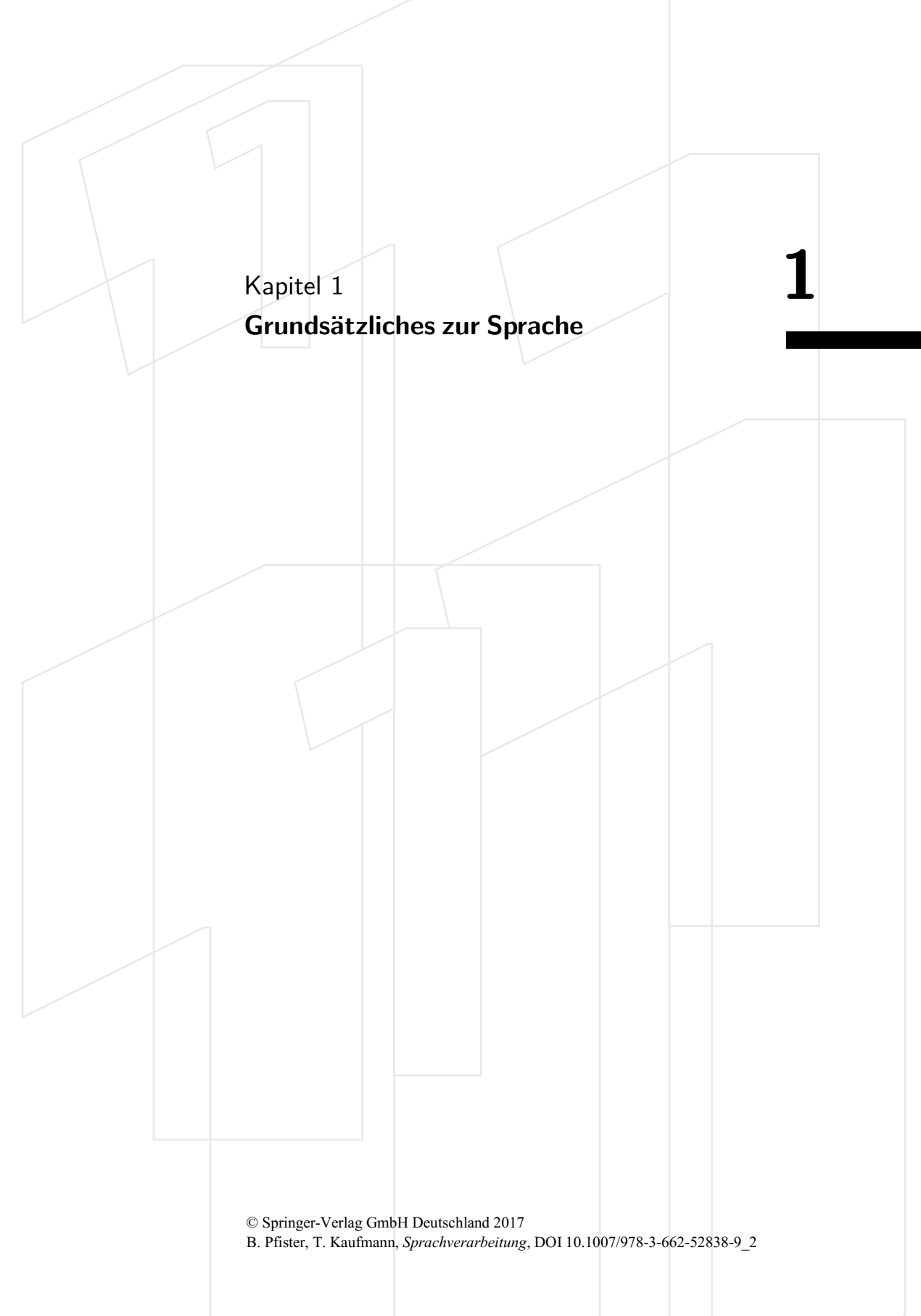




Kapitel 1

Grundsätzliches zur Sprache

1

1

1	Grundsätzliches zur Sprache	5
1.1	Sprache als Kommunikationsmittel	5
1.1.1	Lautsprachliche Kommunikation.....	5
1.1.2	Geschriebene vs. gesprochene Sprache	6
1.2	Die Beschreibung von Sprache.....	8
1.2.1	Die linguistischen Ebenen	8
1.2.2	Die phonetische Schrift.....	10
1.2.3	Die akustisch-phonetische Ebene	11
1.2.4	Die Prosodie der Sprache	11
1.3	Die menschliche Sprachproduktion.....	12
1.3.1	Übersicht über den Sprechapparat	12
1.3.2	Die Funktion des Sprechapparates	12
1.3.3	Die Artikulation der Sprachlaute	14
1.4	Das menschliche Gehör.....	15
1.4.1	Wahrnehmung der Schallintensität.....	16
1.4.2	Periodizität und Tonhöhe	17
1.4.3	Die Phasenwahrnehmung	18
1.4.4	Der Verdeckungseffekt.....	18
1.4.5	Wahrnehmung der Sprachlaute	19
1.5	Verarbeitung natürlicher Sprache.....	21

1 Grundsätzliches zur Sprache

In diesem Kapitel werden allgemeine Grundlagen und Begriffe eingeführt, auf denen die weiteren Kapitel aufbauen.

1.1 Sprache als Kommunikationsmittel

1.1

➤ 1.1.1 Lautsprachliche Kommunikation

Kommunikation bezeichnet allgemein die Übermittlung von Information. Dabei können die Art der Information und das Übertragungsmedium sehr unterschiedlich sein.

Bei der direkten lautsprachlichen Kommunikation zwischen Menschen wird die Information (Botschaft, Gedanke, Idee etc.) in die Form der gesprochenen Sprache umgesetzt, die sich vom Mund des Sprechers via das Medium Luft als Schallwellen ausbreitet, wie in der Abbildung 1.1 schematisch dargestellt. Die Schallwellen gelangen an das Ohr des Zuhörers, der aus dem Sprachschall die Botschaft ermittelt.

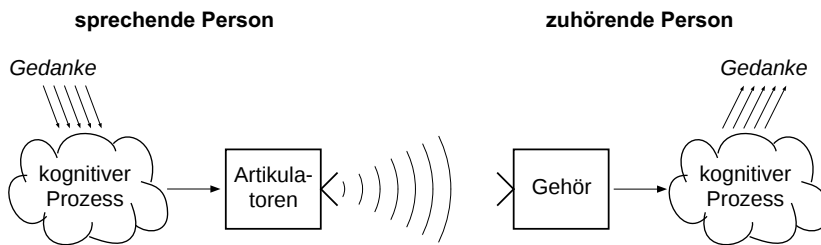


Abbildung 1.1. Symbolische Darstellung der lautsprachlichen Übertragung einer gedanklichen Botschaft von einem Sprecher zu einem Zuhörer

Lautsprachliche Kommunikation kann nicht nur zwischen Menschen, sondern auch zwischen einer Maschine und einem Menschen stattfinden. Im Gegensatz zum menschlichen Sprecher versteht jedoch gewöhnlich die sprechende Maschine den Sinn der ausgegebenen Lautsprache nicht. Die sprechende Maschine ersetzt somit im Allgemeinen nicht einen Sprecher, der Gedanken in Lautsprache umsetzt. Heutzutage geben die meisten der sogenannten sprechenden Maschinen bloss vorgängig aufgezeichnete Sprachsignale aus. Neuerdings kommt auch Sprachsynthese zum Einsatz, also das Umsetzen einer als Text ausformulierten Meldung in ein Sprachsignal. Das Übersetzen von Ideen, Fakten etc. in sprachlich korrekte und sinnvolle Texte wird jedoch weitgehend bei der Konzeption der Maschine vollzogen und ist somit Sache des Menschen.

Sinngemäß trifft dies auch für den Vergleich einer Spracherkennungsmaschine mit einem Zuhörer zu. Während der Zuhörer das Gehörte automatisch auf seinen Sinn hin überprüft, indem er es zum aktuellen Thema und zu allgemeinem Wissen in Bezug setzt, begnügt sich die Spracherkennung in der Regel damit, das Sprachsignal in eine Folge von Wörtern zu transformieren. Ob diese Wortfolge auch sinnvoll ist, überprüft die Spracherkennung nicht.

► 1.1.2 Geschriebene vs. gesprochene Sprache

Die lautliche Form der menschlichen Sprache ist viel älter als die geschriebene. Entwicklungsgeschichtlich gesehen ist das Bestreben, nebst der Lautsprache auch eine schriftliche Form der Kommunikation zu haben, relativ jung, insbesondere im deutschen Sprachraum. Ursprünglich versuchte die (alphabetische) Schrift ein Abbild der Lautsprache zu geben, war also eher eine Art phonetische Schreibweise. Die so entstandene Orthographie der deutschen Sprache war naheliegenderweise stark durch den Dialekt und die persönliche Auffassung des Schreibers geprägt und somit sehr uneinheitlich.¹ Mit der zunehmenden Bedeutung der schriftlichen Kommunikation durch die Erfindung des Buchdrucks Mitte des 15. Jahrhunderts und noch ausgeprägter durch die Einführung der allgemeinen Schulpflicht zu Beginn des 19. Jahrhunderts wurde die Vereinheitlichung der deutschen Rechtschreibung verstärkt. Die Veröffentlichung des “Vollständiges orthographisches Wörterbuch der deutschen Sprache” von Konrad Duden im Jahre 1880 war ein wichtiger Schritt auf diesem Wege. Eine allgemein anerkannte und als verbindlich festgelegte deutsche Orthographienorm entstand jedoch erst 1901 an der Orthographischen Konferenz in Berlin.

Die moderne Zivilisation mit ihrer starken Abhängigkeit von der geschriebenen Sprache verleitet dazu, Sprache in erster Linie in ihrer geschriebenen Form zu betrachten und die Lautsprache zur ungenauen verbalen Realisation eines Textes abzuqualifizieren. Tatsächlich sind jedoch die Ausdrucksmöglichkeiten der Lautsprache vielfältiger als diejenigen der geschriebenen Sprache (vergl. auch Abschnitt 1.2.4), und zwar unabhängig davon, ob es sich um gelesene Sprache oder um sogenannte Spontansprache handelt.

Dieses Buch befasst sich sowohl mit der geschriebenen, als auch mit der gesprochenen Sprache und insbesondere mit der automatischen Umwandlung der einen Form in die jeweils andere. Die Charakterisierung von geschriebener bzw. gesprochener Sprache erfolgt dementsprechend unter einer technischen Perspektive.

¹So sind beispielsweise für das Wort “Leute” in Dokumenten aus dem 14. bis 16. Jahrhundert unter anderem die folgenden Schreibweisen zu finden: Leut, Leüthe, Lude, Lu^ede, Lute, Lüt, Lút, Lu^tte, wobei nicht alle hinsichtlich Bedeutung identisch sind (vergl. [42]).

Unter geschriebener Sprache soll hier Text verstanden werden, der z.B. als Zeichenfolge in einem Computer vorliegt. Mit gesprochener Sprache sind digitalisierte Sprachsignale, sogenannte Zeitreihen gemeint. Die augenfälligsten Unterschiede sind somit (vergl. auch Abbildung 1.2):

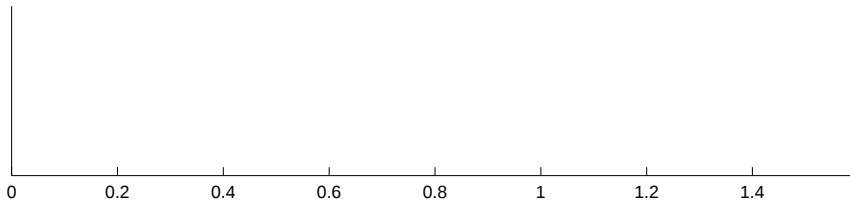
- Text besteht grundsätzlich aus einer Folge diskreter Zeichen aus einem relativ kleinen Zeichensatz, den Buchstaben des Alphabets. Die Zeichen beeinflussen sich gegenseitig nicht und die Wörter sind klar gegeneinander abgegrenzt.
- Lautsprache ist zwar auch eine Abfolge von Elementen, nämlich von Lauten, die wir mehr oder weniger eindeutig wahrnehmen. Die in der Form des Sprachsignals physikalisch erfassbare Lautsprache zeigt jedoch, dass die Laute nicht scharf voneinander abgegrenzt sind, weder in zeitlicher Hinsicht noch bezüglich der charakteristischen Eigenschaften. Die lautlichen Eigenschaften des Sprachsignals verändern sich kontinuierlich von einem Laut zum nächsten, auch über die Wortgrenzen hinweg. Wortgrenzen sind somit in der Regel nicht ersichtlich. Zudem ist die Ausprägung eines Lautes stark von seinen Nachbarlauten (Koartikulation), von seiner Stellung im Wort bzw. Satz und nicht zuletzt vom Sprecher abhängig.

Ein digitalisiertes Sprachsignal ist eine Folge von Abtastwerten, also eine Zahlenreihe und somit auch eine Folge diskreter Zeichen. Im Gegensatz zur geschriebenen Sprache, bei der z.B. jedes ASCII-Zeichen genau einem Buchstaben entspricht, besteht jedoch zwischen einem Signalabtastwert und einem Laut kein direkter Zusammenhang.

Diese rudimentäre Charakterisierung zeigt, dass das Sprachsignal von diversen sehr unterschiedlichen Faktoren geprägt wird. Es kann deshalb kaum erstaunen, dass das Detektieren von Lauten im Sprachsignal und damit die Spracherkennung alles andere als trivial ist.

Text: *“Heinrich kommt nach Hause.”*

Sprachsignal:



1.2 Die Beschreibung von Sprache

In diesem Abschnitt geht es um die Beschreibung von Sprache aus der Sicht der Linguistik. Dabei interessiert, aus welchen Elementen die Sprache besteht, welche Gesetzmässigkeiten zwischen Elementen (oder Folgen von Elementen) vorhanden sind und welche Beziehungen zwischen den Elementen und dem “Rest der Welt” bestehen (siehe Abschnitt 1.2.1). Diese vielschichtigen Zusammenhänge werden auch als Struktur der Sprache bezeichnet und die sprachlichen Elemente werden dabei als Symbole oder Abstrakta betrachtet.

Aus der Sicht der Phonetik geht es bei der Beschreibung von Sprache um die Eigenschaften der konkreten lautsprachlichen Realisierungen der abstrakten Elemente. Dabei wird zwischen der *segmentalen Ebene*, also der Ebene der Laute (vergl. Abschnitt 1.2.3), und der *suprasegmentalen Ebene*, der Prosodie (siehe Abschnitt 1.2.4) unterschieden.

➤ 1.2.1 Die linguistischen Ebenen

Bei der Beschreibung der Struktur der Sprache unterscheidet die Linguistik verschiedene Ebenen. Es ist üblich, die Ebenen nach Komplexität zu ordnen, wobei zur Komplexität hauptsächlich die Vielfalt und die Grösse der Elemente beitragen. Auf der untersten Ebene wird zudem zwischen Lautsprache und geschriebener Sprache unterschieden:

- **Graphemische Ebene:** Sie definiert den Vorrat an Schriftzeichen, das sogenannte Alphabet. Die Elemente des Alphabets werden auch als Grapheme bezeichnet. Wie alle anderen Ebenen ist auch die graphemische Ebene sprachspezifisch. Sie spezifiziert beispielsweise, dass die Grapheme ⟨ä⟩, ⟨ö⟩ und ⟨ü⟩ zum deutschen, nicht aber zum englischen Alphabet gehören.
- **Phonemische Ebene:** Die kleinsten Einheiten der Lautsprache sind die Laute oder Phone, wobei die Linguistik unterscheidet zwischen den *Phonemen* und den *Allophenen*:

- Die *Phoneme* sind die kleinsten bedeutungsunterscheidenden Einheiten einer Sprache. Die Phoneme einer Sprache können mit der Minimalpaaranalyse ermittelt werden. So gibt es im Deutschen beispielsweise das Wortpaar “doch” und “roch” (in phonemischer Darstellung gemäss Abschnitt A.2 geschrieben als /dɔx/ und /rɔx/) und damit die Phoneme /d/ und /r/. Weil im Deutschen auch Wortpaare wie “Rate” /ratə/ und “Ratte” /ratə/ existieren, sind auch /a/ und /aː/ Phoneme.

Zwar nicht im Deutschen, aber in sogenannten Tonsprachen (z.B. Chinesisch) gibt es auch Wortpaare und damit Laute, die sich nur in der Tonhöhe bzw. im Tonhöhenverlauf unterscheiden. Solche Laute sind definitionsgemäss auch Phoneme.

- Die *Allophone* sind zum selben Phonem gehörige Laute oder Lautvarianten. Sie bewirken in der betrachteten Sprache keinen Bedeutungsunterschied. So ist die Bedeutung des Wortes “roch” dieselbe, unabhängig davon, ob es als [rɔx] (mit Zungenspitzen-R) oder als [ʁɔx] (mit Zäpfchen-R) gesprochen wird. Die Laute [r] und [ʁ] werden als Allophone des Phonems /r/ bezeichnet. Sind die Allophone beliebig gegeneinander austauschbar (d.h. unabhängig von der Stellung), dann heissen sie *freie Allophone*. So sind z.B. [ʁ] und [r] freie Allophone des Phonems /r/. Kommen die Laute nur in bestimmten Kontexten vor, dann werden sie als *stellungsbedingte Allophone* bezeichnet. Im Deutschen sind beispielsweise die Laute [ç] in “ich” und [x] in “ach” stellungsbedingte Allophone des Phonems /x/.
- **Morphologische Ebene:** Die Morphologie beschreibt, wie aus Morphemen korrekte Wortformen aufgebaut werden. In der Linguistik werden mit dem Begriff Morphem Abstrakta wie Stamm-, Präfix- und Pluralmorphem bezeichnet. Die konkreten Realisationen sind die Morphe. Morphe wie ⟨haus⟩ und ⟨häus⟩, die zu einer gemeinsamen Grundform {haus} gehören, werden Allomorphe genannt und die Grundform {haus} wird ebenfalls als Morphem bezeichnet.
Morpheme sind die kleinsten bedeutungstragenden Einheiten einer Sprache. Eine wesentliche Eigenschaft der Morphe ist, dass sie nicht anhand eines kompakten Satzes von Regeln beschrieben werden können. Sie lassen sich bloss aufzählen.
- **Syntaktische Ebene:** Die Syntax beschreibt mittels Regeln, wie Wortformen zu Konstituenten (Satzteilen oder Wortgruppen) und diese wiederum zu Sätzen kombiniert werden können. Für das Deutsche gibt es beispielsweise die Regel, dass Artikel, Adjektiv und Nomen in einer Nominalgruppe bezüglich Kasus, Numerus und Genus (Fall, Zahl und grammatisches Geschlecht) übereinstimmen müssen. So ist “*des alten Hauses*” eine korrekte neutrale (sächliche) Nominalgruppe im Genitiv Singular.
- **Semantische Ebene:** Die Semantik befasst sich mit der Bedeutung von Wörtern, Ausdrücken und Sätzen. Die Abgrenzung gegenüber der Syntax zeigt sich einleuchtend anhand von Sätzen wie “*Die Polizei fängt den ausgebrochenen Vulkan wieder ein.*”, der zwar syntaktisch, nicht aber semantisch korrekt (*sinnhaft* im Gegensatz zu *sinnlos*) ist.
- **Pragmatische Ebene:** Diese Ebene wird oft nicht der Linguistik im engeren Sinne zugerechnet. Sie befasst sich mit dem Zweck der Sprache und stellt das Geschriebene oder Gesprochene in Bezug zur schreibenden bzw. sprechenden Person und in einen grösseren Zusammenhang. Beispielsweise kann der Satz “*Du musst dieses Buch lesen!*” eine Empfehlung oder ein Be-

fehl sein, abhängig davon, ob die Kommunikation zwischen Kollegen stattfindet oder ob ein Lehrer den Satz zu einem Schüler spricht. Zudem erzielt die Sprache je nach Wortwahl oder Formulierung (bei der Lautsprache ist auch die Sprechweise² relevant) eine unterschiedliche Wirkung.

► 1.2.2 Die phonetische Schrift

Aus dem Fremdsprachenunterricht weiss man, dass aus der Schreibweise (graphemische Form) von Wörtern oft nicht ohne weiteres auf deren korrekte Aussprache (phonetische Form) geschlossen werden kann. Vielen Menschen fallen diese Unterschiede zwischen Schreib- und Sprechweise in ihrer Muttersprache nicht auf. Sie sind jedoch in jeder Sprache in mehr oder weniger ausgeprägter Form vorhanden, so auch im Deutschen.

Da die Lautvielfalt grösser als die Anzahl der Buchstaben ist, hauptsächlich weil auch Dauer, Tonhöhe, Nasalisierung, Stimmhaftigkeit etc. lautunterscheidende Merkmale sind, ist zur Notation der Aussprache eine spezielle Schrift erforderlich, also eine phonetische Schrift.

Es gibt eine grosse Vielfalt von phonetischen Schriften, die das Lautsystem einer Sprache verschieden differenziert beschreiben. Für den Sprachenunterricht hat sich die als IPA-Alphabet (International Phonetic Association) bezeichnete phonetische Schrift durchgesetzt. Die IPA-Lautschrift wird auch in diesem Buch verwendet. Einige Beispiele solcher IPA-Symbole sind in der Tabelle 1.1 aufgeführt. Die vollständige Liste aller hier verwendeten IPA-Symbole ist im Anhang A.1 zu finden.

Tabelle 1.1. IPA-Symbole mit Beispielen in graphemischer und phonetischer Schreibweise.

a	hat	[hat]	d	dann	[dan]
aː	Bahn	[baːn]	e	Methan	[meˈtʰan]
ɐ	Ober	[ˈoːbɐ]	eː	Beet	[beːt]
ʊ	Uhr	[uːʁ]	ɛ	hätte	[ˈhɛtə]
ai	weit	[vʰait]	ɛː	wähle	[ˈvɛːlə]
au	Haut	[hʰaut]	ə	halte	[ˈhaltə]
b	Ball	[bal]	f	Fass	[fas]
ç	sich	[ziç]	ː		

²Bei der Lautsprache ist selbstverständlich die Prosodie (siehe Abschnitt 1.2.4) ein wichtiges pragmatisches Element. So wird im vorgetragenen Märchen der Wolf mit mächtig bedrohlicher Stimme sprechen, das Rotkäppchen situationsbedingt jedoch mit einem dünnen, ängstlichen Stimmlein antworten.

➤ 1.2.3 Die akustisch-phonetische Ebene

Im Gegensatz zu den Abschnitten 1.2.1 und 1.2.2, in denen die lautsprachlichen Elemente (also die Laute) als Abstrakta betrachtet werden, bezieht sich die akustisch-phonetische Ebene der Sprache auf die konkreten Realisierungen. Es interessiert insbesondere die Frage, wodurch sich die verschiedenen Laute auszeichnen bzw. unterscheiden.

In der Phonetik werden die Sprachlaute anhand verschiedenartiger Merkmale charakterisiert. So sind Laute z.B. stimmhaft oder stimmlos (periodisch oder rauschartig), sind gerundet oder ungerundet, hell oder dunkel, zeichnen sich durch bestimmte Formantfrequenzen aus (siehe Abschnitt 1.4.5), sind vokalisches oder konsonantisches. Die Merkmale können sich somit auf die Lautproduktion bzw. den Sprechapparat (nasal, gerundet etc.), auf das Sprachsignal selbst (messbare physikalische Größen wie Periodizität und Formantfrequenzen), auf die akustische Wahrnehmung (hell / dunkel) oder auf linguistische Konventionen (vokalisches / konsonantisches) beziehen. Da die Laute anhand dieser Merkmale unterschieden werden, werden sie als *distinktive Merkmale* bezeichnet.³

➤ 1.2.4 Die Prosodie der Sprache

Die Laute bilden die segmentale Ebene der gesprochenen Sprache. Die supra-segmentale Ebene der Lautsprache wird als Prosodie bezeichnet. Die Prosodie ist ausschliesslich ein Phänomen der gesprochenen Sprache. Ein Pendant dazu gibt es in der geschriebenen Sprache nicht.⁴

Die *linguistische Funktion* der Prosodie umfasst:

- Kennzeichnung des Satztyps, z.B. von Fragesätzen, die nicht aufgrund der Wortstellung zu erkennen sind, und deshalb mit einer gegen das Satzende ansteigenden Intonation gekennzeichnet werden
- Gewichtung, d.h. Unterscheidung wichtiger von weniger wichtigen Teilen einer Äusserung
- Gliederung längerer Äusserungen in sinnvolle Teile.

³In der Phonetik wird gewöhnlich versucht, die Laute einer Sprache anhand eines minimalen Satzes distinktiver Merkmale zu beschreiben, z.B. in [27], Seite 128 ff. Dadurch lassen sich relevante Unterschiede von irrelevanten trennen, was grundsätzlich eine gute Grundlage für die Lautunterscheidung in der Spracherkennung liefern könnte. Weil jedoch gewisse Merkmale nur auf linguistischen Konventionen beruhen (z.B. vokalisches / konsonantisches) und nicht einer physikalischen Eigenschaft des Sprachsignals entsprechen, sind diese Merkmale in der Spracherkennung nicht sehr hilfreich.

⁴Zwar bietet die textliche Form der Sprache die Möglichkeit, Zeichen (Frage-, Ausruferzeichen, Anführungsstriche etc.) oder Markierungen (z.B. das Unterstreichen) anzubringen, oder den Text hinsichtlich Layout zu gestalten. Aber diese Möglichkeiten sind bei weitem nicht gleichwertig.

Die linguistischen Abstrakta der Prosodie sind die Akzente (Betonungen) und die Phrasen (Sprechgruppen), wobei sowohl Akzente als auch Phrasengrenzen verschiedene Stärkegrade aufweisen.

Nebst der linguistischen hat die Prosodie auch eine *ausserlinguistische Funktion*: Sie bestimmt, ob eine Stimme gehetzt, zaghaft, traurig, langweilig, wütend usw. wirkt. Damit spielt die Prosodie eine wesentliche Rolle auf der pragmatischen Ebene der Sprache (vgl. Abschnitt 1.2.1).

Diesen komplexen, die sinnliche Wahrnehmung betreffenden Eigenschaften der Prosodie stehen die im Sprachsignal messbaren physikalischen Grössen gegenüber, die auch als prosodische Grössen bezeichnet werden: der zeitliche Verlauf der Grundfrequenz und der Signalleistung und die Dauer der Laute und der Pausen.

Wie in Abschnitt 1.2.1 erwähnt ist, können je nach Sprache die Dauer und die Tonhöhe auch phonemischen Charakter haben, also bedeutungsunterscheidend sein. So ist etwa im Deutschen die Dauer der Vokale teils phonemisch, weil es beispielsweise Wortpaare wie /va:l/ und /val/ (Wahl und Wall) oder /vɛ:lə/ und /vɛlə/ (wähle und Welle) gibt. Hingegen ist die Tonhöhe in der deutschen Sprache ein rein suprasegmentales Phänomen, ganz im Gegensatz zu den sogenannten Tonsprachen, zu denen beispielsweise Chinesisch gehört.

Bemerkenswert ist, dass sich sowohl die segmentalen, als auch die suprasegmentalen Merkmale auf dieselben physikalischen Grössen auswirken. So wird der an einer konkreten Stelle eines Sprachsignals messbare Wert der Grundfrequenz beeinflusst durch den Laut an dieser Stelle (die Art des Lautes und, bei Tonsprachen, die Art des Tones), durch den Betonungsgrad der Silbe und durch die Position im Satz.

1.3 Die menschliche Sprachproduktion

➤ 1.3.1 Übersicht über den Sprechapparat

Die Gesamtheit der menschlichen Organe, die an der Produktion von Lautsprache beteiligt sind, wird als Sprechapparat bezeichnet. Dazu gehören im Wesentlichen die Lunge, die Luftröhre, der Kehlkopf mit den Stimmlippen (oder Stimmbändern), das Gaumensegel, die Zunge, die Zähne und die Lippen. Wichtig im Zusammenhang mit der Sprachproduktion sind auch die Hohlräume, insbesondere Rachen, Mund und Nasenraum. Eine Übersicht über den Aufbau des menschlichen Sprechapparates zeigt Abbildung 1.3.

➤ 1.3.2 Die Funktion des Sprechapparates

Hinsichtlich der Funktion beim Sprechprozess können beim menschlichen Sprechapparat zwei Komponenten unterschieden werden:

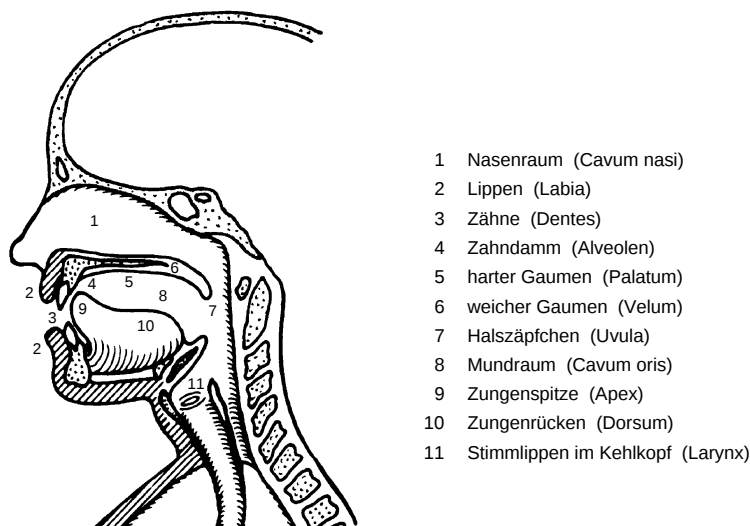


Abbildung 1.3. Mittelschnitt (Sagittalschnitt) durch den menschlichen Sprechapparat

a) **Schallproduktion:** Während des Sprechens wird aus der Lunge Luft ausgestossen, welche die aneinander liegenden Stimmlippen in Schwingung versetzt, wodurch der Luftstrom periodisch unterbrochen wird. Die Schwingfrequenz der Stimmlippen beträgt bei Männern im Mittel etwa 120 Hz, bei Frauen etwa 220 Hz. Die Schwingfrequenz variiert beim Sprechen in einem etwa eine Oktave umfassenden Bereich. Der pulsierende Luftstrom ist nichts anderes als ein akustisches Signal. Da die Stimmlippen im Luftstrom nicht frei schwingen, sondern periodisch aneinander schlagen und so den Unterbruch des Luftstromes bewirken, ist das entstehende akustische Signal stark oberwellenhaltig.

Liegen die Stimmlippen beim Ausstossen der Luft nicht aneinander, dann entweicht die Luft gleichförmig durch Mund und/oder Nase. Wird der Luftstrom durch eine Engstelle behindert, dann entstehen Turbulenzen, die sich akustisch als zischendes Geräusch äussern.

b) **Klangformung:** Die mehr oder weniger neutralen akustischen Signale von den Stimmlippen und von den Engstellen werden durch Rachen, Mund und Nasenraum (sie bilden zusammen den sogenannten Vokaltrakt), die als akustisches Filter wirken, klanglich verändert. Wesentlich ist, dass sich mit der Bewegung der Artikulatoren (Zunge, Lippen, Kiefer und Gaumensegel) die Übertragungsfunktion und damit auch die Resonanzfrequenzen bzw. die Formanten (vergl. Abschnitt 1.4.5) dieses akustischen Filters verändern. So entstehen aus dem Signal von den Stimmlippen recht verschiedene Klänge,

nämlich die Laute, insbesondere die stimmhaften.

Auch die von den Luftturbulenzen herrührenden Geräusche werden durch den Vokaltrakt klanglich geformt, wobei hier in erster Linie der Ort der Engstelle massgebend ist.

Der Zusammenhang zwischen der Stellung der Artikulatoren und den Elementen der Lautsprache, also den Lauten, wird in Abschnitt 1.3.3 erläutert.

➤ 1.3.3 Die Artikulation der Sprachlaute

Die beiden folgenden Abschnitte charakterisieren die Laute der deutschen Sprache anhand der Stellung der Artikulatoren. Die Laute werden dazu in zwei grosse Gruppen unterteilt, die Vokale und die Konsonanten.

➤ Die Vokale

Vokale sind Laute, bei denen die Stimmlippen im Kehlkopf schwingen und die Atemluft ungehindert durch den Mund bzw. durch Mund und Nase (bei nasalierten Vokalen) ausströmt. Die beweglichen Artikulatoren (Lippen, Zungenspitze, Zungenrücken etc.) dürfen sich deshalb nur so verschieben, dass nirgends im Vokaltrakt zu enge Stellen auftreten. So darf auch der Zungenrücken eine gewisse Grenzlinie nicht überschreiten.

Die Klangfarbe des durch den Vokaltrakt produzierten Signals hängt vor allem von der Stellung des Zungenrückens und der Lippen ab. Bezüglich der Lippenstellung werden gerundete und ungerundete Vokale unterschieden:

ungerundet: [i ɪ e ə a ɐ]
gerundet: [y ʏ ø œ u ʊ o ɔ]

Der Zusammenhang zwischen der Stellung des Zungenrückens und des produzierten Lautes wird durch das Vokalviereck in Abbildung 1.4 beschrieben.

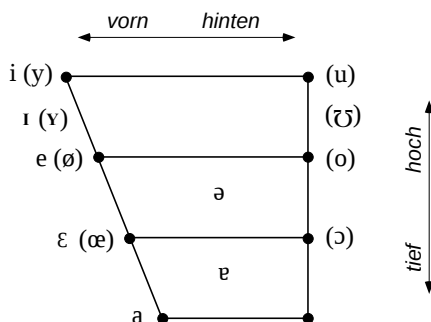


Abbildung 1.4. Das Vokalviereck gemäss [9] beschreibt grob den Zusammenhang zwischen der Stellung des Zungenrückens und den Vokalen der deutschen Sprache. Die gerundeten Laute sind in Klammern gesetzt. Das Zentrum des Vokalvierecks bedeutet die entspannte Mittellage, in welcher der sogenannte Schwa-Laut artikuliert wird.

⊗ **Die Konsonanten**

Konsonanten sind Laute, bei denen ausströmende Atemluft durch eine Verengung im Vokaltrakt behindert oder durch einen Verschluss während einer gewissen Zeit gestoppt wird. Je nach Laut schwingen oder ruhen dabei die Stimmlippen. Im Deutschen werden die Konsonanten üblicherweise nach Artikulationsart und -ort gegliedert, was zu einer Tabelle führt, wie sie in Abbildung 1.5 dargestellt ist. Da die Affrikaten keine eigenständigen Laute sind, sondern durch enge Verbindung eines Verschlusslautes mit dem homorganen Reibelaut (homorgane Laute haben denselben Artikulationsort) entstehen, sind sie in dieser Tabelle nicht aufgeführt.

Ort Art								
	bilabial	labiodental	alveolar	palatoalveolar	palatal	velar	uvular	glottal
plosiv	p b		t d			k g		
nasal	m		n			ŋ		
frikativ		f v	s z	ʃ	ç j	x		
lateral			l					
vibriert			r				R	
aspiriert								h

Abbildung 1.5. Die Konsonanten der deutschen Sprache (ohne Affrikaten) geordnet nach Artikulationsort und -art gemäss [9]

1.4 Das menschliche Gehör

Im Zusammenhang mit der Sprachverarbeitung ist die Anatomie des menschlichen Gehörs nur von untergeordnetem Interesse, weil sich auf diesem Wege die sinnliche Wahrnehmung akustischer Ereignisse nur zu einem kleinen Teil erklären lässt. Nicht selten kommt es vor, dass aus der Anatomie gezogene Schlüsse der Perzeption widersprechen. Ein derartiges Phänomen ist beispielsweise, dass der Mensch die Höhe eines Tons von 1000 Hz auf etwa 0,1 % genau hört, also viel genauer als aufgrund des Aufbaus des Innenohres zu schliessen wäre. Ebenso ist das gute Richtungshören mit dem Aufbau des Ohres nicht erklärbar.

In diesem Abschnitt werden deshalb weder der physiologische Aufbau des Ohrs, noch Hypothesen über die Arbeitsweise des Gehirns behandelt. Vielmehr sollen ein paar für die Sprachverarbeitung bedeutsame Gehörphänomene dargelegt werden, die sich über psychoakustische Experimente feststellen lassen.

► 1.4.1 Wahrnehmung der Schallintensität

Schallintensität I und Schalldruck p sind physikalische Größen⁵ und damit messtechnisch erfassbar. Die Lautheit bezeichnet hingegen eine subjektive Wahrnehmung, die über Experimente mit Versuchspersonen in Beziehung zur Schallintensität gebracht werden kann. So lässt sich beispielsweise feststellen, dass die wahrgenommene Lautheit von Sinustönen stark frequenzabhängig ist, wie Abbildung 1.6 zeigt.

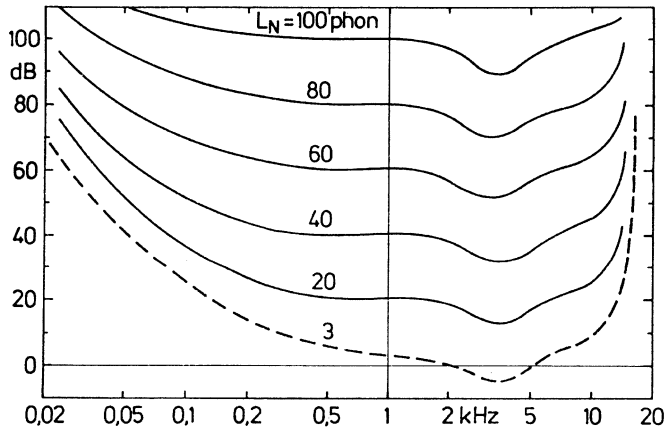


Abbildung 1.6. Kurven gleicher Lautheit nach [50], wobei für 1 kHz die Lautheit in Phon so definiert ist, dass sie gleich der Schallintensität in dB ist. Die gestrichelte Kurve ist die mittlere Ruheschwelle.

Diese Frequenzabhängigkeit ist insbesondere in der Nähe der sogenannten Ruheschwelle (gestrichelte Kurve) ausgeprägt. Die Ruheschwelle gibt in Funktion der Frequenz die minimale Schallintensität an, die das Ohr noch zu hören vermag.⁶

⁵Der Zusammenhang zwischen Schallintensität und Schalldruck ist für eine sinusförmige Schallwelle gegeben als: $I = p_{\text{eff}}^2 / (\rho c)$, wobei ρ die spezifische Dichte des Gases (Luft: 1.293 kg/m^3) und c die Schallgeschwindigkeit ist.

⁶Interessanterweise wird in älteren Publikationen die Ruheschwelle bei 1 kHz mit einem Schalldruck von $p_{\text{eff}} = 2 \cdot 10^{-5} \text{ N/m}^2$ angegeben, was der Schallintensität von etwa 10^{-12} W/m^2 entspricht. Dieser als Ruheschwelle bei 1 kHz ermittelte Schalldruck wurde deshalb als 0 dB definiert. Heute scheint sich die Ruheschwelle

Wichtig zu erwähnen ist auch, dass das Gehör Schallintensitäten relativ wahrnimmt und zwar über einen enorm grossen Bereich. Dieser erstreckt sich von der Ruhehörschwelle bis zur Schmerzgrenze (ca. 130 Phon) über gut sechs Grössenordnungen.

➤ 1.4.2 Periodizität und Tonhöhe

Das menschliche Ohr hört die Frequenz von Sinusschwingungen in einem Bereich von etwa 20 Hz bis 18 kHz als Tonhöhe, wobei mit zunehmendem Alter die obere Grenze etwas abnimmt. Ähnlich wie im Fall der Lautheit nimmt das Ohr auch die relative Tonhöhe wahr. So hört beispielsweise das Ohr dann einen positiven Halbtonschritt zwischen zwei Tönen, wenn das Frequenzverhältnis $f_2/f_1 = 2^{1/12}$ beträgt.

Die meisten natürlichen Schallquellen geben jedoch keine reinen Töne (Sinusschwingungen) ab, auch der menschliche Sprechapparat nicht. Mit den periodisch schwingenden Stimmlippen als Schallquelle entsteht aus dem Vokaltrakt ein periodisches Signal, aus dem das Ohr eindeutig eine Tonhöhe wahrnimmt, die als Grundfrequenz F_0 bezeichnet wird und umgekehrt proportional zur Periode T_0 ist.

Diese Formulierung scheint auf den ersten Blick übertrieben kompliziert zu sein. Sie trägt jedoch der Tatsache Rechnung, dass das Ohr aus einem Signal mit harmonischen Komponenten *nicht* alle diese Komponenten mit ihren Tonhöhen hört, sondern einen einzigen Klang mit einer Tonhöhe, die gleich der Frequenz der Grundwelle ist. Wird jedoch die Grundwelle aus dem Signal ausgefiltert, dann hört das Ohr trotzdem noch die ursprüngliche Tonhöhe. Dies zeigt sich beispielsweise beim Telefon, wo ja nur die Frequenzkomponenten zwischen 300 Hz und 3400 Hz übertragen werden. Bei Männerstimmen gehen somit in der Regel die Grundwelle und die erste Oberwelle verloren. Trotzdem tönt die Stimme durch das Telefon nicht ein oder zwei Oktaven höher als beim direkten Hinhören. Vielfach wird deshalb gesagt, dass das Ohr die reziproke Periode als Tonhöhe wahrnimmt.⁷

Aus einem oberwellenhaltigen Signal hört das Ohr nebst der Tonhöhe auch eine Klangfarbe, welche durch die Amplitudenverhältnisse der Frequenzkomponenten bestimmt wird. Oberwellenhaltige Schallsignale werden deshalb auch als Klänge bezeichnet.

für Versuchspersonen unter 25 Jahren um etwa 3 dB erhöht zu haben, wahrscheinlich infolge der allgemeinen akustischen Reizüberflutung.

⁷Häufig ist dieses Modell der Tonhöhenwahrnehmung zweckmässig, insbesondere auch in der Sprachverarbeitung. Es darf jedoch nicht mit der wirklichen Tonhöhenwahrnehmung verwechselt werden, die nach wie vor unbekannt ist, wie sich mit Signalbeispielen zeigen lässt.

Periodische und rauschartige Teile von Sprachsignalen unterscheiden sich im Wesentlichen durch die Aktivität der Stimmlippen und werden deshalb stimmhaft bzw. stimmlos genannt.

Signale, für welche das Ohr keine Tonhöhe feststellen kann, sind rauschartig. Dies trifft z.B. für das Sprachsignal des Lautes [s] zu, beim dem die Stimmlippen inaktiv sind.

➤ 1.4.3 Die Phasenwahrnehmung

Stark vereinfachend wird das menschliche Gehör oft als phasentaub bezeichnet. Richtig ist, dass beliebige, frequenzabhängige Laufzeitveränderungen im Bereich von wenigen Millisekunden (Kurzzeitphase, wie sie mit einem nicht allzu langen Allpassfilter erzeugt wird) nur sehr schlecht hörbar sind. Grössere Phasenveränderungen werden aber als Hall oder sogar als Echo wahrgenommen.

Eine wichtige Funktion hat die Phase beim binauralen Hören, nämlich in Bezug auf die Schallortung. Insbesondere der Laufzeitenunterschied des Schalls von der Quelle zu den beiden Ohren wird im Gehirn ausgewertet um die Einfallsrichtung des Schalls zu detektieren. Dementsprechend kann durch Verändern der Phase eines oder beider Hörsignale die Ortung der Schallquelle erschwert, verunmöglicht oder sogar irreführt werden, je nach Art der Phasenveränderung.

➤ 1.4.4 Der Verdeckungseffekt

Eine im Zusammenhang mit der Sprachverarbeitung wichtige Eigenschaft des menschlichen Gehörs ist der Maskier- oder Verdeckungseffekt. Er besteht qualitativ darin, dass eine Schallkomponente mit grösserer Leistung eine andere mit kleinerer Leistung verdeckt, also unhörbar macht, wobei der Maskierungsbereich vor allem von der Frequenz und der Intensität des maskierenden Signals abhängt.

Messtechnisch kann der Maskierungsbereich mit verschiedenen Methoden und Signalarten erfasst werden, wobei die Resultate leicht unterschiedlich ausfallen. In [50] ist die als *Mithörschwelle* bezeichnete Grenze des Maskierungsbereiches für Schmalbandrauschen mit der Mittenfrequenz f_m bestimmt worden. Die Mithörschwelle gibt die Intensität eines Sinussignals in Funktion der Frequenz an, das durch das Maskierungsrauschen gerade noch verdeckt wird. In Abbildung 1.7 sind die Mithörschwellen für Schmalbandrauschen (als maskierendes Signal) mit verschiedenen Mittenfrequenzen dargestellt. Abbildung 1.8 zeigt die gemessenen Mithörschwellen, wenn Schmalbandrauschen gleicher Mittenfrequenz aber mit verschiedener Intensität als maskierendes Signal verwendet wird.

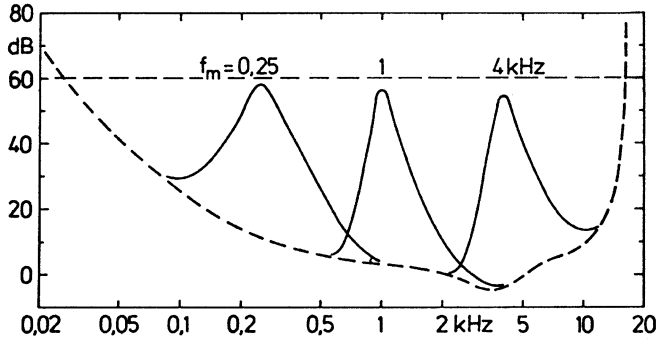


Abbildung 1.7. Mithörschwellen für Schmalbandrauschen mit den Mittenfrequenzen 0,25 kHz, 1 kHz und 4 kHz (nach [50])

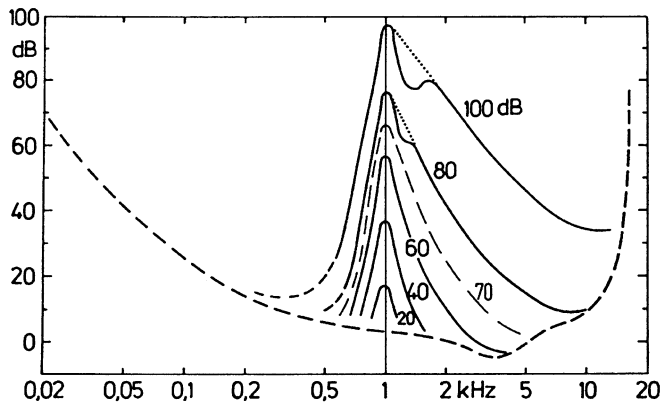


Abbildung 1.8. Mithörschwellen für Schmalbandrauschen mit der Mittenfrequenz 1 kHz und verschiedenen Schallpegeln (nach [50])

➤ 1.4.5 Wahrnehmung der Sprachlaute

In der Phonetik werden die Laute nicht nur anhand artikulatorischer Kriterien eingeteilt, sondern vor allem auch aufgrund akustischer Merkmale. So ist es beispielsweise gebräuchlich, Vokale mittels der Formanten zu beschreiben. Als Formanten werden die lokalen Maxima des Spektrums des Sprachsignals bezeichnet, welche von den Resonanzen des Vokaltraktes herrühren. Da der Vokaltrakt eine komplizierte Form aufweist, treten beim Artikulieren eines Lautes gewöhnlich mehrere Formanten gleichzeitig auf, die als F_1 , F_2 , F_3 etc. bezeichnet werden. Ein Formant wird mit den Parametern Mittenfrequenz (oder Formantfrequenz), Bandbreite und Amplitude beschrieben. Die Bezeichnungen F_1 , F_2 , F_3 etc. werden oft auch für die Formantfrequenzen verwendet. Für die Formant-

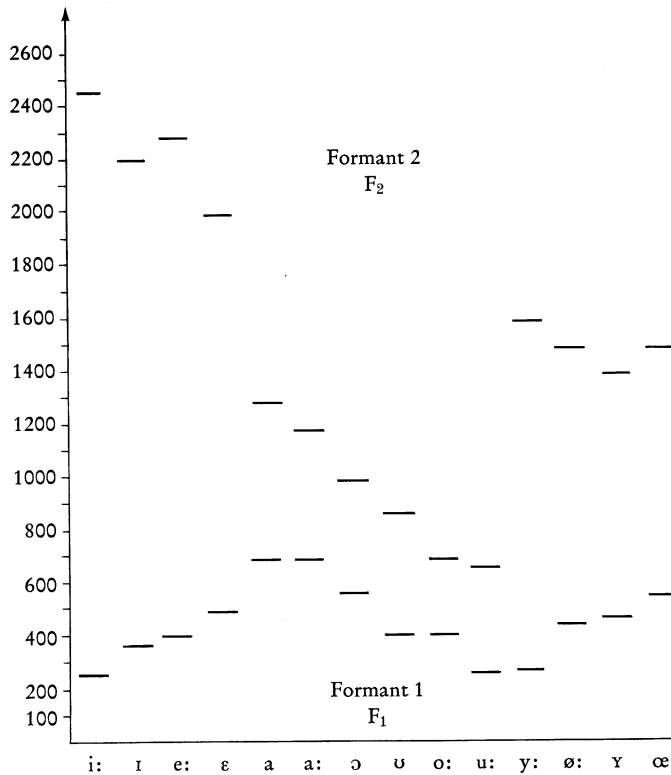


Abbildung 1.9. Durchschnittswerte der zwei untersten Formanten deutscher Vokale (aus [27], Seite 54)

frequenzen eines Lautes gilt: $F_1 < F_2 < F_3$ etc. Der tiefste Formant oder die tiefste Mittenfrequenz ist also stets F_1 .

Die Mittenfrequenz und die Bandbreite der Formanten sind von der Stellung der Artikulatoren abhängig. Wie in Abschnitt 1.3.3 erläutert, bestimmt die Stellung der Artikulatoren den gesprochenen Laut und somit gibt es auch einen Zusammenhang zwischen den Lauten und den Formantfrequenzen. Dieser Zusammenhang ist für die beiden tiefsten Formanten der deutschen Vokale in Abbildung 1.9 veranschaulicht.

Da im Vokaltrakt nur bei den Vokalen ausgeprägte Resonanzen auftreten, wird der Begriff der Formanten nur auf die Vokale angewendet.

1.5 Verarbeitung natürlicher Sprache

1.5

Die Tatsache, dass der Mensch Lautsprache ohne nennenswerte Anstrengung produzieren und verstehen kann, verleitet zur Annahme, dass es sich dabei um eine einfache Aufgabe handle, die auch von einem Computer leicht zu bewältigen sein müsse. Vergleicht man jedoch beispielsweise die heute mit modernsten Mitteln erreichbare Spracherkennungsleistung mit den menschlichen Fähigkeiten, dann stellt sich das bisher Erreichte als noch recht bescheiden heraus.

Ende der sechziger Jahre herrschte die euphorische Ansicht vor, dass im Hinblick auf die rasante Entwicklung der Computertechnik das Spracherkennungsproblem in wenigen Jahren gelöst sein werde. Heute wird die Situation allgemein realistischer eingeschätzt. Viele Forscher sind der Ansicht, dass trotz grosser Forschungsanstrengungen die maschinelle Spracherkennung auch in 20 Jahren noch nicht den Stand der menschlichen Sprachwahrnehmungsfähigkeit erreicht haben wird.

Ein wesentlicher Grund dafür ist sicher, dass die Handhabung natürlicher Sprache der grossen Komplexität wegen enorm schwierig ist.⁸ Während heute die maschinelle Beherrschung des Wortschatzes einer natürlichen Sprache⁹ halbwegs gelingt, ist ein Computer nicht in der Lage, für beliebige Sätze zu entscheiden, ob sie syntaktisch korrekt sind oder nicht. Noch viel schwieriger wird es beim Entscheid über die Bedeutung von Sätzen, die sich je nach Situation stark ändern kann. Genau dies macht der Mensch jedoch beim Verstehen von Sprache: er untersucht fortwährend, ob das Gehörte Sinn macht und kann damit z.B. sehr effizient ähnlich klingende Wörter mit unterschiedlicher Bedeutung gut auseinander halten.

Eine weitere Schwierigkeit liegt darin, dass die beiden Erscheinungsformen natürlicher Sprachen sehr unterschiedlich sind (siehe Abschnitt 1.1.2) und folglich mit völlig anderen Mitteln und Methoden beschrieben bzw. verarbeitet werden müssen. Deshalb sind viele Aufgaben der Sprachverarbeitung, insbesondere auch die Sprachsynthese und die Spracherkennung nicht rein linguistische Probleme. Sie sind mit verschiedenen Wissensbereichen verknüpft (vergl. Abbildung 1.10).

⁸Im Gegensatz zu natürlichen Sprachen sind formale Sprachen (z.B. Programmiersprachen) relativ kompakt und insbesondere genau definiert. Die maschinelle Verarbeitung ist deshalb vergleichsweise unproblematisch. Es kann beispielsweise einfach entschieden werden, ob ein Programm syntaktisch korrekt ist oder nicht.

⁹Die Linguistik bezeichnet den Wortschatz der deutschen Sprache zwar als nicht begrenzt, weil durch Bildung von Komposita stets neue Nomen, Verben und Adjektive gebildet werden können. Im mathematischen Sinne ist der Wortschatz aber begrenzt, weil Wörter praktisch nicht beliebig lang sein können.

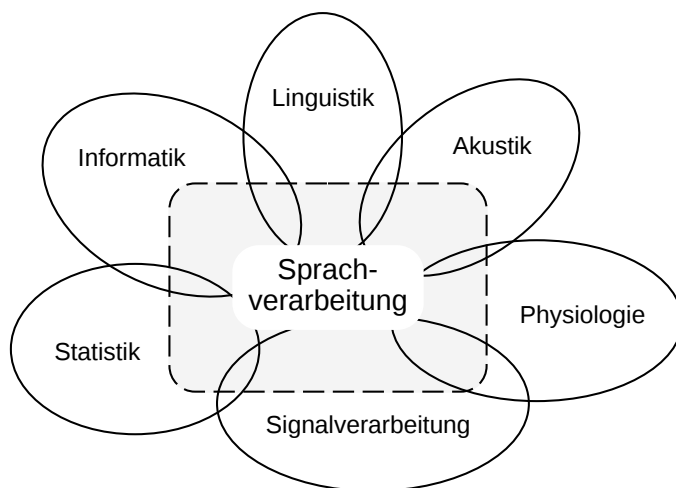


Abbildung 1.10. Die Sprachverarbeitung stützt sich auf Wissen und Methoden verschiedener Disziplinen: sie ist interdisziplinär.

In der Sprachverarbeitung ist es deshalb erforderlich, sich mit Wissen aus recht verschiedenen Disziplinen zu beschäftigen, wobei hier selbstverständlich nur ein paar Schwerpunkte behandelt werden können, die im Zusammenhang mit der Spracherkennung und der Sprachsynthese wichtig sind.

Sprachverarbeitung

Grundlagen und Methoden der Sprachsynthese und
Spracherkennung

Pfister, B.; Kaufmann, T.

2017, XIV, 507 S. 175 Abb. Mit Online-Extras., Softcover

ISBN: 978-3-662-52837-2