1.1.3. Die Repräsentation des Sprachsignals

Bei der automatischen Erkennung gesprochener Sprache kommt es darauf an, im Sprachsignal bestimmte, für einzelne Laute (oder Lautverbindungen) charakteristische Muster zu finden. Diese lassen sich sowohl verbal als auch optisch/graphisch als auch elektronisch beschreiben. Letzteres bedeutet, daß man die Muster elektronisch als Referenz für die Erkennung von Lauten realisieren kann.

Für die Sprachsynthese sind diese Repräsentationen als physikalische Eingabegrößen für den akustischen Synthetisator unmittelbar einzusetzen. Allerdings kann man die Synthese auch dadurch realisieren, daß die Laute durch einen Sprecher erzeugt, dann entsprechend adaptiert (Übergänge in verschiedenen Varianten) und so gespeichert werden.

Vokoide:
Sie sind durch die Lage ihrer Formanten gekennzeichnet. Diese können als konstante Funktionen angenommen werden.

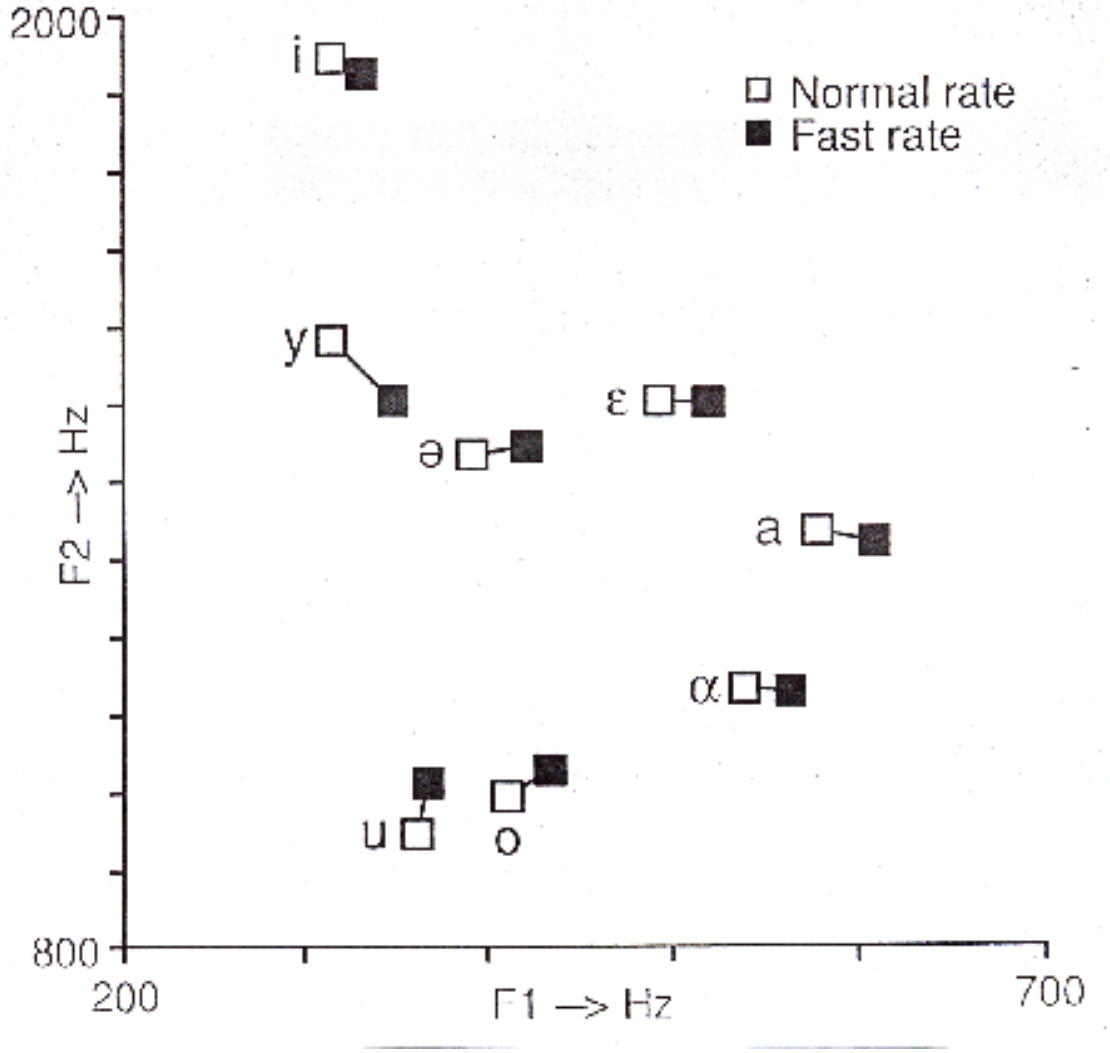
Kontoide:
Sie sind durch den zeitabhängigen Verlauf ihrer Formanten gekennzeichnet.

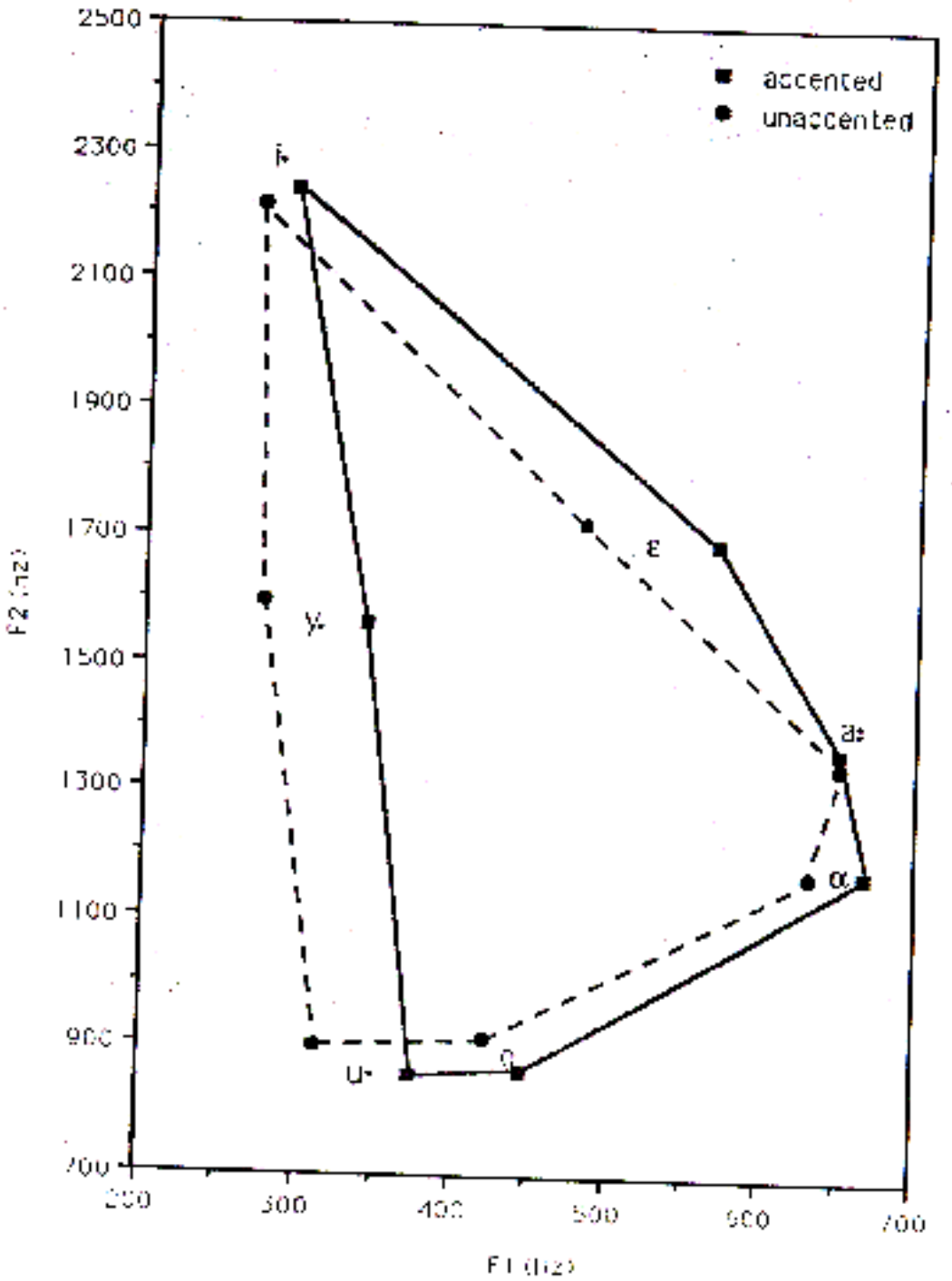
Für beide Klassen hat man die Übergänge zwischen den einzelnen Lauten zusätzlich zu modellieren. Bei Kontoiden (speziell Verschlußlauten) sind relativ kurze Veränderungen charakteristisch, bei denen die Formanten eine hohe Intensität haben.

Beispiele für Vokoide:

Die Formanten der Vokoide hängen zusätzlich von weiteren Faktoren ab, etwa dem Sprechtempo und der Betonung (von dem phonologischen Merkmal der Länge ohnehin, da dies generell mit Qualitätsunterschieden verbunden ist).

Die folgenden Seiten zeigen die Differenzen der Formanten F1 und F2 für die genannten Faktoren (Phoneme des Engl.). Dabei zeigt sich (überraschenderweise?), daß schnelles Sprechen die gleiche Tendenz wie Betonung hervorruft, nämlich einen höheren Wert für F1. Die Tabellen stammen aus verschiedenen Quellen (s. die absoluten Werte!).


2000
800
200
700
F2 —> Hz
F1 —> Hz
Normal rate
Fast rate
i
y
ə
ε
a
α
u
o


accented
unaccented
2500
2300
2100
1900
1700
1500
1300
1100
900
700
200
300
400
500
600
700
F2 (Hz)
F1 (Hz)
i
ɛ
a
α
ɔ
u

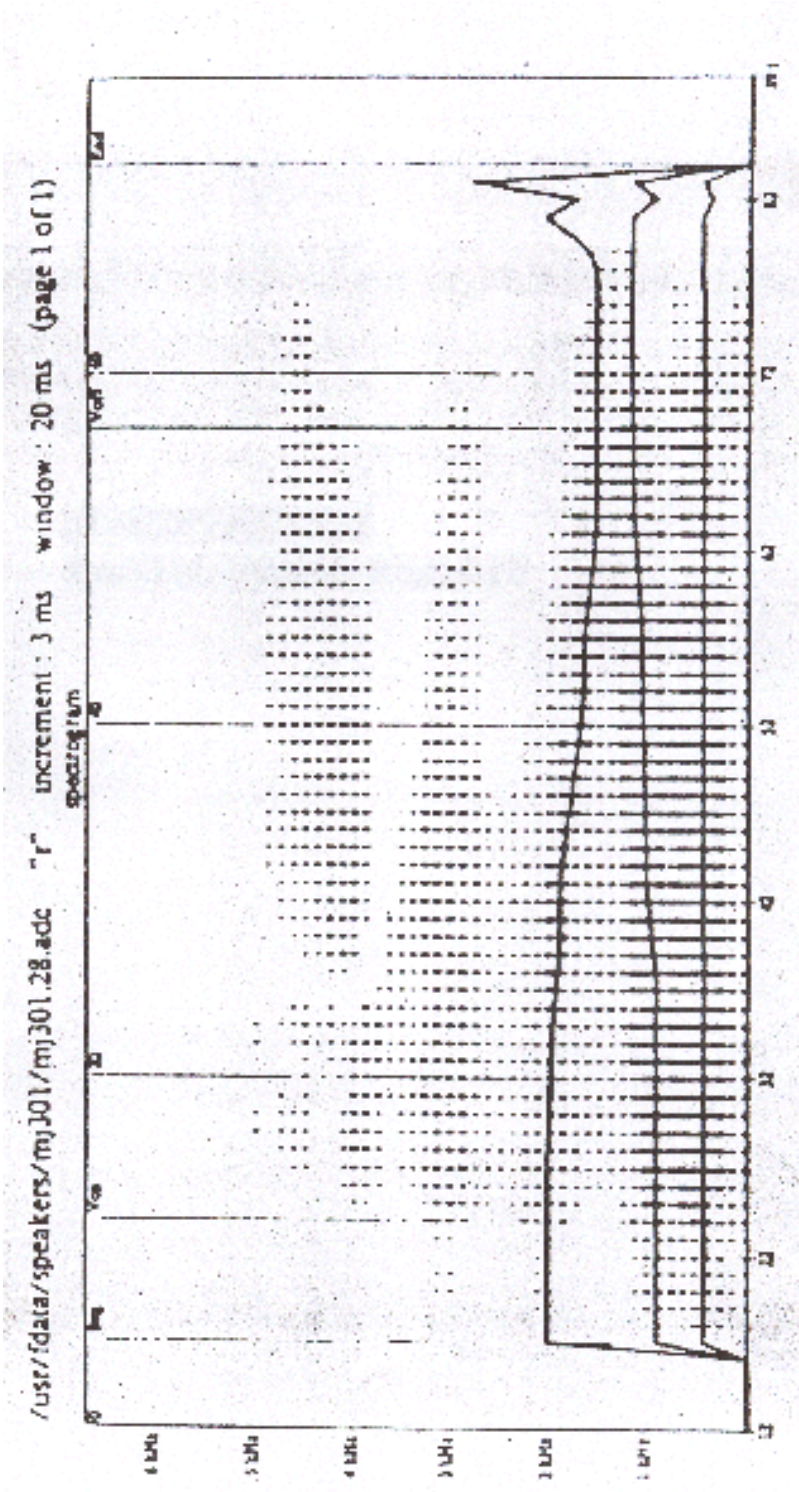
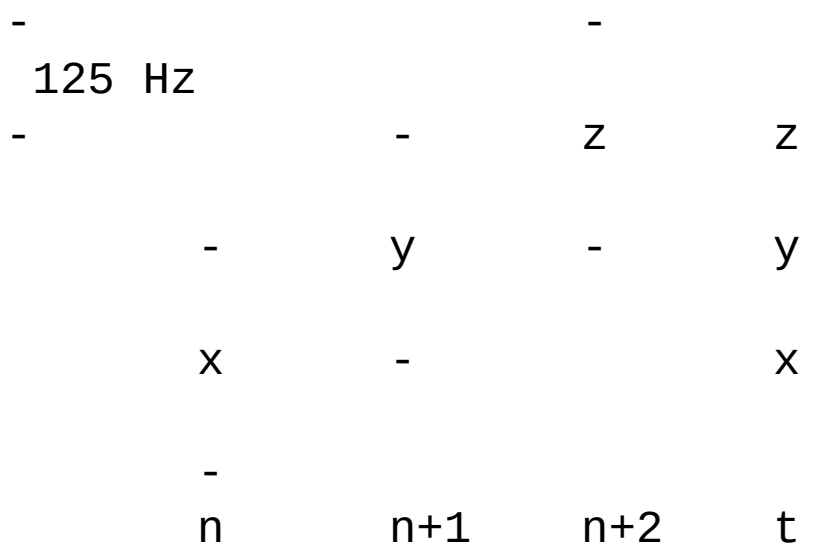


FIG. 15.2. Spectrogram of the letter "R" using digital spectral representation. The lines are drawn automatically using the formant extraction algorithms.

Beispiele für Vokoide:

Zur Darstellung für das engl. Phonem /r/ auf der vorangehenden Folie:

Die Spektraldarstellung beruht auf 54 Koeffizienten, die das Intervall von 63 Hz bis 6093 Hz abdecken. Jeder Koeffizient bezieht sich auf eine Bandbreite von 250 Hz, mit einer Überlappung von 125 Hz:



Ein besonderer Algorithmus `errät' die Formanten und ermittelt ihren Verlauf in der Zeit (durchgezogene Linien).

Für das ndl. Phonem \l\ in initialer Position werden z.B. folgende Parameter als `default specifications' zur Synthese angenommen (VAN HEUVEN/POLS 1993: 166f.):

Gesamtdauer := 65 ms
Dauer des Anfangsübergangs
- für die meisten Parameter := 10 ms
- für Formant F1 := 10 ms

Dauer des Schlußübergangs
- für die meisten Parameter := 10 ms
- für Formant F2 := 10 ms

/l/: /[+voc, +front, +unrounded]

F1: Frequenz F1: 300 Hz
Bandbreite F1 := 120 Hz

F2: Frequenz F2 := 1700 Hz
Bandbreite F2 := 120 Hz

F3: Frequenz F3 := 2500 Hz
Bandbreite F3 := 200 Hz

F4: Frequenz F4 := 3300 Hz
Bandbreite F4 := 400 Hz

Daraus wird abgeleitet:

/l/: ___/[+voc, +back]

F1: Frequenz F1: 300 Hz
Bandbreite F1 := 120 Hz

F2: Frequenz F2 := 1500 Hz
Bandbreite F2 := 120 Hz

F3: Frequenz F3 := 1700 Hz
Bandbreite F3 := 200 Hz

F4: Frequenz F4 := 3300 Hz
Bandbreite F4 := 400 Hz

/l/: ___/[+voc, +round]

F1: Frequenz F1: 300 Hz
Bandbreite F1 := 120 Hz

F2: Frequenz F2 := 1500 Hz
Bandbreite F2 := 120 Hz

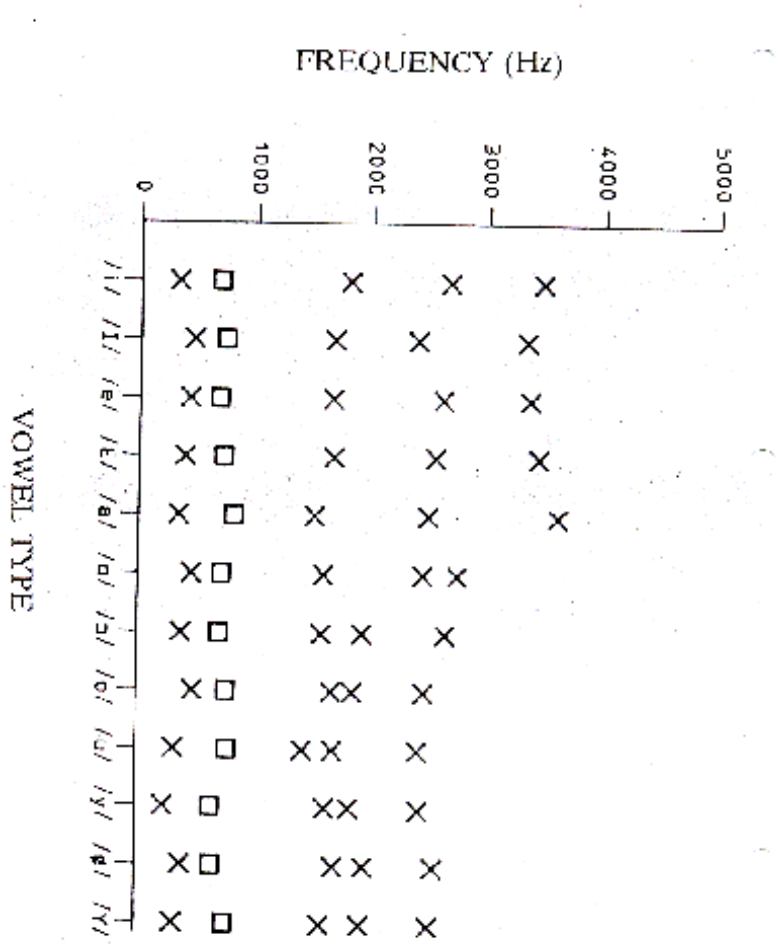
F3: Frequenz F3 := 1700 Hz
Bandbreite F3 := 200 Hz

F4: Frequenz F4 := 2500 Hz
Bandbreite F4 := 400 Hz

Diese Allophone kann man als "Maschinen-

phoneme" auffassen.

Für das ndl. Phonem /l/ ergeben sich stark unterschiedliche Allophone je nach den umgebenden Vokalen, die durch Koartikulation bewirkt werden (VAN HEUVEN/POLS 1993: 303):



Fazit:

Das Sprachsignal ist physikalisch eine in der Zeit veränderliche Frequenz-Energie-Verteilung: Zu jedem Zeitpunkt t ist eine Verteilung von "Intensitäten" (physikalisch: Energien) auf die Frequenzen s gegeben:

$$F(f,t) = A$$

Die Frequenz f hat zum Zeitpunkt t die Intensität A.

Die einzelnen Laute einer Sprache haben charakteristische Verteilungsabläufe dieser Art. Eine mit dem Sprachsignal gegebene Funktion F über einem Zeitintervall T ist daher in solche Abläufe zu zerlegen, wodurch man dann eine Lautkette erhalten kann. Dabei gilt jedoch:

1. Die Abläufe sind zeitlich relativ zu betrachten, man muß eine Gummi-Zeitachse verwenden, damit man Sprachsignal und Abläufe aufeinander abbilden kann. Es ist notwendig, die Abläufe als Folgen von Zuständen darzustellen und die zeitlichen Verhältnisse großzügiger zu handhaben. Die "vor-nach"-Beziehung bleibt natürlich bestehen. Im Sprachsignal sind also gewisse Zustände zu suchen, die in gewisser Weise aufeinander folgen.

2. Die Funktion F braucht nur für einen Ausschnitt des Frequenzbereichs betrachtet zu werden (etwa bis 6 kHz). Außerdem wird dieser Frequenzbereich in Bänder aufgeteilt, und die Intensität wird pro Band ermittelt. Ein Band ist noch keine Lokalisierung bestimmter Abläufe, sie können je nach SprecherIn weiter oben oder unten (d.h. in verschiedenen Bändern) liegen. Außer den Bandfiltern gibt es noch andere (mathematische) Verfahren zur Ermittlung der einzelnen Merkmale. Insgesamt wird die Frequenz-Energie-Verteilung durch einen Vektor erfaßt, der aus 10 bis 30 Zahlen besteht, und die Veränderung dieser Zahlen in der Zeit ist das, was vom Sprachsignal bleibt.

3. Die Funktion F kann natürlich auch nicht für jedes einzelne t ermittelt werden. Man führt daher Abtastintervalle ein, die so klein sind, daß sich in jedem einzelnen von ihnen keine wesentliche Veränderung vollzieht, die nicht auch bei Mittlung über diese Intervalle erkennbar wäre. Es liegt somit ein der Aufgabe angepaßtes Zeitraster vor. Als Länge dieser Intervalle wird etwa 10 ms angenommen, d.h. pro Sekunde werden etwa 100 Messungen durchgeführt.

Die eigentliche Funktion F wird in dem Meßverfahren durch eine andere Funktion approximiert, bei der statt der Frequenzen Bänder und statt der Zeitpunkte kleine Intervalle auftreten. Auf dieser Funktion beruht dann die Berechnung der Vektoren. Insgesamt findet eine mehrfache Reduktion der Daten statt.

Aus den Folgen der Vektoren werden die Merkmale gewonnen, die etwas über die akustischen Ereignisse in Folgen von Zeitintervallen besagen. Diese werden dann mit gegebenen Mustern verglichen, die die Merkmalsabläufe für bestimmte Einheiten repräsentieren (subword units). Dieser Prozeß (pattern matching oder pattern recognition) ist ein ganz wesentlicher Schritt bei der Spracherkennung, der über die Effektivität des Gesamtverfahrens entscheidet. Insbesondere liegt hier ein Einsatzfeld für Lernalgorithmen über statistische Daten.

Wie man die skizzierten Gegebenheiten exakt behandelt, wird bei den HMM's beschrieben.

1.1.4. Perzeptive (auditive) Phonetik

Im Gegensatz zur artikulatorischen Phonetik beschäftigt sich die auditive Phonetik mit dem `Eindruck', den das Sprachsignal beim Hörer hervorruft. Schon die dafür verwendeten Termini (meist Metaphern) zeigen, daß darauf aufbauende Klassifikationen nicht das leisten, was die artikulatorische Beschreibung ermöglicht.

Engl. Beispiel: /l/ hat zwei Allophone:

auditiv:	artikulatorisch:
dunkel	velar
hell	palatal

Wie grenzt man auditiv den *ach-/ich*-Laut ab?

Ziemlich hilflos ist man erst recht bei der auditiven Unterscheidung der Vokale.

Man kann daher die artikulatorische Klassifizierung nicht durch eine "ohrenphonetische" ersetzen, zumal die Vorgänge in den Hörorganen weit schlechter zugänglich sind als die der Artikulationsorgane, sie sind außerdem nicht gesteuert/aktiv. Gelegentlich

kann man die auditiven Merkmale jedoch mit Nutzen verwenden.

1.2. Einiges zur Phonologie

Während die Phonetik die Sprachlaute prinzipiell unabhängig von den Einzelsprachen betrachtet, geht es bei der Phonologie um die einzelsprachenbezogene Untersuchung der Laute. Im ersten Fall wird auch von **Phonen** gesprochen, während die sprachbezogene Klassifizierung der Laute zu **Phonemen** führt.

Der Begriff **Phonem** ist schon über 100 Jahre alt (Baudouin de Courtenay 1895: "psychologisches Äquivalent des Sprachlautes", worin schon der trans-akustische Inhalt zum Ausdruck kommt). Die Phonologie in ihrer heutigen Ausprägung ist jünger (Prager Schule, etwa 1930; strukturalistische Ansätze, Chomsky/Halle 1968). Nach moderner Auffassung wir das Phonem funktional definiert in dem Sinne, daß durch verschiedene Phoneme Bedeutungen (innerhalb einer Sprache) differenziert werden (s.u.). Eine artikulatorische oder auditive Ähnlichkeit/Abgrenzung ist dabei zunächst nicht im Spiel (abgesehen von der scheinbaren Trivialität, das man relevante Unterschiede auch hören können muß, doch s.u.!).

Im Dt. sind [v] und [b] verschiedene Phoneme, da es Paare wie
Wein vs. *Bein*; *Wall* vs. *Ball* etc.
gibt.

Im Span. dagegen spielt dieser Unterschied keine Rolle, das Auftreten von [â] (in etwa [v]) und [b] ist positionsbedingt:
[â]: intervokalisch in Wort und Satz
[b]: im absoluten Anlaut und nach Nasal
Die Schreibung *v* oder *b* ist irrelevant.

[â]	[b]
noventa	
La Habana	
y Barcelona	*en Barcelona*
y Valencia	*en Valencia*
	Barcelona
	Valencia

Somit gehören [â] und [b] zu einem Phonem.

Im Dt. gibt es das dunkle /l/ [»] höchstens in Dialekten, das Phonem wird durch das helle [l] realisiert.

Im Russ. ergibt dieser Unterschied zwei Phoneme:
[dal]: `Weite, Ferne'
[da»]: `(er) gab
'

Im Engl. handelt es sich um zwei positionsbedingte Varianten eines Phonems:
[l]: prävokalisch (*little*)
[»]: präkonsonantisch und final
(*colt, little*)

Zungenspitzen-/r/ (gerollt oder ein Schlag) und Zäpfchen-/r/ bewirken im Dt. keinen Bedeutungsunterschied, obwohl sie phonetisch sehr verschieden sind. Es handelt sich nicht um positionsbedingte Varianten des Phonems /r/, wenngleich das gerollte /r/ im Anlaut häufiger vorkommen mag als sonst. Es liegen fakultative (oder freie) Varianten vor.

[x] und [ç] sind im Dt. positionsbedingte Allophone eines Phonems. Dieses Beispiel zeigt auch, daß man die oben angedeutete Definition des Phonems präzisieren muß: Einerseits kann man die Bedingung für die Auswahl zwischen [x] und [ç] nur unter Berücksichtigung der Morphemgrenzen zutreffend formulieren, zum anderen soll sich die bedeutungsdifferenzierende Funktion auf die **kleinsten** bedeutungstragenden Einheiten - eben die Morpheme - beziehen.

Frau + ch en vs. *fau ch + en*
d 6 * h 7 h d 6 d * 7 h

Die bisherigen dt. Beispiele sind nicht mit

einer Überschreitung der Phonemgrenze verbunden. Demgegenüber waren (z.B. bei den Assimilationen) auch Variationen angegeben worden, für die dies nicht gilt (u.a. Auslautverhärtung).

Wie bei der Phonetik schon gezeigt, können die Phone durch eine Reihe von Merkmalen beschrieben werden. Durch den Phonembegriff stellt sich die Frage, welche dieser Merkmale in einer Einzelsprache relevant sind, um die Phoneme zu unterscheiden. Zur Klärung dessen verwendet man **Oppositionen** (Prager Schule), bei denen minimale Differenzierungen festgestellt werden. So kann man durch

Wein	-	*fein*	[v]	-	[f]
Bein	-	*Pein*	[b]	-	[p]

zeigen daß stimmhaft/stimmlos und Frikativ/Plosiv **distinktiv** sind. Durch

Wein - *sein*

ergibt sich, daß auch der Artikulationsort unterscheidend wirkt. Auf diese Weise kommt man zu einem Inventar von Phonemen und distinktiven Merkmalen für eine Einzelsprache.

Diese Prozedur ist jedoch nicht so einfach, wie es auf den ersten Blick erscheint. Man benötigt eine Reihe von Zusatzannahmen, die hier nicht behandelt werden können und mit denen man der Auslautverhärtung oder der **komplementären Verteilung** von [õ] und [h] im Dt. (nicht ein Phonem!) beikommen kann.

Zur Illustration zwei Beispiele:
1. Bei vielen Sprechern ist kein Unterschied zwischen *scharrt* und *Schacht* zu hören, die Realisierung ist [÷]. Durch *scharren* und *Schächte* (Sprecherwissen!) ist jedoch klar, daß hier zwei Phoneme vorliegen, sie hätten jedoch (wenigstens) teilweise gleiche Realisierungen.

2. Durch *vereisen* (mit [?]) vs. *verreisen* (ohne [?]) könnte man auf die Idee kommen, [?] Phonemstatus zuzuweisen. In *verheißen* hat man [h]. [h] und [?] haben die gleiche Distribution (nicht final, keine Verbindung mit anderen Konsonanten innerhalb eines Morphems). Andererseits ist [?] vorhersagbar (nach dem Grundsatz "Wo nichts ist, ist [?]"). Wie löst man dieses Puzzle?

Für die Verarbeitung gesprochener Sprache ist die Berücksichtigung der skizzierten Phänomene von größter Wichtigkeit, insbesondere gilt dies für die Bedingungen zur Auswahl der Allophone. Da bisher jedes Verfahren zur Generierung oder Erkennung für Einzelsprachen funktioniert, ist dies auch ohne weiteres möglich.