Wort



# Merkmalsmuster, Ausschnitt

0	1	2	2	3	3	...													
0	2	1	1	3	4	...													
0	4	5	5	6	6	...													
1	4	3	3	6	8	...													
2	5	7	7	8	5	...													
2	5	9	9	9	7	...													
7	8	9	9	9	9	...													

# Spracherkennung: (Vereinfachtes) Schema



# Statistische Modellierung

- Aufgabe: Ermittle für ein Eingangesignal, das durch eine Folge von Beobachtungen/ Vektoren  $O = o_1 o_2 \dots o_m$  charakterisiert ist:

$$\max_W P(W|O) = P(w_1 w_2 \dots w_n | o_1 o_2 \dots o_m)$$

- Wir suchen die wahrscheinlichste Wortfolge gegeben die Beobachtungen  $O$ .  $\rightarrow$  sparse-data Problem!
- Erster Schritt: Verwendung des Bayes-Theorems.

# Das Bayessche Theorem

- Das Bayessche Theorem oder die Bayes-Regel:

$$P(E | F) = \frac{P(F | E) \cdot P(E)}{P(F)}$$

- Die Bayes-Regel ist ein elementares Gesetz der Wahrscheinlichkeitstheorie. Sie ist überall da nützlich, wo der Schluss von einer Größe F auf eine andere Größe E bestimmt werden soll (typischerweise von einem Symptom auf eine relevante Eigenschaft/ die Ursache), die Abhängigkeit in der anderen Richtung (von der Ursache auf das Symptom) aber besser zugänglich ist.

## Wie bestimmen wir $P(W|O)$ ?

- **Symptom:** Folge von akustischen Beobachtungen  $O = o_1 o_2 \dots o_m$
- **Ursache:**  
vom Sprecher geäußerte, intendierte Wortkette  $W = w_1 w_2 \dots w_n$
- Mit Bayes-Regel : 
$$P(W | O) = \frac{P(O | W) \cdot P(W)}{P(O)}$$



## Wie bestimmen wir $P(W|O)$ ?

- **Symptom:** Folge von akustischen Beobachtungen  $O = o_1 o_2 \dots o_m$
- **Ursache:**  
vom Sprecher geäußerte, intendierte Wortkette  $W = w_1 w_2 \dots w_n$
- Mit Bayes-Regel : 
$$P(W | O) = \frac{P(O | W) \cdot P(W)}{P(O)}$$
- Die wahrscheinlichste Wortkette: 
$$\begin{aligned} \max_W P(W | O) &= \max_W \frac{P(O | W) \cdot P(W)}{P(O)} \\ &= \max_W P(O | W) \cdot P(W) \end{aligned}$$
- $P(W)$  ist die globale, "a priori"-Wahrscheinlichkeit der Wortkette  $W$ .
- $P(O)$ , die Wahrscheinlichkeit des Merkmalsmusters, wird nicht mehr benötigt.

# Akustisches Modell und Sprachmodell

$$\max_W P(W | O) = \max_W P(O | W) \cdot P(W)$$

- $P(O|W)$  ist die Wahrscheinlichkeit, dass eine Wortfolge in einer bestimmten (durch den Merkmalsvektor bezeichneten) Weise ausgesprochen wird: **Akustisches Modell**
- $P(W)$  ist die Wahrscheinlichkeit, dass eine bestimmte Wortfolge geäußert wird: „**Sprachmodell**“

# Sprachmodelle

$$\max_W P(W | O) = \max_W P(O | W) \cdot P(W)$$

- Wie berechnen wir  $P(W) = P(w_1 w_2 \dots w_n)$  ?
- Grundlage ist die Frequenz von Wortfolgen in Korpora.
- Sparse-Data-Problem: Ganze Sätze kommen viel zu selten vor.
- **Kettenregel** erlaubt die Reduktion von  $P(w_1 w_2 \dots w_n)$  auf bedingte Wahrscheinlichkeiten:

$$P(w_1 w_2 \dots w_n)$$

$$= P(w_1) * P(w_2 | w_1) * P(w_3 | w_1 w_2) * \dots * P(w_n | w_1 w_2 \dots w_{n-1})$$

*aber:*

- $P(w_n | w_1 w_2 \dots w_{n-1})$ : Sparse-Data-Problem ist nicht beseitigt!

# n-Gramme

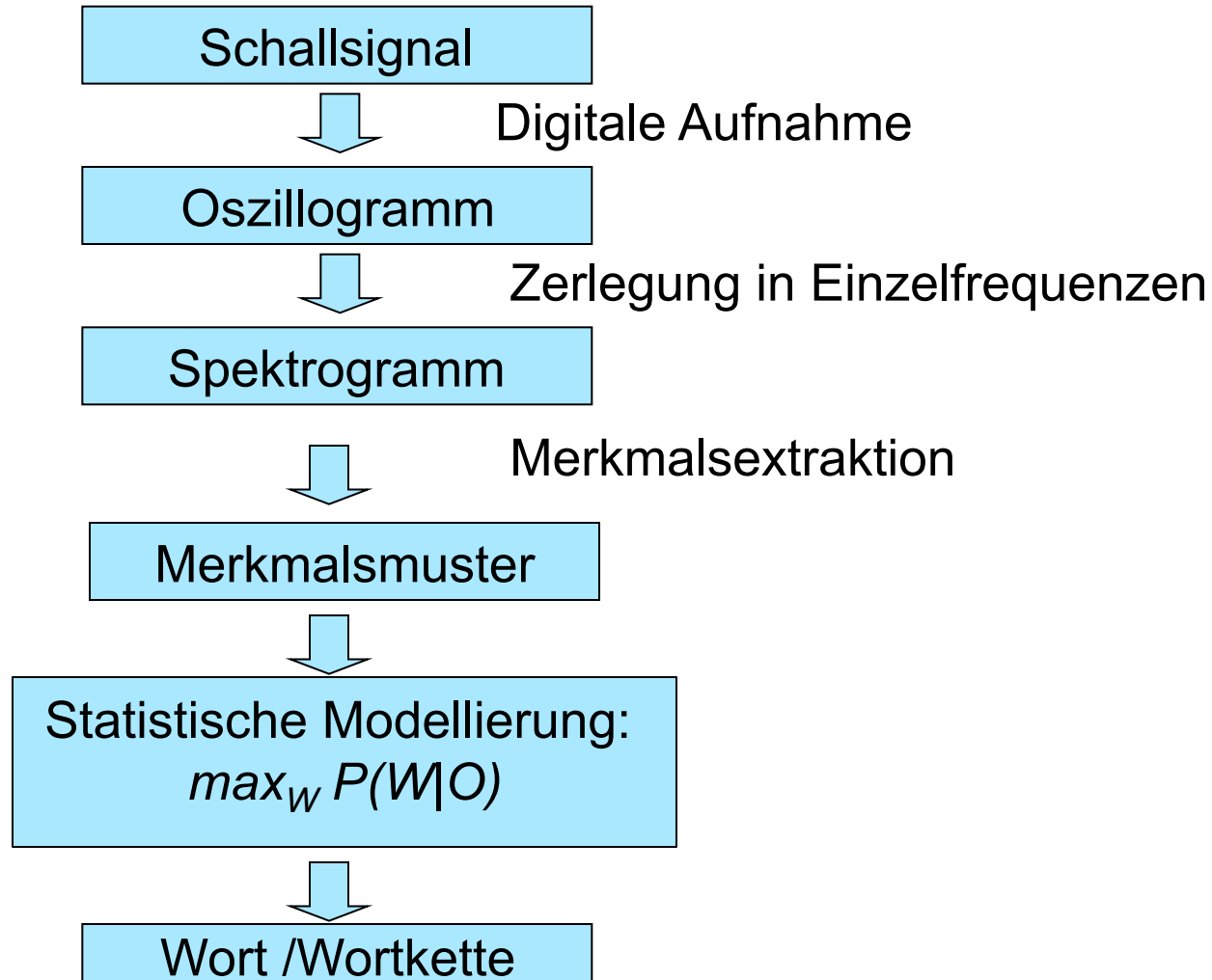
- n-Gramm-Methode:
  - Wir approximieren die Wahrscheinlichkeit, dass ein Wort  $w$  im Kontext einer beliebig langen Wortfolge auftritt, durch die relative Häufigkeit, mit der es in einem auf  $n$  Wörter begrenzten Kontext auftritt ("**Markov-Annahme**")
  - Dabei wird das Wort selbst mitgezählt. n-Gramm-Wahrscheinlichkeit berücksichtigt also einen Vorkontext von  $n-1$  Wörtern.
- Meistens wird mit Bigrammen und Trigrammen gearbeitet.
- Beispiel Bigramm-Approximation:
  - $P(w_n|w_1w_2\dots w_{n-1}) \approx P(w_n|w_{n-1})$
  - $P(w_1w_2 \dots w_n) \approx P(w_1) * P(w_2|w_1) * P(w_3|w_2) * \dots * P(w_n|w_{n-1})$

# Akustische Modelle

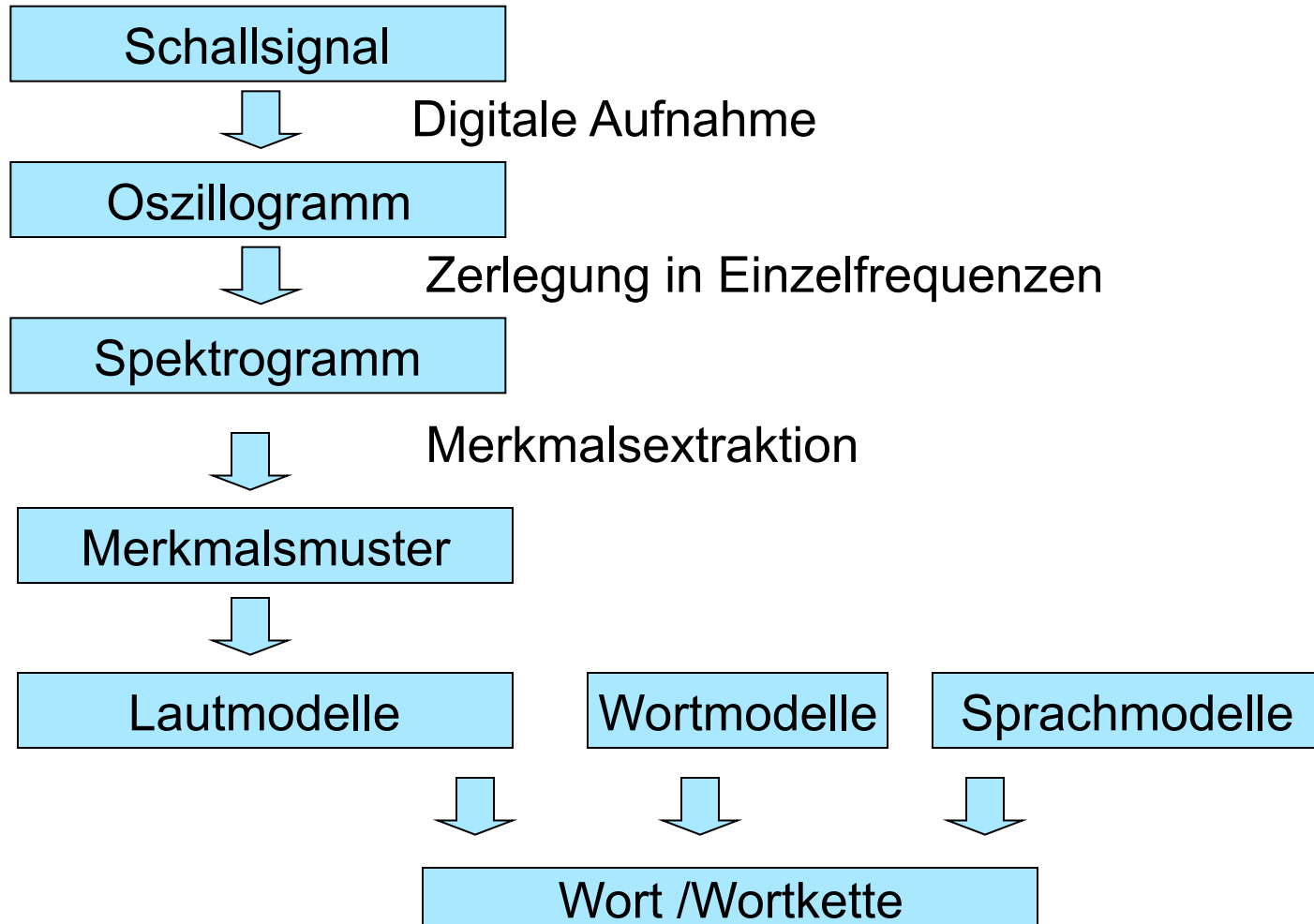
$$\max_W P(W | O) = \max_W P(O | W) \cdot P(W)$$

- Training von „Lautmodellen“ auf Datensammlungen für gesprochene Sprache: Aufnahmen von Sprachlauten mit ihrer phonetischen Kategorie/ Umschrift: Liefert die Wahrscheinlichkeit, mit der bestimmte Laute durch Merkmalsmuster realisiert werden.
- Aussprachewörterbuch, das für jedes Wort die phonetische Umschrift enthält
  - Genauer: Die Umschrift für alternative Aussprachen, die in einem **gewichteten endlichen Automaten** kodiert sind.
- Für die statistische Zuordnung von Merkmalsmustern und Wörtern wird die HMM-Methode („Hidden Markov Models“) verwendet.

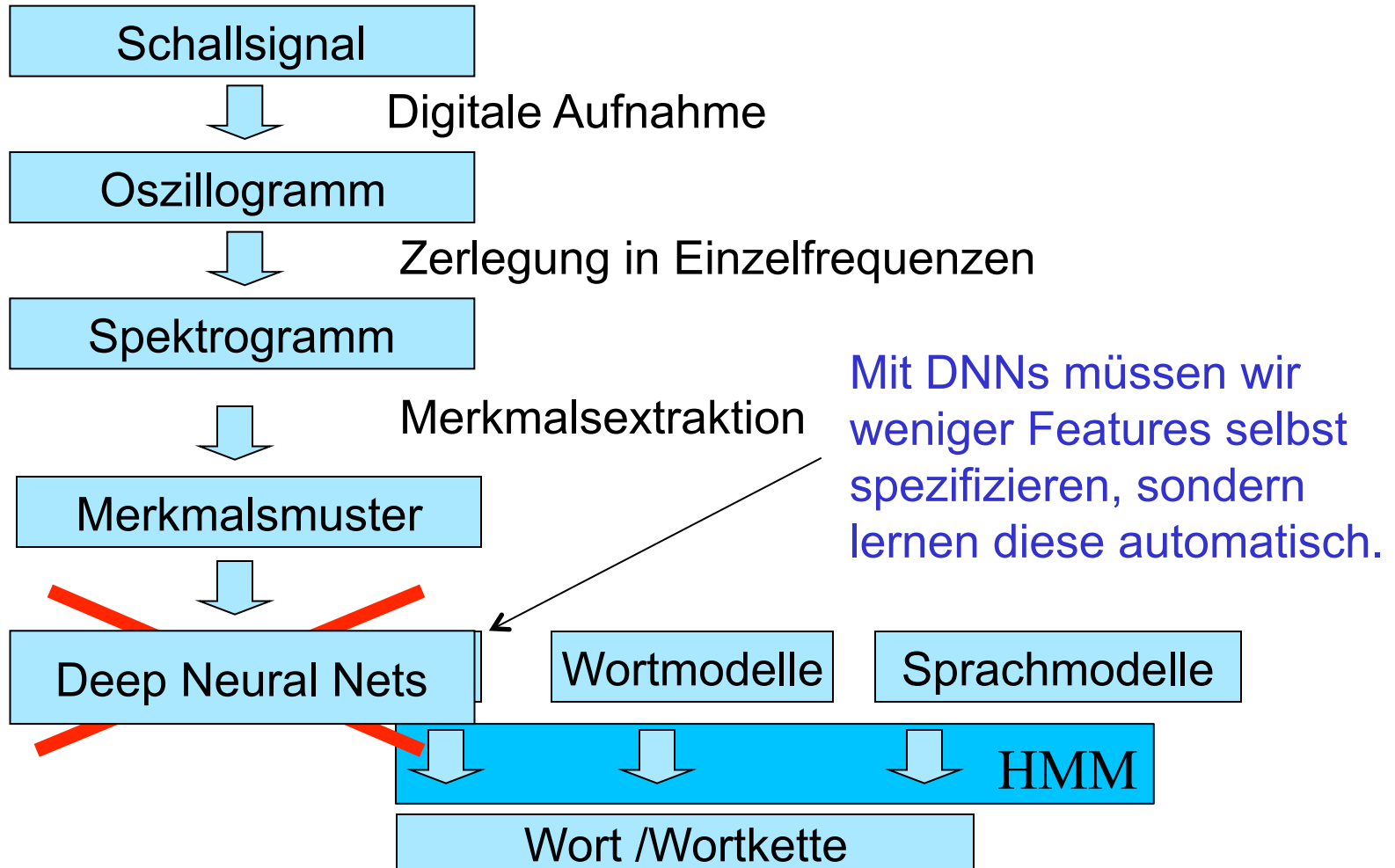
# Spracherkennung: (Vereinfachtes) Schema



# Spracherkennung: Schema

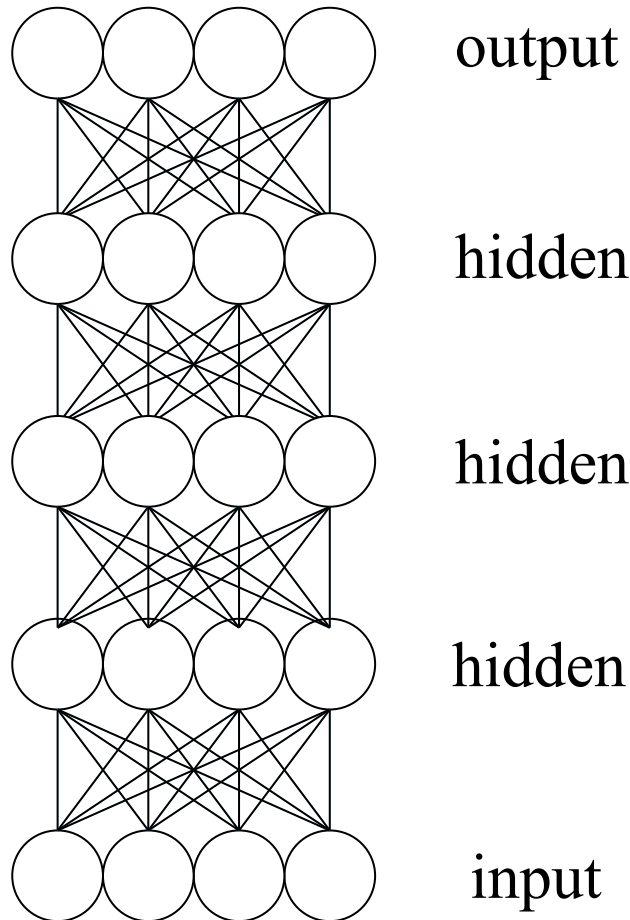
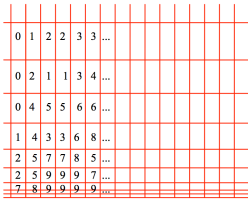


# Spracherkennung: Schema





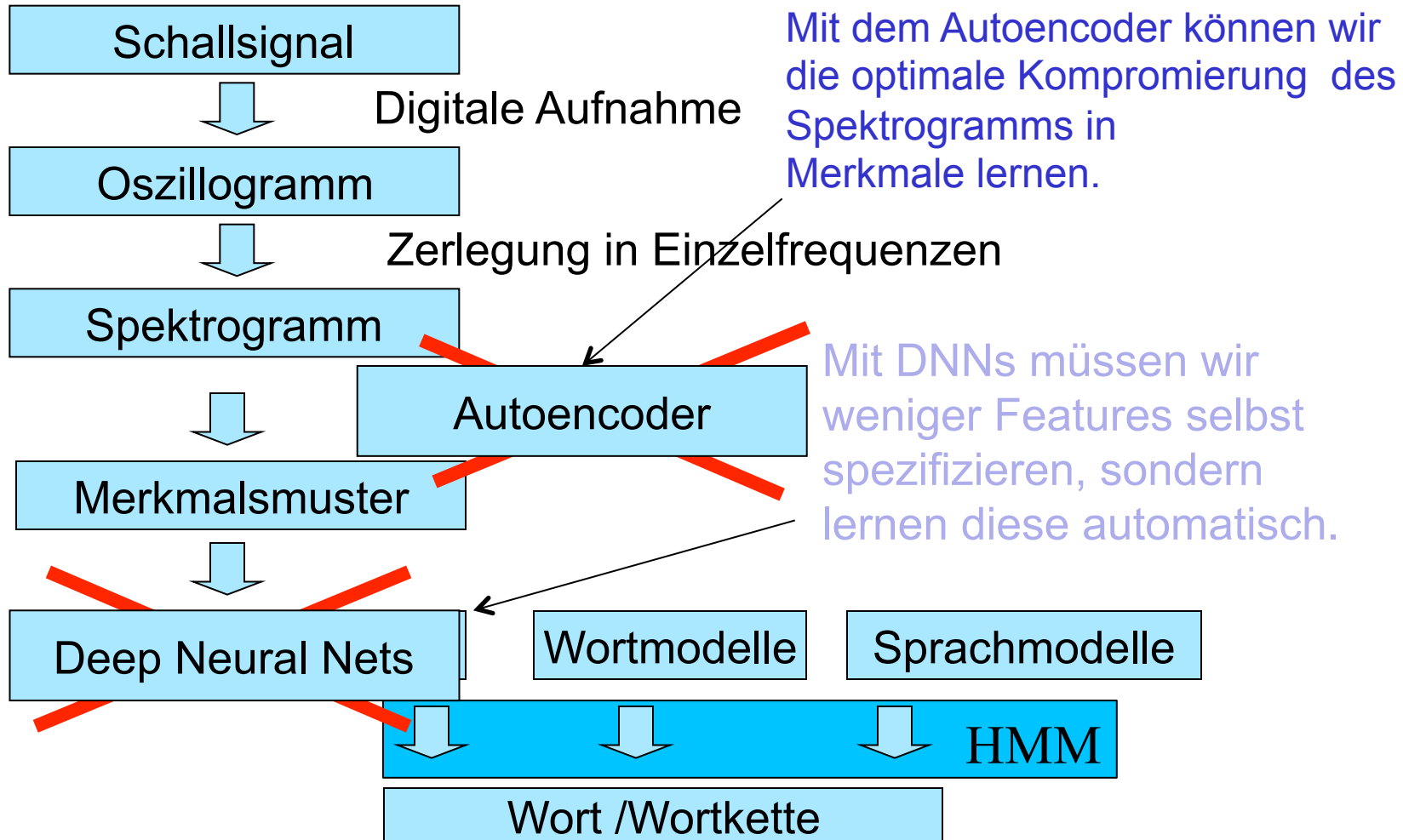
# Deep Neural Nets for Speech Recognition



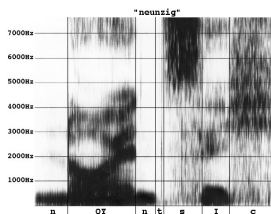
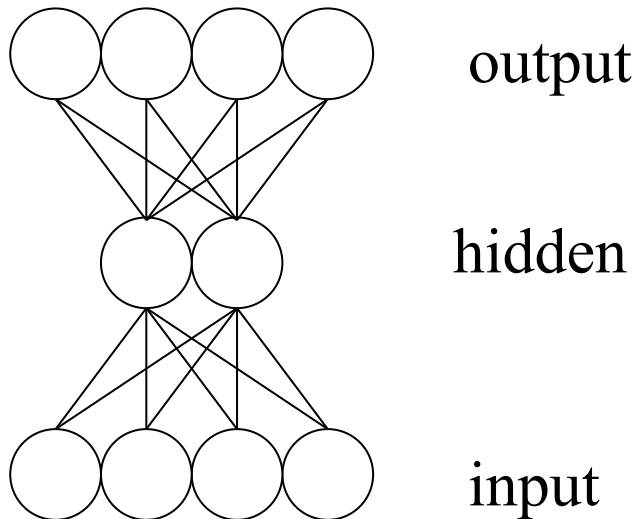
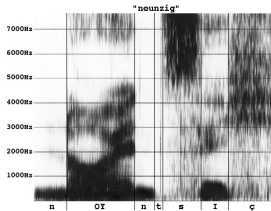
/ydy/

- Die Abbildung von Phonemketten auf Spektrogrammmerkmale ist sehr komplex.
- Mit DNNs kann dies mit größerem Erfolg als bei vorigen Ansätzen gelernt werden.
- Vorteil: bessere Ausnutzung von Kontextinformation.

# Spracherkennung: Schema



# Autoencoder für Merkmalsmusterextraktion



- Der Autoencoder soll als Output den Input möglichst akkurat reproduzieren.
- Durch die kleineren hidden Layers (hier nur schematisch als ein hidden Layer dargestellt, können aber mehr sein) wird die relevante Information mit nur minimalem Informationsverlust komprimiert.
- Vorteil: geringerer Informationsverlust als bei traditioneller Merkmalsextraktion mit Zeit-Frequenz-fenster.

## Erkennerperformanz ist abhängig von:

- Sprechmodus: Einzelwort, kontinuierlich, spontan
- Sprecherbindung: abhängig, unabhängig, adaptiv
- Größe des Lexikons:

Einfache Sprachsteuerungssysteme: 100-200 Wortformen

Dialogsysteme: 500-1000 Wortformen (+ spezieller Wortschatz)

Diktiersysteme: ab 50000 Wortformen

- **Perplexität**: Maß für die Uniformität der Eingabe  
beschränkte Domäne, gesteuerter Dialog: niedrige Perplexität  
keine Domänenbeschränkung, freie Rede: hohe Perplexität
- Eingabequalität
- Verarbeitungszeit

# Stand der Spracherkennungstechnik

- Maß für die Erkennerperformanz: **Wortfehlerrate** (wie viele Wörter der „besten Kette“ wurden falsch verstanden/gar nicht verstanden/hinzuphantasiert?)
- Wortfehlerrate hängt von der verfügbaren Verarbeitungszeit und verschiedenen externen Faktoren ab.
- Gängige Systeme analysieren in Echtzeit (Verarbeitungszeit  $\leq$  Sprechzeit) und sind in der Wortfehlerrate in einem akzeptablen Bereich.