LINGUISTISCHE Phonetik

Jörg Mayer
Universität Stuttgart, 2010

Universität Stuttgart
Institut für Maschinelle Sprachverarbeitung
Azenbergstraße 12
70174 Stuttgart
Email: joerg.mayer@ims.uni-stuttgart.de

Inhaltsverzeichnis

Abbildungsverzeichnis

Einleitung

Es gibt zwei Disziplinen, die sich mit den lautlichen Aspekten der Sprache befassen: Phonetik und Phonologie. Gegenstand der Phonologie ist die Beschreibung von Lautsystemen und von systematischen Prozessen innerhalb von Lautsystemen. Die Phonetik interessiert sich dagegen mehr für die 'materiellen' Aspekte der Lautsprache: Wie werden Laute produziert, wie unterscheiden sich Laute akustisch und wie werden akustische Ereignisse wahrgenommen.

Ein Beispiel: Die Lautkette /lift/ bedeutet im Deutschen etwas anderes als die Lautkette /luft/.[1] Dieser Bedeutungsunterschied wird nur dadurch hergestellt, dass ein Laut — nämlich der Vokal — ausgetauscht wird; alle anderen Laute sind identisch. Dies weist auf eine Eigenschaft des deutschen Lautsystems hin: Im Deutschen scheint es zwei Laute zu geben, /i/ und /u/, die sich kategorial unterscheiden, d.h. sie können einen Bedeutungsunterschied ausdrücken. Es gibt im Deutschen natürlich sehr viel mehr Laute, die sich kategorial unterscheiden, doch mit diesem Test (dem sog. Minimalpaartest) konnten zunächst einmal zwei Lautkategorien identifiziert werden. Bis hierher haben wir Phonologie betrieben. Betrachten wir nun den Laut /l/, der, wie oben gesagt, in beiden Wörtern identisch ist. Stimmt das? Vom phonologischen Standpunkt betrachtet durchaus: /lift/ und /luft/ werden nicht durch den Austausch des initialen Konsonanten unterschieden, sondern durch den Austausch des Vokals. Vom phonetischen Standpunkt betrachtet gibt es jedoch einen erheblichen Unterschied zwischen den beiden /l/–Lauten: In /lift/ wird das /l/ mit gespreizten Lippen produziert, in /luft/ dagegen mit gerundeten Lippen. Der Grund hierfür ist die Koartikulation, d.h. die artikulatorische Beeinflussung eines Lautes durch benachbarte Laute. Im vorliegenden Fall setzt die /i/–typische Lippenspreizung bzw. die /u/–typische Lippenrundung

Koartikulation

[1] Zeichen zwischen Schrägstrichen repräsentieren Laute (nicht Buchstaben!). Das Wort *mein* würde entsprechend als /main/ transkribiert. Näheres zur symbolischen Repräsentation von Lauten (Transkription) in Abschnitt 1.2.1.

schon während der Produktion des /l/ ein; die akustische Charakteristik von /l/ wird dadurch um Nuancen verändert. Diese Veränderung ist zwar messbar und evtl. auch (zumindest von geübten Hörern) hörbar, ein 'naiver', d.h. nicht an phonetischen Feinheiten interessierter Hörer wird den Unterschied jedoch nicht wahrnehmen und in beiden Fällen den selben Laut (genauer: das selbe Phonem) identifizieren.

Hier einige weitere Beispiele, die den Unterschied zwischen phonologischen und phonetischen Fragestellungen aufzeigen: Der Minimalpaartest /lift/ – /luft/ zeigt die phonologische Opposition zwischen /i/ und /u/ und identifiziert zwei Vokalphoneme des Deutschen. Analysiert man jedoch beispielsweise die /i/–Produktion zwei verschiedener Sprecher des Deutschen, wird man schon bei genauem Hinhören erhebliche Unterschiede zwischen den /i/–Realisierungen der zwei Sprecher finden. Vergleicht man z.B. einen Sprecher mit einer Sprecherin, liegt eine ganz offensichtliche Differenz in der Höhe des Stimmtons (Sprachgrundfrequenz), die /i/–Laute der Sprecherin werden generell mit höherer Sprachgrundfrequenz produziert als die des Sprechers. Daneben lassen sich jedoch auch subtilere Unterschiede in der Klangqualität der Vokale wahrnehmen; so werden auch die /i/–Realisierungen zweier männlicher Sprecher nicht genau identisch klingen. Solche subtilen Eigenheiten der Lautproduktion individueller Sprecher sind z.B. ein wichtiges Thema der forensichen Phonetik im Rahmen der Sprechererkennung. Aber auch in der allgemeinen Phonetik spielen solche Unterschiede unter dem Gesichtspunkt eines allgemeinen Erkenntnisinteresses an den Mechanismen der Lautproduktion und des Einflusses individueller Vokaltraktkonfigurationen eine Rolle. Konzentriert man sich bei der Analyse von /i/–Lauten auf einen Sprecher, so wird man auch hier Unterschiede finden. Verantwortlich für solche Varianz sind z.B. Betonungsstatus, Sprechgeschwindigkeit und Sprechstil, aber auch soziale und emotionale Faktoren wie formelle vs. informelle Sprechsituation oder wütende vs. traurige Äußerungen. Obwohl es also aus phonologischer Sicht nur ein /i/–Phonem im Deutschen gibt, wird man bei der Analyse konkreter /i/-Realisierungen prinzipiell unendlich viele /i/–Varianten finden. Diese Varianz der Sprachlaute und die Untersuchung der verantworlichen Faktoren ist ein zentrales Thema der Phonetik: Welche Faktoren beeinflussen auf welche Weise die Artikulation von Sprachlauten und welche Konsequenzen hat dies für die akustische Qualität der Sprachlaute? Die Phonetik beschäftigt sich jedoch nicht nur mit den produktiven Aspekten der Lautsprache, sondern auch mit der Perzeption: Weshalb und unter welchen Umständen werden diese unendlich vielen Varianten eines Lautes stets als ein und das selbe Phonem wahrgenommen? Gibt es bestimmte invariante Eigenschaften

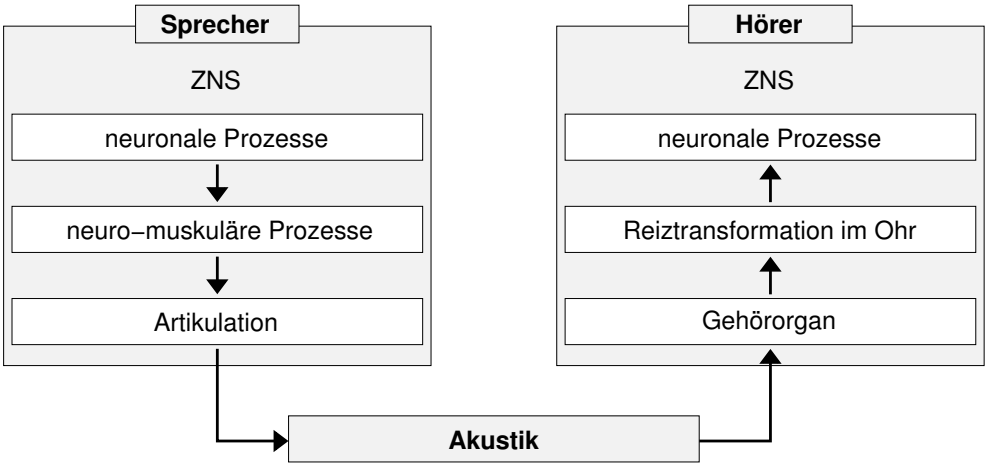


Abbildung 1: *Das 'signalphonetische Band' (nach Pompino-Marschall, [13]).*

z.B. eines /i/–Lautes und welche sind das? Weitere Themen der perzeptiven Phonetik sind z.B. auch das Zusammenspiel der auditiven und der visuellen Wahrnehmung bei der Lautidentifikation (der sog. 'McGurk–Effekt') oder das Verhältnis zwischen *bottom–up–* und *top–down*–Prozessen bei der Lautwahrnehmung.

Ohrenphonetik

Zu phonetischen Fragestellungen gibt es verschiedene Zugänge. Die älteste Art, Phonetik zu betreiben, ist die so genannte Ohrenphonetik. Das bedeutet, dass der Phonetiker sich Äußerungen anhört und versucht, diese z.B. zu transkribieren, d.h. die geäußerten Laute zu identifizieren und mit entsprechenden Symbolen zu beschreiben (daher auch die Bezeichnung Symbolphonetik). Oder er versucht, aus dem Gehörten Rückschlüsse auf artikulatorische Vorgänge zu ziehen, das Gehörte analysierend zu beschreiben (deshalb auch deskriptive Phonetik). Eine andere Art der Phonetik, die sich aufgrund der benötigten technischen Hilfsmittel etwas später entwickelt hat, ist die Instrumentalphonetik. Hierbei werden die physikalischen Aspekte der Lautsprache in Form von Signalen gemessen (deshalb auch Signalphonetik), analysiert und zu dem vorhandenen Wissen über Sprachproduktion und –perzeption in Beziehung gesetzt. In Abbildung 1 sind die einzelnen Komponenten der lautsprachlichen Kommunikation aufgeführt, die einen signalphonetischen Zugang erlauben.

Signalphonetik

Der am einfachsten zugängliche und daher auch am weitesten entwickelte signalphonetische Bereich ist die Akustik. Schon mit einem normalen Computer und Programmen, die häufig kostenlos zu bekommen sind, sind sehr detaillierte akustische Analysen möglich. Andere Bereiche erfordern dagegen

einen erheblich größeren technischen Aufwand; z.B. die Untersuchung artikulatorischer Prozesse mit Hilfe der Elektropalatographie (EPG) oder der Elektromagnetischen mediosagittalen Artikulographie (EMMA) oder die Untersuchung neuromuskulärer Prozesse mit Hilfe der Elektromyographie (EMG).

phonetische Ereignisse

Die Daten des Ohrenphonetikers sind grundsätzlich anderer Art als die des Instrumentalphonetikers. Der Ohrenphonetiker untersucht Lautkategorien, während sich der Instrumentalphonetiker mit physikalischen Signalen beschäftigt. Lautkategorien oder phonetische Ereignisse sind der auditiven Wahrnehmung unmittelbar zugänglich. Diese Ereignisse sind es, die für die Gesprächspartner die Basis einer lautsprachlichen Kommunikation bilden: Wir nehmen Laute wahr, setzen diese zusammen zu Silben, Wörtern und Sätzen und erschließen daraus die 'Botschaft', die man uns mitteilen wollte (sehr vereinfacht ausgedrückt). Allerdings interessieren wir uns dabei in der Regel nicht für die phonetischen Details einer Äußerung (z.B. ob ein /l/ mit gespreizten oder mit gerundeten Lippen produziert wurde). Genau dies ist jedoch die Aufgabe des Ohrenphonetikers; ihn interessiert weniger was gesagt wurde als vielmehr wie es gesagt wurde. Der Ohrenphonetiker unterscheidet sich also prinzipiell nicht von einem Hörer in einer normalen Kommunikationssituation — beide nehmen Sprachlaute wahr —, nur die Aufmerksamkeit richtet sich auf verschiedene Dinge: Den Phonetiker interessieren die phonetischen Nuancen der wahrgenommenen Laute, den normalen Hörer deren kommunikative Funktion. Hier zeigt sich unter anderem die große Relevanz der Untersuchung des Verhältnisses zwischen *bottom–up–* und *top–down–*Prozessen. Der ohrenphonetische Zugang ist idealerweise ein reiner *bottom–up–*Prozess: Der Phonetiker nimmt Laute wahr, ohne sich um deren kommunikative Funktion zu kümmern, und — stark vereinfacht ausgedrückt — analysiert das Wahrgenommene mit seinem Gehör. Dies ist eine ungemein schwierige Aufgabe, wenn der Ohrenphonetiker seine eigene Muttersprache oder eine andere ihm bekannte Sprache untersucht, da sich unwillkürliche *top–down–*Prozesse kaum unterdrücken lassen. Bei Realisierung des Syntagmas *in Berlin* ist das wahrscheinlichste Perzept eines deutschen Muttersprachlers die Lautabfolge /ɪmbɛʁliːn/; die tatsächlich realisierte Lautabfolge ist jedoch mit größter Wahrscheinlichkeit /ɪmbɛrliːn/[2]. D.h. ein Perzept entspricht nicht unbedingt nur dem, was wir wahrnehmen, sondern setzt sich zusammen aus dem Wahrgenommenen und dem, was wir erwarten. Im genannten Bespiel speist sich

[2]Der Artikulationsort des Nasals im Auslaut der Präposition (zugrundeliegend alveolar) assimiliert an den Artikulationsort des nachfolgenden Plosivs (bilabial). Solche Assimilationsprozesse sind insbesondere bei schnellem, informellem Sprechen zu beobachten.

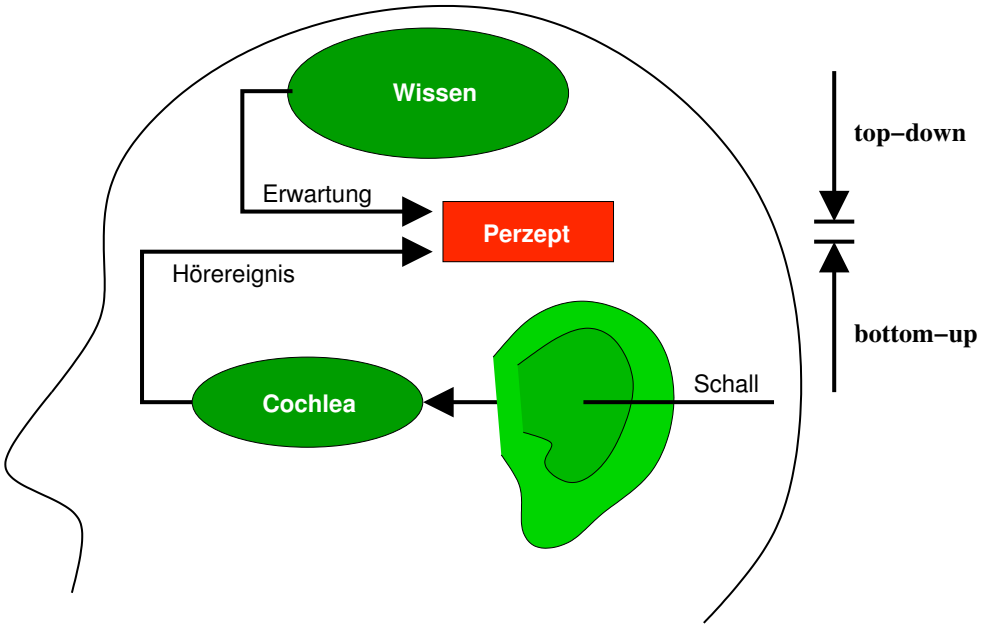


Abbildung 2: *Top–down und bottom–up–Verarbeitung bei der Sprachwahrnehmung.*

diese Hörerwartungen aus unserem lexikalisch–grammatischen Wissen (vgl. Abb. 2).

physikalische Vorgänge

Die Signale des Instrumentalphonetikers sind der Wahrnehmung *nicht* unmittelbar zugänglich. Er untersucht physikalische Vorgänge, die während einer lautsprachlichen Kommunikation ablaufen und die den Teilnehmern der Kommunikation verborgen sind. So sind z.B. die elektrischen Potentiale der Muskeln, die wir benötigen, um die Lippen vorzustülpen (z.B. um ein /u/ zu produzieren), weder für den Sprecher noch für den oder die Hörer wahrnehmbar, können jedoch mit Hilfe der EMG als Signal gemessen und dargestellt werden. Rein logisch besteht zunächst überhaupt kein Zusammenhang zwischen einem Muskelpotential und dem deutschen Vokal /u/; beides sind grundsätzlich unterschiedliche Dinge. Empirisch gesehen besteht jedoch ein Zusammenhang: Wenn ein Sprecher etwas produziert, was als /u/ wahrgenommen werden kann, lassen sich die entsprechenden Muskelpotentiale ableiten, d.h. beide Phänomene sind korreliert, sie hängen — empirisch — zusammen. Die systematische Untersuchung dieser Zusammenhänge ist Gegenstand eines dritten phonetischen Ansatzes, der Experimentalphonetik. Sie versucht

Experimentalphonetik

die Signale des Instrumentalphonetikers mit den wahrgenommenen phonetischen Ereignissen des Ohrenphonetikers in Beziehung zu setzen.

Gegenstandsbereiche der Phonetik

Neben der Unterteilung der Phonetik nach dem methodischen Ansatz ist es üblich, die phonetischen Teildisziplinen nach ihrem Gegenstandsbereich zu unterteilen. Die Teildisziplin, die sich mit den Produktionsaspekten von Sprachlauten beschäftigt, heißt artikulatorische Phonetik, das 'Übertragungssignal' zwischen Sprecher und Hörer ist Gegenstand der akustischen Phonetik und mit der Wahrnehmung von Sprachlauten beschäftigt sich die perzeptive Phonetik (da der auditive Kanal bei der Wahrnehmung von Sprachlauten zwar nicht die einzige aber doch eine zentrale Rolle spielt, wird diese Teildisziplin oft auch auditive Phonetik genannt). An dieser Systematik wird sich das Skript weitgehend orientieren.

Kapitel 1

Artikulatorische Phonetik

1.1 Die Grundlagen der Sprachproduktion

Das grundsätzliche Prinzip der Produktion von Lautsprache ist die Modulation eines Luftstroms, d.h. wenn keine Luft bewegt wird, können auch keine hörbaren Laute produziert werden. Normalerweise wird der für das Sprechen notwendige Luftstrom durch das Ausatmen erzeugt. Prinzipiell ist es jedoch auch möglich, während des Einatmens zu sprechen, allerdings nur relativ leise und relativ kurz. Neben dem durch die Atmung erzeugten Luftstrom, dem sogenannten pulmonalen Luftstrommechanismus, gibt es noch einige andere Möglichkeiten, Luft in Bewegung zu versetzen, die jedoch bei der Lautproduktion eine untergeordnete Rolle spielen (mehr dazu am Ende dieses Abschnitts). Neben der Atmung lassen sich noch zwei bzw. — je nach Sichtweise — drei weitere Komponenten der Sprachproduktion unterscheiden: die Stimmgebung (Phonation), die Artikulation und die Resonanz, die manchmal auch unter die Artikulation subsumiert wird. Diese Komponenten können im funktionalen Modell des Sprechapparates zusammengefasst werden (Abbildung 1.1). Die folgenden Abschnitte behandeln die einzelnen Komponenten der lautsprachlichen Produktion genauer.

Komponenten der Sprachproduktion

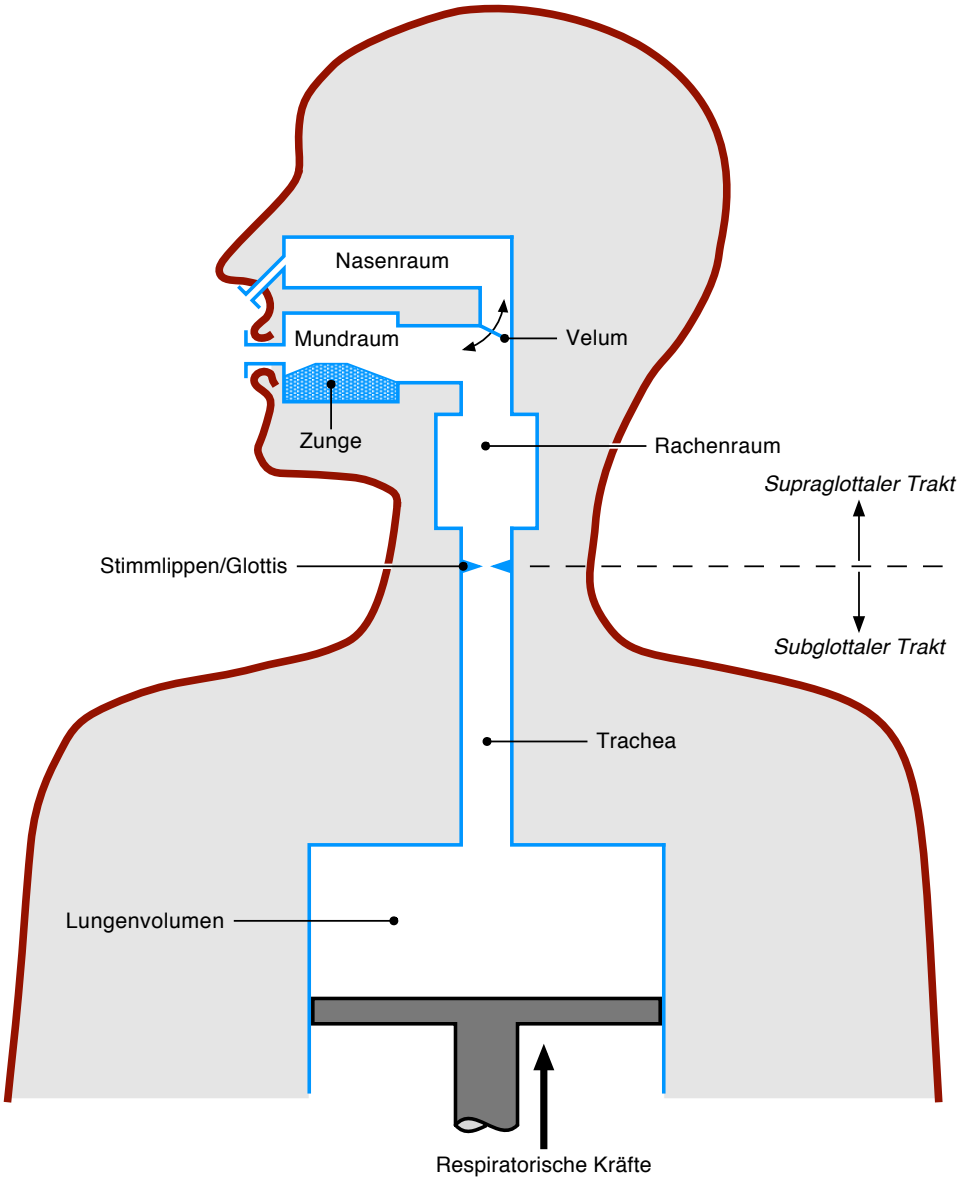


Abbildung 1.1: *Der Sprechapparat als funktionales Modell der lautsprachlichen Produktion: Die respiratorischen Käfte erzeugen einen pulmonalen Luftstrom. Stimmhafte Laute entstehen durch die periodische Unterbrechung des Luftstroms an der Glottis (Phonation). Durch die Veränderung von Form und Größe des Mundraums können unterschiedliche phonetische Lautqualitäten erzeugt werden (Artikulation). Das Absenken des Velums hat eine zusätzliche nasale Resonanzkomponente zur Folge.*

1.1.1 Atmung

Die Atmung vollzieht sich, indem der Brustkorb (Thorax) ausgedehnt (Einatmung) bzw. verengt wird (Ausatmung). Durch die Erweiterung des Brustkorbs entsteht in den beiden Lungenflügeln ein Unterdruck, infolge dessen die Luft passiv durch die Luftröhre (Trachea) in die Lungen einströmt. Verantwortlich für die Erweiterung des Brustkorbs während der Ruheatmung sind im wesentlichen die exterioren intercostalen Muskeln[1] und das Zwerchfell ("inspiratorische Muskeln"). Bei besonders tiefem Einatmen sind zusätzlich weitere Muskelgruppen im Brust–, Hals–, Schulter– und Rückenbereich beteiligt. Die Ausatmung, d.h. die Verengung des Brustraums, ist dagegen ein weitgehend passiver Prozess. Aufgrund verschiedener Rückstellkräfte (z.B. zieht die Schwerkraft die angehobenen Rippen nach unten; die elastischen Lungen, die mit Muskelkraft erweitert wurden, ziehen sich passiv wieder zusammen) verkleinert sich der Brustraum auf seine ursprüngliche Größe und die Luft wird aus den Lungen gepresst. Dies gilt zumindest für die Ruheatmung; bei forcierter Atmung kann auch die Ausatmung muskulär unterstützt werden (durch abdominale und interiore intercostale Muskeln, die sog. "expiratorischen Muskeln"). Der zeitliche Anteil des Einatmens bei einem Ruheatmungszyklus beträgt etwa 40%, der Anteil der Ausatmung entsprechend etwa 60% (vgl. Abb. 1.2, oben).

Ruheatmung

Dieses Verhältnis kann sich bei der dem Sprechen angepassten Atmung, der sog. Sprechatmung, sehr stark verändern: Die Ausatmung kann hier bis zu 90% eines Atemzyklus beanspruchen. Um einen gleichbleibenden subglottalen Luftdruck zu gewährleisten, unterliegt die Ausatmung bei der Sprechatmung einer komplexen muskulären Kontrolle. In einer ersten Phase werden Muskelgruppen aktiv, die den natürlichen Rückstellkräften entgegen wirken, um ein zu schnelles Entweichen der Luft zu verhindern (die sog. "inspiratorischen Muskeln"). In einer zweiten Phase werden andere Muskeln aktiviert, die eine zusätzliche Kompression des Brustkorbs bewirken, um so den entweichenden Luftstrom länger aufrecht zu erhalten ("expiratorische Muskeln") (vgl. Abb. 1.2, unten).

Sprechatmung

subglottaler Luftdruck

1.1.2 Phonation

Für die Phonation, d.h. die Erzeugung von Stimme, ist der Kehlkopf (Larynx) verantwortlich. Genauso wie auch die Atmung primär die Funktion hat, den Organismus mit Sauerstoff zu versorgen, und uns sozusagen nur sekundär das

Kehlkopf/Larynx

[1] Intercostal: zwischen den Rippen; exterior: zur Körperoberfläche hin.

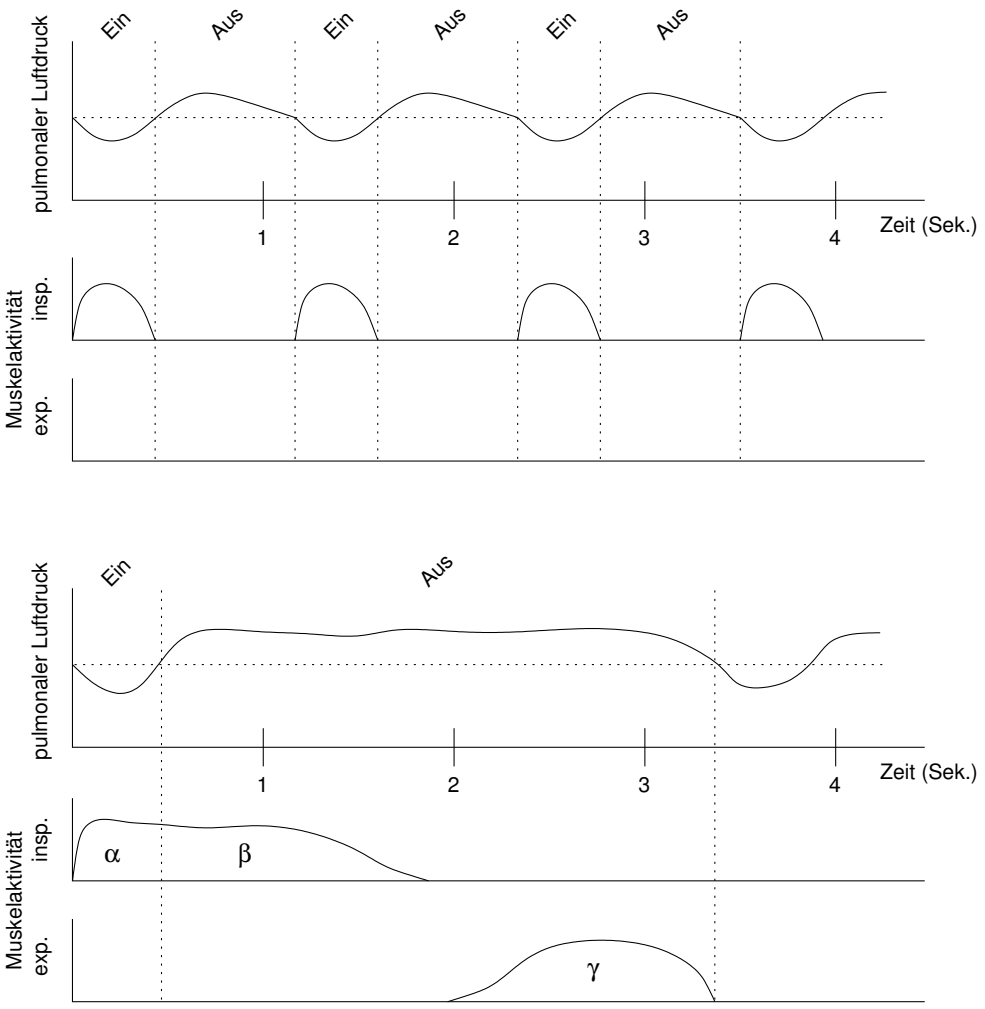


Abbildung 1.2: *Oben: Atemzyklen bei Ruheatmung; Muskelaktivität nur während der Inspirationsphase. Unten: Sprechatmung; die Expirationsphase ist stark gelängt; Aktivität der "inspiratorischen" Muskeln zur Erweiterung des Brustkorbes während des Einatmens* (α) *und als Gegenkraft zu den natürlichen Rückstellkräften während der kontrollierten, verzögerten Ausatmung* (β)*; Aktivität der "expiratorischen" Muskeln zur verlängerten Aufrechterhaltung des subglottalen Luftdrucks* (γ).

Sprechen ermöglicht, ist auch die sprechspezifische, phonatorische Funktion des Larynx 'nur' sekundär; primär dient der Kehlkopf, der den oberen Abschluss der Luftröhre bildet, als Ventil, das verhindern soll, dass z.B. bei der Nahrungsaufnahme feste oder flüssige Substanzen in die Lunge gelangen.

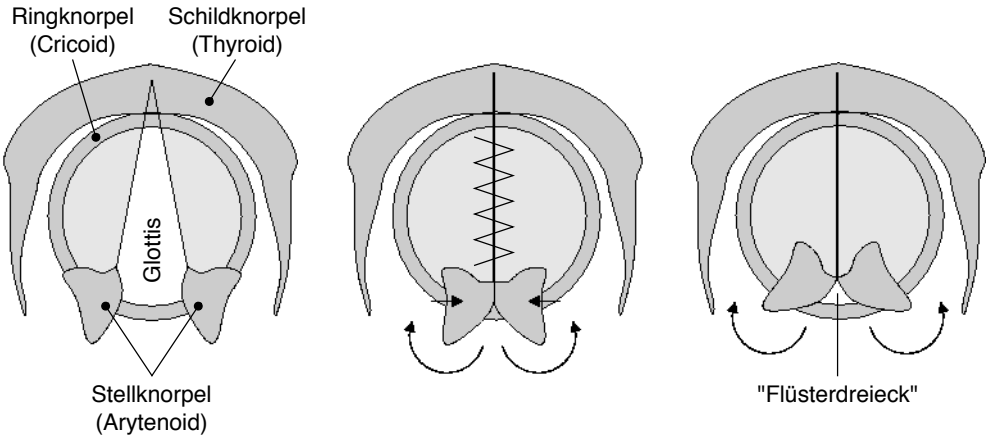


Abbildung 1.3: *Schematische Darstellung unterschiedlicher laryngaler Konfigurationen (Draufsicht; oben ist vorne, unten ist hinten): Produktion stimmloser Laute (links), Phonation (mitte) und Flüstern (rechts).*

Der Kehlkopf besteht aus gelenkig miteinander verbundenen Knorpelstrukturen, Muskeln und Bändern sowie Schleimhäuten. Den unteren Abschluss des Kehlkopfs am Übergang zur Trachea bildet der Ringknorpel (Cricoid). Darüber sitzt der Schildknorpel (Thyroid), dessen zwei seitliche Platten vorne miteinander verbunden sind (”Adamsapfel”). Hinten über dem Ringknorpel liegen die beiden Stellknorpel (Arytenoid). Den oberen Abschluss bildet ein deckelförmiger Knorpel, die Epiglottis. Zwischen den beiden Stellknorpeln und der vorderen Spitze des Schildknorpels spannen sich die Stimmbänder (ligamentum vocale) und die Vocalismuskeln, umgeben von Schleimhäuten und einer Membran. Bänder, Muskeln und Schleimhäute werden zusammen als Stimmlippen bezeichnet[2]. Der Spalt zwischen den Stimmlippen heißt Glottis. Durch entsprechende Konstellation der Stellknorpel kann die Glottis (ganz oder teilweise) geschlossen oder geöffnet werden (Abbildung 1.3). Zum Atmen wird die Glottis ganz geöffnet. Für die Phonation werden die Stimmlippen dagegen adduziert, die Glottis ist komplett geschlossen. Auch beim Flüstern sind die Stimmlippen adduziert, durch eine Drehung der Stellknorpel entsteht jedoch eine Öffnung (das ”Flüsterdreieck”) im hinteren Teil des Kehlkopfs, durch die Luft entweichen kann. Zur Produktion stimmloser Laute befindet sich die Glottis in der Regel in einer Mittelposition, sie ist halb geöffnet.

Stimmlippen und Glottis

[2]Der Begriff ”Stimmlippen” (engl. *vocal folds*, ”Stimmfalten”) ist dem Begriff ”Stimmbänder” vorzuziehen.

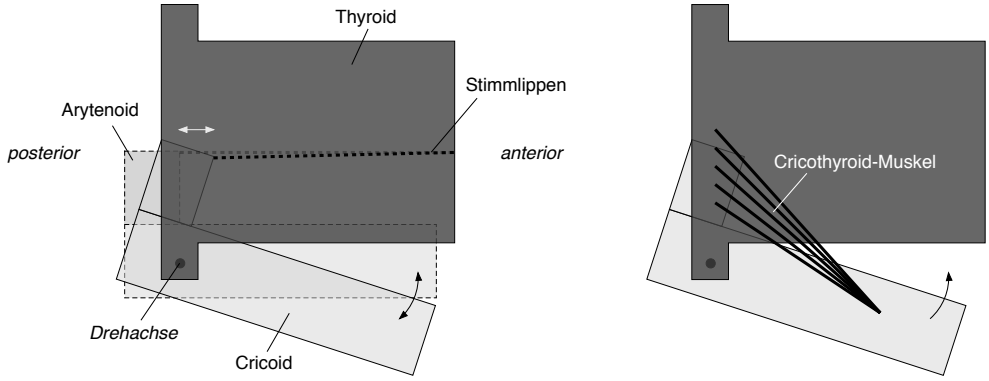


Abbildung 1.4: *Dehnung und Entspannung der Stimmlippen durch Neigung des Ringknorpels (Cricoid). Verantwortlich für die Dehnung ist der Cricothyroid–Muskel, der (beidseitig) vom posterioren (hinteren) Teil des Schildknorpels zum anterioren (vorderen) Teil des Ringknorpels verläuft.*

intrinsische Larynxmuskulatur

Die Initiation und Aufrechterhaltung verschiedener laryngaler Konfigurationen ist im wesentlichen die Aufgabe der intrinsischen Larynxmuskulatur. Diese Muskulkatur verbindet die beweglichen Teile des Kehlkopfes miteinander (im Gegensatz zur extrinsischen Larynxmuskulatur, die den Kehlkopf mit benachbarten Strukturen verbindet und stabilisiert). Diese Muskeln sorgen z.B. für die Drehung und Seitwärtsbewegung der Stellknorpel, für die innere Spannung der Stimmlippen oder für die Dehnung der Stimmlippen (u.a. durch die Neigung des Ringknorpels; vgl. Abb. 1.4).

Phonation als myoelastisch–aerodynamischer Prozess

Bei der Phonation werden die Stimmlippen in regelmäßige Schwingungen versetzt. Dieser Vorgang kann als myoelastisch–aerodynamischer Prozess[3] beschrieben werden. Zur Initiation der Phonation werden zunächst die Stimmlippen adduziert und gespannt. Danach beginnt ein zyklischer Prozess, der solange anhält, bis die Glottis wieder geöffnet wird oder keine Luft mehr in den Lungen vorhanden ist.

Der Phonationszyklus (vgl. Abbildung 1.5)

Phonationszyklus

Druckaufbau: Unterhalb der geschlossenen Glottis entsteht ein Druck auf die Stimmlippen, der subglottale Luftdruck.

[3] ”myo” bedeutet ”die Muskeln betreffend”; ”myoelastisch” bezieht sich also auf die Elastizität muskulärer Strukturen; ”aerodynamisch” deutet darauf hin, dass Luftstrommechanismen eine Rolle spielen.

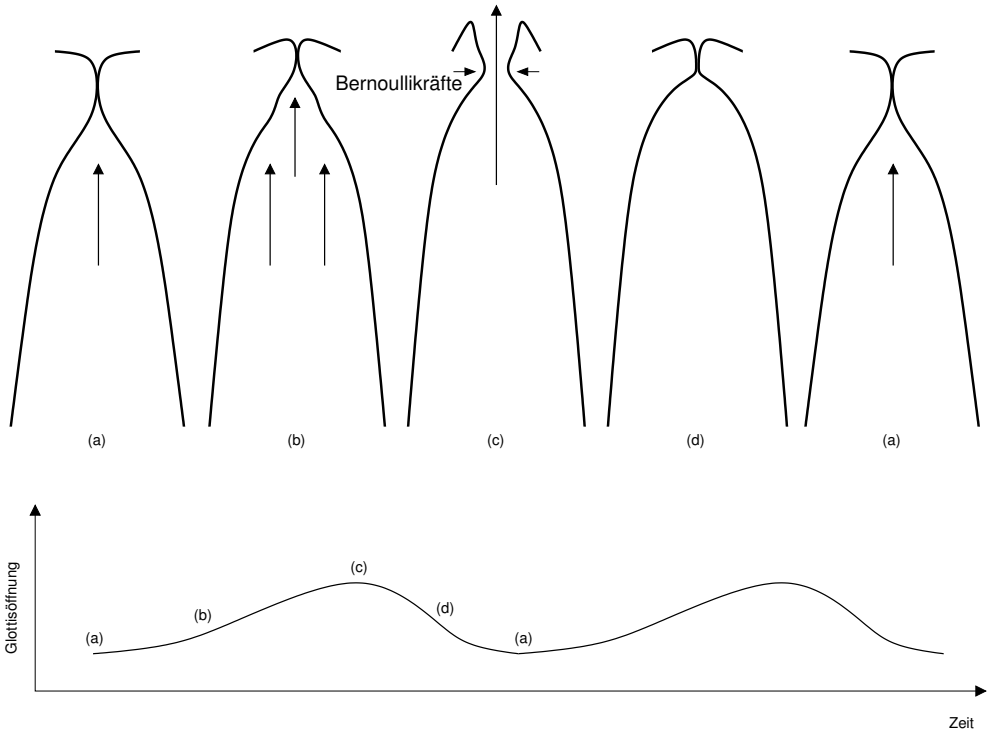


Abbildung 1.5: *Schematische Darstellung des Phonationszyklus (oben) und des resultierenden Anregungssignals (unten).*

Sprengung: Bei ausreichendem Druck werden die Stimmlippen auseinander gedrückt, die Glottis wird gesprengt.

Geöffnete Glottis: Aufgrund des in der Lunge herrschenden Überdrucks (relativ zum atmosphärischen Druck der Umgebung) strömt Luft durch die Glottis.

Bernoulli–Effekt: Da der glottale Spalt eine Verengung der Durchflussöffnung darstellt, erhöht sich an dieser Stelle die Fließgeschwindigkeit der Luft und es entsteht ein Unterdruck. Infolgedessen wirken an der Glottis die sog. Bernoulli–Kräfte senkrecht zur Fließrichtung und die elastischen Stimmlippen bilden erneut einen Verschluss. Der Zyklus beginnt von vorn.

Bernoulli–Effekt

Die Phonation ist das Resultat der komplexen Interaktion von aerodynamischen und aerostatischen Kräften sowie Muskel– bzw. Gewebekräften. Die aerostatischen Kräfte enstehen dadurch, dass der pulmonale Luftdruck gegen

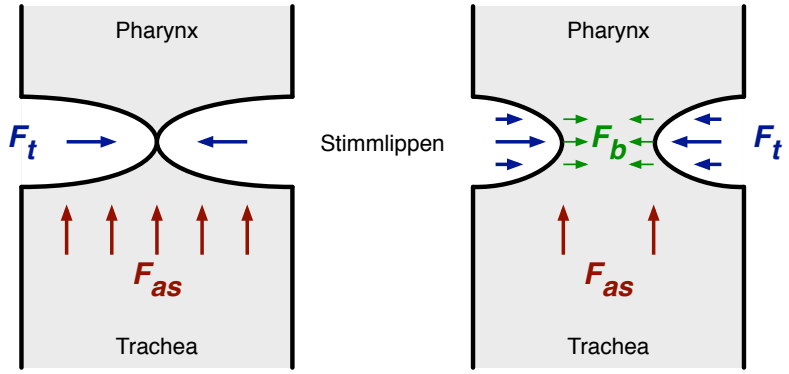


Abbildung 1.6: *Schematische Darstellung der Glottispassage während der Phonation, links mit geschlossener, rechts mit geöffneter Glottis. Aerostatische Kräfte (F_{as}) wirken öffnend, Gewebekräfte (F_t) und Bernoulli–Kräfte (F_b) wirken schließend.*

die untere Oberfläche der adduzierten Stimmlippen drückt; sie wirken öffnend. Die Gewebekräfte sind eine Funktion der Elastizität des Stimmlippengewebes; sie wirken schließend und zwar unterschiedlich stark, abhängig vom Maß der Elastizität. Die aerodynamischen Bernoulli–Kräfte schließlich sind eine Folge des Unterdrucks zwischen den göffneten Stimmlippen, sie wirken ebenfalls schließend (Abb. 1.6). Abbildung 1.7 verdeutlicht die Interaktion und koordinierte Veränderung dieser Kräfte während der Phonation.

Das Zusammenspiel der phonatorischen Kräfte reagiert sehr empfindlich auf kleinste Veränderungen. Solche Veränderungen können willkürlich oder unwillkürlich auftreten und betreffen insbesondere die Gewebekräfte; aber natürlich können über die Variation des pulmonalen Luftdrucks auch die aerostatischen Kräfte verändert werden. Durch die willkürliche Veränderung der Elastizität der Stimmlippen während des Sprechens wird z.B. der sprechmelodische Verlauf (Sprach– bzw. Satzmelodie) einer Äußerung gesteuert. Spannung und Dehnung der Stimmlippen führt zu verringerter Elastizität, die Gewebekräfte nehmen zu, der Glottisverschluss erfolgt schneller, wodurch der Phonationszyklus beschleunigt und ein höherer Stimmton produziert wird. Umgekehrt führt die Entspannung der Stimmlippen letzlich zu einem tieferen Stimmton. Unwillkürliche Veränderungen ergeben sich z.B. infolge des Wachstums bei Kindern und Jugendlichen oder infolge von Kehlkopfentzündungen, die die Viskosität der die Stimmlippen umschließenden Schleimhäute verändern können, was erheblichen Einfluss auf die Elastizität und das Schwingungsverhalten der Stimmlippen hat.

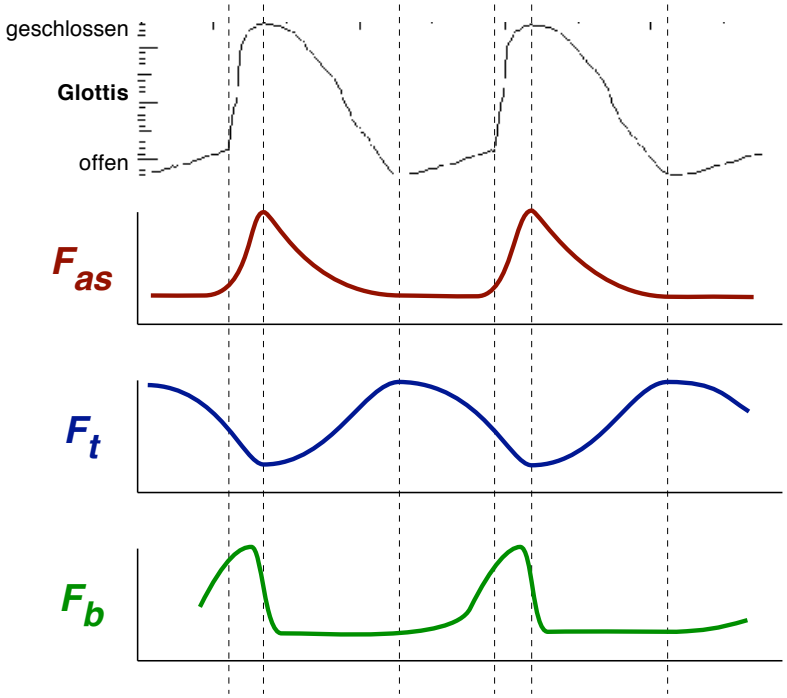


Abbildung 1.7: *Die oberste Kuve zeigt den Status der Glottis (offen/geschlossen) über der Zeit. Die übrigen Kurven zeigen zeitlich koordiniert den schematischen Verlauf der phonatorischen Kräfte. Ist die Glottis vollständig geschlossen, ist F_{as} maximal, F_t und F_b sind gering. Ist die Glottis vollständig geöffnet, ist F_t maximal, F_{as} und F_b sind gering. Nähern sich die Stimmlippen an, wird F_t geringer, F_b steigt an bis zum Maximum kurz vor dem erneuten Verschluss.*

Anregungssignal

Das akustische Resultat der Stimmlippenschwingungen ist das sog. Anregungssignal, das wir leider nicht direkt mit einem Mikrophon messen können, da es auf seinem Weg durch den Vokaltrakt sehr stark verändert wird.

Frequenz und Tonhöhe

Die Geschwindigkeit, mit der der Phonationszyklus abläuft, d.h. die Frequenz der Stimmlippenschwingung bzw. des Anregungssignals, korreliert mit der wahrgenommenen Tonhöhe. Ein wichtiger Faktor für die Höhe des Stimmtons ist die natürliche Länge der Stimmlippen: Kurze Stimmlippen schwingen schneller als lange (bei identischer Steifheit), und schnellere Schwingungen führen zu einem höheren Stimmton. Die Stimmlippen von Männern sind ungefähr 17 bis 24 mm lang, die von Frauen etwa 13 bis 17 mm. Daher ist die männliche Stimme tiefer (ca. 120 Hz) als die weibliche (ca. 230 Hz). Die Stimmlippen von Säuglingen sind ungefähr 5 mm lang, ihre Stimmlage liegt bei etwa 400 Hz. Neben diesem Faktor, der sich unserem

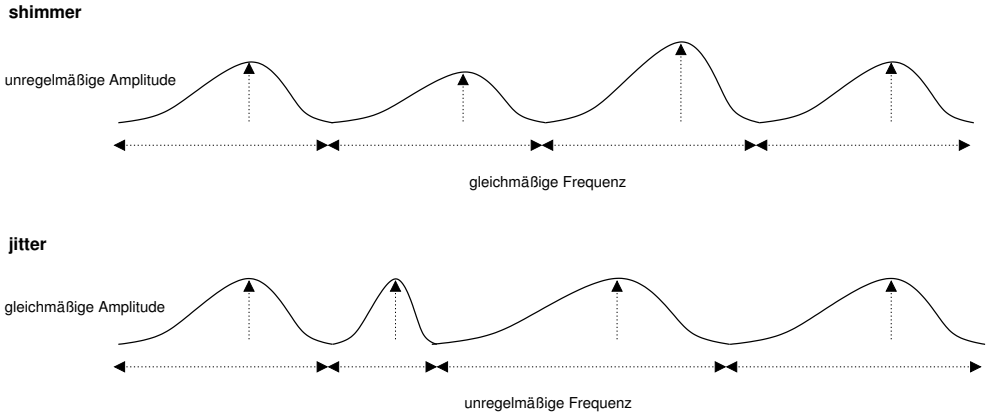


Abbildung 1.8: *Irreguläre Stimmlippenschwingungen; schematisch dargestellt ist das resultierende Anregungssignal bei shimmer (oben) und jitter (unten).*

Einfluss entzieht, gibt es jedoch einige weitere Faktoren, die die Höhe des Stimmtons beeinflussen und die sich gezielt steuern lassen. Dazu zählen z.B. die Steifheit der Stimmlippen, die Stärke der medialen Kompression (Gegeneinanderdrücken der Stimmlippen; dadurch kann die Länge des schwingungsfähigen Teils der Stimmlippen verändert werden), sowie die Stärke des subglottalen Luftdrucks. Auch die Masse des schwingenden Teils der Stimmlippen spielt eine Rolle. Generell gilt, der Stimmton wird höher

mediale Kompression

- je steifer die Stimmlippen
- je kürzer der schwingungsfähige Teil der Stimmlippen
- je stärker der subglottale Luftdruck
- je dünner (masseärmer) die Stimmlippen

Neben der Höhe des Stimmtons kann durch die Steuerung der laryngalen Konfiguration auch die Lautstärke des Stimmtons sowie die Stimmqualität beeinflusst werden. Die Lautstärke des Stimmtons hängt — neben einem erhöhten subglottalen Druck — vor allem davon ab, wie abrupt der transglottale Luftstrom durch den glottalen Verschluss abgeschnitten wird. Das glottale Schließverhalten hängt wiederum von der Steifheit und der geometrischen Konfiguration der Stimmlippen ab, von Parametern also, die wir 'bewusst' steuern können. Je lauter die Stimmgebung, desto abrupter wird der transglottale Luftstrom unterbrochen. In der Akustik zeigt sich dies in schärferen und stärkeren Impulsen im Anregungssignal.

Stimmqualität

Die Stimmqualität hängt z.B. davon ab, ob die Glottis bei der Phonation komplett geschlossen ist, oder ob die Stimmlippen ausreichend steif sind,

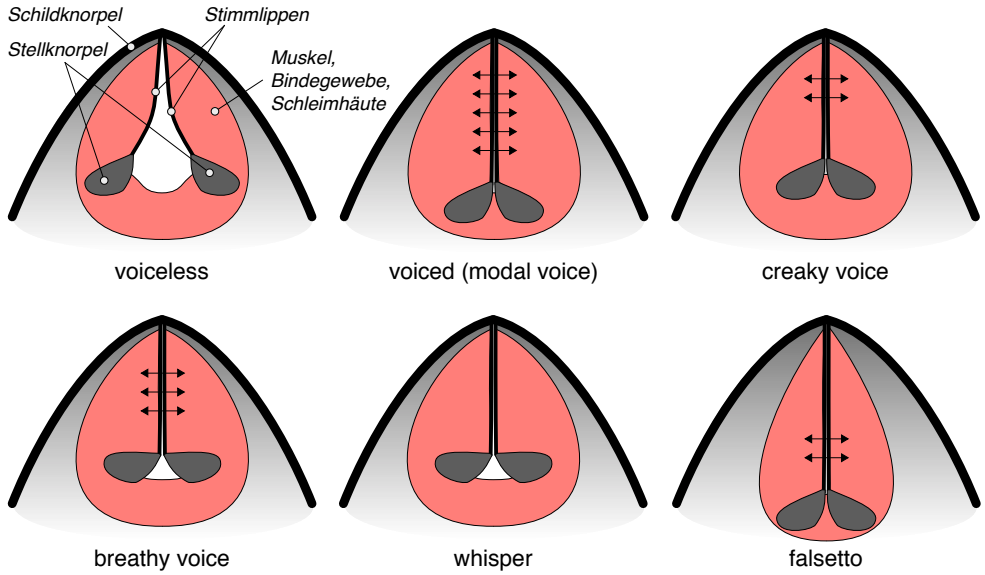


Abbildung 1.9: *Die wichtigsten Phonationsmodi mit schematischer Darstellung der zugrundeliegenden laryngalen Konfigurationen. Die waagerechten Doppelpfeile zeigen Stimmlippenschwingungen an.*

um dem subglottalen Druck genügend Widerstand entgegenzusetzen. Sowohl ein unvollständiger Verschluss als auch eine ungenügende Steifheit führen zu behauchter Stimme. Starke Unregelmäßigkeiten der Stimmlippenschwingungen, sowohl in der Amplitude (engl. *shimmer*) als auch in der Frequenz (engl. *jitter*), führen zum Eindruck der rauhen Stimme (vgl. Abbildung 1.8).

Für die sprachsystematische (phonologische) Beschreibung der Sprachen der Welt werden zumindest 5 verschiedene Phonationsmodi unterschieden: Stimmlos (*voiceless*), stimmhaft (*voiced*), behauchte Stimme (*breathy voice*), Flüsterstimme (*whisper*) und Knarrstimme (*creaky voice*). In sehr vielen (aber nicht in allen) Sprachen gibt es eine systematische Unterscheidung zwischen stimmhaften und stimmlosen Lauten. Behauchung, Flüster– und Knarrstimme bilden in einigen wenigen Sprachen einen phonologischen Kontrast mit der modalen Stimmhaftigkeit (z.B. gibt es im Hindi einen phonologischen Kontrast zwischen stimmhaft und stimmhaft/behaucht). Abbildung 1.9 fasst die Phonationsmodi und die dazugehörigen laryngalen Konfigurationen zusammen, ergänzt um das Gesangsregister *Falsett* (oder *Kopfstimme*), bei der u.a. die Stimmlippen stark gespannt werden, so dass nur ein relativ kleiner Teil davon schwingt, dieser dafür sehr schnell.

Phonationsmodus

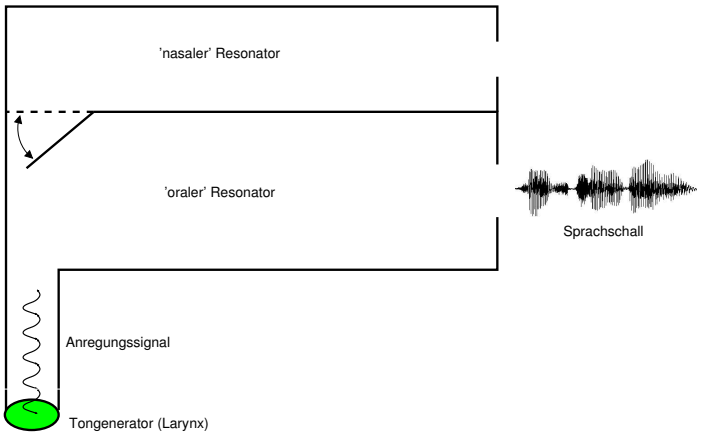


Abbildung 1.10: *Das Ansatzrohr.*

1.1.3 Resonanz

Ansatzrohr

Ein kurzer Ausflug in die akustische Phonetik: Wie bereits erwähnt, wird das bei der Phonation erzeugte Anregungssignal auf dem Weg durch den Vokaltrakt stark verändert. Der Vokaltrakt wirkt dabei als eine Art Filter, der — je nach Konfiguration (z.B. Zungenposition) — bestimmte Frequenzen des Anregungssignals verstärkt oder dämpft. Diese Konstellation kann mit dem aus der Instrumentenkunde entlehnten Begriff des Ansatzrohres beschrieben werden: Ein Primärschall (das Anregungssignal) wird durch einen Resonator (den Vokaltrakt) geleitet und verlässt diesen mit einer spezifischen Klangqualität. Dieses Modell ist stark vereinfacht in Abbildung 1.10 dargestellt.

Resonanzräume

Velum

Wie aus der schematischen Darstellung deutlich wird, verfügen wir über zwei Resonatoren, den konstanten oralen Resonator (Mundraum) und den zuschaltbaren nasalen Resonator (Nasenraum) (vgl. Abbildung 1.11). Zwei Gründe sprechen dafür, die Resonanzkomponente als unabhängige Komponente innerhalb des Sprachproduktionsprozesses zu betrachten (anstatt sie unter die artikulatorische Komponente zu subsumieren; s.o.): Erstens ist es von erheblichem Einfluss auf die Klangqualität aller stimmhaften Laute, ob der nasale Resonator zugeschaltet ist oder nicht. Zweitens kann die nasale Resonanzkomponente relativ unabhängig von artikulatorischen Konfigurationen gesteuert werden. Verantwortlich hierfür ist der weiche Gaumen (Velum).

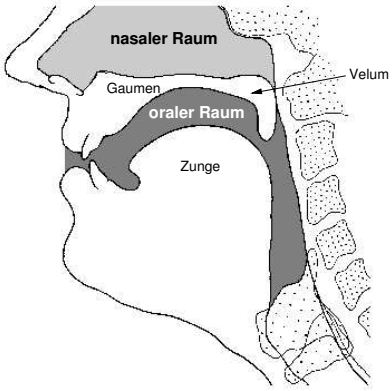


Abbildung 1.11: *Die supraglottalen Resonanzräume.*

1.1.4 Artikulation

Der Begriff ”Artikulation” wird manchmal in einem sehr weiten Sinne verwendet, nämlich als Bezeichnung für den gesamten lautsprachlichen Produktionsprozess (z.B. auch dann, wenn wir diese Teildisziplin der Phonetik als ”artikulatorische Phonetik” bezeichnen). Artikulation im engeren Sinne meint jedoch nur eine bestimmte Komponente im Produktionsprozess: Die Variation des Vokaltrakts während des Sprechens. Die Variationsmöglichkeiten des Vokaltrakts verfügen über einen räumlichen (Artikulationsort oder –stelle) und über einen modalen Aspekt (Artikulationsmodus oder –art). Der räumliche Aspekt kann beschrieben werden als Positionsveränderung der beweglichen Teile des Vokaltrakts in Bezug auf die anatomischen Fixpunkte. Die beweglichen Teile heißen Artikulatoren. Hierzu zählen (vgl. Abbildung 1.12):

Artikulationsort und Artikulationsart

Artikulatoren

- der Unterkiefer (Mandibulum)
- die Lippen (Labia)
- die Zunge (Lingua)
  - Zungenspitze (Apix)
  - Zungenblatt (Lamina)
    (Laute, die mit der Zungenspitze oder mit dem Zungenblatt gebildet werden, bezeichnet man auch als koronale Laute)
  - Zungenrücken (Dorsum)
  - Zungenwurzel (Radix)
- der weiche Gaumen/das Gaumensegel (Velum)

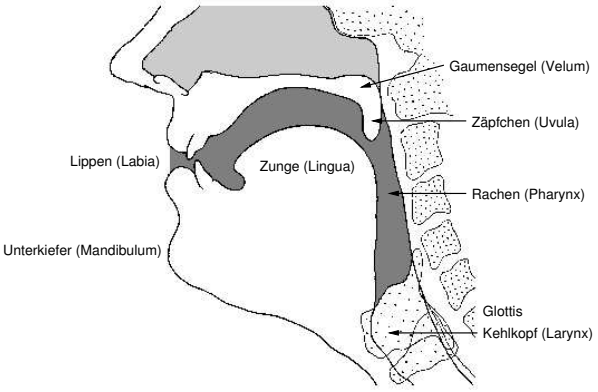


Abbildung 1.12: *Die Artikulatoren.*

- das Zäpfchen (Uvula)

und mit Einschränkung:

- der Rachen (Pharynx)
- der Kehlkopf (Larynx) mit Glottis

Die einzelnen Artikulatoren unterscheiden sich aufgrund der anatomischen Gegebenheiten in ihrer Beweglichkeit. Dies betrifft sowohl die Bewegungsgeschwindigkeit als auch die Möglichkeiten der Formveränderung. Abgesehen von der Glottis, deren sehr schnelle Bewegungen auf einem anderen Mechanismus beruhen (myoelastisch–aerodynamisch, s.o.), ist die Apix (Zungenspitze) zu den schnellsten rein muskulär gesteuerten Bewegungen fähig. Über die geringste Beweglichkeit unter den Artikulatoren verfügt der Pharynx. Die Bewegungsparameter der einzelnen Artikulatoren:

Bewegungsparameter der Artikulatoren

Unterkiefer: horizontal (nach vorne, nach hinten), vertikal (nach oben, nach unten)
Lippen: verschließen, runden/spreizen
Zungenkörper: horizontal, vertikal, konvex/konkav, spitz/breit
Apix, Lamina: horizontal, vertikal, flach/gefurcht
Velum: vertikal
Pharynx: verengen, versteifen

Im folgenden die anatomischen Fixpunkte des Mundraums (Artikulationsorte, vgl. Abbildung 1.13); in Klammer jeweils das gebräuchliche Adjektiv, mit dem Laute, die an der entsprechenden Stelle gebildet werden, bezeichnet werden:

Artikulationsorte

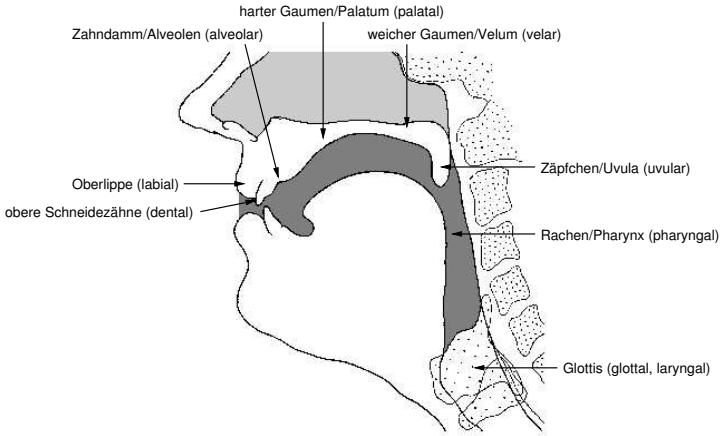


Abbildung 1.13: *Die Artikulationsorte.*

- Oberlippe (labial)
- obere Schneidezähne (dental)
- Zahndamm/Alveolen (alveolar)
- zwischen Zahndamm und hartem Gaumen (post–alveolar/retroflex[4])
- harter Gaumen/Palatum (palatal)
- weicher Gaumen/Velum (velar)
- Zäpfchen/Uvula (uvular)
- Rachenwand/Pharynx (pharyngal)
- Epiglottis (epiglottal)
- Glottis (glottal, laryngal)

vokalischer und konsonantischer Artikulationsmodus

Bei den Artikulationsmodi lassen sich zunächst zwei Grundkonstellationen unterscheiden: Der vokalische Modus und der konsonantische Modus. Der vokalische Modus ist insbesondere dadurch gekennzeichnet, dass die Luft den Vokaltrakt ungehindert passieren kann. Das Anregungssignal wird ausschließlich durch globale Veränderungen des Ansatzrohres moduliert. Diese können die Länge des Ansatzrohres — z.B. durch Vorstülpen der Lippen —

[4]Laute die mit zurückgebogenem Zungenblatt artikuliert werden, heißen retroflexe Laute. Die Bezeichnung ”retroflex” wird häufig unter die Artikulationsstellen subsumiert (z.B. im *internationalen phonetischen Alphabet*; s.u.), obwohl es sich dabei nicht um einen anatomischen Fixpunkt handelt. Die Artikulationsstelle dieser Laute ist eigentlich am Übergang der Alveolen zum harten Gaumen, also post–alveolar. Um jedoch die retroflexen Laute von den ’normal’ (d.h. mit flachem Zungenblatt) gebildeten post–alveolaren Lauten zu unterscheiden, hat sich die Artikulationsstellenbezeichnung ”retroflex” durchgesetzt.

oder dessen Querschnitt betreffen — z.B. durch Absenken bzw. Anheben des Kiefers/der Zunge oder durch Vor– bzw. Zurückbewegen der Zunge.

Beim konsonantischen Modus kommt es dagegen stets zu einer lokalen Enge– bzw. Verschlussbildung im Ansatzrohr, wodurch der Luftstrom durch den Vokaltrakt behindert bzw. blockiert wird. Je nach Grad, Dauer oder Form der Engebildung werden die folgenden Lautklassen unterschieden (bei allen Klassen, außer bei den Nasalen, ist der nasale Raum geschlossen, d.h. es kann keine Luft durch die Nase entweichen):

konsonstische Lautklassen

Plosive (Verschlusslaute): kompletter oraler (und velarer) Verschluss.

Nasale: kompletter oraler Verschluss, das Velum ist abgesenkt (d.h. der Luftstrom wird im Mundraum blockiert, kann jedoch durch die Nase entweichen).

Stops: Im Englischen werden Plosive und Nasale beide als *stops* bezeichnet; Plosive sind demnach *oral stops*, Nasale *nasal stops*.

Vibranten (gerollte Laute, Trills): intermittierende orale Verschlüsse (2-3 in fließender Rede); dieser Artikulationsmodus beruht auf demselben Mechanismus wie die Phonation (myoelastisch–aerodynamisch).

Geschlagene Laute (Taps/Flaps): extrem kurzer oraler Verschluss.

Frikative: starke zentrale Enge; durch die starke Verengung kommt es zur Geräuschbildung infolge von Turbulenzen.

Laterale Frikative: zentraler oraler Verschluss, starke seitliche Enge mit Geräuschbildung.

Approximanten: schwache zentrale Enge ohne Geräuschbildung; da der Luftstrom die Verengung nahezu ungehindert passieren kann, werden Approximanten auch als ”Halbvokale” oder ”Vokoide” bezeichnet.

Laterale Approximanten: zentraler oraler Verschluss, schwache seitliche Enge ohne Geräuschbildung.

Affrikaten: Affriktion ist im strengen Sinne kein eigener Artikulationsmodus, sondern eine Kombination aus Plosiv und homorganem (d.h. an etwa derselben Stelle gebildetem) Frikativ.

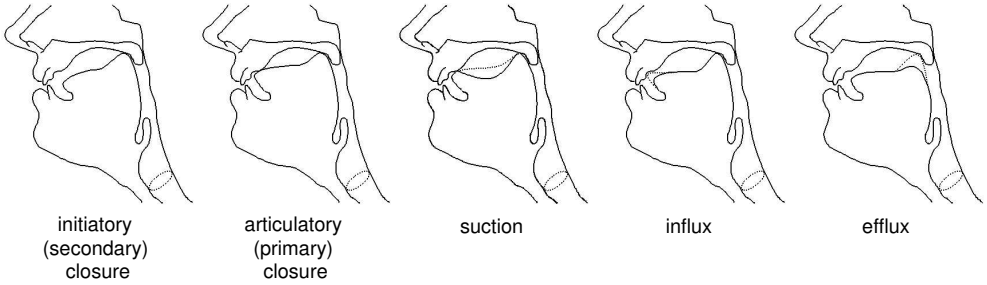


Abbildung 1.14: *Die Artikulationsphasen bei der Produktion von Clicks.*

Obstruenten, Sonoranten, Liquide

Plosive, Frikative und Affrikaten werden häufig unter dem Begriff Obstruenten zusammengefasst; Vibranten und Approximanten bezeichnet man als Liquide. Als Sonoranten bezeichnet man alle Laute außer den Obstruenten (also Liquide, Nasale und Vokale).

Die bisher aufgeführten Konsonantenklassen werden mit dem pulmonalen Luftstrommechanismus gebildet. Daneben gibt es noch drei Klassen von nicht-pulmonalen Konsonanten:

nicht-pulmonalen Konsonanten

Clicks (Schnalzlaute) — velar ingressiv: kompletter oraler Verschluss im vorderen Mundraum (primär/artikulatorisch) plus velarer Verschluss (sekundär/initiatorisch); zwischen primärem und sekundärem Verschluss wird die Zunge abgesenkt, wodurch ein Unterdruck entsteht. Bei Lösung des primären Verschlusses (Influx) entsteht ein Schnalzlaut (vgl. Abb. 1.14).

Implosive — glottal ingressiv: kompletter oraler Verschluss; durch schnelles Absenken des Larynx (bei geschlossener Glottis) entsteht ein Unterdruck, der Verschluss wird nach innen gesprengt.

Ejektive — glottal egressiv: kompletter oraler Verschluss; durch schnelles Anheben des Larynx (bei geschlossener Glottis) entsteht ein Überdruck, der Verschluss wird nach außen gesprengt.

1.2 Lautschriftsysteme

1.2.1 Das internationale phonetische Alphabet (IPA)

Das internationale phonetische Alphabet ist ein an artikulatorischen Merkmalen orientiertes System zur symbolischen Repräsentation aller Laute, die in

den Sprachen der Welt vorkommen[5]. Es wurde in erster Linie zu praktischen Zwecken entwickelt (nicht als 'theoretisches Modell'), z.B. für den Fremdsprachenunterricht oder zur Verschriftung von Sprachen. Prinzipiell sollte für jeden vorkommenden Laut ein Symbol vorhanden sein. Wo dies nicht sinnvoll erscheint, werden Diakritika ("Zusatzzeichen") verwendet. So sind z.B. nasalierte Vokale wie sie im Französischen vorkommen nicht durch eigene Symbole repräsentiert; stattdessen wird das jeweilige Symbol für den nichtnasalierten Vokal mit dem Diakritikum für Nasaliertheit kombiniert: /a/ vs. /ã/.

Diakritika

Diakritika können jedoch auch dazu verwendet werden, Aussprachevarianten detailliert zu beschreiben. Ein Beispiel: Im Deutschen sind Vokale normalerweise nicht nasaliert. Wenn ein Vokal jedoch von zwei nasalen Konsonanten umgeben ist wie das /a/ im Wort *Mama*, kann es vorkommen, dass der Vokal durchgehend nasaliert wird (das Velum verbleibt während der Vokalartikulation in der abgesenkten Position). Tut ein Sprecher dies und will man solche Feinheiten beschreiben, könnte man im vorliegenden Fall [mãma] transkribieren. Eine solche Transkription steht nun nicht mehr zwischen Schrägstrichen, sondern zwischen eckigen Klammern. Der Grund dafür ist, dass es sich hier um eine phonetische Transkription handelt, d.h. um die symbolische Repräsentation einer 'tatsächlichen' Äußerung (wir hatten ja angenommen, dass ein Sprecher das Wort *Mama* tatsächlich auf diese Art realisiert). Im Gegensatz hierzu handelt es sich bei Transkriptionen zwischen Schrägstrichen um phonematische Transkriptionen. Phonematische Transkriptionen repräsentieren die Lautstruktur eines Wortes gemäß den phonologischen Gesetzmäßigkeiten einer bestimmten Sprache. Betrachten wir das Wort *Hund*. Im Deutschen gibt es einen phonologischen Prozess — die sog. Auslautverhärtung — der stimmhafte Plosive am Wortende in stimmlose 'umwandelt'. Die phonematische Transkription lautet entsprechend /hʊnt/, nicht /hʊnd/[6].

phonetische vs. phonematische Transkription

Nun zur Unterscheidung zwischen phonematischer und phonetischer Transkription. Die phonematische Transkription von *Pendler* könnte in etwas so aussehen: /pɛndlɐʀ/. Die meisten Sprecher des Deutschen werden jedoch die Endung *-er* in der normalen Umgangssprache nicht wie in der phonematischen Transkription realisieren (e-Schwa plus r-Laut), sondern als soge-

[5]Die IPA–Symbole gibt es natürlich auch für den Computer, als Truetype–Font (`www.sil.org/computing/catalog/encore_ipa.html`) und für TEX (`www.ctan.org/tex-archive/fonts/tipa/`).

[6]Zur genauen Bedeutung der Symbole, speziell zur Bedeutung des Symbols für den u-Laut, siehe weiter unten.

nannten ”a–Schwa”. Die phonetische Transkription einer solchen Äußerung müsste also so aussehen: [pɛndlɐ].

Phonetische Transkriptionen können unterschiedlich stark ins Detail gehen. Eine sehr detaillierte Transkription, die möglichst viele Nuancen einer Äußerung festhält (und entsprechend ausgiebig von Diakritika Gebrauch macht), nennt man enge phonetische Transkription; eine weniger detaillierte Transkription heißt weite oder breite phonetische Transkription; der Übergang ist fließend. Im obigen Beispiel (*Pendler*) werden die meisten Sprecher den alveolaren Verschluss des /d/ nicht wie üblich zentral lösen, sondern seitlich — ein koartikulatorischer Effekt aufgrund des nachfolgenden lateralen Approximanten /l/. Eine solche Äußerung könnte folgendermaßen mit dem Diakritikum für laterale Verschlusslösung transkribiert werden: [pɛndˡɐ]. Ob eine eher enge oder eine eher weite Transkription gewählt wird, hängt immer davon ab, zu welchem Zweck eine Transkription angefertigt wird.

enge vs. weite phonetische Transkription

Die Systematik des IPA

Das IPA ist in 6 Bereiche eingeteilt:

1. Pulmonale Konsonanten
2. Nicht-Pulmonale Konsonanten
3. Sonstige Konsonanten
4. Vokale
5. Diakritika
6. Suprasegmentalia (zur Transkription prosodischer Merkmale)

Die Systematik der pulmonalen Konsonanten ist folgendermaßen aufgebaut: Von links nach rechts stehen die Artikulationsorte (und die Artikulationsorgane); es beginnt links mit der vordersten Artikulationsstelle (bilabial) und endet rechts mit der hintersten (glottal). Von oben nach unten sind die Laute nach Artikulationsmodus sortiert; es beginnt oben mit dem Modus der stärksten Engebildung (Verschlusslaute) und endet mit dem Modus der geringsten Verengung (Approximanten). Innerhalb der Tabellenfelder stehen (sofern vorhanden) jeweils links die stimmlosen Varianten, rechts die stimmhaften.[7] Leere Tabellenfelder stehen für Laute, deren Artikulation zwar prinzipiell möglich ist, die jedoch in keiner der bisher bekannten Sprachen der Welt verwendet werden (z.B. der labio-dentale Plosivlaut). Schraffierte Felder kennzeich-

[7] Stimmlose Konsonanten werden manchmal auch ”Fortis-Laute” genannt, stimmhafte Konsonanten ”Lenis-Laute”.

nen dagegen unmögliche Artikulationen.[8] Zur Identifikation eines pulmonalen Konsonanten reicht es in der Regel aus, den Artikulationsort, die Stimmbeteiligung und den Artikulationsmodus (in dieser Reihenfolge) zu benennen: /t/ ist ein alveolarer stimmloser Plosiv, /v/ ist ein labio-dentaler stimmhafter Frikativ etc. Bei detaillierterer Betrachtung können auch das Artikulationsorgan sowie weitere Lauteigenschaften mit angegeben werden (z.B. wird das /t/ im Deutschen normalerweise als apiko-alveolarer, stimmloser, aspirierter Plosiv realisiert).

Die Vokale werden zunächst nach horizontaler und vertikaler Zungenposition differenziert. Im Vokalviereck steht die obere linke Ecke für hohe vordere Vokale (/i/), die rechte obere Ecke für hohe hintere (/u/), die untere linke Ecke für tiefe vordere (/a/) und die untere rechte Ecke für tiefe hintere Vokale (/ɑ/).[9] Diese vier Vokale, die die Extrempositionen innerhalb des Vokalvierecks einnehmen, heißen Kardinalvokale. Außer nach Zungenlage und Zungenhöhe können Vokale noch nach der Lippenrundung unterschieden werden. Bei Symbolpaaren im IPA-Vokalviereck steht immer links die ungerundete, rechts die gerundete Variante. Zur Identifikation eines Vokals wird die Zungenhöhe, die Zungenlage und der Grad der Lippenrundung angegeben: /u/ ist ein hoher hinterer gerundeter Vokal, /a/ ist ein tiefer vorderer ungerundeter Vokal, /ə/ ist ein mittlerer zentraler ungerundeter Vokal.

Kardinalvokale

Die Diakritika dienen, wie bereits erwähnt, der Beschreibung artikulatorischer Details in einer engen phonetischen Transkription, so z.B. auch der Beschreibung pathologischer Lautrealisationen (sog. ”phonetische Fehler”): /t/ wird im Deutschen in der Regel apiko-alveolar gebildet. In einer breiten oder gemäßigt engen Transkription wird die Apikalität normalerweise nicht vermerkt: [taːl] (*Tal*), während eine sehr enge Transkription dies mit dem entsprechenden Diakritikum kennzeichnen kann: [t̻aːl]. Ein phonetischer Fehler könnte nun z.B. darin bestehen, dass /t/ nicht mit der Zungenspitze sondern mit dem Zungenblatt realisiert wird: [t̺aːl]; oder der Verschluss wird nicht am Zahndamm, sondern weiter vorne, an den oberen Schneidezähnen gebildet: [t̪aːl].

Die Suprasegmentalia dienen schließlich der Kennzeichnung prosodischer Merkmale wie Wortbetonung, Vokaldauer, Melodieverläufe etc. Häufig verwendet werden die Zeichen für Haupt– und Nebenbetonung (*Unterrichtsstun-*

[8] So ist es z.B. prinzipiell nicht möglich, einen pharyngalen oder glottalen Nasallaut zu bilden. Der notwendige Verschluss des Vokaltraktes muss vor dem Zäpfchen gebildet werden, damit der pulmonale Luftstrom durch die Nase entweichen kann.

[9] Hohe Vokale werden manchmal auch als ”geschlossene”, tiefe Vokale als ”offene” Vokale bezeichnet.

de /ˈʊntɐʀɪçt͡s̩ʃtʊndə/), für Langvokale (*Tal* /taːl/) und für fehlende Grenzen (vor allem zur Kennzeichnung von Diphtongen: *Taufe* /taʊ̯fə/).

Zur erweiterten Beschreibung der Stimmqualität und zur Transkription gestörter Sprache z.B. im Rahmen einer Dysarthrophonie existieren seit einiger Zeit erweiterte Inventare, die *Voice Quality Symbols* (VoQS) und das *extended IPA* (extIPA). ExtIPA stellt z.B. Diakritika bereit für linguolabial realisierte Konsonanten, für inadäquate Lippenspreizung oder für Denasalierung. In Ball, Rahilly & Tench (1996) ([2]) wird die Anwendung von extIPA detailiert beschrieben.

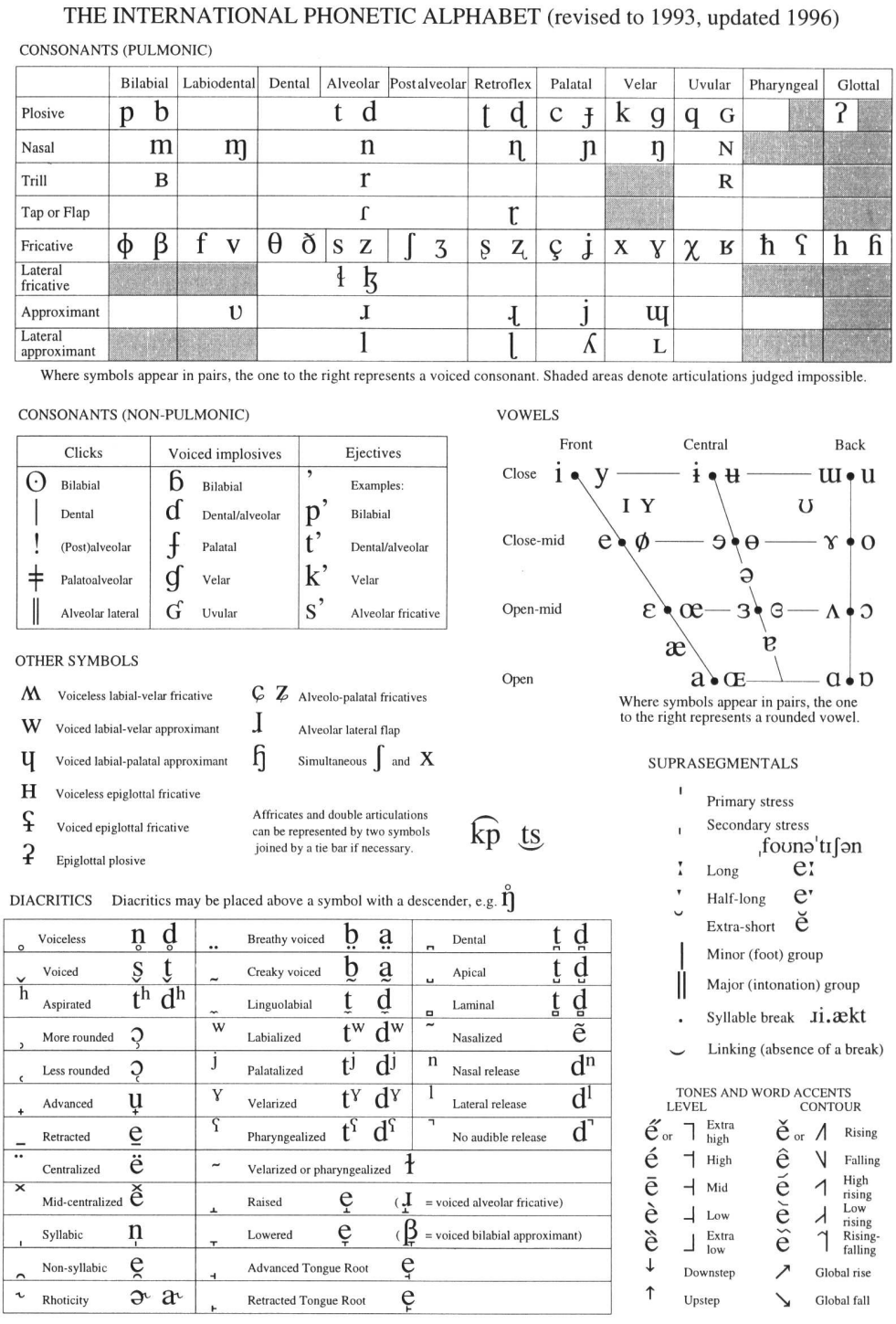


Abbildung 1.15: *Das Internationale Phonetische Alphabet (vgl. [1]).*

1.2.2 SAM Phonetic Alphabet (SAMPA)

ASCII–basierte Lautschrift

Da auf älteren Computersystemen der Umgang mit IPA–Symbolen sehr problematisch war, wurde Ende der 80er Jahre auf EU–Ebene eine Initiative ins Leben gerufen, um eine ASCII–basierte Lautschrift zu entwickeln. Der ASCII–Zeichensatz kann auf jedem Computersystem verarbeitet werden, jeder Drucker kann die entsprechenden Zeichen ausgeben und auch die Tastatureingabe stellt kein Problem dar. Einer der wichtigsten Gründe für die Entwicklung einer solchen Lautschrift war das Aufkommen computerlinguistischer Methoden, wie z.B. das Erstellen von Korpora, die in der zweiten Hälfte der 80er Jahre angesichts der massenhaften Verbreitung von bezahlbaren PCs ihren Durchbruch erlebten. So bemühte sich das europäische SAM-Projekt (***S**peech **A**ssessment **M**ethods*) um Qualitätsstandards im Zusammenhang mit der Erstellung von Korpora gesprochener Sprache. Ein wesentlicher Vorteil solcher Korpora gegenüber einfachen Aufnahmesammlungen auf Tonbändern ist die Möglichkeit der Annotation, d.h. die Sprachdaten können auf vielfältige Weise maschinenlesbar beschrieben werden. Solche Annotationen sind dann wiederum die Grundlage für Abfragesysteme, Korpusanalysen und statistische Auswertungen. Im Zusammenhang mit Korpora gesprochener Sprache ist die wichtigste Beschreibungsebene und Basis jeder weiteren Annotation natürlich die phonetische Transkription. Insofern war die Entwicklung einer geeigneten Lautschrift eine zentrale Aufgabe des SAM-Projekts. Das Ergebnis dieser Entwicklung ist das SAM Phonetic Alphabet, kurz SAMPA, ein Lautschriftsystem das ausschließlich die 256 Zeichen des ASCII-Zeichensatzes verwendet. Um mit diesem begrenzten Zeichenvorrat den vielfältigen Lautsystemen gerecht zu werden, sind die SAMPA–Konventionen in der Regel sprachspezifisch, d.h. ein bestimmtes Zeichen repräsentiert im deutschen SAMPA unter Umständen einen etwas anderen Laut als im SAMPA einer anderen Sprache. Außerdem führt der begrenzte Zeichenvorrat dazu, dass sich SAMPA eher für eine breite phonetische Transkription eignet (oder für phonematische Transkriptionen), weniger dagegen für enge, detaillierte Transkriptionen.

Trotz der Entwicklung von graphischen Benutzeroberflächen, leistungsfähigen Fonts und modernen Zeichenkodierungen wie Unicode hat SAMPA nach wie vor seine Berechtigung. Sofern für den gewünschten Verwendungszweck eine breite phonetische Transkription ausreicht, ist SAMPA immer noch das mit Abstand zuverlässigste und unkomplizierteste Lautschriftsystem. Informationen zu SAMPA und eine aktuelle Liste mit Sprachen, für die ein SAMPA existiert, findet man auf der folgenden Homepage:

http://www.phon.ucl.ac.uk/home/sampa/home.htm. Hier nun eine Liste mit den wichtigsten SAMPA–Symbolen für das Deutsche:

IPA	SAMPA	IPA	SAMPA	IPA	SAMPA
Plosive		Affrikaten		Sonoranten	
b	b	p͡f	pf	m	m
d	d	t͡s	ts	n	n
g	g	t͡ʃ	tS	ŋ	N
p	p	d͡ʒ	dZ	l	l
t	t			j	j
k	k			ʁ/ʁ	R
				r/ɾ	r
Frikative		gesp. Vokale		ungesp. Vokale	
f	f	i	i	ɪ	I
v	v	y	y	ʏ	Y
s	s	e	e	ɛ	E
z	z	ø	2	œ	9
ʃ	S	aː	a:	a	a
ʒ	Z	o	o	ɔ	O
ç	C	u	u	ʊ	U
x	x				
χ	X				
h	h				
Diphtonge		zentr. Vokale		Diakritika	
a͡ɪ	aI	ə	@	Dehnung	
a͡ʊ	aU	ɐ	6	iː	i:
ɔ͡ʏ	OY			silbische Kons.	
				n̩	=n

Und einige Beispiele:

	IPA	SAMPA
Pfeffer	p͡fɛfɐ	pfEf6
Löcher	lœçəʁ	l9C@R
Genie	ʒɛniː	ZEni:
Laugen	la͡ʊgŋ̩	laUg=N

1.3 Das Lautinventar des Deutschen

1.3.1 Plosive (Verschlusslaute, Explosive)

	stimmlos	stimmhaft
bilabial	/p/	/b/
alveolar	/t/	/d/
velar	/k/	/g/

Die stimmlosen Plosive sind im Deutschen meist aspiriert, außer nach einem silbeninitialen Frikativ (*Tal* [tʰaːl] vs. *Stahl* [ʃtaːl]) oder vor einem silbischen Nasal/Lateral (*Seite* [zaɪ̯tʰə] vs. *Seiten* [zaɪ̯tn̩] oder *Tante* [tʰantʰə] vs. *Mantel* [mantl̩]).

Die stimmhaften Plosive sind im Deutschen nicht immer vollständig stimmhaft; manchmal sind sie teilweise oder vollständig entstimmt. Dies hängt u.a. von der Position im Wort und der lautlichen Umgebung, aber auch vom Dialekt ab. Die Wahrscheinlichkeit eines vollständig stimmhaften Plosivs ist am größten zwischen zwei stimmhaften Lauten (*Laden* [laːdən]). Am Wortanfang sind stimmhafte Plosive dagegen häufig vollständig entstimmt (*Dame* [d̥aːmə]). Am Wortende werden stimmhafte Plosive (und Frikative) grundsätzlich durch die stimmlose Variante ersetzt (Auslautverhärtung: *Hunde* [hʊndə] vs. *Hund* [hʊnt]).

Beispiele:[10]

	initial	medial	final
/p/	*Panne* [panə]	*Lappen* [lapm̩]	*Lump* [lʊmp]
/t/	*Tanne* [tanə]	*Ratten* [ʀatən]	*Glut* [gluːt]
/k/	*Kanne* [kanə]	*räkeln* [ʀɛːkəln]	*Glück* [glʏk]
/b/	*Bad* [baːt]	*Rabe* [ʀaːbə]	—
/d/	*Dame* [daːmə]	*Laden* [laːdən]	—
/g/	*geben* [geːbm̩]	*Trage* [tʀaːgə]	—

[10]Die Transkription dieser und der folgenden Beispiele gibt jeweils eine mögliche Aussprachevariante wieder. Dies muss nicht notwendigerweise die ”Standardaussprache” sein. So ist es z.B. vom Stil bzw. von der Sprechgeschwindigkeit abhängig, ob die Endung *–en* als [ən] oder als [n̩] realisiert wird.

1.3.2 Nasale

	stimmhaft
bilabial	/m/
alveolar	/n/
velar	/ŋ/

Beispiele:

	initial	medial	final
/m/	*malen* [maːln̩]	*rammen* [ʀamən]	*Lamm* [lam]
/n/	*Nase* [naːzə]	*Henne* [hɛnə]	*reden* [ʀeːdən]
/ŋ/	—	*Anker* [aŋkəʀ]	*lang* [laŋ]

1.3.3 Vibranten

	stimmhaft
alveolar	/r/
uvular	/ʀ/

Der vordere gerollte r–Laut (/r/) tritt v.a. in süddeutschen Dialekten auf, z.B. im Bairischen oder in einigen Varianten des Schwäbischen. Er ist jedoch auch Teil der deutschen ”Bühnensprache”. Der uvulare Vibrant wird dagegen eher von norddeutschen Sprechern realisiert (zu den Varianten des deutschen r–Lautes siehe Abschnitt 1.4).

Beispiele:

	initial	medial	final
/ʀ/	*Rad* [ʀaːt]	*Lehre* [leːʀə]	*starr* [ʃtaʀ]

1.3.4 Frikative

	stimmlos	stimmhaft
labio–dental	/f/	/v/
alveolar	/s/	/z/
post–alveolar	/ʃ/	/ʒ/
palatal	/ç/	—
velar	/x/	—
uvular	/χ/	/ʁ/
glottal	/h/	—

Frikative unterliegen im Deutschen der Auslautverhärtung, d.h. am Wortende (bzw. im Morphemauslaut vor Konsonanten) tritt jeweils nur die stimmlose Variante auf.

Der stimmhafte uvulare Frikativ /ʁ/ ist eine Realisierungsvariante des deutschen r–Lautes; /ç/, /x/ und /χ/ sind Varianten des deutschen ch–Lautes (s. Abschnitt 1.4). /ʒ/ tritt nur in Lehnwörtern auf und wird häufig durch /ʃ/ ersetzt.

Beispiele:

	initial	medial	final
/f/	*Vogel* [foːgəl]	*kaufen* [kaʊ̯fən]	*Suff* [zʊf]
/v/	*Waage* [vaːgə]	*Lavendel* [lavɛnd̩l]	—
/s/	*Skala* [skaːla]	*Kapsel* [kapsəl]	*Riss* [ʀɪs]
/z/	*Sahne* [zaːnə]	*Käse* [keːze]	—
/ʃ/	*Stadt* [ʃtat]	*Asche* [aʃə]	*lasch* [laʃ]
/ʒ/	*Genie* [ʒeniː]	*Blamage* [blamaːʒə]	—
/ç/	*China* [çiːna]	*Licht* [lɪçt]	*mich* [mɪç]
/x/	—	*Frucht* [fʁʊxt]	*Tuch* [tuːx]
/χ/	—	*Fracht* [fʁaχt]	*Fach* [faχ]
/ʁ/	*Rolle* [ʁɔlə]	*Dorf* [dɔʁf]	—
/h/	*Hieb* [hiːp]	*daheim* [daha͡ɪm]	—

1.3.5 Approximanten

	stimmhaft
palatal	/j/

Dieser Laut kann im Deutschen auch als stimmhafter palataler Frikativ realisiert werden: /ʝ/.

Beispiele:

	initial	medial	final
/j/	*jodeln* [joːdəln]	*Mayonnaise* [majɔneːzə]	—

1.3.6 Laterale Approximanten

	stimmhaft
alveolar	/l/

Beispiele:

	initial	medial	final
/l/	*Laden* [laːdən]	*Quelle* [kvɛlə]	*toll* [tɔl]

1.3.7 Affrikaten

	stimmlos	stimmhaft
labio–dental	/p͡f/	—
alveolar	/t͡s/	—
post–alveolar	/t͡ʃ/	/d͡ʒ/

/d͡ʒ/ tritt nur in Lehnwörtern auf und wird häufig durch /t͡ʃ/ ersetzt.

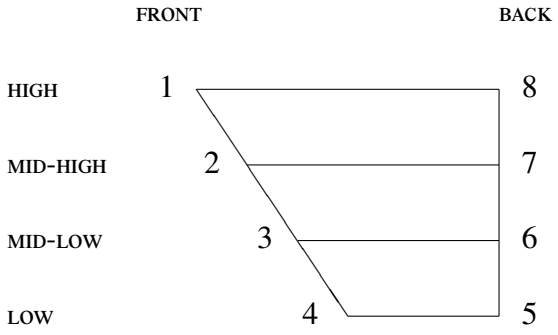
Beispiele:

	initial	medial	final
/p͡f/	*Pfanne* [p͡fanə]	*schlüpfen* [ʃlʏp͡fən]	*Napf* [nap͡f]
/t͡s/	*Zahl* [t͡saːl]	*Witze* [vɪt͡sə]	*Latz* [lat͡s]
/t͡ʃ/	*Tschechien* [t͡ʃɛçɪən]	*Latschen* [laːt͡ʃən]	*Matsch* [mat͡ʃ]
/d͡ʒ/	*Gin* [d͡ʒɪn]	*Manager* [mɛnɪd͡ʒɐ]	—

1.3.8 Vokale

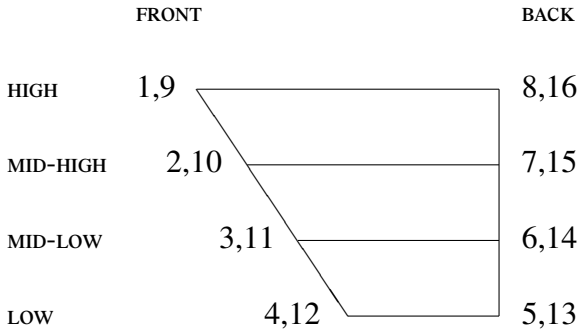
Die artikulatorische Beschreibung von Vokalen ist stark durch das von Daniel Jones Anfang des 20. Jahrhunderts entwickelte System der Kardinalvokale geprägt. Das Kardinalvokalsystem basiert auf artikulatorisch definierten Referenzpunkten im universalen Vokalraum. Diese Referenzpunkte entsprechen Zungenpositionen, die (1) limitierend sind, d.h. ein Überschreiten würde zu einer so starken Verengung führen, dass Friktion entstünde, und (2) relativ einfach zu definieren sind. Die Referenzpositionen sind vorne/oben für den Kardinalvokal 1 (bei Überschreitung, d.h. weiter vorne und/oder weiter oben, entstünde ein alveo–palataler Frikativ) und hinten/unten für den Kardinalvokal 5 (bei Überschreitung entstünde ein pharyngaler Frikativ). Die weiteren Kardinalvokale sind weniger eindeutig zu definieren: Ausgehend von Kardinalvokal 1 geht es in drei äquidistanten Schritten nach unten, bis mit Kardinalvokal 4 die vordere untere Extremposition erreicht ist. Umgekehrt geht es ausgehend von Kardinalvokal 5 in 3 drei äquidistanten Schritten nach oben bis zur hinteren oberen Extremposition des Kardinalvokals 8.

Kardinalvokale

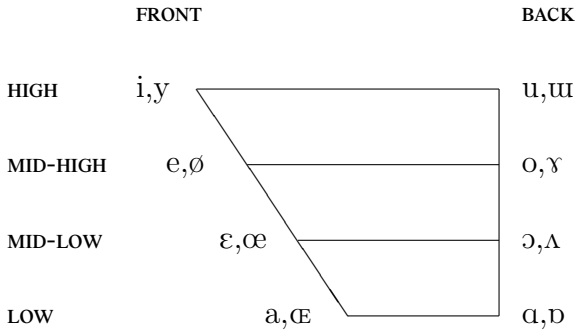


Obwohl die Bezeichnungen der vertikalen Zwischenschritte der artikulatorischen Begriffswelt entstammen (und ursprünglich wohl auch so gemeint waren), ist es sinnvoller, diese als auditorische Qualitäten zu interpretieren. So unterscheiden sich z.B. die Kardinalvokale 6 und 7 kaum hinsichtlich ihrer vertikalen Zungenposition; zudem liegt ihre Zungenposition viel näher bei Kardinalvokal 5 als bei Kardinalvokal 8. Bei den Abstufungen geht es also weniger um artikulatorische als vielmehr um auditorische Äquidistanz und das Kardinalvokalsystem wird heute besser als auditorischer Referenzrahmen für Vokalqualitäten verstanden (die Terminologie hat sich dieser veränderten Interpretation mangels Alternativen (noch) nicht angepasst). Der Vollständigkeit halber sei noch erwähnt, dass es nicht nur 8, sondern 16 Kardinalvokale

gibt: Von den ungerundeten Kardinalvokalen 1–4 gibt es gerundeten Varianten (Kardinalvokale 9–12) und von den gerundeten Kardinalvokalen 5–8 gibt es ungerundete Varianten (Kardinalvokale 13–16).



Das folgende Diagramm zeigt die IPA–Kardinalvokalsymbole. Streng genommen sollten einzelsprachliche Vokalinventare (wie z.B. das Vokalsystem des Deutschen), die praktisch nie wirkliche Kardinalvokale beinhalten, mithilfe der Kardinalvokalsymbole plus Diaktitika beschrieben werden. In der Praxis verwendet man jedoch fast durchgängig Kardinalvokalsymbole auch für Vokalqualitäten, die nicht identisch mit Kardinalqualitäten sind, aber zumindest 'in deren Nähe' liegen. Dieser Praxis schließe ich mich im folgenden an. Daneben wurden für einige häufig benötigten Vokalqualitäten eigene Symbole eingeführt, die die Kombination aus Kardinalsymbol+Diakritikum ersetzten, was einerseits den Transkriptionsaufwand verringert und andererseits die Konsistenz von Transkriptionen erhöht. Diese Symbole sind mittlerweile fester Bestandteil des IPA und werden im folgenden selbstverständlich verwendet.



Monophtonge im Deutschen

	vorne		zentral	hinten
	ungerundet	gerundet	ungerundet	gerundet
hoch	/i/	/y/	—	/u/
halbhoch	/ɪ/	/ʏ/	—	/ʊ/
obermittelhoch	/e/	/ø/	—	/o/
mittel	—	—	/ə/	—
untermittelhoch	/ɛ/	/œ/	—	/ɔ/
halbtief	—	—	/ɐ/	—
tief	—	—	/a/, /aː/	—

Neben der IPA–Terminologie kursieren für die deutschen Vokale noch diverse andere Bezeichnungen: Statt ”hoch” und ”tief” findet man manchmal die Bezeichnungen ”geschlossen” und ”offen”, außerdem spricht man gelegentlich von ”gespannten” und ”ungespannten” bzw. von ”langen” und ”kurzen” Vokalen. Als gespannt bzw. lang gelten die hohen und obermittelhohen Vokale, also /i/, /y/, /u/, /e/, /ø/ und /o/. Als ungespannt bzw. kurz gelten die halbhohen und untermittelhohen Vokale /ɪ/, /ʏ/ /ʊ/, /ɛ/, /œ/ und /ɔ/. Die Verknüpfung des Gespanntheitsparameters mit dem Dauerparameter ist jedoch problematisch. So gibt es z.B. in der deutschen Standardaussprache auch ein langes (ungespanntes) /ɛː/ (nicht jedoch in den norddeutschen Dialekten; dort wird /ɛː/ stets durch /eː/ ersetzt: *Käse* [kɛːzə] vs. [keːzə]). Außerdem sind die gespannten Vokale nur in betonter Position (d.h. in einer Silbe, die die Wortbetonung trägt) lang; in unbetonter Position werden auch gespannte Vokale meist kurz realisiert.

Schwa vs. Vollvokale

/ə/ und /ɐ/ werden als Schwa–Laute bezeichnet (”e–Schwa” bzw. ”a–Schwa”). Schwa–Laute gelten als Reduktionsformen; z.B. tritt /ə/ nur in unbetonten Silben auf und kann beispielsweise bei schneller Sprechgeschwindigkeit ganz verschwinden ([laːdən] → [laːdn̩]); /ɐ/ ist eine vokalische Realisierungsvariante des r–Lautes bzw. das Resultat der *-er*–Reduktion in unbetonter Position ([liːdəʀ] → [liːdɐ]). Alle anderen Vokale heißen Vollvokale.

Beispiele:

	initial[a]	medial	final
/i/	*ihm* [iːm]	*Miete* [miːtə]	*Ski* [ʃiː]
/ɪ/	*im* [ɪm]	*Mitte* [mɪtə]	—
/y/	*Übel* [yːbl̩]	*hüten* [hyːtn̩]	*früh* [fʁyː]
/ʏ/	*üppig* [ʏpɪç]	*Hütten* [hʏtn̩]	—
/e/	*eben* [eːbm̩]	*beten* [beːtn̩]	*See* [zeː]
/ɛ/	*essen* [ɛsn̩]	*Betten* [bɛtn̩]	—
/ɛː/	*äsen* [ɛːzn̩]	*bärtig* [bɛːɐtɪç]	(*säh* [zɛː])
/ø/	*Öfen* [øːfn̩]	*Söhne* [zøːnə]	*Bö* [bøː]
/œ/	*öffnen* [œfnən]	*Töpfe* [tœ͡pfə]	—
/u/	*Unding* [uːndɪŋ]	*Mut* [muːt]	*Schuh* [ʃuː]
/ʊ/	*unter* [ʊntɐ]	*Mutter* [mʊtɐ]	—
/o/	*Ofen* [ofn̩]	*Schrot* [ʃʁoːt]	*Po* [poː]
/ɔ/	*offen* [ɔfn̩]	*Schrott* [ʃʁɔt]	—
/aː/	*Ahle* [aːlə]	*Kahn* [kaːn]	*sah* [zaː]
/a/	*alle* [alə]	*kann* [kan]	—
/ə/	—	*Tages* [taːgəs]	*Wanne* [vanə]
/ɐ/	—	*Wirt* [wɪɐt]	*Uhr* [uːɐ]

[a] Vor anlautenden Vokalen wird im Deutschen stets der glottale Verschlusslaut /ʔ/ produziert. Daher gibt es — zumindest aus phonetischer Sicht — eigentlich keine wortinitialen Vokale im Deutschen. Da es sich hierbei jedoch um einen sehr regelmäßigen und stabilen Vorgang handelt, wird der glottale Verschlusslaut in der Transkription meist weggelassen (außer vielleicht in einer sehr engen Transkription). Und aus phonologischer Sicht ist es selbstverständlich, dass es auch im Deutschen wortinitiale Vokale gibt.

Dynamik der Vokalartikulation

Von einem idealisierenden Standpunkt aus betrachtet, kann man sagen, dass für die Produktion eines Monophtongs ein stabiles artikulatorisches Ziel angesteuert wird. Dieser Idealvorstellung kommen wir nahe, wenn wir z.B. einen Vokal längere Zeit anhalten; in diesem Fall bleibt die artikulatorische Zielposition tatsächlich über einen längeren Zeitraum unverändert. Beim normalen, zusammenhängenden Sprechen (*connected speech*) ist dies eher die Ausnahme. Am Beginn eines Vokals benötigen die Artikulatoren eine gewisse Zeit, um sich von der Konfiguration des vorangehenden Konsonanten weg und zur vokalischen Zielposition hin zu bewegen. Zum Ende des Vokals hin

wird die artikulatorische Konfiguration des nachfolgenden Konsonanten antizipiert und die Artikulatoren beginnen, sich von der vokalischen Position weg und zur konsonatischen Position hin zu bewegen. Dies führt zu artikulatorischen Übergangsphasen zu Beginn und am Ende eines Vokals; diese Übergangsphasen heißen Transitionen.[11]

Transitionen

Transitionen können als mehr oder weniger automatisierter, universaler und für die flüssige Artikulation notwendiger Adaptionsprozess des sprechmotorischen Systems gelten, nämlich als Reaktion des sprechmotorischen Systems auf die biomechanischen Eigenschaften des Sprechapparates (z.B. Masseträgheit der Artikulatoren). Wobei zu betonen ist, dass Transitionen nicht als 'Störung' eines ansonsten idealen Artikulationsprozesses zu deuten sind. Einerseits verhindert das Vorhandensein von Transitionen keinesfalls die Wahrnehmung einer einheitlichen Vokalidentität, andererseits hat sich gezeigt, dass Transitionen eine wichtige Rolle in der Perzeption und Spracherkennung spielen.

Anders verhält es sich mit dem On– und Offglide. Im Gegensatz zu den unwillkürlichen Transitionen handelt es sich hier um eine willkürliche, dynamische Veränderung der Vokalqualität zu Beginn (Onglide) oder am Ende eines Vokals (Offglide). Glide–Phänomene treten sprachspezifisch auf (z.B. im Englischen, nicht jedoch im Deutschen) und werden auch als signifikante Qualität wahrgenommen, während Transition für gewöhnlich nicht bewusst wahrgenommen werden. Glides starten (Onglide) bzw. enden (Offglide) meist in der neutralen, zentralen Zungenposition (= Schwa). Beispielsweise wird in einigen Varietäten des Englischen das Wort *four* mit Offglide produziert: [fɔᵊ].

Onglide/Offglide

Charakteristisch für Glides ist, dass es eine primäre, eindeutig identifizierbare Vokalqualität gibt (im Beispiel oben die ɔ–Qualität), während die Ausgangs– bzw. Zielqualität des Glides sekundär ist (deshalb wird sie in der Transkription auch nur durch ein hochgestelltes Diakritikum repräsentiert). Von Diphtongen spricht man dagegen, wenn zwei prominente, gleichwertige Vokalqualitäten vorliegen, d.h. bei Diphtongen gibt es zwei gleichwertige vokalische Zielkonfigurationen, die das Ausmaß und die Richtung der Gleitbewegung determinieren. Dabei kann der Anteil der beiden Targets an der Gesamtdauer des Diphtongs durchaus variieren, d.h. die Verteilung ist nicht zwingend 50:50. Ebenso kann das Verhältnis zwischen Bewegungsphase und stabiler Phase innerhalb eines Diphtongs variieren.

Diphtonge

[11] Ob dazwischen tatsächlich das artikulatorische Ziel (*target*) erreicht wird, ist nicht selbstverständlich und hängt von zahlreichen Faktoren ab. So kann man z.B. davon ausgehen, dass bei schnellem, informellem Sprechen artikulatorische Ziele nur annäherungsweise erreicht werden; man spricht in diesem Fall von *target undershoot*.

Abschließend sei darauf hingewiesen, dass die hier vorgestellte Klassifikation zwar üblich, aber nicht immer unproblemtisch anzuwenden ist. Weder bei der Frage, ob etwas als On– bzw. Offglide oder als Diphtong zu werten ist, noch bei der Frage, ob eine Sequenz aus zwei Vokalqualitäten als ein Diphtong oder als zwei Monophtonge zu werten ist, herrscht immer Einigkeit. Der Status der Diphtonge im Deutschen ist jedoch wenig umstritten.

Diphtonge im Deutschen

/a͡ɪ/, /a͡ʊ/, /ɔ͡ɪ/ (oder auch /ɔ͡ʏ/)

Beispiele:

	initial	medial	final
/a͡ɪ/	*Eisen* [a͡ɪzən]	*Saiten* [za͡ɪtən]	*Schrei* [ʃʁa͡ɪ]
/a͡ʊ/	*außen* [a͡ʊzən]	*klauen* [kla͡ʊən]	*Schau* [ʃa͡ʊ]
/ɔ͡ɪ/	*Eule* [ɔ͡ɪlə]	*träumen* [tʁɔ͡ɪmn̩]	*scheu* [ʃɔ͡ɪ]

1.4 Phone und Phoneme: Von der Phonetik zur Phonologie

Im Abschnitt 1.3 (Das Lautinventar des Deutschen) wurde einige Male darauf hingewiesen, dass manche Laute mit unterschiedlichen Varianten realisiert werden können. Die bekanntesten Beispiele hierfür sind der deutsche ch–Laut und der deutsche r–Laut.

Die ch–Variation ist abhängig vom vorangehenden Laut:[12] Nach einem vorderen Vokal oder einem Konsonanten folgt der palatale Frikativ /ç/, nach einem hohen hinteren Vokal folgt der velare Frikativ /x/, nach dem tiefen Vokal /a/ folgt der uvulare Frikativ /χ/. /ç/ wird auch als ich–Laut bzw. ch1 bezeichnet, /x/ und /χ/ werden unter der Bezeichnung ach–Laut bzw. ch2 zusammengefasst. Die Auswahl des entsprechenden ch–Lautes ist also nicht frei, sondern durch die jeweilige lautliche Umgebung vorgegeben, d.h. die ch–Variation ist kontextabhängig.

kontextabhängige Variation

Dies gilt nicht für die Variation des r–Lautes. Ein orthographisches <r> kann als Vibrant (/r/ oder /ʀ/), als stimmhafter (/ʁ/) oder stimmloser Frikativ (/χ/; z.B. nach stimmlosen Obstruenten), als Approximant (/ʁ̞/; vor allem intervokalisch) oder vokalisiert auftreten (/ɐ/; vor allem postvokalisch vor Kon-

[12] Außerdem gibt es eine Positionsabhängigkeit: wort– bzw. morpheminitial wird stets /ç/ realisiert (*China* /çiːnaː/ oder *Tauchen* (ein kleines Tau) /taû̯çən/).

sonant oder final). Diese Varianten können von Sprechern des Deutschen relativ frei gewählt werden, d.h. die r–Variation ist überwiegend frei.

freie Variation

Neben kontextabhängigen und freien Variationen gibt es auch positionsabhängige Variationen. Dies sind lautliche Prozesse, deren Auftreten von der Position eines Lautes im Wort (bzw. Morphem) abhängt. Darunter fällt z.B. die Auslautverhärtung (Entstimmung von Obstruenten am Wortende) oder die Produktion des glottalen Verschlusslautes vor wortinitialen Vokalen.

Im ersten Absatz dieses Abschnitts haben wir von *dem* deutschen ch–Laut und von *dem* deutschen r–Laut gesprochen, obwohl es doch eigentlich *mehrere* ch–Laute und *mehrere* r–Laute gibt. Dennoch sind beide Aussagen richtig — das Problem liegt darin, dass die Aussagen zu unterschiedlichen Beschreibungsebenen gehören, und dass der Begriff "Laut" zu ungenau ist. Wenn wir von *dem* deutschen ch–Laut sprechen, meinen wir eine abstrakte lautliche Einheit der deutschen Sprache. Diese abstrakte Einheit wird, abhängig vom lautlichen Kontext, unterschiedlich realisiert — diese Realisierungsvarianten meinen wir, wenn wir von *mehreren* ch–Lauten sprechen. Im ersten Fall reden wir von Phonemen und befinden uns auf der phonologischen Beschreibungsebene, im zweiten Fall reden wir von Phonen und befinden uns auf der phonetischen Beschreibungsebene. Phone sind die kleinsten segmentierbaren Einheiten der Lautsprache, d.h. sie sind nicht weiter analysierbar[13] und können in verschiedenen Umgebungen als eine (mehr oder weniger) invariante Einheit identifiziert werden. Wenn ein (Ohren–) Phonetiker eine unbekannte Sprache erforscht, ist das Ziel seiner Arbeit die Erstellung eines Phoninventars dieser Sprache, d.h. die Auflistung aller in dieser Sprache verwendeten Laute. Das im vorhergehenden Abschnitt besprochene "Lautinventar des Deutschen" ist demnach exakt formuliert ein Phoninventar.

Phonem vs. Phon

Phoninventar

Das Phoninventar einer Sprache bildet die Basis für deren phonologische Beschreibung. Die Phonologie ist jedoch nicht an der phonetischen Identifizierbarkeit von Lauten interessiert, sondern an deren kommunikativer Funktion, d.h. ob ein bestimmtes Phon in einer bestimmten Sprache dazu verwendet wird, Bedeutungen zu unterscheiden oder nicht. Mit diesem Kriterium wird das Phoninventar einer Sprache analysiert, und nur diejenigen Phone, die dem Kriterium entsprechen, die also bedeutungsunterscheidend sind, sind Teil des Phoneminventars dieser Sprache. Phone, die dem Kriterium nicht entsprechen, heißen "phonetische Varianten".

Phoneminventar

[13] Daher ist es umstritten, ob Affrikaten tatsächlich Phone sind, oder ob es sich nicht um komplexe, aus zwei Phonen zusammengestzte Einheiten handelt.

1.5 Übungsaufgaben

1. Welche der folgenden Faktoren sind mitverantwortlich für die Höhe des Stimmtons beim Sprechen?

   ☐ Stärke des subglottalen Drucks

   ☐ Grad der Kieferöffnung

   ☐ Geschwindigkeit der Stimmlippenschwingungen

   ☐ Stärke der medialen Kompression

   ☐ Öffnungsgrad des Velums

2. Ordnen Sie die Lautklassen den entsprechenden Artikulationsmodi zu.

Lautklassen	**Artikulationsmodi**
Plosiv ∘	∘ zentrale Enge ohne Geräuschbildung
Nasal ∘	∘ kompletter Verschluß, Velum abgesenkt
Vibrant ∘	∘ seitliche Enge ohne Geräuschbildung
Frikativ ∘	∘ kompletter Verschluß, Velum angehoben
Approximant ∘	∘ zentrale Enge mit Geräuschbildung
Lateraler Apprx. ∘	∘ intermittierende orale Verschlüsse

3. Geben Sie jeweils das IPA–Symbol an und ein Beispielwort, das mit dem entsprechenden Laut beginnt.

   Bsp.: stimmhafter bilabialer Plosiv: /b/ *Baum*

   stimmloser velarer Plosiv: / / _____

   stimmloser post–alveolarer Frikativ: / / _____

   stimmhafter labio–dentaler Frikativ: / / _____

   stimmh. alveolarer lateraler Approximant: / / _____

4. Beschreiben Sie folgende Laute (nach Stimmbeteiligung, Artikulationsort und Artikulationsart).

   Bsp.: /p/ *stimmloser bilabialer Plosiv*

   /s/ _____
   /ŋ/ _____
   /ʀ/ _____
   /ç/ _____

5. Welche Wörter wurden hier transkribiert?

   Bsp.: [baʊ̯m] *Baum*

   [ʃat͡s] _____
   [ʀɔ̯ɪbɐ] _____
   [zeːlə] _____
   [klaʊ̯n] _____
   [ˈnɔɐtˌvɪnt] _____

6. Sie geben einer Italienerin Deutschunterricht. Ihre Schülerin möchte vorallem ihre Aussprache verbessern. Das Problem ist, dass es im Italienischen einige Laute (und Buchstaben) nicht gibt, die es im Deutschen gibt.

   - Beschreiben Sie ihr ausführlich die Laute, die sich hinter »ö« und »eu« verbergen.
   - Beschreiben Sie ihr bittte auch das sogenannte 'Zäpfchen–R'.
   - Außerdem möchte sie gerne das »ch« lernen; was können Sie ihr erklären?

7. Im Deutschen wird angeblich so geschrieben, wie gesprochen. Erläutern Sie anhand des folgenden Textes, dass diese Aussage problematisch ist. Nennen Sie eine Regel, mit der eines der Phänomene erklärt werden kann.

   Ich soll die Zecke in die Pfanne lotsen. Dass das Spaß macht liegt auf der Hand.

8. Sie erhalten den folgenden Ausschnitt aus dem Nachsprechteil des Aachener Aphasie Tests (Zielwörter und transkribierte Äußerungen des Patienten). Beschreiben Sie den Befund mit Hilfe artikulatorisch–phonetischer Merkmale.

Ast	—	[aft]
Floh	—	[floː]
Mund	—	[mʊnt]
Glas	—	[glaːs]
Stern	—	[ʃtɛʁn]
Fürst	—	[fʏɐt]
Spruch	—	[ʃpʀʊx]
Knirps	—	[kniɐp]
Zwist	—	[vɪft]
Strumpf	—	[ʃtʀʊm͡pf]

9. Sie arbeiten mit einer Stimmpatientin auf Textebene. Die Patientin trägt Ihnen einen zu Hause bearbeiteten Text vor (vgl. Transkription). Korrigieren Sie die Patientin und begründen Sie Ihre Einwände.

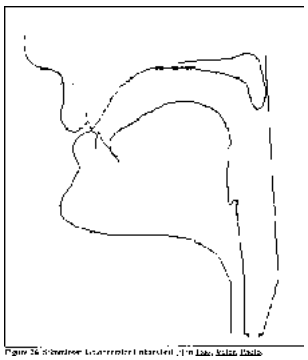
   da ʃtɔlpeːʁtə deːʁ moːnd ɪm gəɛstʰ deʁ βa͡ɪdən ʊnd fiːl a͡ʊf das dɪxtə gʁaːs

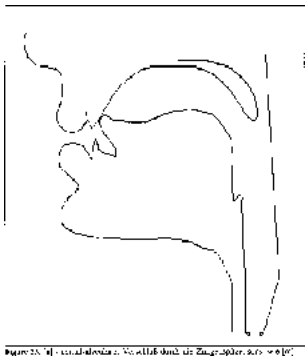
   (Da stolperte der Mond im Geäst der Weiden und fiel auf das dichte Gras.)

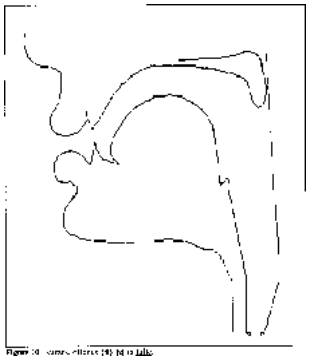
10. Für die Wörter auf der linken Seite der folgenden Liste findet man in der informellen 'normalen' Alltagssprache häufig die rechts transkribierten Aussprachevarianten. Beschreiben Sie, ausgehend von der Standardauslautung, die phonetischen Prozesse, die zu den beschriebenen Formen führen.

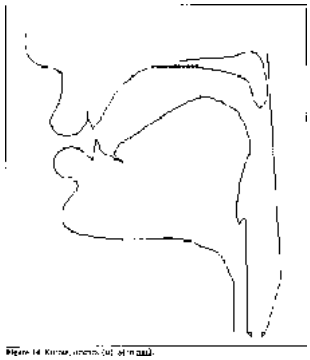
Schlüssel	—	[ʃlʏsl̩]
haben	—	[haːbn̩] oder [haːbm̩]
klagen	—	[klaːgŋ]

11. Begründen Sie ausführlich, um welche Laute es sich bei den folgenden Sagittalschnitten handeln könnte.

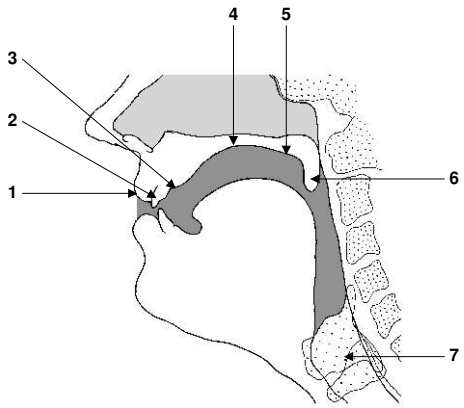
(a) 

(b) 

(c) 

(d) 

12. Benennen Sie die folgenden Artikulationsorte. Nennen Sie je einen Laut des Deutschen, der an diesem Ort gebildet wird, und beschreiben Sie diesen vollständig (nach der Systematik des IPA).



13. Ein Kollege entwirft einen Lautprüfungsbogen und möchte ihn mit Ihnen besprechen. Der Test soll dazu dienen, bei Kindern (Muttersprache Deutsch) zu überprüfen, welche Laute an welchen Wortpositionen bereits erworben wurden und korrekt gebildet werden können.

    - Auf welche phonetischen Unkorrektheiten können Sie ihren Kollegen aufmerksam machen? Begründen Sie Ihre Kritik.
    - Welche Lautkombinationen werden im Schwäbischen eher nicht zu überprüfen sein?

Ziellaut	Testwort	Anlaut	Inlaut	Auslaut
/m/	Mond	m–		
	Hammer		–m–	
	Baum			–m
/b/	Banane	b–		
	Gabel		–b–	
	Korb			–b
/v/	Wurst	v–		
	Löwe		–v–	
	Calw			–v
/d/	Dach	d–		
	Nadel		–d–	
	Pfad			–d
/t/	Tisch	t–		
	Auto		–t–	
	Bett			–t
/ʀ/	Roller	R–		
	Burgen		–R–	
	Mutter			–R
/sp/	Spinne	sp–		
	Kasper		–sp–	
/st/	Stuhl	st–		
	Kiste		–st–	
	Nest			–st

Kapitel 2

Anmerkungen zur perzeptiven Phonetik

2.1 Einleitende Bemerkungen

Wie in der Einleitung erwähnt, beschäftigt sich die perzeptive Phonetik mit der Wahrnehmung von Lautsprache. Dieser Wahrnehmungsprozess kann zunächst vereinfacht in zwei Stufen unterteilt werden:

primäre Wahrnehmungsereignisse

- Das Erleben primärer, insbesondere auditiver Wahrnehmungsereignisse.
- Die Interpretation dieser Ereignisse, d.h. die Integration der primären Wahrnehmungsereignisse in das Sprachsystem.

Die weitere Differenzierung und Beschreibung dieser beiden Stufen ist Aufgabe einer umfassenden Theorie der Sprachwahrnehmung. Die perzeptive Phonetik im engeren Sinne konzentriert sich dagegen auf die erste Stufe, also auf die objektiven und subjektiven Aspekte der primären Wahrnehmung. Naturgemäß stehen dabei Hörereignisse im Mittelpunkt des Interesses, insofern ist eine Konzentration auf die auditiven Aspekte der lautsprachlichen Wahrnehmung und die alternative Bezeichnung auditive Phonetik zu verstehen. Wie gleich zu sehen sein wird, spielen jedoch auch andere, beispielsweise visuelle Wahrnehmungsereignisse eine gewisse Rolle.

Doch zunächst noch einmal zurück zu den beiden Stufen der Perzeption. Anhand eines kleinen Beispiels soll der komplexe Zusammenhang zwischen einem primären Wahrnehmungsereignis und dem Perzept, also dem

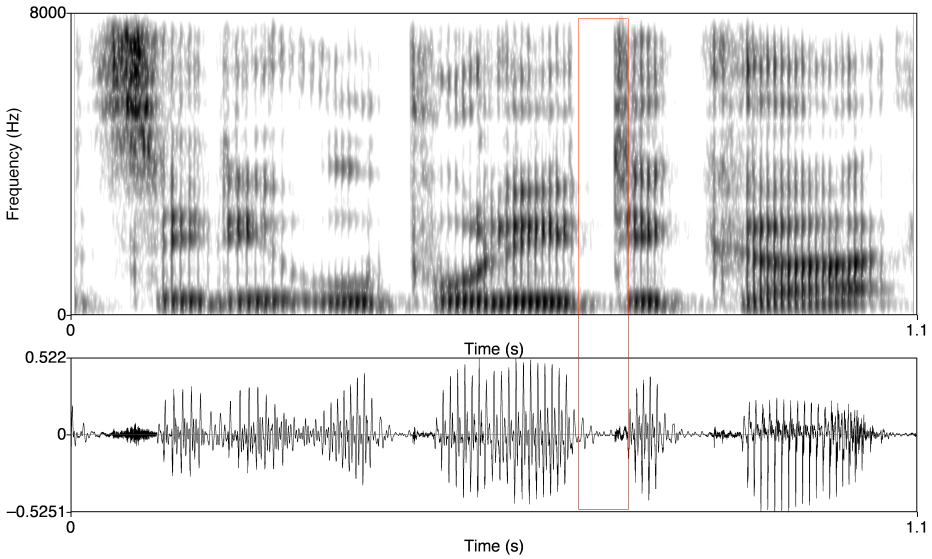


Abbildung 2.1: *Signal (unten) und Spektrogramm (oben) der Äußerung "CDU–Politiker" (tse:de:u:pOlItIk@R).*

Endergebnis des Wahrnehmungsprozesses, verdeutlicht werden. Abbildung 2.1 zeigt die Signaldarstellung und das Spektrogramm der Äußerung "CDU–Politiker" (SAMPA–Transkription: tse:de:u:pOlItIk@R) aus der Aufnahme einer Radionachricht. Beide Darstellungsformen – Signal und Spektrogramm – visualisieren akustische Ereignisse, wie sie von einem menschlichen Hörer wahrgenommen werden können,[1] und zumindest als geübter Betrachter solcher Abbildungen kann man erkennen, dass es sich um ein vollständiges, unversehrtes Signal handelt, das keinem Hörer Schwierigkeiten bereiten dürfte.

Abbildung 2.2 zeigt die selbe Aufnahme nach einer Signalmanipulation. Der stimmlose alveolare Plosiv [t] im Onset der dritten Silbe von /pɔlɪtɪkəʀ/ wurde in einem Signaleditor durch Stille ersetzt. Das bedeutet, dass alle primären akustischen Merkmale des Plosivs aus dem resultierenden Signal gelöscht wurden. Solche Signalmanipulationen werden z.B. zur Erstellung von

[1] Die Signaldarstellung (Oszillogramm) gibt den Schalldruckverlauf über die Zeit wieder. Das Spektrogramm zeigt die Intensitätsvariation im Frequenzbereich: Dunkleres Grau bedeutet hohe Intensität, helleres Grau geringe Intensität. Stille stellt sich entsprechend als weiße Fläche über den gesamten Frequenzbereich dar; mehr dazu im Abschnitt 3.4.2.

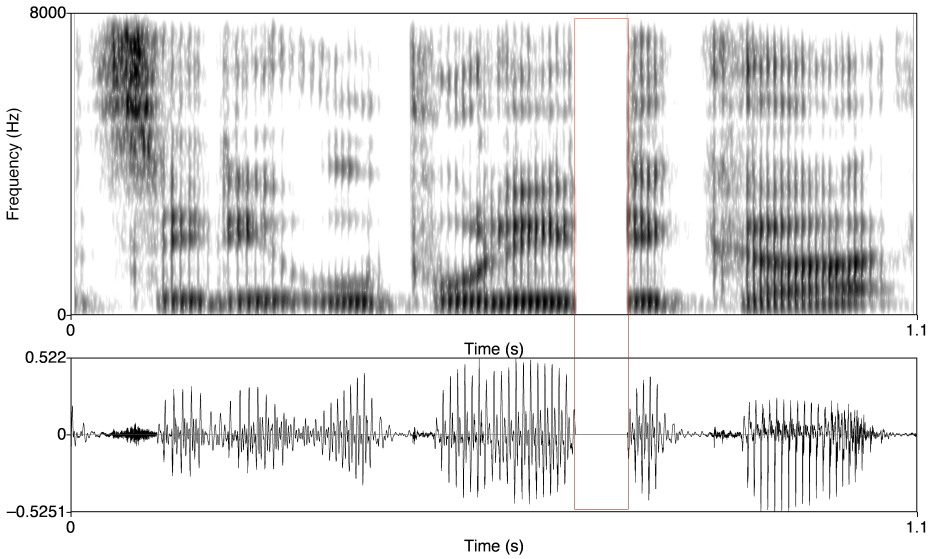


Abbildung 2.2: *Signal (unten) und Spektrogramm (oben) der Äußerung "CDU–Politiker". Der alveolare Plosiv wurde mit Hilfe eines Signaleditors herausgeschniten und durch Stille ersetzt.*

Stimuli für Perzeptionsexperimente angewendet. Und tatsächlich zeigen Experimente mit derartigen Stimuli, dass Hörer keinerlei Probleme haben, solche unvollständigen Signale zu verstehen. Das Fehlen von Lauten fällt einem naiven Hörer in der Regel gar nicht auf und selbst geübte Ohrenphonetiker haben Schwierigkeiten, dies zu erkennen. Auf die zwei Stufen der Wahrnehmung übertragen heißt das, dass trotz eines defizitären primären Wahrnehmungsereignisses, dem ganz eindeutig wichtige Informationen zur Lautidentifikation fehlen, am Ende ein vollständiges, unversehrtes Perzept wahrgenommen wird. Man nennt dieses Phänomen 'phonemic restauration', d.h. ein hörendes Subjekt stellt im Verlauf des Wahrnehmungsprozesses einen objektiv nicht vorhandenen Laut wieder her. Woher nehmen wir aber die fehlende Information?

Perzept

bottom–up vs. top–down

Hier kommen wir auf die bereits in der Einführung erwähnte Unterscheidung zwischen *bottom–up–* und *top–down–*Prozessen zurück. Die Generierung einer phonemischen Repräsentation des Gehörten auf der Basis primärer Wahrnehmungsereignisse ist ein *bottom–up–*Prozess: Aus kleinsten Informationseinheiten, wie z.B. den akustischen Merkmalen von Lauten, wird ein

größeres, abstrakteres Bild aufgebaut. Wenn wir einer uns vertrauten Sprache zuhören, sind jedoch parallel dazu auch ständig unsere Wissensrepräsentationen aktiv und erzeugen bestimmte Erwartungen. Aufgrund unseres situativen Wissens bzw. unseres Weltwissens haben wir z.B. eine recht gute Vorstellung davon, worüber in Radionachrichten gesprochen wird, und entwickeln in Kombination mit unserem lexikalischen Wissen nach der Sequenz [t͡seː deː uː pɔˈlɪ] eine sehr starke Erwartung, wie es weitergeht. Da diese Erwartungen aus Wissensrepräsentationen auf einer höheren kognitiven Ebene gespeist werden, spricht man hier von *top–down*–Prozessen (s. Abb. 2 auf Seite 11).

Stark vereinfacht kann man sich das Wechselspiel zwischen diesen beiden Prozessen folgendermaßen vorstellen: Zu Beginn einer Äußerung werden *bottom–up*–Informationen stärker gewichtet, eventuell vorhandene Erwartungen sind eher unspezifisch. Mit zunehmender Äußerungsdauer werden die *top–down* generierten Informationen jedoch immer spezifischer, bis sie irgendwann u.U. sogar stark genug sind, um *bottom–up*–Informationen zu überschreiben. Tatsächlich konnte in Perzeptionsexperimenten gezeigt werden, dass Hörer bei einer entsprechend starken Erwartung nicht nur fehlende Laute ersetzen, sondern auch unpassende Laute 'überhören' und durch den korrekten Laut ersetzen (in unserem Beispiel würde in diesem Fall beispielsweise statt eines tatsächlich dargebotenen Pseudowortes [pɔˈlɪpɪkəʁ] das Wort *Politiker* wahrgenommen).

Bevor wir nach diesem kurzen Exkurs über allgemeine Fragen der Sprachperzeption zurückkehren zum Kernbereich der perzeptiven Phonetik, nämlich der Beschäftigung mit der auditiven Wahrnehmung, soll hier noch ein Beispiel angeführt werden, dass auch die Ebene der primären Wahrnehmungsereignisse komplexer ist, als man vielleicht zunächst annehmen würde. Schon hier kann es zu konkurrierenden Informationen kommen, die vom wahrnehmenden Subjekt zu einem einheitlichen Perzept fusioniert werden müssen. Es ist offensichtlich, dass wir in einer gewöhnlichen *face–to–face*–Situation Sprache nicht nur hören sondern auch sehen: Während unser Gegenüber Schall erzeugt, den wir hören, nehmen wir gleichzeitig über den visuellen Kanal die dazu notwendigen Sprechbewegungen wahr. Um zu untersuchen, ob wir wirklich beide Informationsquellen nutzen, und wenn ja, wie diese miteinander verrechnet werden, haben McGurk & MacDonald (1976)[2] ein mittlerweile berühmtes und oft wiederholtes Perzeptionsexperiment durchgeführt. Dabei wurden den Probanden über den auditiven und den visuellen Kanal

auditiver vs. visueller Kanal

McGurk–Effekt

[2]McGurk, H. & MacDonald, J., 1976, Hearing lips and seeing voices. *Nature* **264**, 746–748.

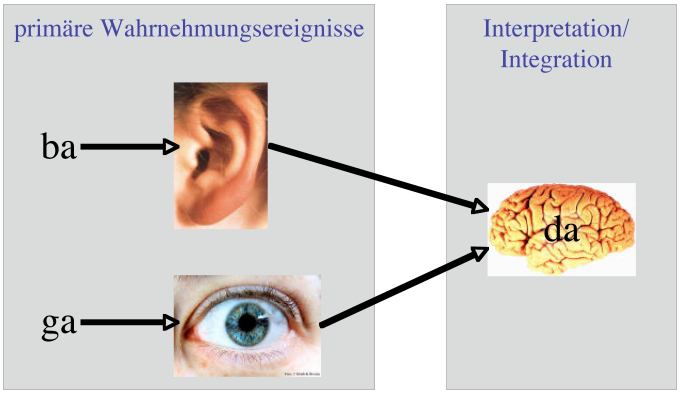


Abbildung 2.3: *Der McGurk–Effekt: Auditiv wahrgenommenes* [ba] *plus visuell wahrgenommene Sprechbewegung von "*ga*" führt zum Perzept* /da/.

widersprüchliche Informationen angeboten. So bekamen sie z.B. über Kopfhörer ein [ba] zu hören, während sie synchron dazu die Videonahaufnahme eines Sprechers sahen, der ein "ga" artikuliert. Danach befragt, was sie wahrgenommen hätten, gaben die meisten Probanden /da/ an (McGurk–Effekt, Abb. 2.3).

Da es sich hierbei um die Verrechnung von Informationen aus unterschiedlichen Modalitäten (auditiv und visuell) handelt, wird der Vorgang auch als heteromodale Fusion bezeichnet. Ähnliche Effekte findet man aber auch bei Untersuchungen von unimodalen, z.B. dichotischen Fusionen: Hierbei werden über Kopfhörer dem linken und dem rechten Ohr synchron unterschiedliche Stimuli dargeboten.[3]

Es folgen nun einige Anmerkungen zu den anatomischen und physiologischen Grundlagen der menschlichen Hörfähigkeit (auditorisches System) und im nächsten Abschnitt zum Verhältnis zwischen objektiven physikalischen Größen und subjektiven Wahrnehmungsgrößen (Psychoakustik).

2.2 Das auditorische System

Das auditorische System setzt sich zusammen aus dem Gehörorgan und dem auditorischen Nervensystem, welches das Gehörorgan über zahlreiche Verschaltungen hinweg mit dem auditorischen Kortex im Großhirn verbindet.

[3] Siehe z.B.: Pompino, B., 1980, Selective adaptation to dichotic psychacoustic fusions. *Journal of Phonetics* **8**, 379–384.

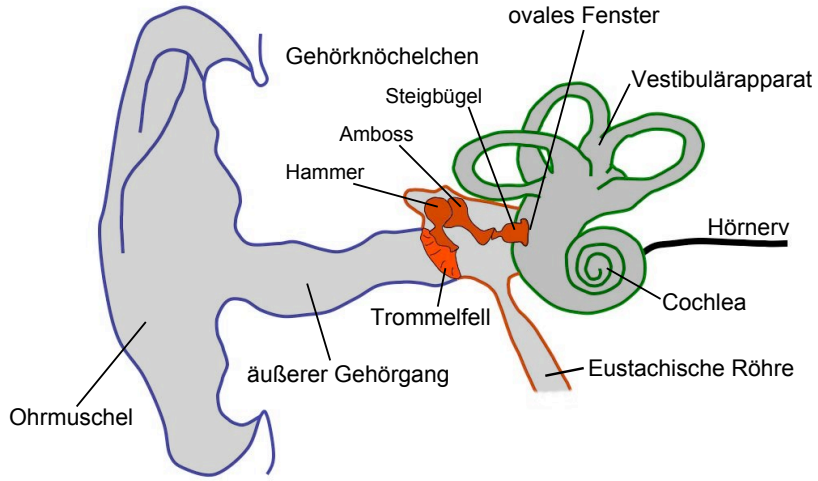


Abbildung 2.4: *Außenohr (blau), Mittelohr (rot) und Innenohr (grün).*

Das Gehörorgan wiederum lässt sich in anatomisch und funktionell differenzierte Abschnitte unterteilen: das Außenohr, das Mittelohr und das Innenohr (Abb. 2.4).

Außenohr

Anatomie: Ohrmuschel und äußerer Gehörgang; das Trommelfell bildet die Grenze zum Mittelohr.
Funktion: Die Ohrmuschel unterstützt das Richtungshören, der äußere Gehörgang verstärkt bestimmte, insbesondere auch für das Sprachverstehen wichtige Frequenzbereiche und beide zusammen schützen das Mittelohr.

Mittelohr

Anatomie: Der zwischen Trommelfell und ovalem Fenster (Grenze zum Innenohr) gelegene Hohlraum (”Paukenhöhle”); die mechanische Verbindung zwischen Trommelfell und ovalem Fenster bilden die Gehörknöchelchen: ”Hammer”, ”Amboss” und ”Steigbügel”; über die Eustachische Röhre ist das Mittelohr mit dem Rachenraum verbunden.
Funktion: Mechanische Signalübertragung zwischen Außen– und Innenohr und Anpassung des Schalldrucks; Normalerweise wird der Schalldruck durch das Hebelsystem der Gehörknöchelchen verstärkt (notwendig, um den unterschiedlichen Schallwiderstand der Luft im Außenohr und der Lymphflüssigkeit im Innenohr auszugleichen); durch Versteifung des Hebelsystems (Kontraktion des Steigbügelmuskels und des

Schutzfunktion

Trommelfellspanners) können hohe Schallintensitäten jedoch auch abgeschwächt werden, um das Innenohr vor Schäden zu schützen.

Innenohr

Anatomie: Schneckenlabyrinth (Cochlea, Hörorgan) und Vorhoflabyrinth (Vestibulärapparat, Gleichgewichtsorgan) im Felsenbein des Schläfenbeins gelegen; die Cochlea, ein mit Lymphflüssigkeit gefülltes Kanalsystem, weist 2 ½ schneckenförmige Windungen auf; im wesentlichen sind 3 Kanäle zu unterscheiden: oben die Scala vestibuli, unten die Scala tympani und dazwischen die Scala media bzw. Ductus cochlearis; Scala tympani und Ductus cochlearis sind durch die ca. 32 mm lange Basilarmembran getrennt, dem Sitz des Corti–Organs, unseres eigentlichen Hörorgans; wichtigster Teil des Corti–Organs sind die in ungefähr 3600 Reihen angeordneten Haarzellen (sekundäre Rezeptorzellen); durch Beugen ihrer Härchen (Cilien) werden in den Zellen des Ganglion spirale (Hörnerv) synaptische Prozesse ausgelöst und durch das auditorische Nervensystem weitergeleitet.

Cochlea

Basilarmembran

Funktion: Im Innenohr findet die entscheidende Signaltransformation statt: Mechanische Schallwellen werden in elektro–chemische Aktionspotentiale – die 'Sprache' unseres Nervensystems – übersetzt; dabei werden die im Schallsignal enthaltenen Frequenzen an unterschiedlichen Positionen entlang der Basilarmembran analysiert; die gleichmäßige Abbildung der Frequenzen auf der Basilarmembran wird auch als tonotope Abbildung bezeichnet (s. Abb. 2.5); für die Analyse tiefer Frequenzen stehen auf der Basilarmembran größere Bereiche zur Verfügung als für die Analyse höherer Frequenzen, dadurch verfügt das menschliche Gehör über ein besseres Auflösungsvermögen für tiefere Frequenzen als für höhere Frequenzen; die Haarzellen sind nicht nur afferent sondern auch efferent innerviert, wodurch das Innenohr durchaus auch akustische Signale generieren kann (otoakustische Emissionen).

tonotope Abbildung

An die Signaltransformation im Corti–Organ des Innenohrs schließt sich die Reizweiterleitung entlang der afferenten Bahnen des auditorischen Nervensystems an. Hier lassen sich zwei Systeme wiederum sowohl anatomisch als auch funktionell unterscheiden:

Ventrale auditorische Bahn

Anatomie: Eine ipsilaterale Verbindung vom Nucleus cochlearis ventralis über die Olivia superior und weitere höhere Kerne (Lemniscus

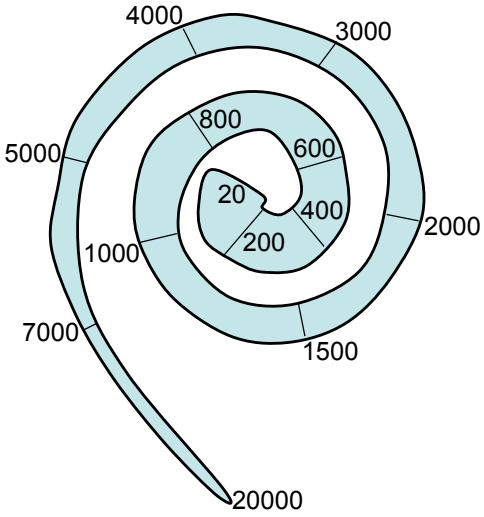


Abbildung 2.5: *Die Basilarmembran in schematischer Darstellung; tonotope Frequenzabbildung.*

lateralis, Colliculus inferior, Corpus geniculatum mediale) zum auditorischen Kortex im Temporallappen der ipsilateralen Hemisphäre.
Funktion: Diese Bahn dient insbesondere der Richtungsbestimmung wahrgenommener Schallereignisse; schon auf der Ebene der Olivia superior werden neuronale Informationen des kontralatralen Ohrs hinzugezogen, um durch den Abgleich z.B. der Laufzeitunterschiede oder der Intensitätsunterschiede zwischen den Ohren die Lage der Schallquelle zu bestimmen.

Dorsale auditorische Bahn

Anatomie: Eine Verbindung vom Nucleus cochlearis dorsalis über den kontralateralen Lemniscus lateralis und weitere Kerne zum kontralateralen auditorischen Kortex.
Funktion: Diese Bahn dient der Analyse und Erkennung auditorischer Ereinisse; aufbauend auf der Frequenzanalyse der Basilarmembran werden hier komplexere Komponenten des akustischen Reizes analysiert, wie z.B. Frequenzkombinationen, Geschwindigkeit von Frequenz– und Intensitätsveränderungen, Rauschkomponenten etc.

2.3 Psychoakustische Grundlagen

Die Psychoakustik [4] beschäftigt sich mit der mathematischen Abbildung physikalischer Größen, mit denen die Parameter des akustischen Signals beschrieben werden, auf psychologische Größen der auditiven Wahrnehmung. Es geht also um das Verhältnis zwischen Schallübertragung und Schallwahrnehmung und die Skalierung der entsprechenden Parameter. Wir konzentrieren uns hier auf die zwei akustischen Parameter Schalldruck und Frequenz, denen auf psychoakustischer Seite die Parameter Lautheit (*loudness*) und Tonhöhe (*pitch*) gegenüber stehen.

Schallwahrnehmung

2.3.1 Schalldruck und Lautheit

Schalldruckvariationen sind die Ursache dafür, dass ein Hörer Schalle als lauter oder leiser empfindet, also Lautheitsunterschiede wahrnimmt. Der Bereich des wahrnehmbaren Schalldrucks reicht von etwa 0,00002 Pa (20 μPa, Hörschwelle) bis 100 Pa (Schmerzgrenze).[4] Um diesen enormen Bereich an das menschliche Lautheitsempfinden angepasst zu skalieren, kann man auf die Dezibel–Skala (*dB*) zurückgreifen.[5] Die Dezibel–Skala ist eine logarithmische Skala, die sowohl relativ als auch absolut genutzt werden kann. Relativ ist z.B. die Angabe, dass sich zwei Schallereignisse um 20 dB unterscheiden; das bedeutet aufgrund der logarithmischen Skalierung, dass das eine Ereignis einen 10mal höheren Schalldruck aufweist als das andere. Wie hoch jedoch der tatsächliche absolute Schalldruck jeweils ist, lässt sich daraus nicht ableiten. Für eine absolute dB–Skala ist daher ein Referenzwert notwendig. Dieser Referenzwert liegt im Bereich der Hörschwelle eines jugendlichen Normalhörers für einen 1000 Hz Sinuston. Ein solcher Ton, über Kopfhörer dargeboten, kann, wie oben erwähnt, ab einem Schalldruck von 20 μPa wahrgenommen werden. Per Definition entspricht dies einem Schalldruckpegel (L) von 0 dB absolut.[6] Im sogenannten freien Schallfeld, also ohne Kopfhörer, liegt

Dezibel

Hörschwelle

[4] Der atmosphärische Druck liegt bei 100000 Pa [Pascal] bzw. 1 b [bar].

[5] Die *Bel*–Skala *(B)* ist nach Alexander Graham Bell benannt. Da sich die Einheit *Bel* jedoch für auditive Zwecke als zu grob erwiesen hat, verwendet man in diesem Bereich ausschließlich die nächst kleinere Einheit *Dezibel (dB)*.

[6] Die Bezeichnung *dB absolut* ist nicht sehr gebräuchlich. Meist wird für absolute dB–Werte ebenso wie für relative Werte die Einheit dB angegeben und es bleibt dem Leser überlassen, aus dem Kontext zu erschließen, ob sich dahinter eine relative oder eine absolute Angabe verbirgt. Bisweilen findet man für absolute Angaben die Einheit *dB(A)*; dies ist eine Messeinheit von Schallpegelmessgeräten, die ihre Messwerte hörbewertet ausgeben. Die Angabe *(A)* verweist auf eine für normalen Sprechschall geeignete Bewertungskennlinie (*(D)*

Schmerz– und Unbehaglichkeits-schwelle

die Hörschwelle eines jugendlichen Normalhörers für einen 1000 Hz Sinuston bei etwa 4 dB absolut. Das obere Ende der Hördynamik ist durch die Schmerzgrenze gegeben. Diese liegt, ebenfalls bezogen auf jugendliche Normalhörer und einen 1000 Hz Sinuston, bei etwa 125 dB absolut. Die z.B. für die Hörgeräteanpassung wichtige Unbehaglichkeitsschwelle liegt definitionsgemäß 15 dB unterhalb der Schmerzgrenze, also im Bereich 110 dB absolut.

Hör–, Schmerz– und Unbehaglichkeitsschwelle verlaufen allerdings nicht linear über den gesamten wahrnehmbaren Frequenzbereich von etwa 20 Hz bis 20 kHz. Ursache hierfür ist, dass das menschliche Lautheitsempfinden frequenzabhängig ist, d.h. außerhalb des 1000–Hz–Bereichs finden sich sowohl Frequenzbereiche mit niedrigerer Hörschwelle als auch Frequenzbereiche mit höherer Hörschwelle, wobei letztere deutlich überwiegen. Die maximale Hörempfindlichkeit liegt etwa zwischen 1,5 und 5 kHz. Eine abfallende Hörempfindlichkeit kann man sowohl unterhalb von 1 kHz als auch oberhalb 10 kHz beobachten. Abbildung 2.6 zeigt den Verlauf der Hörschwellenkurve über den gesamten wahrnehmbaren Frequenzbereich; auf der y–Achse ist der Schalldruckpegel in *dB* abgetragen, auf der x–Achse die Frequenz in *kHz* (logarithmisch skaliert, d.h. tiefe Frequenzen nehmen mehr Raum ein als hohe). Aus der Kurve ist z.B. ersichtlich, dass ein 500–Hz–Ton mit einer Lautstärke von etwa 40 dB dargeboten werden muss, um gerade eben wahrgenommen zu werden, während bei einem 1000–Hz–Ton, wie oben erwähnt, 4 dB ausreichen. Letztendlich verbindet die Hörschwellenkurve also alle Punkte in diesem Koordinatensystem (”Hörfläche”), die zu einem identischen Lautheitsempfinden führen. Um dies auszudrücken, kann die psychoakustische Einheit *phon* verwendet werden, die den dB–Wert bei 1000 Hz übernimmt. Dies bedeutet im Falle der Hörschwellenkurve: Alle Punkte auf dieser Kurve haben die Lautstärke 4 phon, da die Kurve die 1000–Hz–Linie bei 4 dB kreuzt; eine Kurve wie die Hörschwelle heißt deshalb auch Isophone.

Lautheitsempfinden

Die Abbildung 2.7 zeigt weitere Isophonen zwischen 4 und 100 phon. Um einen Eindruck davon zu bekommen, was diese phon–Werte bedeuten, sind in der Tabelle 2.1 einige alltägliche Schallereignisse zusammen mit ihrer ungefähren Lautstärke in phon aufgeführt. Im Übrigen gilt, wie bereits erwähnt, dass phon–Werte im Groben gleichgesetzt werden dürfen mit dB(A)–Werten. Ebenfalls in der Tabelle befinden sich Angaben zur sogenannten Verhältnislautheit, die in der Einheit *sone* angegeben wird. Per Definition ent-

Verhältnislautheit

bezeichnet z.B. eine Bewertungskennlinie, die eher für die Messung von Flugzeugschall angepasst ist). Da es sich bei db(A)–Angaben um hörbewertete Angaben handelt, entspricht diese Einheit in etwa der Einheit *phon* (s.u.).

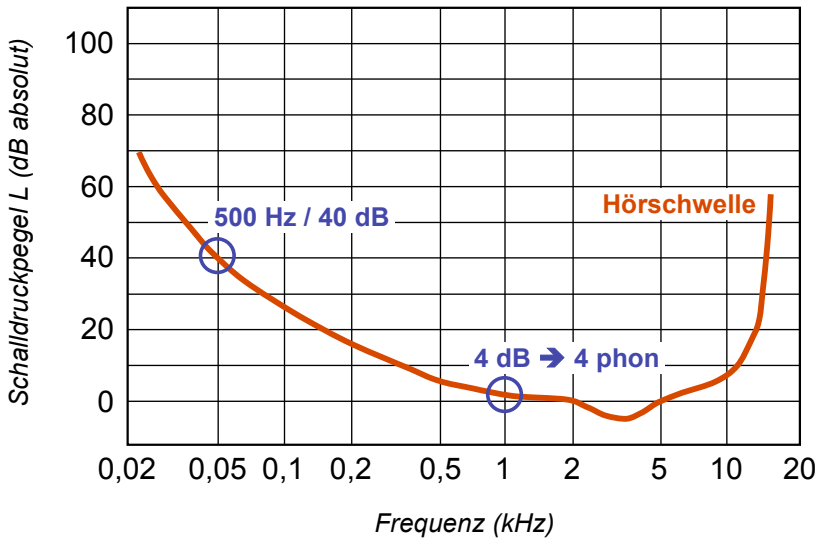


Abbildung 2.6: *Die Hörschwellenkurve im freien Schallfeld; sie entspricht der 4–phon–Isophone.*

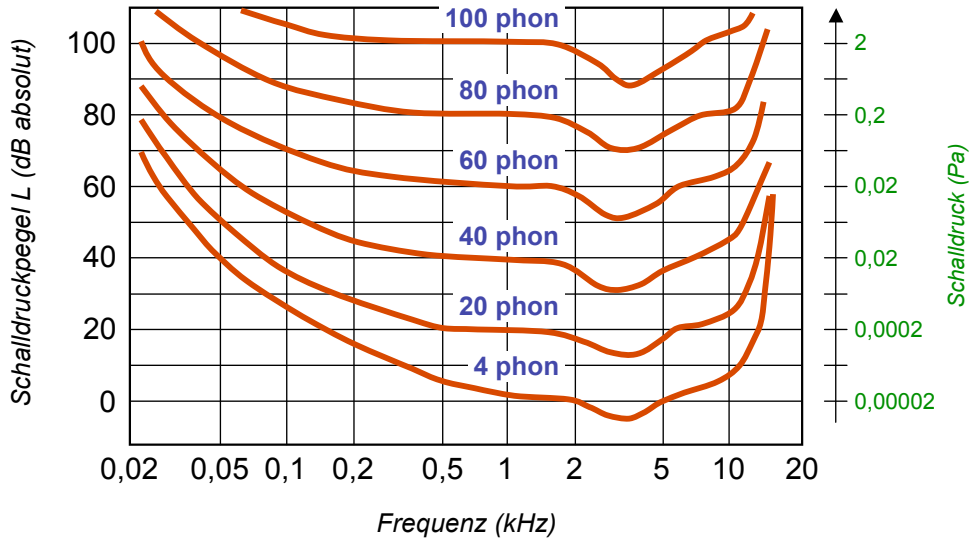


Abbildung 2.7: *Einige Isophonen zwischen 4 und 100 phon. Die Skala rechts verdeutlicht noch einmal den Zusammenhang zwischen dB absolut und effektivem Schalldruck: 0 dB absolut entsprechen per Definition 0,00002 Pa.*

spricht 1 sone = 40 phon. Davon ausgehend wurde mit Hilfe von Versuchspersonen ermittelt, wann ein Schall als doppelt, viermal, achtmal,... so laut,

Tabelle 2.1: *Einige alltägliche Schallereignisse.*

Schallereignis	Lautstärke (*phon*)	Lautheit (*sone*)
knapp unterhalb der Hörschwelle	0	0
sehr leises Flüstern	20	0,25
leises Sprechen, Kühlschrank	40	1
normales Gespräch ohne Störschall	50	2
lautes Sprechen, Staubsauger	60	4
belebte Straße	70	8
starker Straßenverkehr, Schreien	80	16
Preßluftbohrer (nah)	90	32
Discothek	100	64
Düsenflugzeug (nah)	120	256

bzw. halb, ein viertel, ein achtel,... so leise empfunden wird. So wird z.B. ein Schall mit 4 sone viermal lauter empfunden als ein Schall mit 1 sone; entsprechend wird ein Schall mit 0,5 sone nur halb so laut empfunden wie der Referenzschall mit 1 sone. Die Verhältnislautheit ist eine Funktion der Lautstärke: Der Verdoppelung der Verhältnislautheit entspricht jeweils eine Erhöhung der Lautstärke um 10 phon.

Außer von der Frequenz, dem sicherlich wichtigsten Faktor, ist das menschliche Lautheitsempfinden auch von der Schalldauer abhängig. Bei Schalldauern unterhalb von 200 ms steigt der Schalldruckpegel, der notwendig ist, um den gleichen Lautheitseindruck wie für ein Schallereignis mit 200 ms und längerer Dauer zu erzeugen, linear an (Abb. 2.8, oben). Und auch mit zunehmendem Alter verändert sich das Lautheitsempfinden (wie auch die Tonhöhenwahrnehmung (s.u.)). Abbildung 2.8 zeigt unten die altersabhängige Lageveränderung der Hörschwellenkurve in dem besonders betroffenen höheren Frequenzbereich oberhalb 2 kHz.

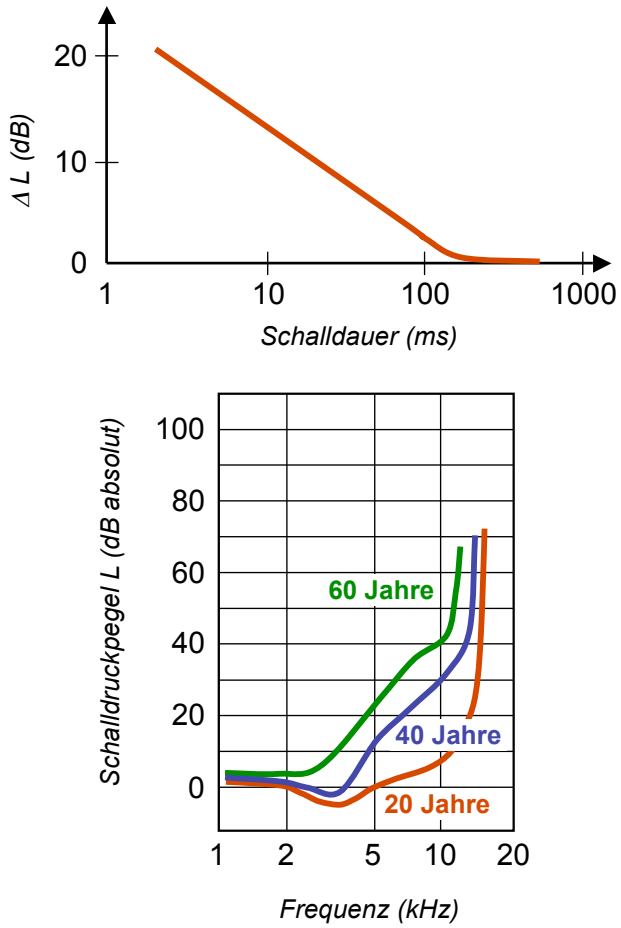


Abbildung 2.8: *Veränderung des Lautheitsempfindens mit der Schalldauer (oben) und mit dem Alter (unten).*

2.3.2 Frequenz und Tonhöhe

Tonhöhendifferenzierung

Wie der Schalldruck mit dem Lautheitsempfinden korreliert ist, so ist die Frequenz mit der wahrgenommenen Tonhöhe korreliert. Und ebenso wie dort ist auch hier das Verhältnis nicht proportional. So kann das menschliche Gehör z.B. minimal unterschiedliche Frequenzen in tieferen Frequenzbereichen viel feiner differenzieren als in höheren Frequenzbereichen. Annäherungsweise gilt, dass wir unterhalb von 500 Hz einen Frequenzunterschied zwischen zwei Tönen ab einer Differenz von 1,8 Hz erkennen können. Über 500 Hz muss der Unterschied dagegen mindestens 0,35% betragen. Ein Beispiel: Zwischen zwei Tönen mit 200 Hz und 202 Hz nehmen wir eine Tonhöhendifferenz wahr, zwischen zwei Tönen mit 2200 Hz und 2202 Hz dagegen nicht, hier muss der Unterschied mindestes 0,35% von 2200, also knapp 8 Hz betragen.

Grundsätzlich können bei der Betrachtung der Tonhöhe zwei Herangehensweisen unterschieden werden, die auch tatsächliche Wahrnehmungsunterschiede reflektieren: Zum einen kann Tonhöhe, von der Musik her kommend, als harmonische Tonhöhe, oder *Tonalität*, betrachtet werden.[7] Zum anderen lassen sich Tonhöhenunterschiede auch 'objektiv', d.h. losgelöst von musikalischen Hörgewohnheiten und harmonischen Gesetzmäßigkeiten analysieren; man spricht dann von melodischer Tonhöhe, oder *Tonheit*.

Tonalität

Oktave

Die auditive Grundeinheit der Tonalität ist die Oktave, wobei zwischen Oktave und Frequenz ein logarithmischer Zusammenhang besteht: Von Oktave zu Oktave aufwärts verdoppelt sich die Frequenz, abwärts halbiert sie sich. Ausgehend von einer beliebigen Frequenz f_1 sind also die Oktavschritte nach der folgenden Formel zu berechnen:

$$f_n = f_1\, 2^{n-1} \quad (n = 1, 2, 3, \ldots, n)$$

Das bedeutet, dass ausgehend von 125 Hz die nächste Oktave bei 250 Hz liegt, die übernächste bei 500 Hz usw. (Abb. 2.9).

Halbtöne

Die Oktave wiederum kann in zwölf Halbtonschritte unterteilt werden (chromatische Tonleiter). Der zu einer beliebigen Frequenz nächst höhere Halbtonschritt ergibt sich durch Multiplikation mit $\sqrt[12]{2}$, also 1,06, der nächst tiefere Halbtonschritt durch Multiplikation mit dem Kehrwert, also 0,94. Angaben mit Hilfe von Halbtönen sind in der Phonetik durchaus gebräuchlich.

[7] Wir beschränken uns hier auf tonale Kategorien der europäischen Musikkultur.

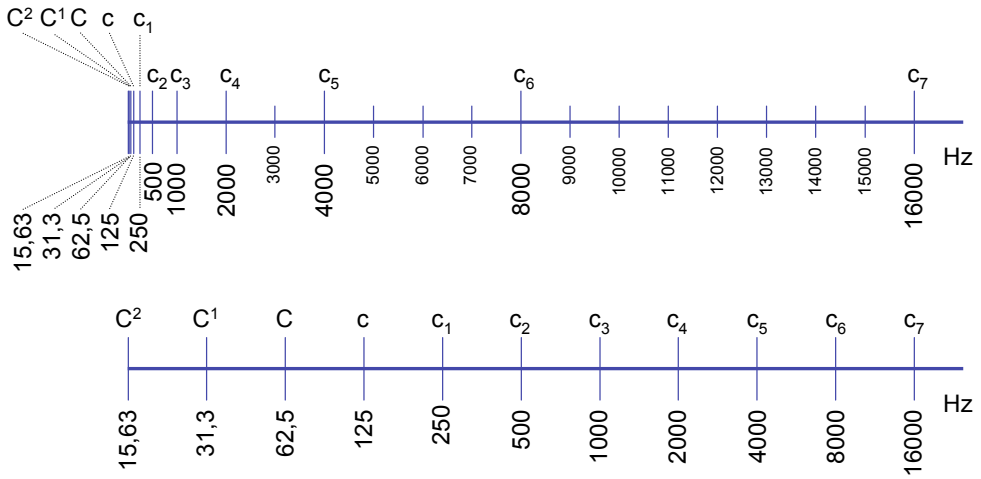


Abbildung 2.9: *Lineare Frequenzskala (oben) mit größer werdenden Abständen für gleiche Intervalle; logarithmische Frequenzskala (unten) mit gleichen Abständen für gleiche Intervalle; als Intervalle wurden Oktaven gewählt. (Die C–Frequenzen entsprechen den in der Audiometrie gebräuchlichen; nach Kammerton–a gestimmte C–Frequenzen liegen etwas höher.)*

Insbesondere die Analyse oder Synthese von Intonationskonturen (s. Abschnitt 3.4.3 ab Seite 103) in Halbtonschritten reflektiert die Tonhöhenwahrnehmung eines menschlichen Hörers adäquater als Frequenzangaben in Hz. Der Umrechnung von Frequenzintervallen in Halbtöne und umgekehrt dienen die folgenden Formeln:

$$st = \frac{12}{ln(2)}\, ln(\frac{f_2}{f_1})$$

$$f_2 = e^{\frac{st\, ln(2)}{12}}\, f_1$$

$$f_1 = \frac{f_2}{e^{\frac{st\, ln(2)}{12}}}$$

f_1: Ausgangsfrequenz (Hz)
f_2: Endfrequenz (Hz)
st: Halbtöne (*semi tones*)
ln: natürlicher Logarithmus, Basis e
e^x: Exponentialwert von x auf Basis der Eulerschen Konstanten

Ein Beispiel: Der Anstieg der Grundfrequenz (s. Seite 74 und 103ff) von f_1 = 120 Hz auf f_2 = 140 Hz am Ende einer Entscheidungsfrage eines männlichen Sprechers entspricht nach der ersten Formel 2,67 Halbtönen. Eine Sprecherin mit generell höherer Stimmlage muss also, damit ein Hörer einen äquivalenten Tonhöhenanstieg wahrnimmt[8], ausgehend von f_1 = 220 Hz ihre Grundfre-

[8] Äquivalent im Sinne der harmonischen Tonalität.

quenz nicht um 20 Hz auf 240 Hz erhöhen (= 1,51 st), sondern ebenfalls um 2,67 st, also um knapp 37 Hz auf f_2 = 256,7 Hz (zweite Formel).

Tonheit

Bei Hörexperimenten zur Wahrnehmung der melodischen Tonhöhe (Tonheit) konnte gezeigt werden, dass die Probanden – nach der Loslösung vom musikalischen Hören – für die musikalischen Intervalle (z.B. Oktave) zu den höheren Frequenzen hin immer geringere Tonhöhendifferenzen feststellten. Diese Diskrepanz zwischen harmonischer und melodischer Tonhöhenwahrnehmung setzt ab etwa 500 Hz ein. Während also die Harmonielehre 'behauptet', es bestehe eine jeweils identische Tonhöhendifferenz zwischen 62,5 Hz (C) und 125 Hz (c), 125 Hz und 250 Hz (c_1) sowie z.B. 4000 Hz (c_5) und 8000 Hz (c_6), da es sich jeweils um Oktavsprünge handelt, konnte in den Experimenten zur Wahrnehmung der melodischen Tonhöhe zwar das Verhältnis zwischen den ersten beiden Intervallen bestätigt werden (da unterhalb 500 Hz), zwischen 4000 Hz und 8000 Hz wurde jedoch eine im Vergleich dazu sehr viel geringere Tonhöhendifferenz wahrgenommen.

Verhältnistonhöhe

Aus solchen Experimenten zur vergleichenden Beurteilung von Tonhöhenverhältnissen (höher/tiefer, doppelt/halb so hoch etc.) entstand die Skala der sogenannten Verhältnistonhöhe mit der Einheit *mel* (*Tonheit in mel*, H_v). Die Skala reicht von 0 mel bis 2400 mel (16 kHz). Unterhalb 500 Hz verlaufen die mel–Skala und die logarithmische Hz–Skala nahezu proportional. 125 Hz entspricht als definierte Ausgangsfrequenz genau 125 mel; die übrigen Werte unterhalb 500 Hz sind zwar nicht exakt identisch, was aber in der praktischen Anwendung keine Rolle spielt. Erst oberhalb 500 Hz wird die Abweichung zwischen Hz–Werten und mel–Werten zunehmend größer und führt auch zu experimentell nachweisbaren Verschiebungen. So entspricht z.B. 1000 Hz 850 mel und 8000 Hz nur noch 2100 mel. Eine Verdoppelung der Tonheit in mel entspricht einer Verdoppelung der wahrgenommenen (melodischen) Tonhöhe, d.h. z.B. ein Stimulus mit 8000 Hz (2100 mel) wird als doppelt so hoch wahrgenommen wie ein 1300–Hz–Stimlus (1050 mel).

Tonheit in Bark

Mithilfe eines anderen psychoakustischen Erhebungsverfahrens (vgl. [4], Kapitel 6) wurde noch eine zweite Tonheitsskala gewonnen: Die Tonheit in Bark (Frequenzgruppenskala, *Critical Band Rate*). Hierbei wird mit Hilfe von Hörtests die Bandbreite des auditorischen Filters bestimmt, wobei ein schmalbandiges Filter (Repräsentation der tieferen Frequenzen auf der Basilarmembran) einer besseren Frequenzauflösung entspricht als ein breitbandiges Filter (Repräsentation der höheren Frequenzen auf der Basilarmembran). Gleichzei-

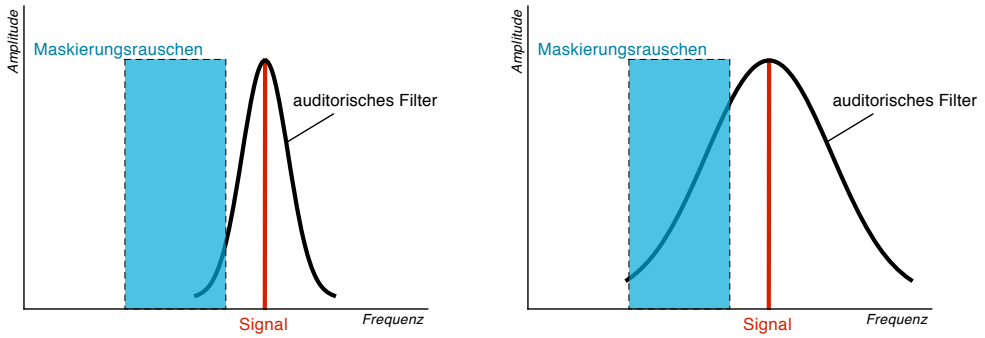


Abbildung 2.10: *Critical band rate. Links: Auditorisches Filter mit geringer Bandbreite und feiner Frequenzauflösung → geringer Maskierungseffekt. Rechts: Auditorisches Filter mit großer Bandbreite und grober Frequenzauflösung → starker Maskierungseffekt. Das Nutzsignal ist rot, das Maskierungsrauschen (jeweils mit gleicher Bandbreite) blau dargestellt.*

tig mit dem Nutzsignal, einem Ton mit einer bestimmten Frequenz, bekommen die Probanden ein schmalbandiges Rauschen präsentiert, welches das Nutzsignal mehr oder weniger stark überdeckt ('maskiert'). Das schmalbandige Rauschen wird als Maskierungssignal bezeichnet und der Maskierungseffekt ist umso stärker, je breitbandiger das auditorische Filter im Frequenzbereich des Nutzsignals ist, da dann der Signal–Rausch–Abstand (*signal–to–noise ration, SNR*) geringer wird (Abb. 2.10).

Trotz der vollkommen unterschiedlichen Erhebungsverfahren ist erstaunlicherweise eine sehr große Ähnlichkeit zwischen der mel–Skala und der Bark–Skala festzustellen. In der Psychoakustik wird dies darauf zurückgeführt, dass beide Skalen mit den selben physiologischen Merkmalen der Cochlea bzw. der Basilarmembran korrelieren (Länge/Anzahl der Haarzellen), und daher diesen physiologischen Merkmalen des Innenohrs anscheinend eine wichtige Rolle bei der menschlichen Tonhöhenwahrnehmung zugesprochen werden kann.

Wie oben bereits besprochen (Abschnitt 2.2 und Abb. 2.5), können bestimmte Abschnitte auf der Basilarmembran bestimmten zu analysierenden Frequenzen zugeordnet werden (tonotope Frequenzabbildung), wobei für tiefere Frequenzen größere Abschnitte (mit entsprechend mehr Haarzellen) zur Verfügung stehen als für höhere Frequenzen. Diese Tatsache spiegelt sich recht genau in den beiden Tonheitsskalen wider (vgl. Tabelle 2.2). Ein gesunder Mensch mit normalem Hörvermögen kann über den gesamten wahrnehmbaren Frequenzbereich etwa 640 Tonhöhenstufen unterscheiden, wofür ihm

Tabelle 2.2: *Die Beziehung zwischen psychoakustischen Tonhöhenskalen und physiologischen Merkmalen der Basilarmembran. Die Spalten von links nach rechts: Tonheitsdifferenz in Bark und Mel, Anzahl der gerade noch wahrnehmbaren Tonhöhenabstufungen (pitch steps), Größe eines entsprechenden Abschnitts auf der Basilarmembran (Distanz) und Anzahl der Haarzellenreihen in einem solchen Abschnitt (Werte aus [4], S. 162).*

Bark		Mel		*pitch steps*		Distanz		Haarzellen-reihen
24	≈	2400	≈	640	≈	32 mm	≈	3600
1	≈	100	≈	27	≈	1,3 mm	≈	150
0,01	≈	1	≈	0,26	≈	13 μm	≈	1,5

etwa 3600 Haarzellenreihen auf der ca. 32 mm langen Basilarmembran zur Verfügung stehen; dies entspricht etwa einer Tonheitsdifferenz von 24 Bark bzw. 2400 mel. Ein wahrgenommener Tonhöhenunterschied von 1 Bark (≈ 100 mel) kann in 27 gerade noch differenzierbare Tonhöhenstufen unterteilt werden und entspricht ungefähr einem 1,3 mm langen Abschnitt auf der Basilarmembran mit ca. 150 Haarzellenreihen[9].

Die Korrelation zwischen Basilarmembran, Haarzellenreihen und Tonheit wird auch in Abbildung 2.11 nochmals deutlich. Der Bezug zum akustischen Parameter Frequenz kann nur mithilfe einer nicht–linearen Skala hergestellt werden.

[9] Alle Werte und insbesondere die Beziehungen zwischen den einzelnen Skalen sind ungefähre Angaben. Im einzelnen sind die Abweichungen jedoch so gering, dass sie in der Praxis vernachlässigt werden können. Beim Vergleich von Bark– und mel–Skala sollte dennoch bedacht werden, dass die beiden Skalen auf vollkommen unterschiedlichen Erhebungsverfahren basieren, und man sollte es daher vermeiden, die Einheit mel als ein Art Zenti–Bark zu betrachten.

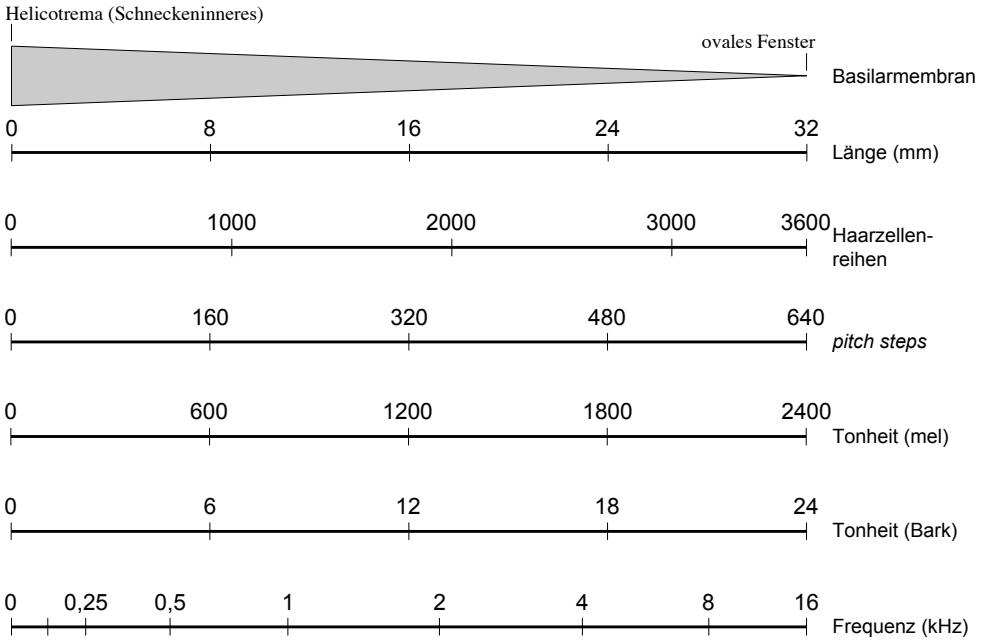


Abbildung 2.11: *Die Korrelation zwischen akustischen, psychoakustischen und physiologischen Dimensionen. Linear skaliert: Länge der Basilarmembran, Anzahl der Haarzellenreihen, pitch steps und Tonheit; nicht–linear skaliert: Frequenz.*

Kapitel 3

Akustische Phonetik

Die Schallübertragung durch die Luft stellt üblicherweise den Kanal zwischen dem Sprecher auf der einen Seite und dem Hörer auf der anderen Seite dar.[1] Dieses Schallsignal ist der Untersuchungsgegenstand der akustischen Phonetik.[2] Aufgrund der Übertragungsfunktion des Schalls bei der lautsprachlichen Kommunikation kann der Gegenstandsbereich der akustischen Phonetik genauer beschrieben werden als die Beziehung zwischen Artikulation im weitesten Sinne und dem Schallsignal einerseits und zwischen dem Schallsignal und dessen Verarbeitung im Gehörorgan andererseits. D.h. die akustische Phonetik interessiert sich sowohl für produktive Aspekte (welcher Zusammenhang besteht zwischen sprechmotorischen Vorgängen und bestimmten Schallformen) als auch für rezeptive Aspekte des Sprachschalls (wie werden bestimmte Schallformen von einem menschlichen Hörer interpretiert, vgl. Kapitel 2). In der folgenden Einführung in die akustische Phonetik werden die produktiven Aspekte im Vordergrund stehen.

3.1 Grundlagen der Akustik

Als Teilgebiet der physikalischen Disziplin der allgemeinen Schwingungslehre beschäftigt sich die Akustik mit Schwingungsvorgängen in elastischen Medien (z.B. Luft). Schall kann als ein solcher Schwingungsvorgang beschrieben werden, nämlich als auditiv wahrnehmbare Luftdruckschwankungen. Um für

[1] (Sprach–) Schall kann natürlich auch durch andere Medien als Luft übertragen werden (andere Gase, Holz, Stein, Flüssigkeiten etc.).

[2] Akustik kann ganz allgemein definiert werden als Lehre vom Ablauf und von der Ausbreitung mechanischer Schwingungen in Gasen, Flüssigkeiten und Festkörpern.

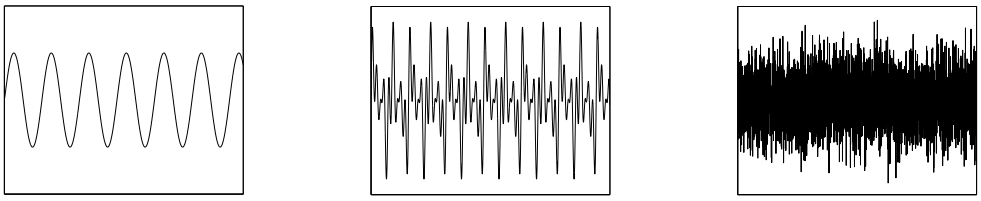


Abbildung 3.1: *Schallformen.*

einen menschlichen Hörer wahrnehmbar zu sein, d.h. um das Trommelfell in Schwingungen zu versetzen, müssen diese Luftdruckschwankungen bestimmten Anforderungen an die Geschwindigkeit und an die Amplitude genügen: Schwingungen, die weniger als 20 mal und häufiger als 20000 mal pro Sekunde ablaufen, sind noch nicht bzw. nicht mehr auditiv wahrnehmbar, d.h. wir können nur Schallwellen mit einer Frequenz zwischen 20 Hz (Hertz) und 20 kHz hören. Der Amplitudenbereich wahrnehmbarer Luftdruckschwankungen ist enorm; er liegt ungefähr zwischen 0,0000000001 b (bar) und der Schmerzschwelle bei etwa 0,001 b (der atmosphärische Druck beträgt 1 b). Die Frequenz einer Schwingung wird in Hertz (Schwingungen pro Sekunde) angegeben, der Schalldruckpegel — die Amplitude — in Dezibel(dB).[3]

Töne, Klänge, Geräusche

Luftdruckschwankungen können verschiedene Formen annehmen. Reine sinusförmige Schwingungen (die in der Natur praktisch nicht vorkommen) nennt man Töne; komplexe Schwingungen, die aus einzelnen Sinusschwingungen zusammengesetzt sind, die zueinander in einem harmonischen Verhältnis stehen (s.u.), nennt man Klänge; aperiodische, stochastische Abfolgen von Amplitudenwerten über die Zeit nennt man Geräusch (vgl. Abbildung 3.1).

Amplitude, Frequenz, Phase

Ein Ton lässt sich durch drei Parameter charakterisieren: Amplitude (Auslenkung auf der y–Achse), Frequenz (Geschwindigkeit der Schwingung) und Phase (Verschiebung des Startpunkts einer Schwingung) (vgl. Abbildung 3.2). Ein Klang lässt sich charakterisieren als die Summe der Töne, aus denen er zusammengesetzt ist. Nur wenn die Frequenzen dieser Töne jeweils ein ganzzahliges Vielfaches einer sog. Grundfrequenzdarstellen, spricht man im strengen Sinne von Klängen. Ein solches Frequenzverhältnis nennt man harmonisch. Generell gilt: Die Grundfrequenz eines Klanges entspricht dem größten gemeinsamen ganzzahligen Nenner der Teilschwingungen, aus denen er

Grundfrequenz

[3] dB ist eine logarithmische Einheit; eine Erhöhung um 6 dB entspricht einer Verdoppelung des Schalldrucks.

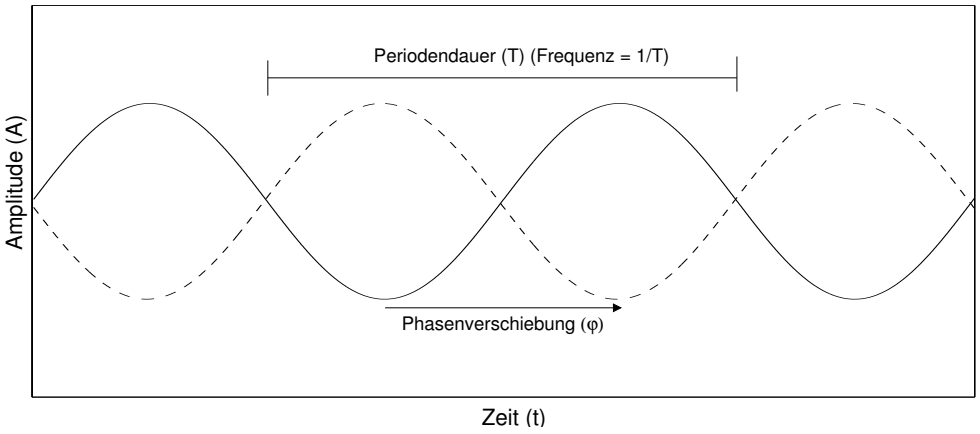


Abbildung 3.2: *Signalparameter.*

zusammengestzt ist. Abbildung 3.3 zeigt dieses Prinzip: In der linken Spalte werden zwei Töne addiert, wobei der zweite Ton die doppelte Frequenz hat. Das Ergebnis ist ein Klang mit einer Grundfrequenz, die der Frequenz des ersten Tons entspricht (die Periodendauer des Klangs ist gleich der Periodendauer des ersten Tons; s. gestrichelte Kurve). In der mittleren Spalte wird ein weiterer Ton, diesmal mit der 4–fachen Frequenz hinzu addiert. Das Ergebnis ist wiederum ein Klang mit einer Grundfrequenz entsprechend der Frequenz des ersten Tons. Das selbe gilt für die rechte Spalte, wo ein vierter Ton mit der 6–fachen Frequenz hinzu addiert wird. Nach dem Entdecker des Prinzips der Analysierbarkeit von Klängen in einzelne harmonische Sinustöne, dem französischen Mathematiker Jean Baptiste Joseph Fourier (1768–1830), wird dieser Vorgang Fouriersynthese genannt.

Fouriersynthese

Ist das Frequenzverhältnis der einzelnen Sinuskomponenten nicht ganzzahlig, entsteht eine aperiodische Schwingung, also ein Geräusch.[4] Dies bedeutet, dass auch Geräusche als Summe sinusförmiger Frequenzkomponenten beschrieben werden können. Allerdings verfügen Geräusche mathematisch gesehen über eine unendliche Periodendauer, die Frequenzkomponenten liegen entsprechend unendlich nahe beieinander. Um einen Spezialfall aperiodischer Schwingungen handelt es sich bei den Transienten. Transienten werden durch plötzlich auftretende, sich nicht wiederholende Luftdruckschwankungen verursacht. Sehr kurze Transienten nennt man auch Impuls.

Transiente und Impuls

[4]Eine Ausnahme hiervon bilden Schwingungen, deren Frequenzkomponenten zwar nicht in einem ganzzahligen, jedoch in aus der musikalischen Harmonielehre bekannten Verhältnissen zueinander stehen. In einem weiteren Sinne werden auch diese als Klang bezeichnet.

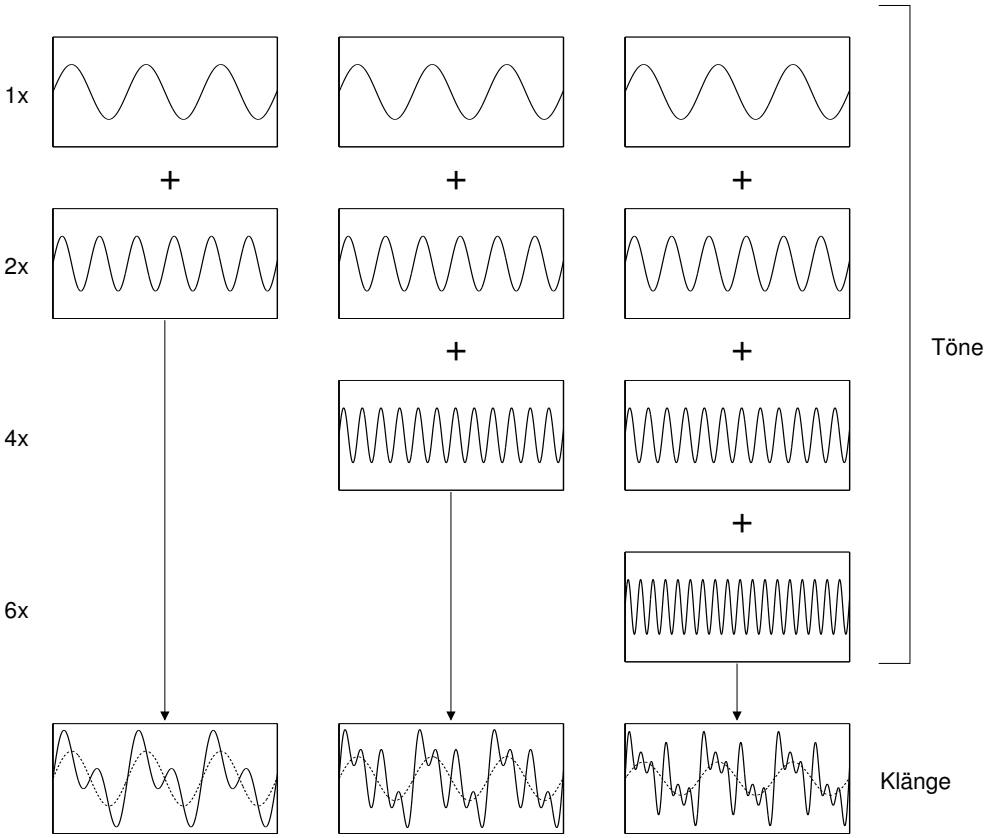


Abbildung 3.3: *Die Addition von Tönen zu Klängen.*

Amplituden– bzw. Leistungsspektrum

Ein wichtiges Darstellungs– und Analysemittel in der akustischen Phonetik ist das Spektrum (genauer: Amplituden– oder Leistungsspektrum). Ein Spektrum ist eine Analyse der Frequenzkomponenten eines Signals bzw. eines Signalausschnitts. Es handelt sich dabei um die Umkehrung des oben beschriebenen Additionsprinzips: Ein gegebenes Signal wird in seine Frequenzkomponenten zerlegt; der Vorgang heißt entsprechend Fourieranalyse. Als Resultat wird die Amplitude jeder Komponente (y–Achse) über der Frequenzachse (x–Achse) dargestellt (vgl. Abbildung 3.4). Informationen über den Zeitverlauf eines Signals sind in dieser Darstellungsform nicht mehr enthalten.

Fourieranalyse

Linienspektrum

Eine Spektraldarstellung wie in Abbildung 3.4 heißt auch Linienspektrum, da die einzelnen Frequenzkomponenten klar voneinander abgrenzbar sind und als vertikale Linien abgetragen werden können. Geräuschspektren werden dagegen als kontinuierliche Spektren dargestellt, da die Frequenzkomponenten

kontinuierliches Spektrum

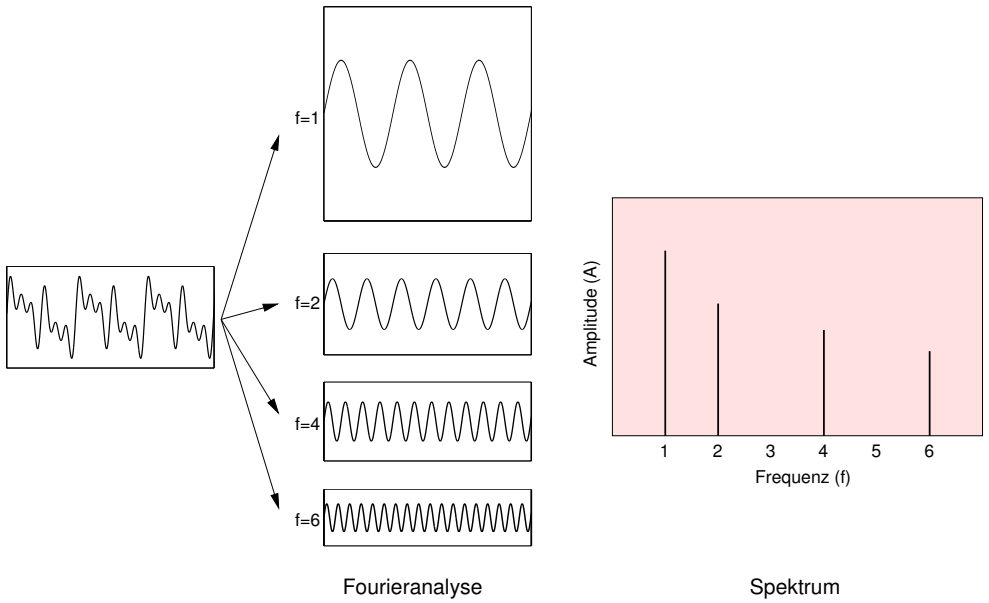


Abbildung 3.4: *Fourieranalyse und Spektraldarstellung.*

unendlich nahe beieinander liegen und daher nicht mehr als diskrete Linien darstellbar sind. In Abbildung 3.5 sind die Spektren von zwei verschiedenen Geräuschen dargestellt. Im oberen Signal sind die Frequenzen zwischen 2 kHz und 4 kHz stärker vertreten, d.h. sie haben eine höhere Amplitude im Spektrum als im unteren Signal. Diese Verteilung ist typisch für die Unterscheidung von post–alveolaren und alveolaren Frikativen; tatsächlich handelt es sich beim oberen Signal um ein [ʃ], beim unteren Signal um ein [s]. Damit kommen wir von den Grundlagen der allgemeinen Akustik zur akustischen Phonetik: der Analyse von Sprachschall.

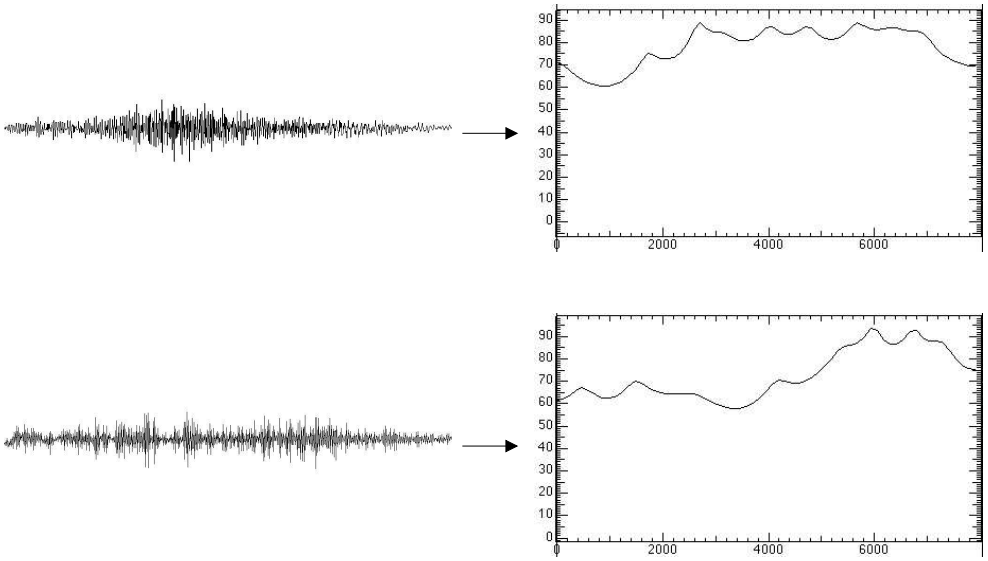


Abbildung 3.5: *Spektraldarstellung von Geräuschen.*

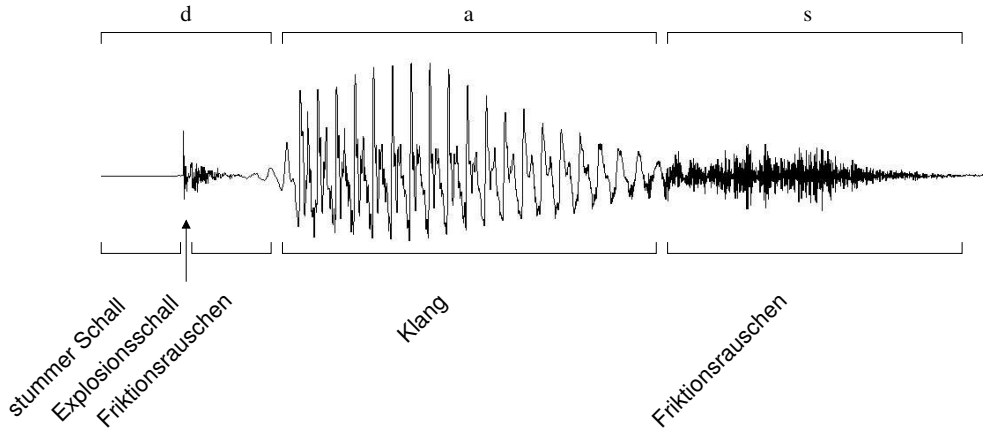


Abbildung 3.6: *Die Grundschallformen. (Äußerung:* [d̥as]*)*

3.2 Sprachschall

Grundschallformen

Stark vereinfachend können im Sprachschallsignal vier Grundschallformen unterschieden werden (Abb. 3.6):

Explosionsschall (Transiente): Entsteht bei der Sprengung eines oralen oder glottalen Verschlusses infolge von Überdruck; von kurzer Dauer; charakteristisch für alle Arten von Verschlusslauten (Plosive, Clicks, Implosive, Ejektive).

Friktionsrauschen: Verursacht durch Turbulenzen, wenn Luft durch eine Engebildung strömt; charakteristisch für alle Frikative, aber z.B. auch — dann von kürzerer Dauer — unmittelbar nach der Verschlusslösung bei Plosiven

Klang: Zurückzuführen auf die Phonation; Schallform der Vokale, Approximanten und Nasale, als zusätzliche Komponente auch bei stimmhaften Frikativen (zusammen mit Friktionsrauschen).

”stummer Schall”: Die Signalamplitude ist nahe Null, d.h. es ist kein Nutzschall hörbar; typischerweise bei stimmlosen Verschlusslauten während der Verschlussphase, gefolgt von einem Explosionsschall.

Als Modellvorstellung für die Transformation von — im weitesten Sinne — artikulatorischen Vorgängen in akustische Ereignisse hat sich das sog. Quelle–Filter–Modell (Gunnar Fant) durchgesetzt. Demnach ist die Erzeugung von

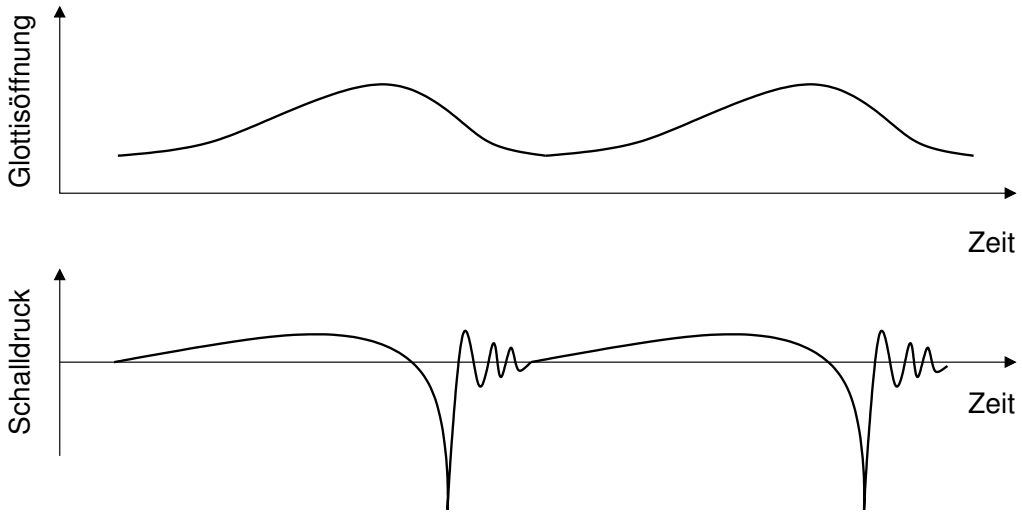


Abbildung 3.7: *Luftdruckschwankungen über der Glottis (unten) während der Phonation (oben); schematische Darstellung.*

Sprachschall ein zweistufiger Prozess: Zunächst wird ein Rohschall erzeugt, welcher dann modifiziert wird. Rohschall kann auf zwei verschiedene Arten erzeugt werden: (1) durch die Phonation (klangförmiger Rohschall, z.B. bei Vokalen)) oder (2) durch Geräuschbildung infolge einer glottalen oder supraglottalen Engebildung (geräuschförmiger Rohschall, z.B. bei stimmlosen Obstruenten).[5] Die Bezeichnung 'Rohschall' kommt daher, dass dieser Schall nie in seiner reinen Form wahrgenommen werden kann; so wird z.B. das Phonationssignal auf seinem Weg durch den Rachen und den Mundraum erheblich verändert. Man kann also sagen, das Phonationssignal ist noch 'roh' und trifft erst nach seiner 'Veredelung' in den Resonanzräumen des Sprechers auf das Ohr des Hörers.

Quelle–Filter–Modell

klangförmiger Rohschall

Die physiologischen Vorgänge bei der Phonation sind aus Abschnitt 1.1 bekannt (siehe auch Abbildung 1.5 auf Seite 19). Eine schematische Darstellung des Schalldruckverlaufs oberhalb der Glottis während der Phonation ist in Abbildung 3.7 dargestellt. Entscheidend für die Rohschallerzeugung ist der negative Druckimpuls bei Verschluss der Stimmlippen. Je prominenter dieser Impuls im Gesamtsignal, desto 'kräftiger', d.h. reicher an Resonanzen, ist die Stimme.

geräuschförmiger Rohschall

Die geräuschhafte Rohschallerzeugung basiert auf einer Luftverwirbelung

[5] Bei stimmhaften Frikativen wird der Rohschall durch eine Kombination dieser beiden Methoden erzeugt.

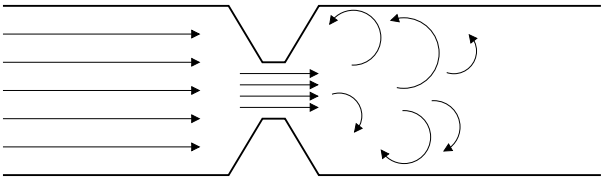


Abbildung 3.8: *Luftverwirbelung an einer Verengung; schematische Darstellung.*

infolge einer Engebildung. Dieser Vorgang ist schematisch in Abbildung 3.8 dargestellt. Die akustische Konsequenz einer solchen Turbulenz ist eine zufällige, aperiodische Veränderung des Luftstroms über die Zeit.

Zur Erzeugung von Sprachschall stehen uns also zwei Arten von Quellsignalen zur Verfügung: der klangförmige Rohschall und der geräuschhafte Rohschall. Die Unterscheidung verschiedener Lautklassen nach dem zugrundeliegenden Rohschall ist in folgender Tabelle zusammengefasst:

Quellsignal	**Lautklassen des Deutschen**
klangförmig (Phonation)	Vokale, (Lateral–) Approximanten, Nasale, Vibranten
geräuschhaft (Friktion)	stimmlose Obstruenten
klangförmig+geräuschhaft	stimmhafte Obstruenten

Nun zur zweiten Phase der Sprachschallerzeugung, der Modifikation des Rohschalls. Wie bereits erwähnt, kann der Hohlraum zwischen Glottis und Lippen — das sog. Ansatzrohr — als akustisches Filter beschrieben werden. Ein solches Filter verändert ein Quellsignal, indem es bestimmte Frequenzen selektiv verstärkt (die sog. Resonanzfrequenzen). Das Phänomen der Resonanzfrequenzen lässt sich am besten anhand eines klassischen Beispiels aus der allgemeinen Schwingungslehre darstellen, der gefederten Masse (Abb. 3.9). Wenn man die Masse nach unten zieht und los läßt, versetzt man das System in eine harmonische Schwingung. Bei gleicher Feder und gleicher Masse hat diese Schwingung immer die gleiche Frequenz, die sog. Eigenfrequenz.[6] Nachdem das System angeregt wurde, bleibt die Frequenz konstant, nicht je-

Resonanzfrequenzen

[6]Die mechanischen Eigenschaften einer Feder werden durch die sogenannte Federkonstante ausgedrückt; nimmt man eine Feder mir einer anderen Federkonstante, verändert sich die Frequenz. Mit einer steiferen Feder wird das System z.B. schneller schwingen, d.h. die Frequenz erhöht sich.

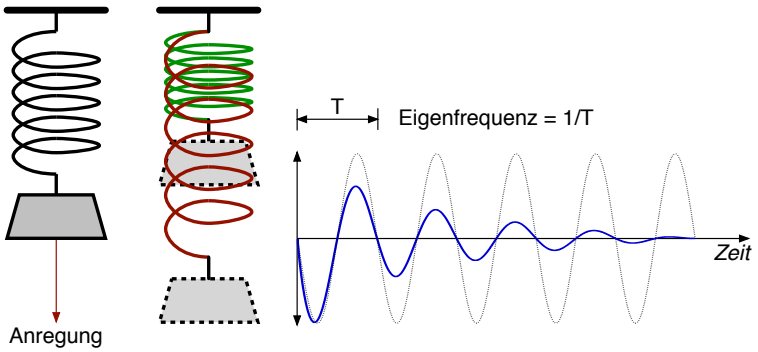


Abbildung 3.9: *Gefederte Masse: Nach Anregung beginnt das System mit seiner Eigenfrequenz zu schwingen, wobei die Frequenz konstant bleibt. Im Gegensatz zu einer idealisierten Schwingung (gepunktete Kurve), sind natürliche mechanische Schwingungen immer gedämpft (blaue Kurve), d.h. die Amplitude nimmt über die Zeit ab.*

doch die Amplitude, d.h. die Auslenkung nimmt aufgrund von Reibungsverlusten etc. immer mehr ab, bis die Schwingung zum Stillstand kommt. Diese Eigenschaft, die im übrigen allen mechanischen Schwingern gemeinsam ist, heißt Dämpfung.

Resonanz

Resonanz kommt ins Spiel, wenn anstatt der gefederten Masse das gesamte System angeregt wird, d.h. wenn man nicht das Gewicht nach unten zieht, sondern die Platte, an der die Feder aufgehängt ist, auf und ab bewegt (Abb. 3.10). Bewegt man die Platte mit der Eigenfrequenz des Systems auf und ab, reagiert das Gewicht darauf, indem es anfängt zu schwingen. Dieses Phänomen heißt Resonanz und die Frequenz, mit der das Gewicht zu schwingen beginnt, heißt Resonanzfrequenz; sie entspricht der Eigenfrequenz. Bewegt man nun die Platte etwas langsamer oder etwas schneller, wird das Gewicht mit der selben Resonanzfrequenz weiterschwingen, allerdings wird die Amplitude kleiner. D.h. die eingebrachte Energie wird nur dann optimal übertragen, wenn ein System im Bereich der Resonanzfrequenz angeregt wird. Je weiter die Frequenz des Anregungssignals von der Resonanzfrequenz abweicht, desto schlechter wird die Energie übertragen. Im Extremfall, wenn wir die Platte sehr langsam oder sehr schnell auf und ab bewegen, wird das Gewicht überhaupt nicht in Schwingung versetzt, d.h. von der Energie, die wir für die Anregung aufwenden, kommt nichts bei dem Gewicht an.

Nun zurück zum Sprechen, wobei wir uns bei der detailierten Beschreibung hier und im nächsten Kapitel im wesentlichen auf die Vokalproduktion beschränken werden. Das Anregungssignal für die Vokalproduktion ist das

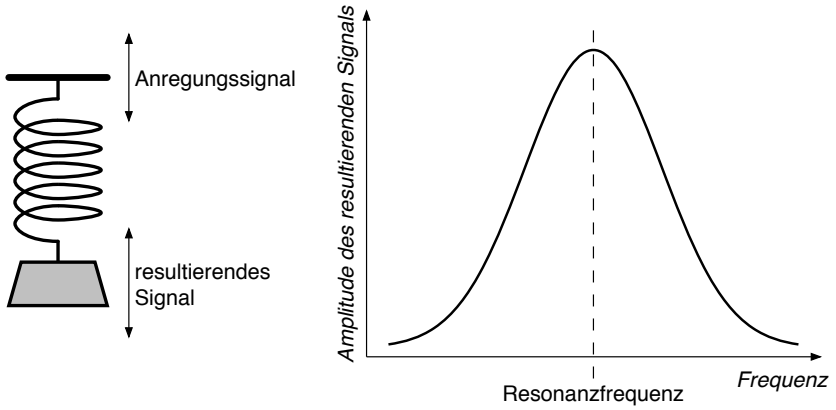


Abbildung 3.10: *Die Resonanzfunktion (rechts) eines mechanischen Schwingers (links). Die Resonanzfunktion beschreibt die Effizienz, mit der die Energie des Anregungssignals übertragen wird. Das Optimum liegt im Bereich der Resonanzfrequenz (= Eigenfrequenz) des angeregten Systems.*

phonatorische Signal — quasiperiodische Luftdruckschwankungen oberhalb der Glottis. Der mechanische Schwinger, der beim Sprechen angeregt wird, ist die unten (Glottis) geschlossene und oben (Lippen) offene Luftsäule im Ansatzrohr (Vokaltrakt). Da dieses System etwas komplexer ist als eine gefederte Masse, verfügt es nicht nur über eine, sondern über mehrere Resonanzfrequenzen (mehr dazu im nächsten Kapitel). Die Lage dieser Resonanzfrequenzen hängt ab von der Form des Ansatzrohrs, d.h. durch die Veränderung der Geometrie des Ansatzrohres (Bewegung der Zunge, des Kiefers, der Lippen etc.) können die Resonanzfrequenzen und damit die Filterwirkung variiert werden.[7] Im Gegensatz zu einem idealisierten Filter haben wir es beim menschlichen Ansatzrohr genau genommen nicht mit einzelnen, exakten Resonanzfrequenzen zu tun, sondern mit Frequenzbändern, d.h. in bestimmten Bereichen des Frequenzspektrums werden immer mehrere benachbarte Obertöne verstärkt. In einem Klangspektrum nennt man diese Bereiche Formanten. *Formanten*
Sie sind charakterisiert durch die Frequenz (Resonanzfrequenz; Lage des Kurvengipfels im Frequenzbereich) und durch die Bandbreite. Diese Zusammenhänge sind in Abbildung 3.11 (oben und unten links) zusammengefasst. In der selben Abbildung, unten in der Mitte und unten rechts, wird noch einmal ein wichtiger Aspekt des Quelle–Filter–Modells verdeutlicht, nämlich dass Quelle und Filter unabhängig voneinander sind und somit auch unabhängig

[7] Ausführlichere Darstellungen der Quelle–Filter–Theorie und insbesondere der Filterfunktion des Ansatzrohres finden sich in [13], S. 102 *ff* und [11], S. 105 *ff*.

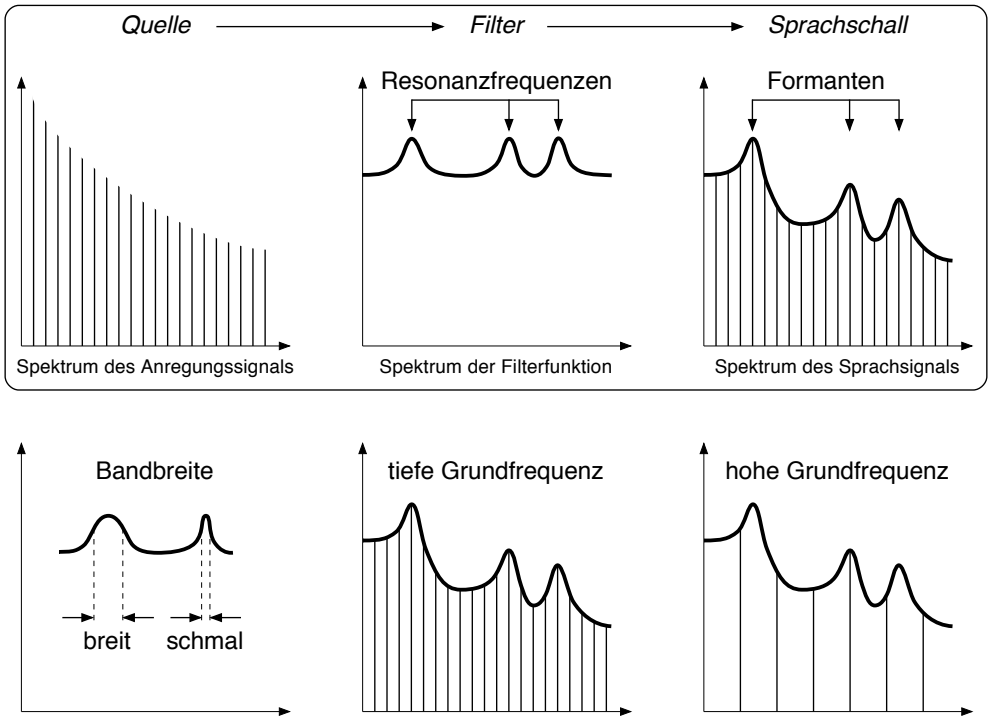


Abbildung 3.11: *Oben: Das Quelle–Filter–Modell dargestellt im Frequenzbereich. Unten, links: Unterschiedliche Bandbreiten. Unten, mitte und rechts: Variation der Grundfrequenz bei gleichbleibender Filterfunktion.*

voneinander verändert werden können. Außerdem wird deutlich, dass bei höheren Grundfrequenzen (z.B. bei Kindern) der Abstand zwischen den Harmonischen zunimmt, was dazu führen kann, dass Formanten an Distinktivität verlieren und Formantfrequenzen u.U. nicht mehr mit den Resonanzfrequenzen des Vokaltrakts übereinstimmen (in Abb. 3.11, unten rechts, bei der ersten und der dritten Resonanzfrequenz).

Die Position der Formanten im Spektrum ist ein entscheidendes akustisches Merkmal zur Unterscheidung von Vokalen. Aber auch Nasale und Liquide weisen eine charakteristische Formantstruktur auf. Bei der akustischen Unterscheidung von Frikativen spielen dagegen lokale Energiemaxima keine besondere Rolle — entscheidend ist hier die globale Form des Spektrums, d.h. die Energieverteilung in relativ breiten Frequenzbändern. Diese ist im wesentlichen zurückzuführen auf die Position, die Form und den Grad der Engebildung im Ansatzrohr, d.h. im Falle der Frikative hat schon das Quellsi-

gnal einen wichtigen Anteil an der spektralen Form des Ausgangssignals, die Filterung spielt hier eine untergeordnete Rolle.

3.3 Digitale Signalverarbeitung

Bevor wir tiefer in die akustische Analyse von Sprachschall einsteigen, folgen hier zunächst einige Anmerkungen zu den Instrumenten, die wir für solche Analysen verwenden: Computer–Hardware und –Software. Die heute übliche Verarbeitung von Sprachschall auf Computern (und auch die Aufnahme von gesprochener Sprache auf DAT–Bänder, CDs oder Festplatten) setzt eine tiefgreifende Manipulation des Untersuchungsgegenstandes voraus, nämlich die Umwandlung eines analogen Signals, wie es vom menschlichen Sprechapparat erzeugt wird, in ein digital repräsentiertes Signal, wie es vom Computer (und anderen digitalen Geräten) verarbeitet werden kann. Analoge Signale sind kontinuierliche Signale, z.B. kontinuierlich variierende Luftdruckschwankungen, die am besten durch eine Linie repräsentiert werden (Abb. 3.12, oben). Die Zeit– und Amplitudenwerte solcher Signale verfügen theoretisch über unendlich viele Nachkommastellen (womit ein Computer nicht umgehen kann). Digitale Signale werden dagegen besser durch eine Sequenz von separaten Punkten repräsentiert; in festgelegten Zeitintervallen auf der horizontalen Achse werden Amplitudenwerte auf der vertikalen Achse abgetragen (Abb. 3.12, unten) — digitale Signale sind also nicht kontinuierlich, sondern diskret. Das bedeutet auch, dass die Anzahl der Nachkommastellen sowohl bei den Zeit– als auch bei den Amplitudenwerten begrenzt ist.

kontinuierliche und diskrete Signale

Damit also ein Computer Schallwellen speichern und verarbeiten kann, muss das kontinuierliche analoge Signal in ein diskretes digitales Signal umgewandelt werden ('AD–Wandlung'; umgekehrt, wenn wir z.B. ein auf dem Computer gespeichertes Signal über Kopfhörer abspielen, spricht man entsprechend von 'DA–Wandlung'). Diese Umwandlung besteht im wesentlichen aus zwei Schritten: (1) Das Signal wird in regelmäßigen Zeitabständen abgetastet (*Sampling*), d.h. die Linie wird in eine Punktsequenz umgewandelt, Zeitwerte mit unendlich vielen Nachkommastellen werden in Zeitwerte mit endlich vielen Nachkommastellen konvertiert. (2) Das Signal wird quantisiert, Amplitudenwerte mit unendlich vielen Nachkommastellen werden in eine festgelegte Anzahl von Amplitudenstufen konvertiert. Wie oft ein Signal pro Zeiteinheit abgetastet wird (Abtastrate, *sampling rate*) und wie akkurat die Amplitudenwerte konvertiert werden (Abtasttiefe, Quantisierung,

AD/DA–Wandlung

Sampling und Quantisierung

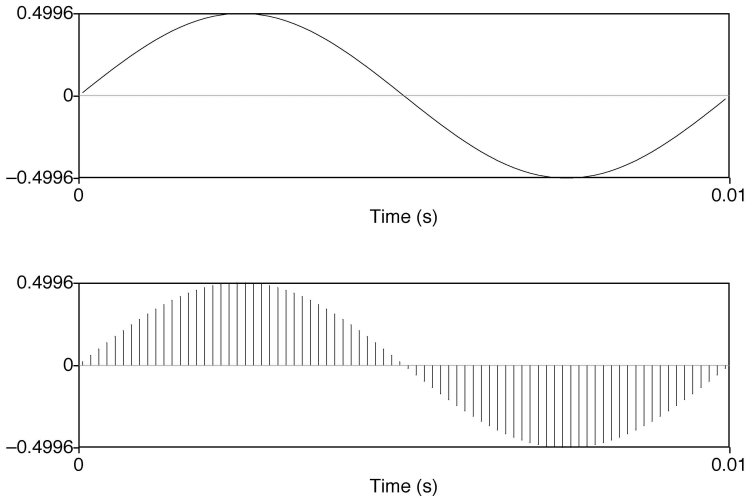


Abbildung 3.12: *Oben: Analoges Signal; kontinuierliche Liniendarstellung. Unten: Digitales Signal; diskrete Punktdarstellung; die einzelnen Abtastpunkte sind durch vertikale Striche mit der x–Achse verbunden.*

quantization) sind die wichtigsten Parameter bezüglich der Qualität der AD–Wandlung.

3.3.1 Abtastrate

Um ein analoges Signal möglichst detailliert digital zu repräsentieren, ist eine möglichst hohe Abtastrate wünschenswert, d.h. die Zeitabstände zwischen den einzelnen Abtastpunkten sollten gering sein. Andererseits bedeutet eine hohe Abtastrate auch hohen Speicherbedarf. Musikaufnahmen auf einer Audio–CD sind mit 44 kHz abgetastet, DAT–Aufnahmen meist sogar mit 48 kHz. Das bedeutet, dass das analoge Signal 48000–mal pro Sekunde abgetastet wird; bei einer Abtasttiefe von 2 Bit (s.u.) ergibt dies schon einen Speicherbedarf von 48000×2, also 96 kBit, bei einer üblichen Abtasttiefe von 16 Bit 48000×16 dementsprechend 768 kBit, was etwa 94 KB entspricht[8] — für eine Monoaufnahme von einer Sekunde[9]! Um Speicherplatz zu sparen wird man

[8] 1 Byte = 8 Bit, 1 KB (Kilobyte) = 1024 Byte; 768000/8 = 96000 Byte, 96000/1024 = 93,75 KB

[9] Eine fünfminütige Monoaufnahme mit 48 kHz und 16 Bit benötigt also 27,5 MB Speicherplatz: $(48000 \times 16 \times 300)/8 = 28800000$ Byte, 28800000/1024 = 28125 KB, 28125/1024 = 27,5 MB.

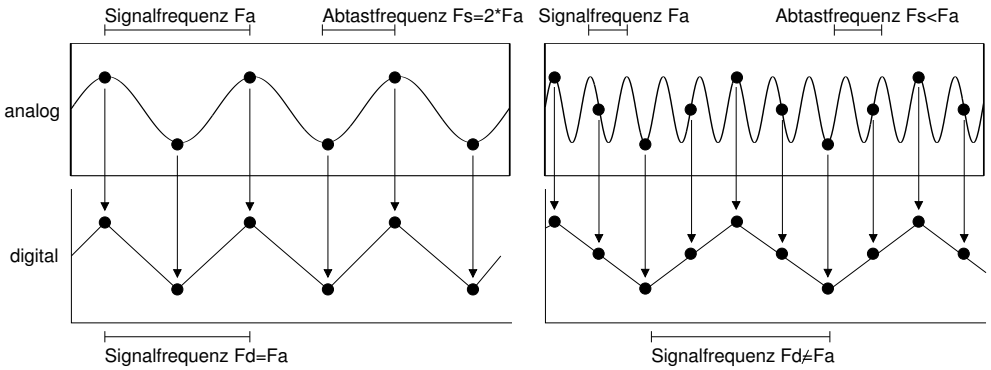


Abbildung 3.13: *Illustration des Abtasttheorems: Um eine Frequenz F_a in einem analogen Signal zu erfassen und digital zu repräsentieren (F_d), muss die Abtastfrequenz mindestens doppelt so hoch gewählt werden (links). Sind im analogen Signal Frequenzanteile enthalten, die oberhalb der Nyquist–Frequenz liegen, kommt es zu Aliasing: Die Frequenz des digital repräsentierten Signals F_d entspricht nicht der Frequenz des analogen Signals F_a (rechts).*

also DAT–Aufnahmen nicht unverändert auf den Computer übernehmen, um sie dort zu analysieren, sondern wird die Abtastrate verringern (*down sampling*). Was aber ist ein sinnvoller Kompromiss zwischen detaillgetreuer Repräsentation und sparsamem Umgang mit Speicherplatz, oder anders gefragt, welche Abtastrate ist für unseren Zweck — akustische Analyse gesprochener Sprache — empfehlenswert?

Nyquist–Frequenz

Zunächst gilt es, das Abtasttheorem bzw. die sogenannte Nyquist–Frequenz zu beachten: Um die Periodizität eines analogen Signals mit einer bestimmten Frequenz digital zu erfassen, muss das Signal mindestens mit der doppelten Frequenz abgetastet werden (vgl. Abbildung 3.13). Um also eine 100–Hz–Komponente in einem analogen Signal zu erfassen, muss mit einer Abtastrate von mindestens 200 Hz gearbeitet werden. Anders ausgedrückt: Wenn wir mit 200 Hz abtasten, liegt die Nyquist–Frequenz bei 100 Hz; die für unsere Analysen nutzbare Bandbreite beträgt 0 – 100 Hz.

relevante Frequenzkomponenten

Zweitens muss man sich klarmachen, welche Komponenten eines komplexen analogen Signals erfasst werden sollen. Da unser Gehör Frequenzen über 20 kHz nicht wahrnehmen kann, brauchen Frequenzkomponenten über 20 kHz im analogen Eingangssignal bei der AD–Wandlung nicht berücksichtigt zu werden, da sie in der lautsprachlichen Kommunikation sicher keine Rolle spielen. Das Telefon überträgt sogar nur Frequenzen bis 4 kHz; das Ergebniss ist zwar ein qualitativ schlechtes, aber durchaus verständliches Sprachsi-

gnal. Wenn man sich also mit Telefonqualität zufrieden gibt und sich auf eine Bandbreite von 0 – 4000 Hz beschränkt, genügt eine Abtastrate von 8 kHz. Möchte man sichergehen, dass alle Frequenzkomponenten bis zur Wahrnehmungsgrenze bei 20 kHz im digitalen Signal repräsentiert sind, muss man eine Abtastrate von min. 40 kHz wählen[10]. Eine empfehlenswerte und im phonetischen Bereich häufig genutzte Abtastrate liegt dazwischen: 22 kHz. Dies ist ausreichend, da die für die akustischen Eigenschaften von Sprachlauten relevante Bandbreite etwa von 50 Hz bis 10 kHz reicht[11].

Für eine akustisch–phonetische Analyse am Computer steht also nur eine Frequenz–Bandbreite zur Verfügung, die der Hälfte der Abtastrate entspricht; über Vorgänge in höheren Frequenzbereichen kann keine Aussage gemacht werden. So reicht z.B. das Rauschspektrum in Abbildung 3.5 bis 8 kHz, d.h. das zu analysierende Signal wurde mit 16 kHz abgetastet; über den spektralen Verlauf oberhalb 8 kHz erfahren wir nichts. Frequenzkomponenten des analogen Signals oberhalb der halben Abtastrate sind jedoch nicht nur einer Computer–Analyse unzugänglich, sondern können darüberhinaus auch die digitale Repräsentation verfälschen. Auf der rechten Seite von Abbildung 3.13 ist dieses als *Aliasing* bezeichnete Phänomen zu sehen: Angenommen die Frequenz des analogen Signals F_a beträgt 15 kHz, die Abtastfrequenz 14 kHz. Die Frequenz des resultierenden digitalen Signals F_d entspricht in diesem Fall 1 kHz und würde selbstverständlich in einem am Computer erzeugten Spektrum auftauchen — obwohl diese Frequenzkomponente im Originalsignal gar nicht vorhanden ist! Um solche Aliasing–Fehler auszuschließen, ist es unbedingt notwendig, vor der AD–Wandlung alle Frequenzkomponenten oberhalb der Nyquist–Frequenz aus dem analogen Signal zu entfernen. Dies geschieht mit Hilfe sogenannter Anti–Aliasing–Filter. Solche Filter lassen Frequenzen unterhalb der Nyquist–Frequenz passieren, während sie höhere Frequenzen blockieren[12]. Erst nach dieser Filterung wird das analoge Signal abgetastet und quantisiert. Es ist jedoch praktisch unmöglich, ein Filter zu konstruieren, welches bis zu einer bestimmten Frequenz alle Signalkomponenten passieren lässt, und ab dieser Frequenz alle Komponenten unterdrückt. Stattdessen gibt es einen Übergangsbereich: Unterhalb des Übergangsbereichs kann alles

Aliasing

Anti–Aliasing–Filter

[10] Daher werden CDs mit 44 kHz gesampelt: 20 kHz nutzbare Bandbreite, plus 2 kHz für das Anti–Aliasing–Filter (s.u.) = 22 kHz, gemäß Abtasttheorem verdoppelt ergibt dies 44 kHz.

[11] Bis vor kurzem war in der Phonetik sogar eine Abtastrate von nur 16 kHz üblich, also eine Nyquist–Frequenz von 8 kHz. Da jedoch Speicherplatz und Rechenleistung immer billiger werden, haben sich mittlerweile 22 kHz weitgehend durchgesetzt.

[12] Anti–Aliasing–Filter gehören daher zur Klasse der Tiefpassfilter (*low pass filter*).

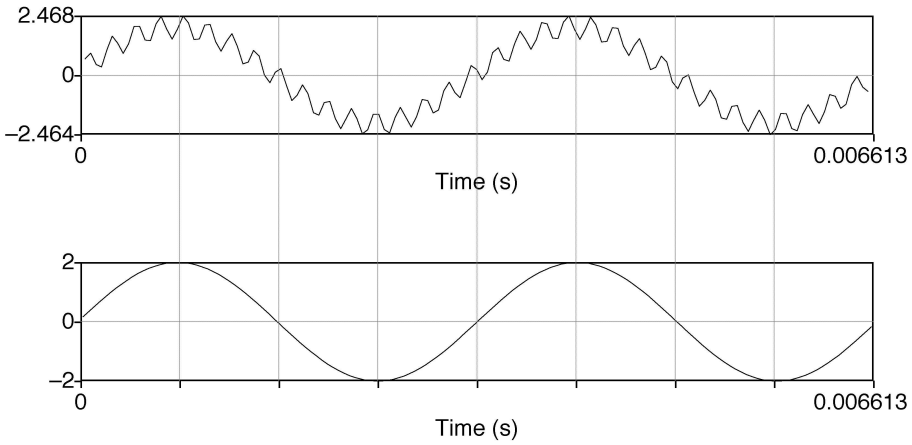


Abbildung 3.14: *Der Effekt eines Tiefpassfilters. Oben: Klang, zusammengesetzt aus einer 300 Hz und einer 5 kHz Sinusschwingung. Unten: Das selbe Signal nach einer Tiefpassfilterung mit Grenzfrequenz 3 kHz.*

passieren, oberhalb wird alles unterdrückt und die Frequenzen im Übergangsbereich selbst werden sozusagen langsam ausgeblendet[13].

Abbildung 3.14 zeigt exemplarisch den Effekt eines Tiefpassfilters. Das Originalsignal (3.14, oben) ist zusammengesetzt aus einer 300 Hz Sinusschwingung und einer 5000 Hz Sinusschwingung. Beide Frequenzkomponenten sind deutlich zu erkennen: Die höherfrequente Kurve 'reitet' auf der niedrigfrequenten Kurve. Dieses Signal wurde mit einem Tiefpassfilter, dessen Grenzfrequenz bei 3000 Hz festgelegt war, gefiltert (mit dem Programm Praat lassen sich solche Filter recht einfach realisieren). Im resultierenden Signal (3.14, unten) ist nur noch die 300–Hz–Schwingung zu sehen, die 5000–Hz–Komponente wurde komplett unterdrückt.

Tiefpassfilter

3.3.2 Quantisierung

Im Zusammenhang mit der Digitalisierung von Audiosignalen bedeutet Quantisierung die Übersetzung einer kontinuierlichen Amplitudenskala in eine diskrete Amplitudenskala mit einer zählbaren, d.h. endlichen Anzahl mög-

[13] Dies erklärt, weshalb CDs mit 44 kHz anstatt mit 40 kHz gesampelt werden. Die nutzbare Bandbreite reicht bis zu den gewünschten 20 kHz, darüber, bis zur Nyquist–Frequenz von 22 kHz, ist 'Platz' für den Übergangsbereich des Anti–Aliasing–Filters.

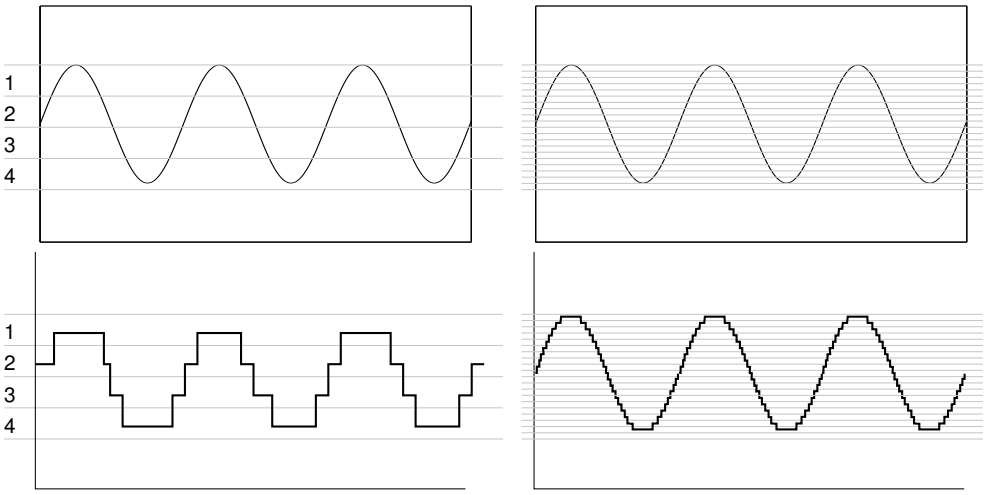


Abbildung 3.15: *Quantisierung: Konversion von kontinuierlichen Amplitudenwerten in eine diskrete 4–stufige Skala (links) bzw. in eine 20–stufige Skala (rechts).*

licher Werte[14]. Die Qualität der Quantisierung und des resultierenden digitalen Signals hängt entscheidend von der Größe der diskreten Amplitudenskala, d.h. von der Anzahl der möglichen Werte ab. Abbildung 3.15 zeigt den Unterschied zwischen einer 4–stufigen Quantisierung (links) und einer 20–stufigen Quantisierung (rechts) des selben analogen Signals (jeweils oben).

Ganz offensichtlich bildet die 20–stufige Skala das Originalsignal akkurater ab als die 4–stufige Skala. Die Größe der Amplitudenskala — die Abtasttiefe — wird in Bit (Binärzahlen) angegeben. Mit einer zweistelligen Binärzahl (2 Bit) kann der Computer vier verschiedene Werte darstellen ($2^2 = 4$), d.h. eine Abtasttiefe von 2 Bit führt zu einer 4–stufigen Skala wie in Abbildung 3.15, links. Im phonetischen Bereich (und auch auf Audio–CDs) ist eine Quantisierung mit 16 Bit üblich; dies ergibt eine sehr fein aufgelöste Amplitudenskala mit 65536 Stufen ($2^{16} = 65536$).

Abtasttiefe

[14]Quantisierung spielt natürlich bei allen Arten der AD–Wandlung eine Rolle. So spricht man z.B. auch beim Scannen von Bildern von Quantisierung. In diesem Zusammenhang bedeutet Quantisierung das Übersetzen einer kontinuierlichen analogen Farbskala in eine diskrete Farbskala mit einer endlichen Anzahl möglicher Farbwerte.

3.3.3 Fast Fourier Transformation

Die zwei vorangehenden Abschnitte behandelten mit der Abtastrate und der Quantisierung zwei grundlegende Konzepte der digitalen Signalverarbeitung. Es sollte deutlich geworden sein, dass zwischen analogen Signalen, wie sie uns in der realen Welt begegnen, und digitalen Signalen, wie sie uns für die Computeranalyse zur Verfügung stehen, ein grundsätzlicher Unterschied besteht und dass es bei der Konvertierung von Signalen zwischen diesen Welten einiges zu beachten gilt, um Fehler bzw. Artefakte bei der späteren Analyse zu vermeiden. Nach diesen notwendigen Anmerkungen zur Vorbereitung der Signale für die Analyse, soll es nun im letzten Abschnitt des Kapitels über digitale Signalverarbeitung um die Grundlagen der computerbasierten Analysemethoden selbst gehen, bevor dann das nächste Kapitel in die Anwendung dieser Methoden einführt.

Fourieranalyse

Eine der wichtigsten Methoden zur akustischen Untersuchung von Sprachschall ist die spektrale Analyse bzw. Fourieranalyse, d.h. die Zerlegung eines komplexen Signals — nämlich des Sprachsignals — in seine Frequenzbestandteile. Das Ergebnis dieser Analyse kann in unterschiedlicher Form dargestellt und interpretiert werden, z.B. als Amplitudenspektrum (s. Abschnitt 3.1) oder als Spektrogramm (s. Abschnitt 3.4.2). Das Standardverfahren zur Durchführung einer Spektralanalyse am Computer ist die *Fast Fourier Transformation (FFT)*; es handelt sich dabei um einen Algorithmus, der die Fourieranalyse diskreter Signale (*Discrete Fourier Transform, DFT*) besonders effizient implementiert.[15]

Zeit– vs. Frequenzauflösung

Für die praktische Anwendung der FFT ist eine Eigenschaft dieses Algorithmus von besonderer Bedeutung, nämlich die gegenseitige Abhängigkeit zwischen zeitlicher Auflösung und Frequenzauflösung. Vereinfacht gesagt: Soll die zeitliche Auflösung einer Analyse verbessert werden, so muss man unweigerlich Einbußen bei der Frequenzauflösung hinnehmen, und umgekehrt, unter einer verbesserten Frequenzauflösung leidet die zeitliche Auflösung. Der Grund dafür ist, dass der Frequenzbereich zwischen 0 Hz und der Nyquist–Frequenz durch eine bestimmte Anzahl diskreter Punkte mit festem Abstand repräsentiert wird. Je mehr Punkte hier zur Verfügung stehen, desto

[15] Außer der FFT gibt es noch andere Verfahren zur spektralen Analyse digitaler Signale, z.B. Wigner–Verteilung oder die *Wavelet*–Analyse. Diese Verfahren werden jedoch (noch?) selten verwendet und sind auch in der gängigen Software nicht implementiert. Etwas anders verhält es sich mit *Linear Predictive Coding (LPC)*; dieses Verfahren findet häufiger Anwendung, und zwar als Alternative zu FFT–Amplitudenspektren insbesondere bei der Analyse von Vokalen. In [6] findet sich ein kurzer Abschnitt, der erklärt, wie LPC funktioniert. Das Ergebnis gleicht einem geglätteten Breitbandspektrum (s.u.).

geringer ist der Abstand zwischen den einzelnen 'Messpunkten', desto besser ist folglich die Frequenzauflösung (vergleichbar der Quantisierung: je mehr Stufen zur Verfügung stehen, desto feiner die Amplitudenauflösung; s. Abb. 3.15). Nun ist jedoch die Anzahl der Punkte, die den Frequenzbereich bei einer FFT–Analyse repräsentieren, kein frei wählbarer Parameter, sondern entspricht genau der Anzahl der Abtastpunkte (*samples*), die in die Analyse eingehen. Die Anzahl der Abtastpunkte wird durch die Größe des sogenannten Analysefensters (*analysis window*) festgelegt. Die Analyse eines einzelnen Abtastpunktes macht offensichtlich keinen Sinn, da es sich bei der Fourier–Transformation um eine Frequenzanalyse handelt, d.h. der Algorithmus benötigt Informationen darüber, wie sich das Signal über die Zeit verändert. Mit der Größe des Analysefensters legen wir fest, wie groß der Signalabschnitt ist, der dem FFT–Algorithmus für die Analyse zur Verfügung steht. Wählt man einen relativ großen Abschnittt, erhält der Algorithmus vergleichsweise viel Information über den Signalverlauf und der Frequenzbereich von 0 Hz bis zur Nyquist–Frequenz kann mit vielen Stufen fein aufgelöst werden. Allerdings wird alles aus diesem Signalabschnitt 'in einen Topf' geworfen — enthält das Analysefenster z.B. einen ganzen Diphtong, so erhalten wir zwar ein sehr fein aufgelöstes Durchschnittsspektrum, erfahren jedoch nichts über die artikulatorische Dynamik, die sich z.B. in ausgeprägten Formantbewegungen zeigt. Um solche dynamischen Aspekte, die in der akustischen Phonetik sehr charakteristisch und wichtig sind, berücksichtigen zu können, müssen kürzere Signalabschnitte analysiert werden, d.h. das Analysefenster muss kürzer gewählt werden. Damit wird man dynamischen Veränderungen im Signal besser gerecht — die zeitliche Auflösung wird feiner —, der Preis ist jedoch, dass dem FFT–Algorithmus nun weniger Analysepunkte zur Verfügung stehen — die Frequenzauflösung wird gröber. Abbildung 3.16 veranschaulicht, was feine (links) bzw. grobe (rechts) Frequenzauflösung in der Praxis bedeuten: Mit der feinen Auflösung ('Schmalbandspektrum') kann man die einzelnen Harmonischen erkennen, während die grobe Auflösung ('Breitbandspektrum') nur die globale Form des Spektrums zeigt.

Analysefenster

Schmalband– vs. Breitbandspektrum

Die Größe des Analysefensters wird meist mit der Anzahl der Abtastpunkte angegeben[16], manchmal jedoch auch als Zeitangabe (z.B. im Programm Praat). Eine feine Frequenzauflösung erhält man mit einem 1024–Punkte–Fenster (Abb. 3.16, links); dies entspricht bei 22 kHz Abtastrate einem Signalabschnittt von 46,6 Millisekunden (bei 16 kHz Abtastrate entsprechen

[16]Normalerweise werden hier 2er–Potenzen verwendet, also z.B. 64, 128, 512 oder 1024 Punkte.

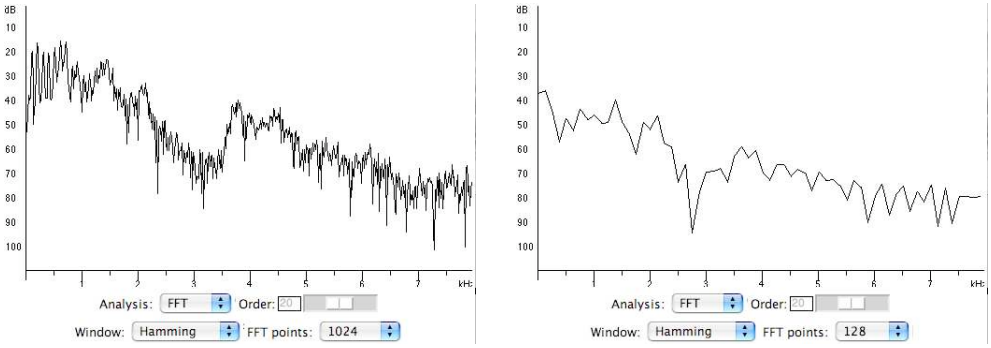


Abbildung 3.16: *Schmalband–FFT–Spektrum (1024 Punkte, links) und Breitband–FFT–Spektrum (128 Punkte, rechts) des Vokals* [a]*, Abtastrate 16 kHz (Programm:* WAVESURFER*).*

1024 Punkte 64 ms). Eine gute Zeitauflösung erhält man etwa mit einem 128–Punkte–Fenster (22 kHz: 5,8 ms; 16 kHz: 8 ms) (Abb. 3.16, rechts). Die Zeitauflösung ergibt sich also aus der Multiplikation der Anzahl der Abtastpunkte mit dem Abstand zwischen diesen Punkten in Sekunden: Bei einer Abtastrate von 22 kHz beträgt der Abstand zwischen zwei Abtastpunkten $1/22000 = 0,046$ Sekunden, multipliziert mit 128 ergibt dies 5,8 ms. Die Frequenzauflösung erhält man, indem man die Abtastrate durch die Anzahl der Abtastpunkte im Analysefenster teilt: Eine 1024–Punkte–FFT bei 22 kHz Abtastrate ergibt eine Frequenzauflösung von $22000/1024 = 21,5$ Hz; eine 128–Punkte–FFT bei gleicher Abtastrate ergibt dementsprechend eine Frequenzauflösung von $22000/128 = 171,9$ Hz.[17]

Form des Analysefensters

Neben der Größe des Analysefensters lässt sich bei den meisten Programmen auch dessen Form bestimmen. Der naiven Vorstellung eines Fensters entspricht am ehesten das sogenannte Rechteckfenster (*rectangle window*). Hierbei werden innerhalb des ausgewählten Signalabschnitts von allen Abtastpunkten die originalen Amplitudenwerte übernommen. Wenn Anfang und Ende des Fensters jedoch nicht zufällig auf Abtastpunkte mit dem Amplitudenwert 0 fallen — was in der Tat sehr unwahrscheinlich ist — kann dies insofern problematisch sein, als dass sich im resultierenden Spektrum Artefakte

[17] Zur Erinnerung: Wir bekommen natürlich immer nur die Hälfte der FFT–Punkte zu sehen — von 0 Hz bis zur Nyquist–Frequenz. Bei der angesprochenen 1024–Punkte–Analyse sehen wir also beispielsweise 512 'Messpunkte' eines Amplitudenspektrums im Abstand von 21,5 Hz, der Frequenzbereich reicht von 0 bis 11 kHz. (Meist werden nicht die einzelnen Punkte dargestellt, sondern eine durch Interpolation gewonnene Linie; vgl. Abb 3.16.)

zeigen[18]. Abhilfe schafft z.B. das häufig verwendete glockenförmige *Hamming*–Fenster, das die Amplitudenwerte zu den Fensterrändern hin langsam 'ausblendet'.

3.4 Grundlagen der akustischen Analyse

In diesem Abschnitt werden einige grundlegende Methoden der akustischen Analyse beschrieben. Ausführliche Anwendungsbeispiele würden den Rahmen dieses Skripts sprengen. Da jedoch die technischen Voraussetzungen zur Durchführung akustischer Analysen heute praktisch überall vorhanden sind, möchte ich Sie ermutigen, mit den hier vorgestellten Methoden selbst zu experimentieren. Während früher aufwendige und teuere Spezialgeräte notwendig waren, lassen sich die meisten Analysen heute mit einem normalen PC durchführen. Daneben benötigt man nur noch eine geeignete Software[19] und ein einigermaßen brauchbares Mikrophon[20].

3.4.1 Signal und Intensität

Oszillogramm

In der Signaldarstellung (Oszillogramm) sind Amplitudenwerte (y–Achse) über der Zeit (x–Achse) abgetragen. Das Signal kann hier — basierend auf den Grundschallformen — segmentiert werden, um z.B. Lautdauern zu messen. Diese einfachste Form der Darstellung bietet damit die Möglichkeit, z.B.

[18]Der Grund hierfür ist, dass sich in diesem Fall die Fensterränder als Transienten darstellen — was sie in Wirklichkeit natürlich nicht sind — und charakteristische spektrale Muster produzieren. Für andere Analyseverfahren (wie RMS, LPC oder *pitch tracking*, s. nächstes Kapitel) ist das Rechteckfenster jedoch durchaus geeignet.

[19]Mittlerweile gibt es sehr viele Programme zur akustischen Analyse von Sprache. Das bekannteste kommerzielle Produkt, das auch im klinischen Bereich häufig eingesetzt wird, ist das *Computerized Speech Lab (CSL)* von Kay Elemetrics. Aber auch kostenlos gibt es zahlreiche, zum Teil sehr mächtige Programme; hier eine kleine Auswahl: *Wavesurfer* bietet relativ wenige Funktionen, ist aber für den Anfang völlig ausreichend; vorallem ist es einfach zu bedienen (`www.speech.kth.se/wavesurfer/`). Sehr viel komplexere Analysen sind mit *Praat* (`www.fon.hum.uva.nl/praat/`) und dem *Speech Filing System (SFS)* (`www.phon.ucl.ac.uk/resource/sfs/`) möglich. Beide Programme haben jedoch ein gewöhnungsbedürftiges Bedienkonzept und setzten eine gewisse Einarbeitungszeit voraus. *Wavesurfer* und *Praat* gibt es für *Windows*, *Mac OS X* und *Linux*, *SFS* gibt es nur für *Windows*.

[20]Eingebaute Mikrophone, z.B. in Laptops, sind nicht brauchbar, da sie hauptsächlich Computergeräusche (Festplatte, Lüfter etc.) aufnehmen. Besser geeignet ist ein Mikrophon mit langem Kabel, sodass die Aufnahme möglichst weit entfernt vom Computer (und anderen Störgeräuschen) gemacht werden kann.

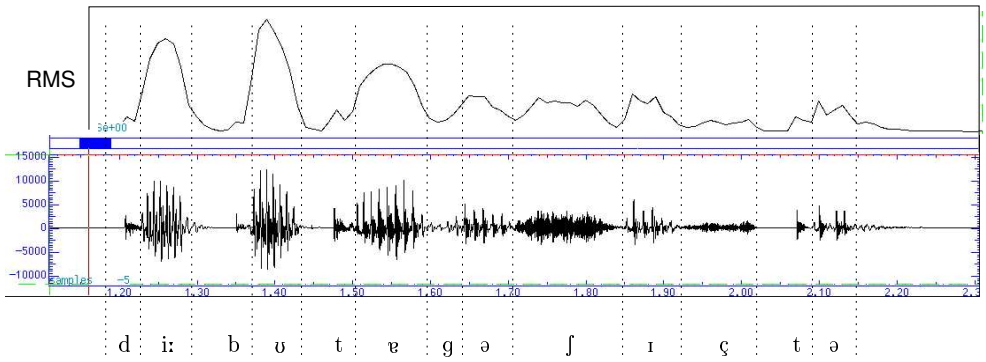


Abbildung 3.17: *Oszillogramm (unten) und RMS–Kurve (oben) der Äußerung* [diːbʊtɐgəʃɪçtə].

die Fähigkeit eines Sprechers zur Unterscheidung von Lang– und Kurzvokalen zu prüfen und zu objektivieren. Oder es kann überprüft werden, ob betonte Silben bzw. Vokale in betonten Silben im Vergleich mit ihrem unbetonten Pendant gelängt werden. Die Signaldarstellung eignet sich auch zur Messung von VOTs (es gibt allerdings exaktere Methoden, die in der einschlägigen Literatur beschrieben sind) und zur Messung der Sprechgeschwindigkeit (z.B. gemessen als Anzahl der Silben pro Zeiteinheit). Nicht zuletzt ist das Oszillogramm auch bestens geeignet, Sprachaufnahmen gezielt und kontrolliert auditiv zu analysieren, da alle Programme, die ein Signal darstellen können, auch das Markieren und wiederholte Abspielen von beliebigen Ausschnitten erlauben. Damit ist der Computer auch ein viel besseres Werkzeug zur Transkription von Aufnahmen als z.B. ein Kassettenrekorder.

Obwohl im Oszillogramm die Amplitude (Schalldruck) über der Zeit abgetragen ist, ist es schwierig, allein anhand des Oszillogramms Aussagen über die Lautintensität zu machen, da es sich bei Sprachaufnahmen in der Regel um komplexe Signale mit positiven und negativen Amplitudenwerten handelt. Besser geeignet hierfür ist eine Verlaufsdarstellung der sog. RMS–Werte (RMS = *root mean square*). Zur Berechnung des RMS–Wertes werden die Amplitudenwerte eines kleinen Signalabschnitts quadriert (damit bekommt man z.B. das Problem der negativen Werte in den Griff), vom Ergebnis wird der Mittelwert gebildet und aus diesem die Wurzel gezogen. Die Werte für aufeinanderfolgende Signalabschnitte werden dann über der Zeitachse abgetragen. Die resultierende Kurve zeigt den Intensitätsverlauf im Sprachsi-

RMS–Werte

Intensität

gnal[21]. In Abbildung 3.17 ist dies am Beispiel der Äußerung *die Buttergeschichte* ([diːbʊtɐgəʃɪçtə]) dargestellt. Es ist zu sehen, dass Vokale die größte Intensität aufweisen; allerdings wird auch deutlich, dass die Intensität zum Äußerungsende abnimmt: Die durchschnittliche Intensität von [ɪ] in der vorletzten Silbe ist kaum größer als die von [ʃ]. Im Vergleich der beiden Frikative zeigt sich, dass [ç], wie oben erwähnt, eine geringere Intensität aufweist als [ʃ]. Das Intensitätsmaximum auf [ʊ] deutet darauf hin, dass der Sprecher die Hauptbetonung in dieser Äußerung auf die erste Silbe des Wortes *Buttergeschichte* gelegt hat.

3.4.2 Spektrographie

Wie bereits erwähnt ist das Spektrum eine statische Darstellungsform, es enthält keine Informationen über den Zeitverlauf. In die Berechnung eines Spektrums gehen alle Signalanteile innerhalb eines definierten Signalabschnitts ein (Analysefenster); eventuelle Veränderungen des Signals innerhalb des Analysefensters sind im Spektrum nicht mehr sichtbar. Aber auch langsame Veränderungen, die sich über längere Zeit erstrecken (z.B. Formanttransitionen) sind in einem einzelnen Spektrum nicht darstellbar. Hierfür ist es notwendig, mehrere Spektren hintereinander zu erzeugen, also das Analysefenster auf der Zeitachse sukzessive nach rechts zu verschieben. Das Resultat dieser Methode kann in Form eines sog. Wasserfalldiagramms dreidimensional dargestellt werden (Abbildung 3.18). Die Spektren sind entlang der Zeitachse (x) aufgereiht; die Frequenz ist auf der y–Achse abgetragen, die Amplitude in z–Richtung.

Spektrogramm

Üblicher, weil besser 'lesbar', ist jedoch die Darstellung in Form eines Spektrogramms (Spektrogramme werden oft auch als Sonagramme bezeichnet). Das Problem, 3 Dimensionen (Zeit, Frequenz und Amplitude) in einer 2–dimensionalen Grafik unterzubringen, wird bei der Spektrographie dadurch gelöst, dass eine Dimension — nämlich die Amplitude — durch Graustufen (bzw. verschiedene Farben) repräsentiert wird: Geringe Amplituden werden durch hellere Grautöne (bzw. blassere Farben) dargestellt, hohe Amplituden durch dunklere Grautöne (bzw. intesivere Farben). Die Entstehung eines Spektrogramms Schritt für Schritt (schematisch dargestellt in den Abbildungen 3.19 und 3.20): (1) Kodierung der spektralen Amplituden in Graustu-

[21] RMS ist tatsächlich ein Intensitätsmaß und im strengen Sinne keine Messung der akustischen Amplitude. Da jedoch die wahrgenommene Lautheit (*loudness*) eher mit der Intensität als mit dem Amplitudenverlauf korreliert, ist im phonetischen Bereich die RMS–Methode üblicher als z.B. die Vermessung von Amplitudenspitzen.

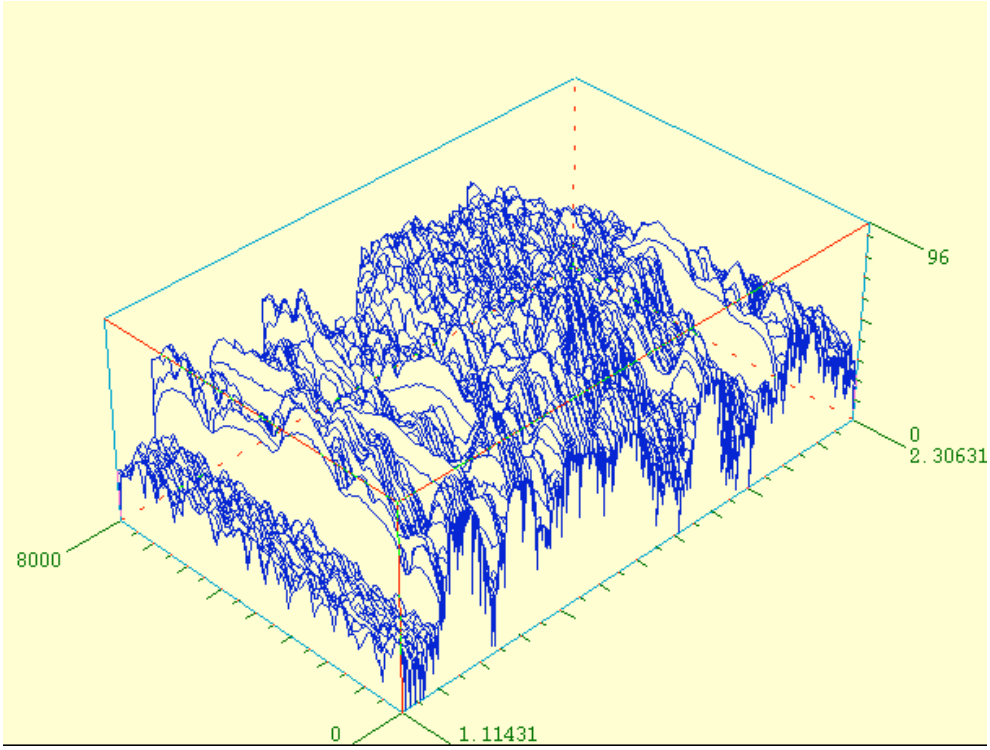


Abbildung 3.18: *3d–Darstellung mehrerer Spektren (Wasserfalldiagramm); Äußerung:* [diːbʊtɐgəʃɪçtə].

fen; (2) Drehen des Spektrums um 90 Grad gegen den Uhrzeigersinn, die Frequenz wird nun auf der y–Achse abgetragen, die x–Achse ist ungenutzt; (3) Aneinanderreihen mehrerer aufeinanderfolgender Graustufenspektren auf der x–Achse (Zeitachse). Formanten und vorallem Formantverläufe stellen sich im Spektrogramm somit als dunkelgraue horizontale Balken dar. Abbildung 3.20 zeigt dies schematisch und in einem 'echten' Spektrogramm am Beispiel des Diphtongs [a͡ɪ]. Abhängig davon, aus welcher Art von Spektren das Spektrogramm erzeugt wurde, unterscheidet man Schmalband– und Breitbandspektrogramme. Spektrogramme erben die Eigenschaften der zugrundeliegenden FFT–Spektren (s. Abschnitt 3.3.3): Schmalbandspektrogramme haben eine gute Frequenzauflösung aber eine schlechte Zeitauflösung; die in der Phonetik im allgemeinen bevorzugten Breitbandspektrogramme haben umgekehrt eine gute Zeitauflösung aber eine schlechtere Frequenzauflösung.

Schmalband– und Breitbandspektrogramme

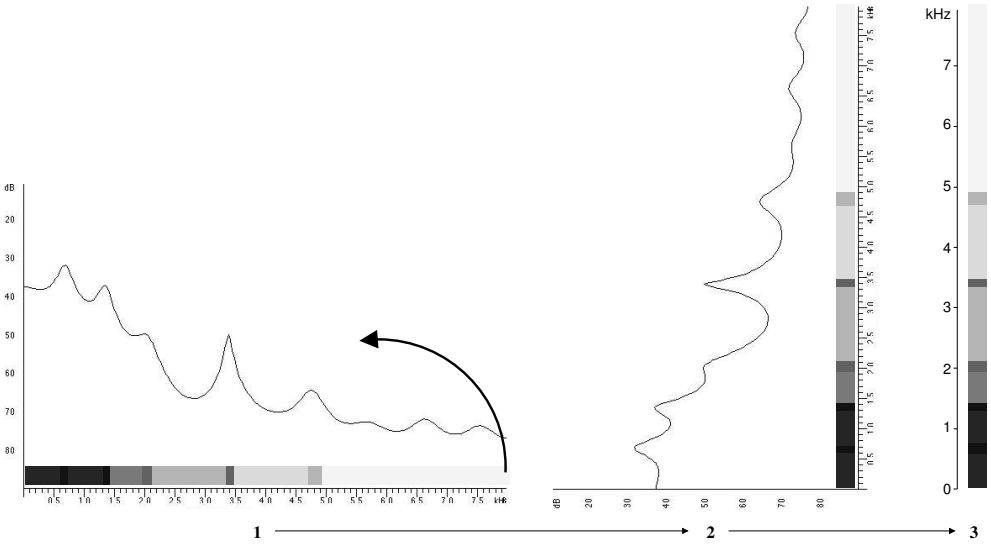


Abbildung 3.19: *Vom Spektrum zum Spektrogramm I: Amplitudenwerte im Amplitudenspektrum werden in Graustufen kodiert (1); das Spektrum wird um 90 Grad gedreht (2); die Frequenz wird auf der y–Achse abgetragen, die x–Achse ist ungenutzt (3).*

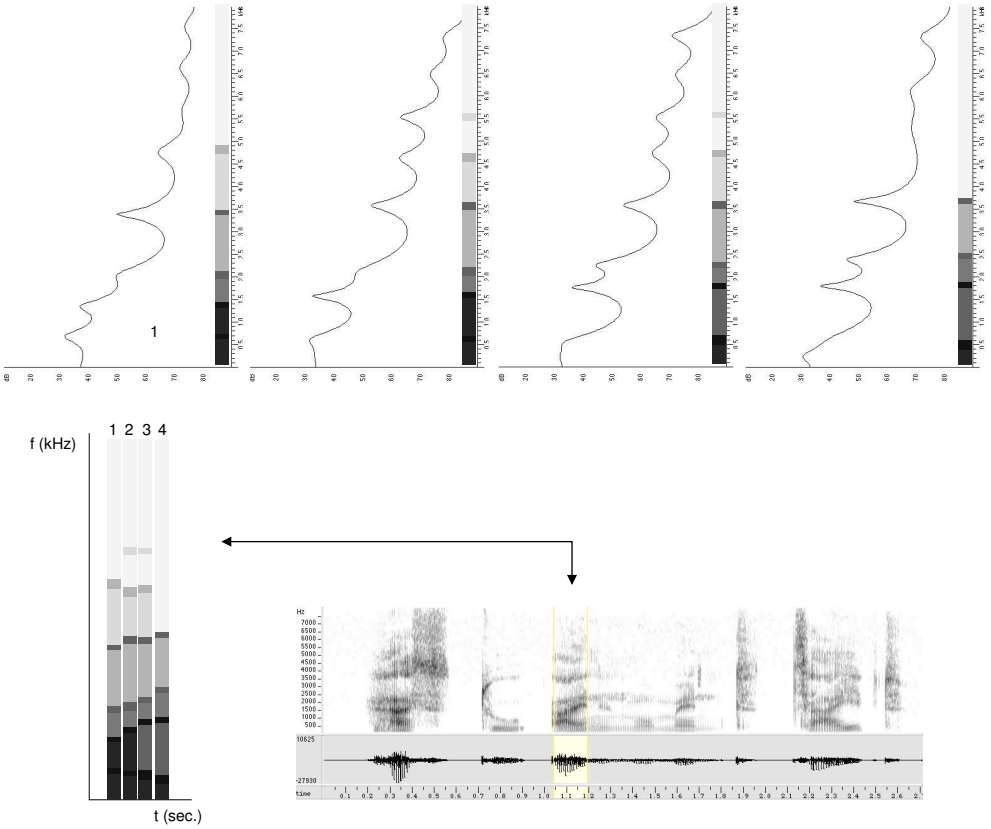


Abbildung 3.20: *Vom Spektrum zum Spektrogramm II. Oben: 4 aufeinanderfolgende Spektren am Anfang, nach dem ersten Drittel, nach dem zweiten Drittel und am Ende des Diphtongs* [aɪ̯] *erstellt; in Graustufen kodiert und um 90 Grad gedreht. Unten: Aneinanderreihung der 4 schematischen Graustufenspektren auf der Zeitachse (links) und der entsprechende Abschnitt markiert in einem Breitbandspektrogramm (Gesamtäußerung:* Hast du einen Moment Zeit?*).*

Spektrogramme lesen

Abbildung 3.21 zeigt das Spektrogramm und das Oszillogramm der Äußerung [diːbʊtɐgəʃɪçtə]. Im Folgenden einige Erläuterungen dazu.

Segmentierung. Lautgrenzen lassen sich relativ gut an einer abrupten Veränderung der Amplituden (= Graufärbung) und der spektralen Struktur erkennen. Während der Verschlussphase am Anfang ist das Spektrogramm weiß; die Verschlusslösung korreliert mit einem plötzlichen Amplitudenanstieg im gesamten Frequenzbereich (senkrechter grauer Balken); die Affriktionsphase stellt sich ebenfalls als grauer Balken dar, allerdings beschränkt auf Frequenzbereiche oberhalb ca. 2000 Hz; danach folgt ein deutlicher Amplitudenabfall oberhalb 5000 Hz und eine vokaltypische harmonische Struktur zwischen 0 und 5000 Hz; die Lage der Formanten kann grob bestimmt werden: F1 unter 500 Hz, F2 über 2000 Hz; der Übergang vom Vokal zur Verschlussphase des nachfolgenden Konsonanten ist wiederum sehr abrupt; durch die Hälfte der Verschlussphase zieht sich ein schwacher horizontaler grauer Balken im Bereich der Grundfrequenz des Sprechers: die *voice bar*. Sehr deutlich zu sehen ist auch die Abgrenzung des Friktionsrauschens der beiden Frikative [ʃ] und [ç] gegenüber der harmonischen Struktur der benachbarten Vokale.

Voice bar

Lautidentifikation. Die Lautklassen lassen sich anhand des Spektrogramms in der Regel gut erkennen: Vokale zeigen eine harmonische Struktur mit schmalen horizontalen Schwärzungen (Formanten); Frikative erkennt man an einer breitbandigen Graufärbung ohne ausgeprägte horizontale Strukturierung; bei Plosiven lassen sich zumeist sogar die einzelnen Phasen (Verschluss, Plosion, Affriktion/Aspiration) unterscheiden. Am Beispiel von [g] wird jedoch auch deutlich, dass bei schwach oder unvollständig gebildeten Verschlusslauten auch die charakteristischen spektrographischen Merkmale 'verschwinden'. Der für die Verschlussphase typische Amplitudenabfall ist sehr kurz, die *voice bar* bricht nicht ab und der bei der Plosion zu erwartende senkrechte graue Balken fehlt völlig. Solche Reduktionsprozesse sind bei flüssigem Sprechen nicht selten, und das Spektrogramm ist eine gute Methode, diese sichtbar zu machen. Dass der Sprecher jedoch zumindest den velaren Artikulationsort angesteuert hat (wenn es auch nicht unbedingt zu einem vollständigen Verschluss gekommen ist), sieht man ebenfalls im Spektrogramm — nämlich an den Transitionen der benachbarten Vokale (s.u.).

Lautklassen

Für den geübteren 'Leser' sind nicht nur die Lautklassen, sondern sogar einzelne Laute im Spektrogramm identifizierbar. Vokale (Nasale und Appro-

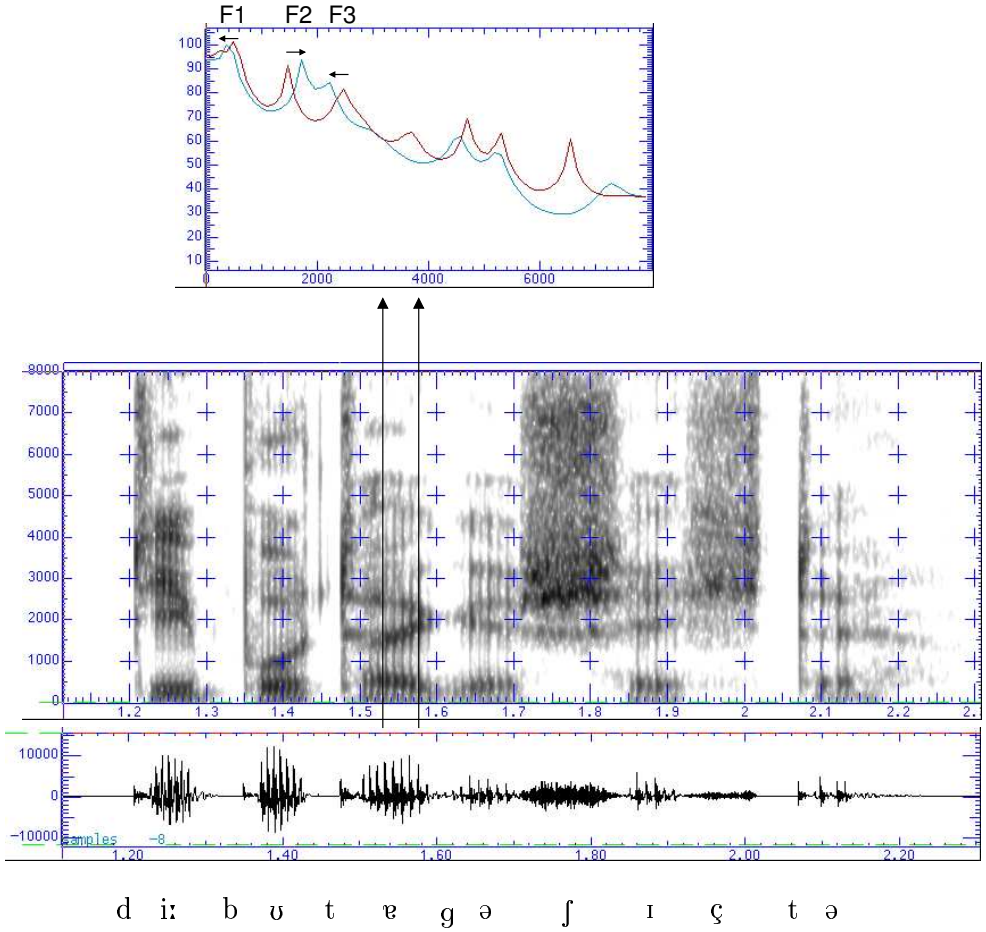


Abbildung 3.21: *Oszillogramm (unten) und Spektrogramm (Mitte) der Äußerung* [diːbʊtɐgəʃɪçtə]*; oben zwei Spektren des Lautes* [ɐ]*, zu verschiedenen Zeitpunkten berechnet.*

ximanten) anhand der Lage der Formanten, Frikative anhand der globalen Energieverteilung, Plosive an den Transitionen (s. Kapitel 4).

Dynamik. In der spektrographischen Darstellung wird deutlich, dass Sprechen ein dynamischer, kontinuierlicher Prozess ist. So gut die einzelnen Laute voneinander abgrenzbar sind, so ist doch offensichtlich, dass beim Sprechen nicht einzelne, unveränderliche Laute aneinander gereiht werden, sondern dass benachbarte Laute 'weich' ineinander übergehen, dass sie miteinander verzahnt sind und dass sich Laute während des Sprechens praktisch ständig

Sprechen als dynamischer, kontinuierlicher Prozess

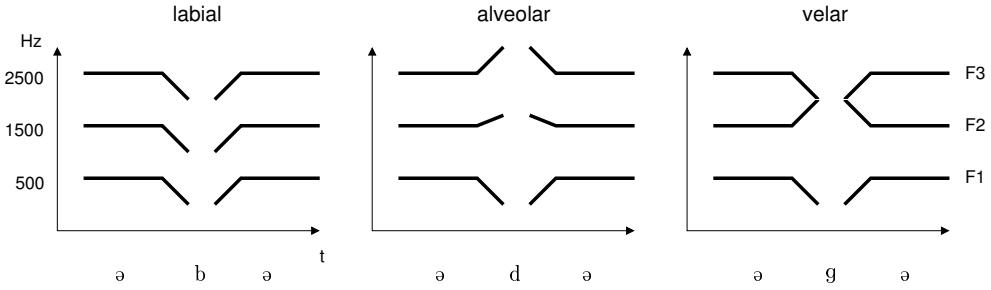


Abbildung 3.22: *Formanttransitionen in Abhängigkeit vom Artikulationsort.*

verändern. Besonders gut zu sehen ist dies am zweiten Formanten von [ɐ]. Dieser entsteht schon während der Aspirationsphase von [t], d.h. der Vokaltrakt bewegt sich unmittelbar nach der alveolaren Verschlusslösung in die für die Vokalproduktion notwendige Position und wird von der glottalen Friktion angeregt. Diese Bewegung des Vokaltrakts resultiert in einer Abwärtsbewegung des Formanten. Kaum ist der Formant an seinem tiefsten Punkt angelangt, beginnt eine Aufwärtsbewegung, d.h. der Vokaltrakt verändert seine Form in Vorbereitung des velaren Verschlusslautes. Diese durchgehende u–förmige Bewegung des Formanten zeigt, dass sich die Artikulatoren während der gesamten Vokalproduktion bewegen; es gibt praktisch keine stabile Phase, in der die Artikulatoren an einer imaginären Zielposition für einige Millisekunden zur Ruhe kämen.

Transitionen

Diese sogenannten Formanttransitionen verdeutlichen nicht nur die Dynamik des Sprechens, sondern sind auch ein wichtiges Merkmal zur Identifikation benachbarter Konsonanten. Besonders ausgeprägt sind die Transitionen vor oder nach Verschlusslauten (vgl. Abb. 3.22). So zeigt sich nach einem labialen Verschlusslaut normalerweise eine Aufwärtsbewegung der drei ersten Formanten (F1, F2 und F3). Nach einem alveolaren Verschlusslaut steigt F1 an, F3 fällt ab; die Bewegung von F2 hängt von dessen Zielposition ab: liegt diese oberhalb etwa 1800 Hz (z.B. bei einem [i]), steigt der Formant an, liegt sie unterhalb 1800 Hz (z.B. bei einem [u]), fällt er ab. Nach einem velaren Verschlusslaut steigt F1 an, F2 und F3 bewegen sich von einem gedachten Punkt (dem sog. 'Lokus') bei etwa 2000 Hz in Richtung ihrer Zielposition. Die Transitionen vor dem Konsonanten verhalten sich spiegelbildlich. In Abbildung 3.21 ist dies besonders gut zu sehen im Kontext des velaren Konsonanten [g]: F1 von [ɐ] fällt ab, F1 von [ə] steigt an; die zweiten und dritten Formanten bewegen sich zu einem gedachten Punkt bei ca. 2000 Hz hin bzw.

von diesem weg (vgl. Spektren von [ɐ] über dem Spektrogramm in Abb. 3.21, berechnet in der Mitte und am Ende des Lautes).

3.4.3 Grundfrequenzkonturen

Um die Sprachmelodie sichtbar zu machen, können aus einem gegebenen Signal sukzessive die Grundfrequenzwerte berechnet und über der Zeitachse dargestellt werden (*pitch tracking*). Anhand einer solchen Darstellung kann z.B. überprüft werden,

- ob ein Sprecher die Unterscheidung zwischen Frageintonation (steigender Ton am Satzende; Abb. 3.23 unten) und Aussageintonation (fallender Ton am Satzende; Abb. 3.23 oben) beherrscht; *Intonation*
- wo in einer Äußerung tonal markierte Akzente realisiert werden (Abb. 3.23);
- welcher Stimmumfang (Umfang der Grundfrequenzvariation) einem Sprecher zur Verfügung steht (Abb. 3.24). *Stimmumfang*

So kann z.B. der auditive Eindruck des 'monotonen Sprechens' objektiviert werden.

Grundfrequenzkonturen müssen mit Vorsicht interpretiert werden. Ein grundsätzliches Problem stellen die Fehler der Berechnungsalgorithmen dar; sie sind in Form unsinniger Extremwerte und abrupter Sprünge im Konturverlauf relativ einfach zu erkennen und dürfen selbstverständlich nicht in eine Analyse eingehen. Die Häufigkeit solcher Fehler hängt unter u.a. von der Aufnahmequalität und der Stimmqualität des Sprechers ab. Problematisch sind z.B. *creaky–voice*–Passagen, aber auch bei Stimmpatienten, beispielsweise mit Diplophonie, ist Vorsicht angeraten. Keine Fehler, aber ein Problem für die Interpretation globaler Konturverläufe (um die es hier geht) stellen sog. mikroprosodische Einflüsse dar. So dürfen z.B. kleine Bewegungen der Grundfrequenzkontur im Kontext von Verschlusslauten nicht als tonaler Akzent interpretiert werden. *Mikroprosodie*

Zu Abbildung 3.24: Die sehr geringe Variationsbreite der Grundfrequenz beim unteren Sprecher (ca. 50 Hz) ist nur eines von mehreren Merkmalen monotonen Sprechens. Die sehr gleichförmigen, immer wieder kehrenden Bewegungsmuster der Grundfrequenzkontur (vergleichen Sie den Melodieverlauf beim oberen Sprecher) tragen ebenso dazu bei wie die kurzen, durch Pausen getrennten Phrasen (zu erkennen im Oszillogramm). *Monotones Sprechen*

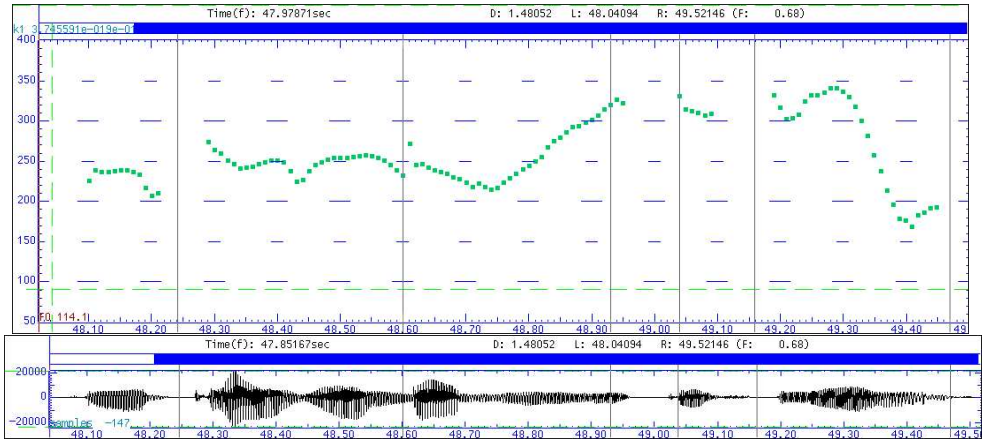


Abbildung 3.23: *Grundfrequenzkonturen von zwei Äußerungen einer weiblichen Sprecherin (Spontansprache); tonal markierte Akzente: zweite Silbe von "Ber***lin***" (steigend), "***links***" (fallend), erste Silbe von "***Non***nenweiher" (steigend/fallend).*

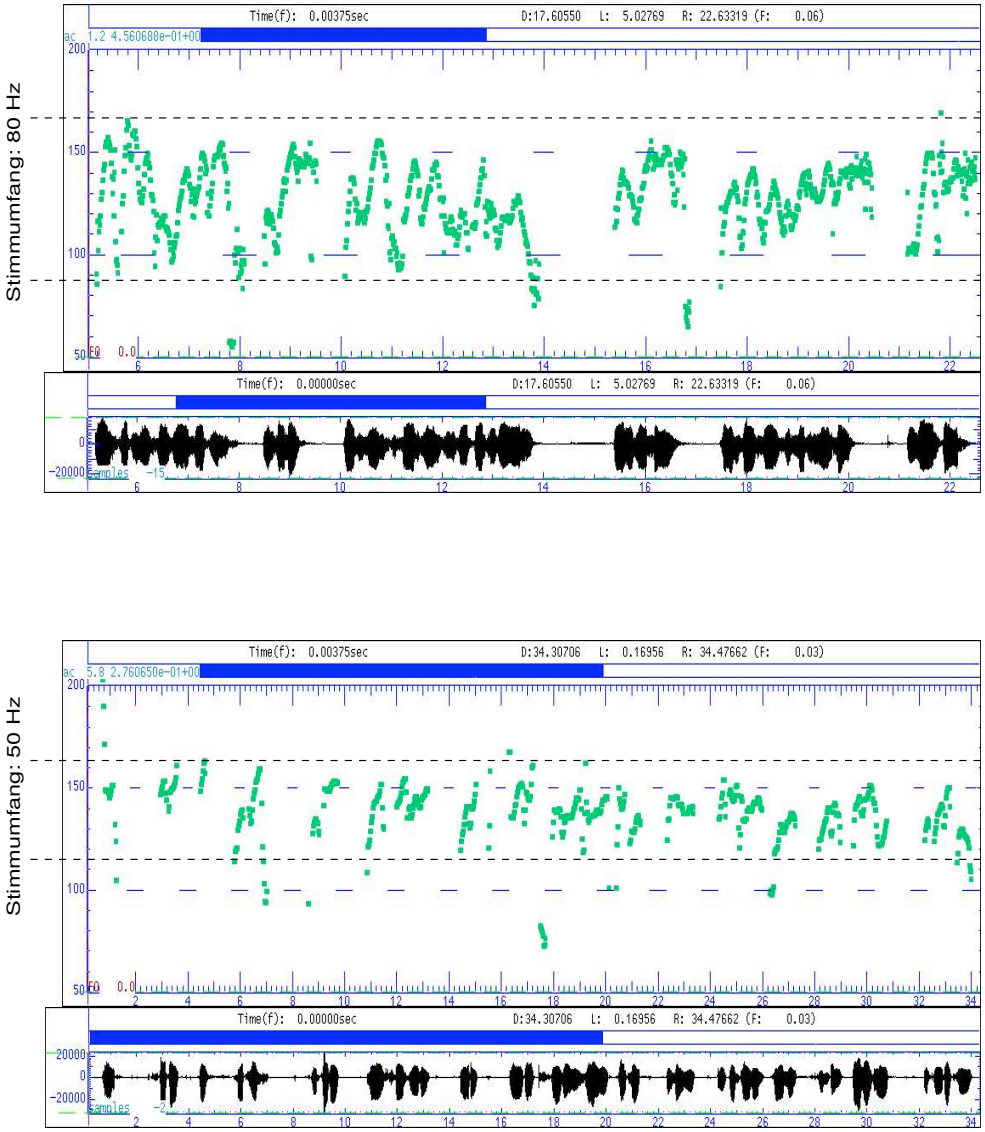


Abbildung 3.24: *Anfangspassage von 'Nordwind und Sonne', gelesen von einem männlichen Patienten mit schlaffer Dysarthrie (oben) und einem männlichen Patienten mit dyskinetisch hyperkinetischer Dysarthrie (Chorea Huntington) (unten).*

Kapitel 4

Akustische Eigenschaften der verschiedenen Lautklassen

4.1 Vokale

Vokalische Spektren verfügen über eine harmonische Struktur, d.h. sie sind als Linienspektrum darstellbar. Die erste Linie entspricht der Grundfrequenz, mit der der Laut produziert wurde (Frequenz der Stimmlippenschwingungen); diese korreliert mit der wahrgenommenen Tonhöhe (vgl. Abschnitt 2.3.2). Die weiteren Linien sind ganzzahlige Vielfache der Grundfrequenz. Entscheidend für die Vokalqualität ist die Verteilung lokaler Energiemaxima und –minima im Spektrum — die Formantenstruktur. Die Formanten in einem Spektrum entsprechen den Resonanzfrequenzen einer bestimmten Vokaltraktkonfiguration (vgl. Abschnitt 3.2). Sie werden durchnummeriert, wobei man mit dem Formanten mit der niedrigsten Frequenz beginnt (F1, F2, F3 etc.). Die wichtigsten Formanten zur perzeptiven Unterscheidung von Vokalen sind der erste (F1) und der zweite Formant (F2), mit Einschränkungen auch noch der dritte (F3). Diese liegen bei erwachsenen Sprechern unterhalb von ca. 3 – 3,5 kHz. Höhere Frequenzen (und Formanten) spielen zwar für die Lauterkennung eine untergeordnete Rolle, enthalten jedoch wichtige Informationen für die Sprechererkennung.

Grundfrequenz

Formanten

Zum besseren Verständnis, wie Formanten entstehen und wie ihre Lage im Spektrum abhängig von artikulatorischen Veränderungen variiert, dient die Modellvorstellung des Ansatzrohres. Da weder die Querschnittsform des Ansatzrohres (rund, viereckig, mehreckig...) noch die Biegung des Ansatzrohres im Bereich des Velums eine besondere Rolle hinsichtlich der Resonanz

Ansatzrohr

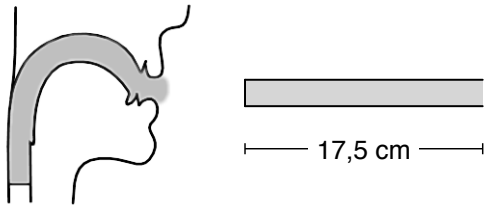


Abbildung 4.1: *Ansatzrohr: Der supraglottale Trakt (links) und das vereinfachte Röhrenmodell (rechts).*

spielen, geht man vereinfachend von einer geraden, kreisrunden Röhre mit schallreflektierenden Wänden aus, die auf einer Seite geschlossen (Glottis) und auf einer Seite offen ist (Lippen). Bei einem männlichen Erwachsenen beträgt die Distanz von der Glottis bis zu den Lippen, also die Länge des Ansatzrohres, etwa 17,5 cm (Abb. 4.1).

Die Luftsäule im Ansatzrohr wird durch das Anregungssignal in Schwingung versetzt. Wenn man annäherungsweise und vereinfachend davon ausgeht, dass bei einem neutralen, zentralen Vokal ([ə]) die Querschnittsfläche über die gesamte Länge des Ansatzrohres gleich bleibt, ist die Länge der entscheidene Parameter für die Lage der Resonanzfrequenzen. Durch Schallreflektion innerhalb des einseitig geschlossenen menschlichen Ansatzrohrs bilden sich sogenannte stehende (Schall–) Wellen aus. Die stehenden Wellen haben ihr Druckschwankungsmaximum am geschlossenen Ende (Glottis), während am offenen Ende Gleichdruck herrscht (Druckschwankungsminimum, atmosphärischer Druck). Der Abfall vom (positiven oder negativen) Druckschwankungsmaximum am einen Ende zum Druckschwankungsminimum (Nulldurchgang) am anderen Ende entspricht mindestens 1/4 Periode, die Wellenlänge dieser längsten stehenden Welle entspricht somit der vierfachen Ansatzrohrlänge: $4 \times 17,5 = 70$ cm (Abb. 4.2).[1] Von der nächst kürzeren Welle passt eine 3/4 Periode in das Ansatzrohr (Wellenlänge = 23,3 cm), dann eine 5/4 Periode (Wellenlänge = 14 cm) usw. (Abb. 4.3).

Stehende Wellen

Wellenlänge

[1] Analog zur Periodendauer, die die Ausbreitung einer Schwingung in der zeitlichen Dimension beschreibt, beschreibt die Wellenlänge die Ausbreitung einer Schwingung in der räumlichen Dimension. Die Wellenlänge entspricht der räumlichen Ausdehnung einer Periode. Sofern die Ausbreitungsgeschwindigkeit (c) einer Welle bekannt ist (in unserem Fall also die Schallgeschwindigkeit) kann aus der Wellenlänge (λ) mit $f = \frac{c}{\lambda}$ die Frequenz einer Schwingung berechnet werden.

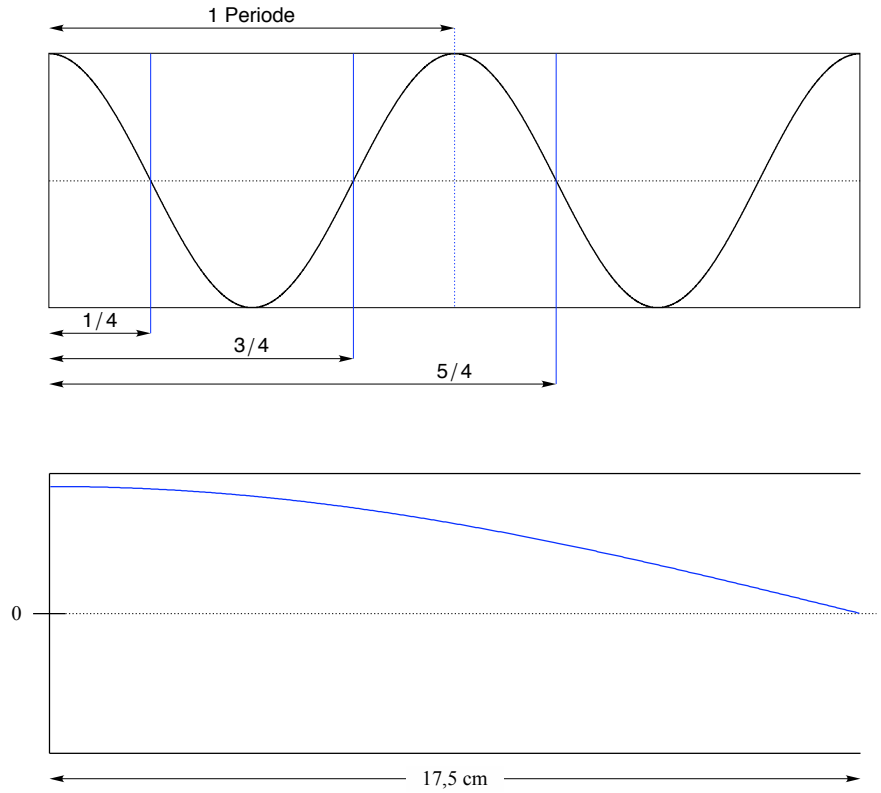


Abbildung 4.2: *Für den Weg von einem positiven oder negativen Maximalausschlag bis zu einem Nullduchgang benötigt eine Sinusschwingung mindestens 1/4 Periode; der nächste Nulldurchgang wird nach einer 3/4 Periode erreicht, dann nach einer 5/4 Periode usw (oben). Wenn 1/4 Periode 17,5 cm zurücklegt, dann beträgt die Wellenlänge der Schwingung* $4 \times 17,5 = 70$ *cm.*

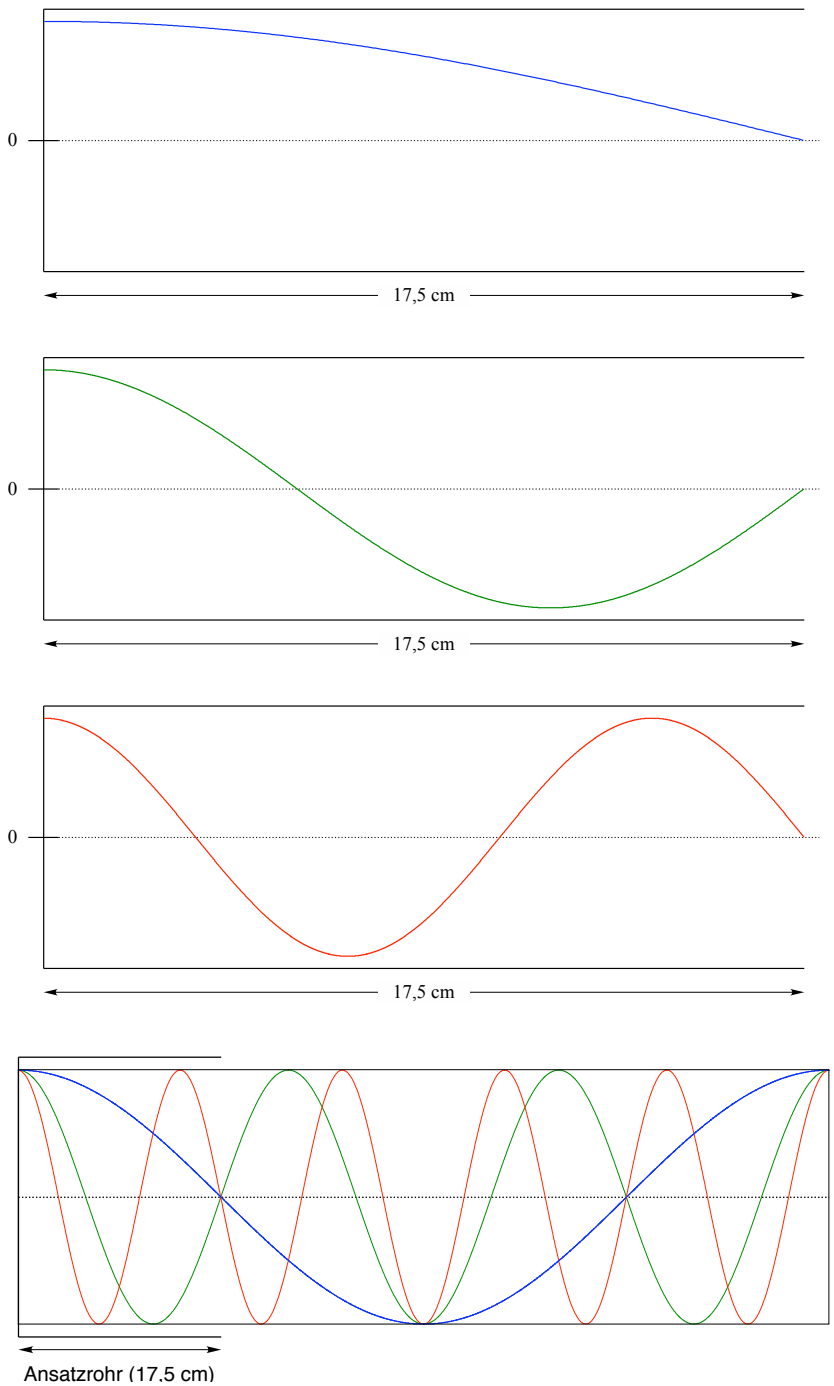


Abbildung 4.3: *Die ersten drei stehenden Wellen im Ansatzrohr. Unten der direkte Vergleich der unterschiedlichen Frequenzen.*

Die den Wellenlängen der stehenden Wellen zugeordneten Frequenzen entsprechen den Resonanzfrequenzen des Ansatzrohres. Die Frequenz f einer Schallwelle ergibt sich aus der Schallgeschwindigkeit c des Mediums geteilt durch die Wellenlänge λ:

$$f = \frac{c}{\lambda}$$

Die Schallgeschwindigkeit in Luft beträgt ca. 340 m/s bei 20° Celsius und ca. 350 m/s bei Körpertemperatur. Damit ergeben sich die folgenden ersten drei Resonanzfrequenzen des Ansatzrohrs (Formanten) in neutraler Stellung, d.h. ohne spezifische Verengung oder Erweiterung (≈ [ə]):

- Ansatzrohr 17,5 cm (= 0,175 m, erwachsener Mann):
  - $f_1 = \frac{c}{\lambda_1} = \frac{350m/s}{\frac{1}{0,25} \times 0,175m} = 500\frac{1}{s} = 500$ Hz
  - $f_2 = \frac{c}{\lambda_2} = \frac{350m/s}{\frac{1}{0,75} \times 0,175m} = 1500$ Hz
  - $f_3 = \frac{c}{\lambda_3} = \frac{350m/s}{\frac{1}{1,25} \times 0,175m} = 2500$ Hz
- Ansatzrohr ca. 14,6 cm (erwachsene Frau):
  - $f_1 = \frac{c}{\lambda_1} = \frac{350m/s}{\frac{1}{0,25} \times 0,146m} = 600$ Hz
  - $f_2 = \frac{c}{\lambda_2} = \frac{350m/s}{\frac{1}{0,75} \times 0,146m} = 1800$ Hz
  - $f_3 = \frac{c}{\lambda_3} = \frac{350m/s}{\frac{1}{1,25} \times 0,146m} = 3000$ Hz
- Ansatzrohr ca. 8,75 cm (kleines Kind):
  - $f_1 = \frac{c}{\lambda_1} = \frac{350m/s}{\frac{1}{0,25} \times 0,0875m} = 1000$ Hz
  - $f_2 = \frac{c}{\lambda_2} = \frac{350m/s}{\frac{1}{0,75} \times 0,0875m} = 3000$ Hz
  - $f_3 = \frac{c}{\lambda_3} = \frac{350m/s}{\frac{1}{1,25} \times 0,0875m} = 5000$ Hz

Daraus lässt sich schon die erste wichtige Regel für die Korrelation zwischen Vokaltraktkonfiguration (Artikulation) und Lage der Formantfrequenzen (Akustik) herleiten, die Längeregel:

Längeregel

> Die Lage der Formantfrequenzen ist umgekehrt proportional zur Länge des Ansatzrohrs: Je kürzer das Ansatzrohr desto höher die Formantfrequenzen.

Hier spielt natürlich zunächst einmal Alter und Geschlecht eine wichtige Rolle, aber es gibt durchaus auch die Möglichkeit, die Länge des Ansatzrohres im Rahmen der Artikulation willkürlich zu verändern. So führt die Verlängerung des Ansatzrohrs durch Vorstülpen der Lippen oder Absenken des Kehlkopfs zu einem leichten Absenken der Formantfrequenzen.

Der Sprechapparat bietet jedoch sehr viel mehr Möglichkeiten, durch Verengung oder Erweiterung die Querschnittsfläche des Ansatzrohres zu verändern als durch Längung oder Kürzung. Daher kommt den folgenden Verengungsregeln mehr Bedeutung zu als der Längeregel. Im folgenden wird der Einfluss von artikulatorischen Verengungen und Erweiterungen auf den ersten und den zweiten Formanten diskutiert, wobei immer von der neutralen artikulatorischen Ausgangslage ([ə]) ausgegangen wird. Der Einfluss einer artikulatorischen Konfiguration auf F1 und F2 wird dann jeweils mit Bezug auf die oben hergeleiteten Werte als Abweichungen nach oben oder unten angegeben. Die allgemeine Regel für Veränderungen der Querschnittsfläche gegenüber der neutralen Stellung lautet:

Allgemeine Verengungsregel

> Bei Verengung nahe eines Druckschwankungsmaximums wird die Formantfrequenz höher, bei Verengung nahe eines Druckschwankungsminimums niedriger.
>
> Bei Erweiterung nahe eines Druckschwankungsmaximums wird die Formantfrequenz niedriger, bei Verengung nahe eines Druckschwankungsminimums höher.

Damit lassen sich bei Bedarf auch weitere Regeln für die höheren Formanten F3, F4 etc. herleiten.

Um die Auswirkung der allgemeinen Verengungsregel auf den ersten Formanten einschätzen zu können, betrachten wir noch einmal die erste Resonanzfrequenz des Ansatzrohres, also die stehende Welle mit der größten Wellenlänge. Diese hat genau ein Druckschwankungsmaximum am geschlossenen Ende und genau ein Druckschwankungsminimum am offenen Ende.

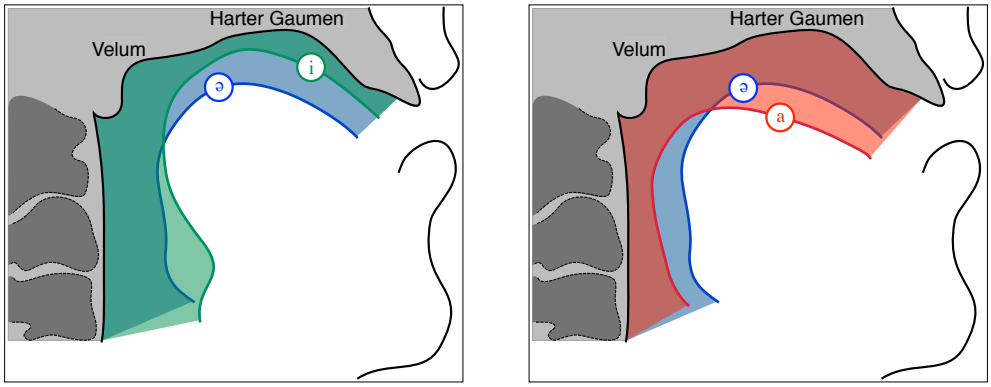


Abbildung 4.4: *Querschnittsfläche des Ansatzrohres im Vergleich mit der neutralen Stellung ([ə]): Hohe Vokale ([i], links) und tiefe Vokale ([a], rechts).*

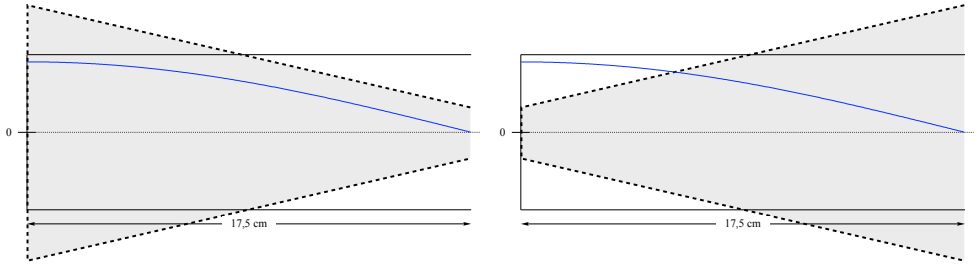


Abbildung 4.5: *Röhrenmodell des Ansatzrohres: Hohe Vokale (links) und tiefe Vokale (rechts).*

Nun zur Vokalartikulation: Bei hohen Vokalen wie [i] oder [u] ist die Querschnittsfläche in der vorderen Hälfte des Ansatzrohres (Druckschwankungsminimum) geringer als bei der neutralen Stellung, in der hinteren Hälfte (Druckschwankungsmaximum) dagegen größer (Abb. 4.4, links). Bei tiefen Vokalen wie [a] ist es umgekehrt, die Querschnittsfläche in der vorderen Hälfte des Asatzrohres ist größer, in der hinteren Hälfte geringer (Abb. 4.4, rechts). Abbildung 4.5 zeigt diese Zusammenhänge im vereinfachten Röhrenmodell.

Damit läßt sich aus der allgemeinen Verengungsregel eine spezifische Regel für den ersten Formanten herleiten:

Der erste Formant ist im Vergleich zur Neutralstellung niedriger

Verengungsregel F1

- bei Verengung des vorderen Teils des Ansatzrohres
- bei Erweiterung des hinteren Teils des Ansatzrohres

Der erste Formant ist im Vergleich zur Neutralstellung höher

- bei Erweiterung des vorderen Teils des Ansatzrohres
- bei Verengung des hinteren Teils des Ansatzrohres

Hohe Vokale haben demnach von allen Vokalen den tiefsten ersten Formanten. Bei halbhohen Vokalen ist die Abweichung vom F1–Wert der Neutralstellung entsprechend geringer und tiefe Vokale haben den höchsten F1 (vgl. Tabelle 4.1).

Beim zweiten Formanten sind die Verhältnisse etwas komplizierter, weil die zweite stehende Welle (3/4 Periode) bereits jeweils zwei Druckschwankungsmaxima und –minima aufweist (vgl. Abb. 4.3). Daher reicht es nun nicht mehr aus, für die Herleitung der Abweichungen von der Neutralstellung das Ansatzrohr in zwei Hälften zu teilen. Betrachten wir zunächst die artikulatorischen Veränderungen am Beispiel [ə] vs. [i] bzw. [ə] vs. [u] (Abb. 4.6). Bei beiden Vokalen ist der Rachenraum etwas erweitert. Die Verengung befindet sich beim vorderen Vokal [i] im vorderen Teil des Mundraums, beim hinteren Vokal [u] dagegen im hinteren Teil. Eine weitere Verengung befindet sich beim gerundeten Vokal [u] im Bereich der Lippen (in der Abbildung nicht zu sehen).

Was dies für den zweiten Formanten bedeutet wird erkennbar, wenn man die artikulatorischen Gegebenheiten in das Röhrenmodell mit der zweiten stehenden Welle überträgt (Abb. 4.7). Die erweiterte Querschnittsfläche ganz hinten (Druckschwankungsmaximum) sorgt zunächst in beiden Fällen für ei-

Tabelle 4.1: *Typische F1–Werte eines männlichen Erwachsenen in Hertz (Hz), wobei zu beachten ist, dass es durchaus individuelle Unterschiede gibt, und dass je nach Sprechstil durchaus andere Werte gemessen werden können. Bei weiblichen Erwachsenen sind etwas höhere Werte zu erwarten, bei Kindern deutlich höhere.*

Artikulatorische Dimension		Akustische Dimension (F1 in Hz)
hoch	[i,y,u]	250
mittelhoch	[e,ø,o]	350
neutral		500
mitteltief	[ɛ,œ,ɔ]	550
tief	[a,ɑ]	700

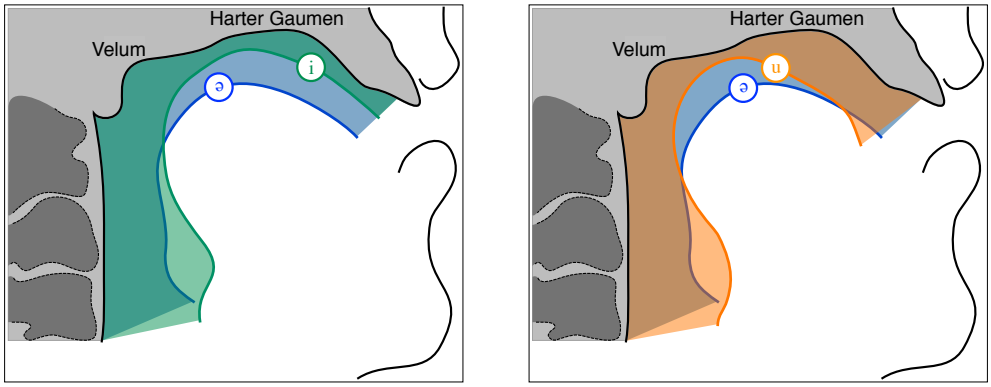


Abbildung 4.6: *Querschnittsfläche des Ansatzrohres im Vergleich mit der neutralen Stellung ([ə]): Vordere Vokale ([i], links) und hintere Vokale ([u], rechts).*

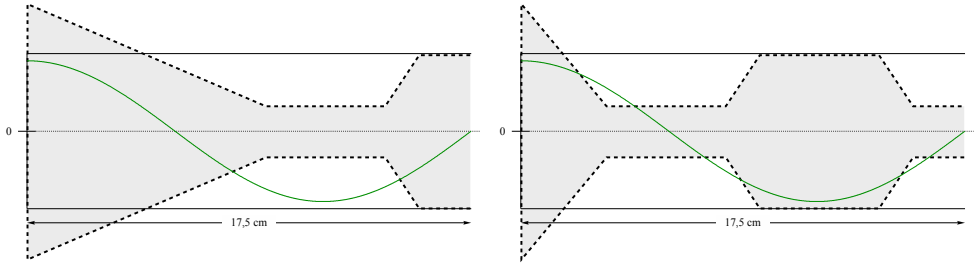


Abbildung 4.7: *Röhrenmodell des Ansatzrohres: Vordere Vokale (links) und hintere, gerundete Vokale (rechts).*

ne leichte Absenkung von F2. Bei [u] (Abb. 4.7, rechts) wird dies aufgrund von Verengungen im Bereich von zwei Druckschwankungsminima verstärkt: einmal durch die Verengung im hinteren Teil des Mundraums und einmal durch die Verengung an den Lippen. Bei [i] (Abb. 4.7, links) wird der Absenkungseffekt dagegen aufgehoben und aufgrund der Verengung im vorderen Teil des Mundraums (Druckschwankungsmaximum) ins Gegenteil verkehrt: F2 liegt bei vorderen Vokalen wie [i] deutlich über dem F2–Wert der Neutralstellung. Daraus lassen sich nun zwei Regeln herleiten:

Verengungsregel F2

Der zweite Formant ist im Vergleich zur Neutralstellung niedriger bei Verengung im hinteren Teil des Mundraums und höher bei Verengung im vorderen Teil des Mundraums.

Regel der Lippenrundung

Lippenrundung führt zu einer Absenkung der Formanten, insbesondere des zweiten Formanten.

Tabelle 4.2: *Typische F2–Werte eines männlichen Erwachsenen in Hertz (Hz), wobei zu beachten ist, dass es durchaus individuelle Unterschiede gibt, und dass je nach Sprechstil durchaus andere Werte gemessen werden können. Bei weiblichen Erwachsenen sind etwas höhere Werte zu erwarten, bei Kindern deutlich höhere.*

	vorne		→		→		→		hinten	
	[i]	[y]	[e]	[ø]	[ɛ]	[œ]	[a]	[ɔ]	[o]	[u]
F2 in Hz	2200	1600	2000	1500	1800	1400	1200	1000	700	600

Die Lage des zweiten Formanten korreliert also hautptsächlich mit der horizontalen Zungenposition: Vordere Vokale haben einen hohen, hintere Vokale einen tiefen F2. Bei gerundeten Varianten von vorderen Vokalen liegt F2 tiefer als bei den ungerundeten Varianten (vg. Tabelle 4.2).

Die Abbildungen 4.8 und 4.9 zeigen die Spektren einiger deutscher Vokale (von einem männlichen Sprecher). Hier gilt dasselbe wie bei den Tabellenwerten: Die Formantfrequenzen unterliegen einer gewissen Variation (sie sind z.B. abhängig von der Form des Ansatzrohres eines individuellen Sprechers, vom phonetischen Kontext (Koartikulation), vom Sprechstil (formell/informell) etc.), d.h. die genaue Lage einzelner Formanten soll nur als ungefährer Anhaltspunkt verstanden werden.

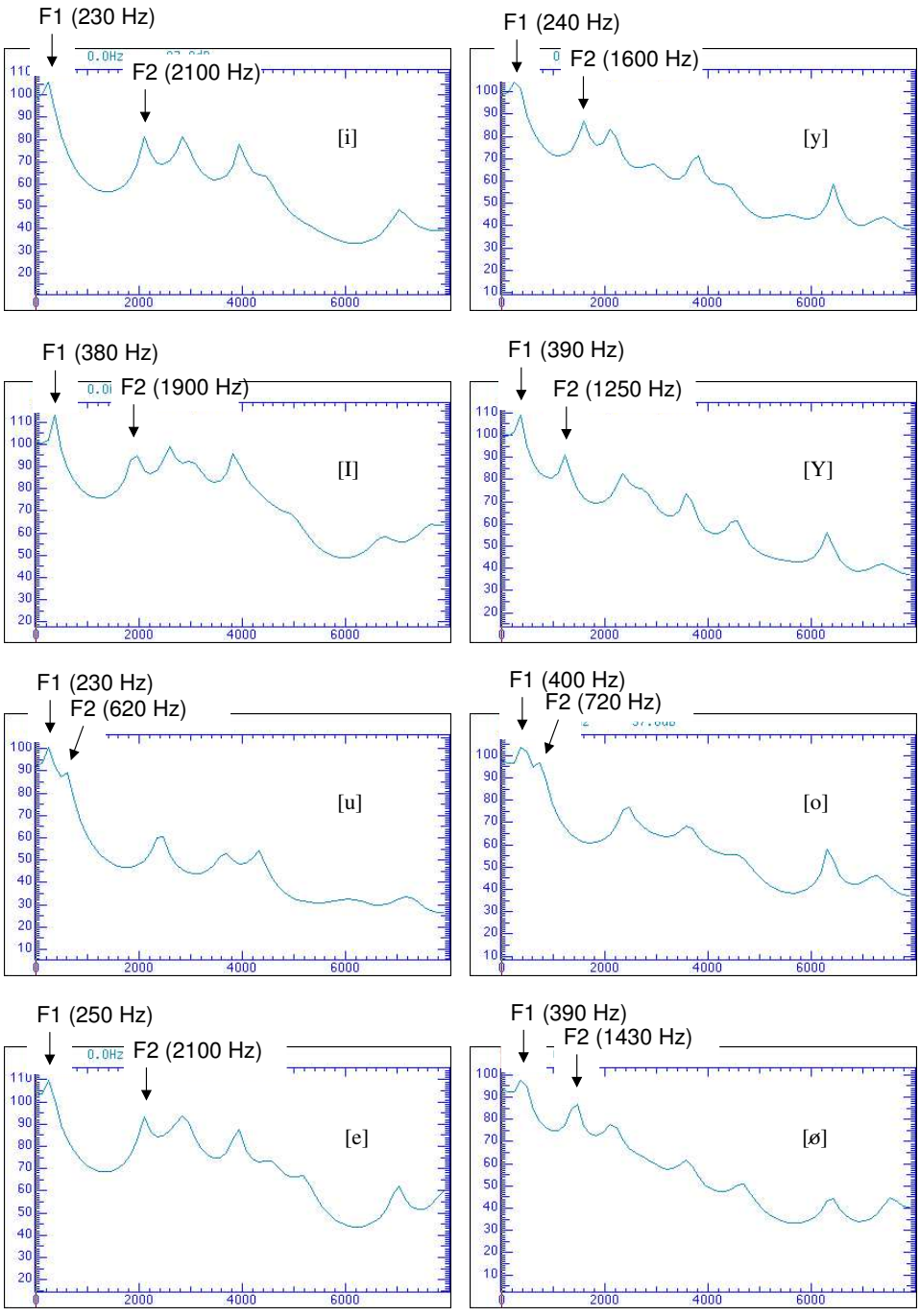


Abbildung 4.8: *Geglättete Vokalspektren eines männlichen Sprechers; hohe Vokale.*

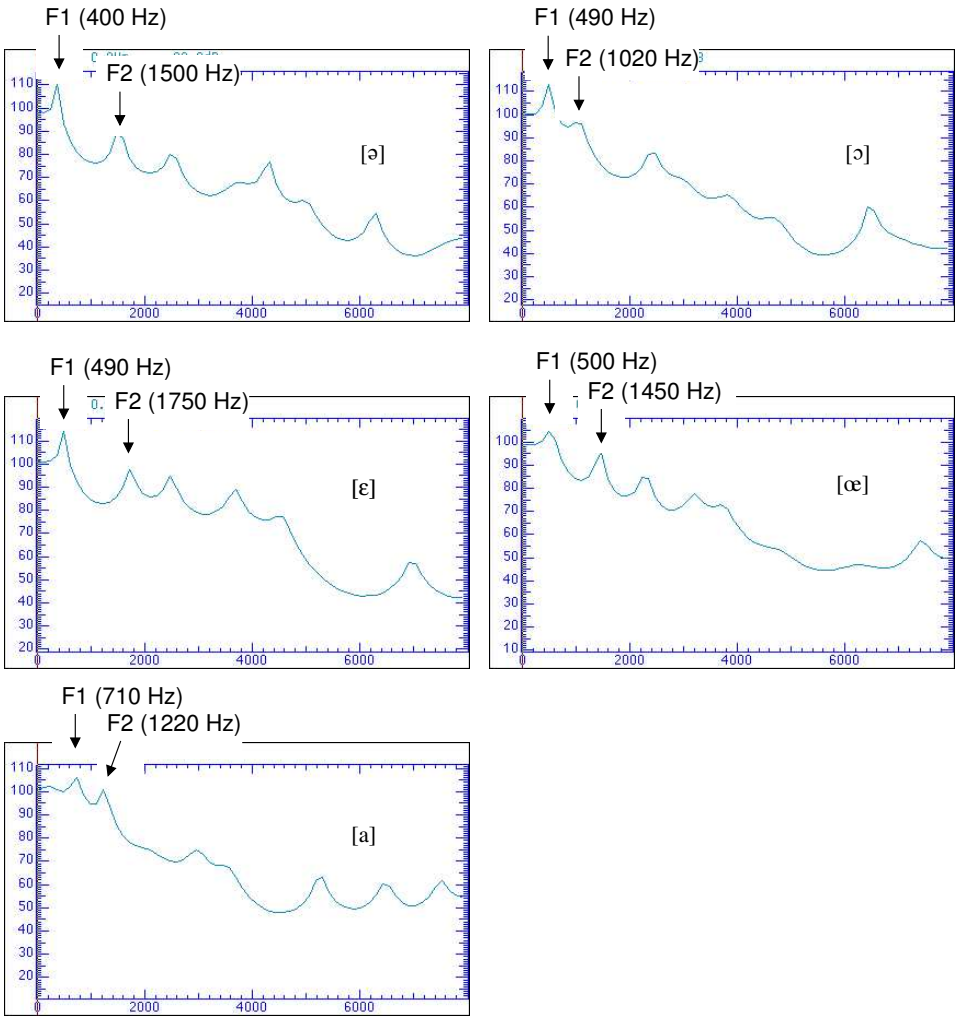


Abbildung 4.9: *Geglättete Vokalspektren eines männlichen Sprechers; mittlere und tiefe Vokale.*

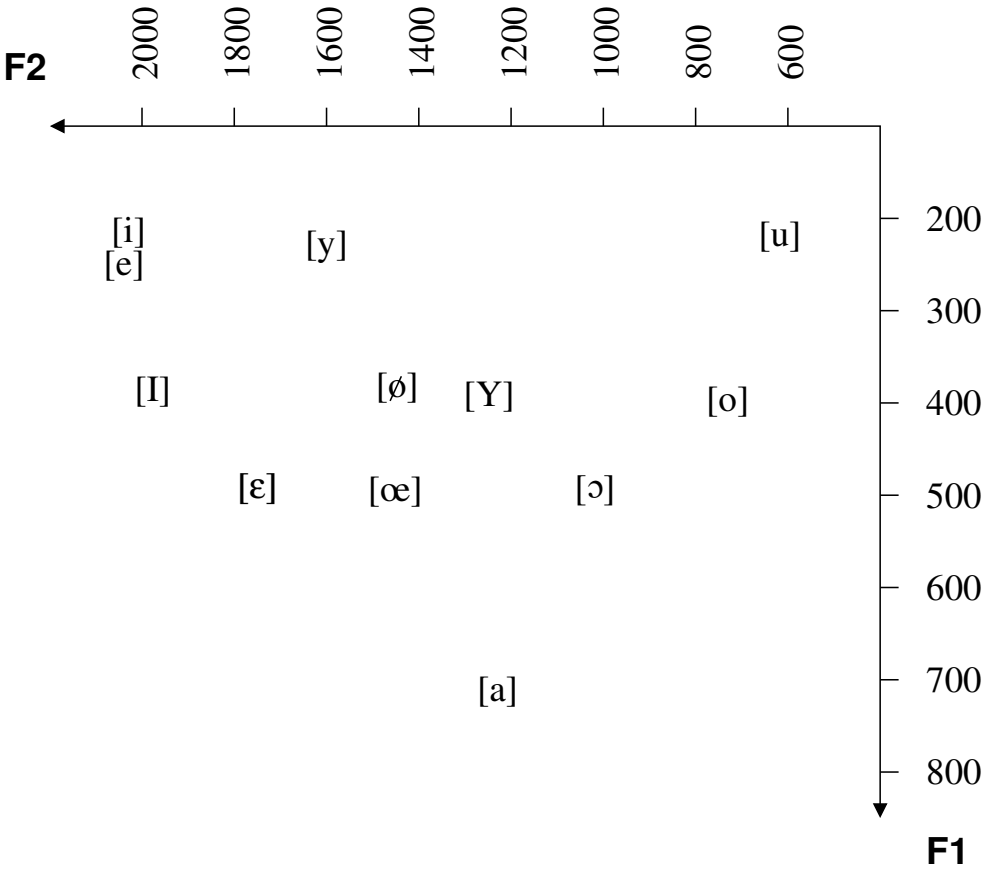


Abbildung 4.10: *Vokalraum eines männlichen Sprechers.*

Wichtiger als die absoluten Werte ist für die Vokalidentifikation jedoch das Verhältnis von F1 und F2 zueinander und die Lage der Formanten im Vokalraum des Sprechers oder der Sprecherin, wie in Abbildung 4.10 dargestellt. Aufgrund des Zusammenhangs zwischen Formantfrequenzen und Vokaltraktkonfiguration (also u.a. Zungenlage und –höhe) ähnelt der Vokalraum bei entsprechender Darstellung dem aus dem IPA bekannten Vokalviereck bzw. –dreieck. Die hier angegebenen Formantwerte wurden bei flüssiger Sprachproduktion gemessen (vorgelesener Text). Werden die Vokale sehr deutlich und isoliert produziert, grenzen sich die Werte im Vokalraum noch klarer voneinander ab. Bei flüssiger Spontansprache hingegen nähern sich die Werte aneinander an, Extremwerte werden kaum noch erreicht und einzelne Kategorien können zusammenfallen (im Beispiel ist diese Tendenz bei [i] und [e] zu beobachten).

Vokalraum

Formantenstruktur und Sprechstil

4.2 Konsonaten I: Sonoranten

4.2.1 Nasale

Nasalformanten

Auch die Spektren von Nasalen zeigen eine harmonische Struktur. Die supraglottale Filterung des Phonationssignals (im Deutschen sind Nasale wie in den meisten Sprachen stets stimmhaft) ist jedoch deutlich komplexer als bei der Vokalproduktion, da bei der Nasalproduktion zwei Luftsäulen zum Schwingen gebracht werden: Die Luftsäule im nach vorne geschlossenen Mundraum sowie die Luftsäule im Nasenhohlraum. Das Resultat dieser Filterung ist zwar ebenfalls eine Formantstruktur, die Lage der Formanten in Nasalspektren ist jedoch deutlich variabler (und schwieriger vorherzusagen). Dies gilt sowohl für den Vergleich unterschiedlicher Sprecher als auch für den Vergleich von unterschiedlichen Realisierungen des gleichen Nasallautes eines Sprechers. Daneben sind die sog. Nasalformanten auch stärker vom lautlichen Kontext abhängig als Vokalformanten.

Prinzipiell gilt, dass der erste Nasalformant (FN1) im Bereich der Eigenfrequenz des Nasenhohlraums liegt, d.h. bei einem bestimmten Sprecher ist die Frequenz des ersten Nasalformanten relativ konstant und für alle Nasale gleich; FN1 liegt ungefähr im Bereich von 200 – 250 Hz. Der zweite Nasalformant ist dagegen abhängig von der Position des oralen Verschlusses. Bei labialem Verschluss ([m]) liegt FN2 etwa im Bereich von 1000 bis 1200 Hz, bei alveolarem Verschluss ([n]) im Bereich von 1500 Hz und bei velarem Verschluss ([ŋ]) im Bereich von 2300 Hz. Zwischen FN1 und FN2 zeigen Nasalspektren meist ein sehr ausgeprägtes Minimum (sog. 'Antiformant'); ebenfalls sehr charakteristisch ist ein sehr steiler Abfall der Amplitudenwerte oberhalb von FN2 (Abbildung 4.11).

Antiformant

Vokale vor oder nach einem Nasallaut sind häufig teilweise, manchmal sogar ganz nasaliert (assimilatorische Nasalierung). Dies ist dadurch zu erklären, dass der Zugang zum Nasenhohlraum schon vor der oralen Verschlussbildung geöffnet wird (durch Absenken des Velums) bzw. erst nach Lösen des oralen Verschlusses verschlossen wird. Dieser Vorgang ist normalerweise nicht hörbar, schlägt sich jedoch in der Akustik nieder: Das wichtigste akustische Merkmal nasalierter Vokale ist eine Dämpfung der Frequenzen im Bereich des nasalen Antiformanten, also oberhalb von etwa 250 Hz. Da viele Vokale ihren ersten Formanten in diesem Bereich haben, ist bei nasalierten Vokalen F1 oft nur sehr schwach ausgeprägt.

assimilatorische Nasalierung

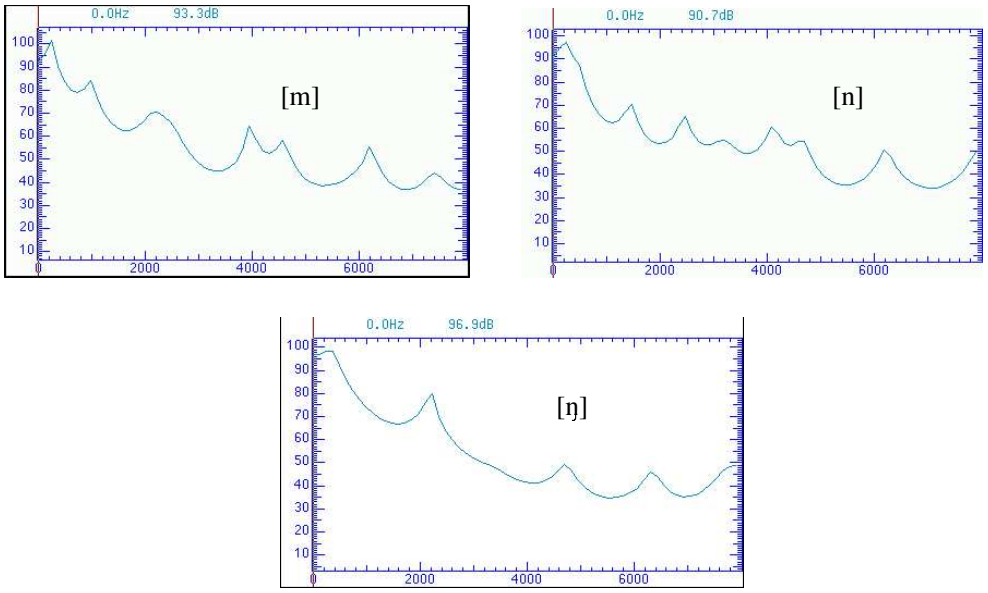


Abbildung 4.11: *Spektren der deutschen Nasallaute.*

4.2.2 Approximanten und Vibranten

Das wichtigste akustische Merkmal von Approximanten ist eine sehr ausgeprägte Bewegung der Formanten, insbesondere von F2 und F3. Start– und Endfrequenzen dieser Bewegungen hängen sehr stark vom lautlichen Kontext ab. Da ein Spektrum keine Informationen über den zeitlichen Verlauf enthält, ist es zur Darstellung von Formantbewegungen nicht geeignet. Ein besseres Darstellungsmittel ist in diesem Fall das Spektrogramm (siehe Abschnitt 3.4).

Vibranten sind insbesondere durch eine niedrigfrequente Amplitudenmodulation gekennzeichnet. Dies ist zurückzuführen auf die kurzzeitigen Unterbrechungen des Luftstroms durch intermittierende Verschlüsse. Auch dies lässt sich besser im Spektrogramm zeigen.

4.3 Konsonanten II: Obstruenten

4.3.1 Frikative

breitbandige Rauschmaxima

Das akustische Hauptunterscheidungsmerkmal der Frikative ist die Lage breitbandiger Maxima im Rauschspektrum. Außerdem unterscheiden sich die Frikative in ihrer Gesamtintensität; dieses Merkmal hängt jedoch von zahl-

reichen Faktoren ab (Einzellautproduktion vs. flüssige Produktion, Sprache, Sprecher, Situation) und ist daher nicht besonders zuverlässig. In flüssiger Sprachproduktion gemessen, gilt in etwa, dass [s] und [ʃ] eine relativ große Gesamtintensität aufweisen, während [f] und [h] eher schwach ausgeprägt sind; [ç], [x] und [χ] liegen dazwischen.

Nun zur Energieverteilung im Rauschspektrum der Frikative (Beispiele in Abbildung 4.12):

labiodental: relativ gleichmäßige Energieverteilung im gesamten Frequenzbereich, insgesamt von geringer Intensität und zu den höheren Frequenzen hin leicht abfallend.

alveolar: unterhalb von ca. 3000 Hz sehr geringe Intensität, nach einem steilen Anstieg um 3500 Hz folgt ein breitbandiges Maximum.

post–alveolar: sehr ähnliche spektrale Form wie alveolare Frikative, allerdings erfolgt der Anstieg früher, bei ca. 2000 Hz.

palatal: geringe Intensität bis etwa 2000 Hz, danach ein steiler Anstieg und ein flacherer, aber deutlicher Abfall.

velar: ähnlich palatalen Frikativen, das Maximum befindet sich jedoch unterhalb 2000 Hz.

uvular: breitbandiges Maximum um 1000 Hz, gefolgt von einem steilen Abfall und einem steilen Anstieg; ab etwa 2500 Hz flacher Abfall.

glottal: sehr geringe Energie; die Lage der Energiemaxima im Rauschspektrum ist sehr stark vom Kontext abhängig.

Voice bar

Stimmhafte Frikative weisen im Prinzip das gleiche Rauschspektrum wie die entsprechenden stimmlosen Frikative auf. Daneben zeigt sich im unteren Frequenzbereich mindestens ein schmalbandiges Maximum, die Voice bar (Abbildung 4.13). Man spricht in diesem Fall von einem Mischspektrum, bestehend aus harmonischen und nichtharmonischen Teilschwingungen. In höheren Frequenzbereichen werden die harmonischen Teilschwingungen in der Regel von den nichtharmonischen überlagert. Ist die Rauschkomponente jedoch schwach, kann die harmonische Struktur im Spektrum sichtbar werden.

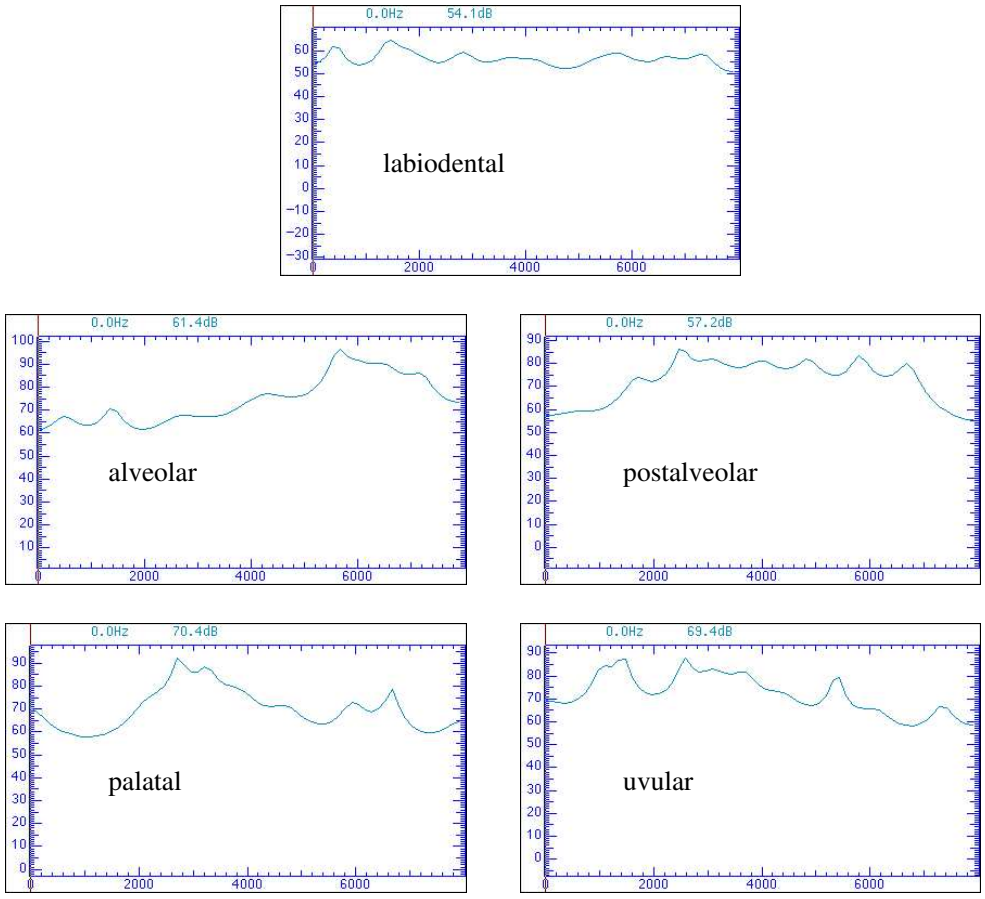


Abbildung 4.12: *Frikativspektren.*

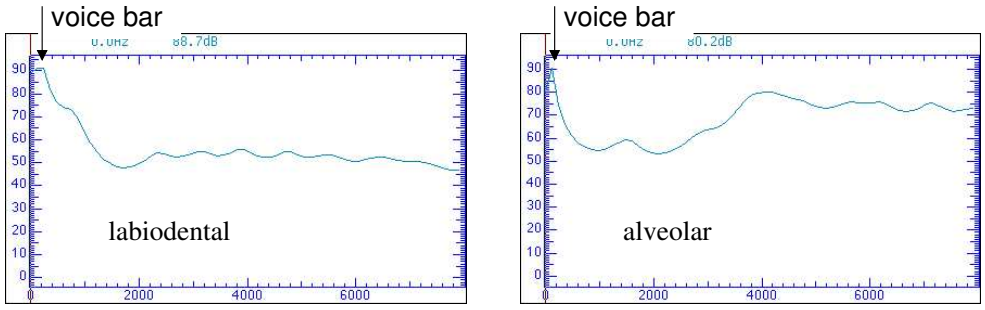


Abbildung 4.13: *Spektren stimmhafter Frikative.*

4.3.2 Plosive

Plosive können aus bis zu vier Phasen bestehen, die sich akustisch klar unterscheiden lassen:

Verschlussphase: Die Verschlussphase ist bei stimmlosen Plosiven stets durch stummen Schall gekennzeichnet. Auch bei stimmhaften Plosiven, insbesondere wortinitial, ist die Verschlussphase häufig stumm. Wortmedial in stimmhafter Umgebung wird die Phonation jedoch während der Produktion stimmhafter Plosive nicht ausgesetzt; dadurch kommt es in der (dann sehr kurzen) Verschlussphase zu periodischen Schwingungen geringer Intensität, die sich im Spektrum als voice bar darstellen.

Verschlusslösung/Plosion: Impulsförmige, d.h. sehr rasch ansteigende und wieder abfallende Amplitudenveränderung. Die Stärke des Impulses ist laut– und positionsabhängig; wortmedial ist der Impuls in der Regel kleiner als initial, bei stimmlosen Plosiven ist der Impuls meist deutlicher ausgeprägt als bei stimmhaften Plosiven.

Affriktion: Die (sehr kurze) Phase unmittelbar nach der Verschlusslösung; die Luft strömt durch die nach der Verschlusslösung enstehende Verengung an der Artikulationsstelle des Plosivs, dabei entsteht ein Geräusch wie bei einem an der entsprechenden Stelle gebildeten Frikativ.

Aspiration: Im Deutschen sind stimmlose Plosive normalerweise aspiriert (außer z.B. im Silbenonset nach einem Frikativ). Während dieser Phase besteht keine Verengung mehr an der Artikulationsstelle des Plosivs, der Öffnungswinkel des Kiefers nähert sich schon dem für die Vokalproduktion notwendigen Maß an, allerdings setzt die Phonation noch nicht ein. Akustisch gesehen ist diese Phase also vergleichbar mit dem glottalen Frikativ [h].

Stimmansatzzeit

Ein sehr verlässlicher akustischer Parameter zur Unterscheidung stimmhafter und stimmloser Verschlusslaute ist die Stimmansatzzeit (engl. *voice onset time, VOT*). Die Stimmansatzzeit ist die Zeit, die zwischen Verschlusslösung und Beginn der Phonation vergeht. Dieses Intervall ist bei stimmlosen Plosiven relativ lang — etwa 40 bis 100 Millisekunden bei wortinitialen Plosiven —, während die Stimmansatzzeit bei stimmhaften Plosiven sogar negativ sein kann, d.h. die Phonation beginnt schon vor der Plosion in der Verschlussphase; VOTs von wortinitialen stimmhaften Plosiven sind kleiner als 30 Millisekunden. Wortmedial oder bei flüssigem Sprechen und wenig ausgeprägten

Wortgrenzen auch wortinitial werden VOTs generell kürzer, die Wertebereiche für stimmlose und stimmhafte Plosive bleiben jedoch normalerweise klar voneinander abgegrenzt.

Transitionen

Ein akustisch und perzeptiv sehr wichtiges Merkmal zur Erkennung und Differenzierung unterschiedlicher Verschlusslaute sind die Transitionen (Formantbewegungen) vor oder nach einem Plosiv (siehe Abschnitt 3.4.2 und Abbildung 3.22 auf Seite 102). Abbildung 4.14 zeigt Oszillogramme von verschiedenen Verschlusslauten des Deutschen.

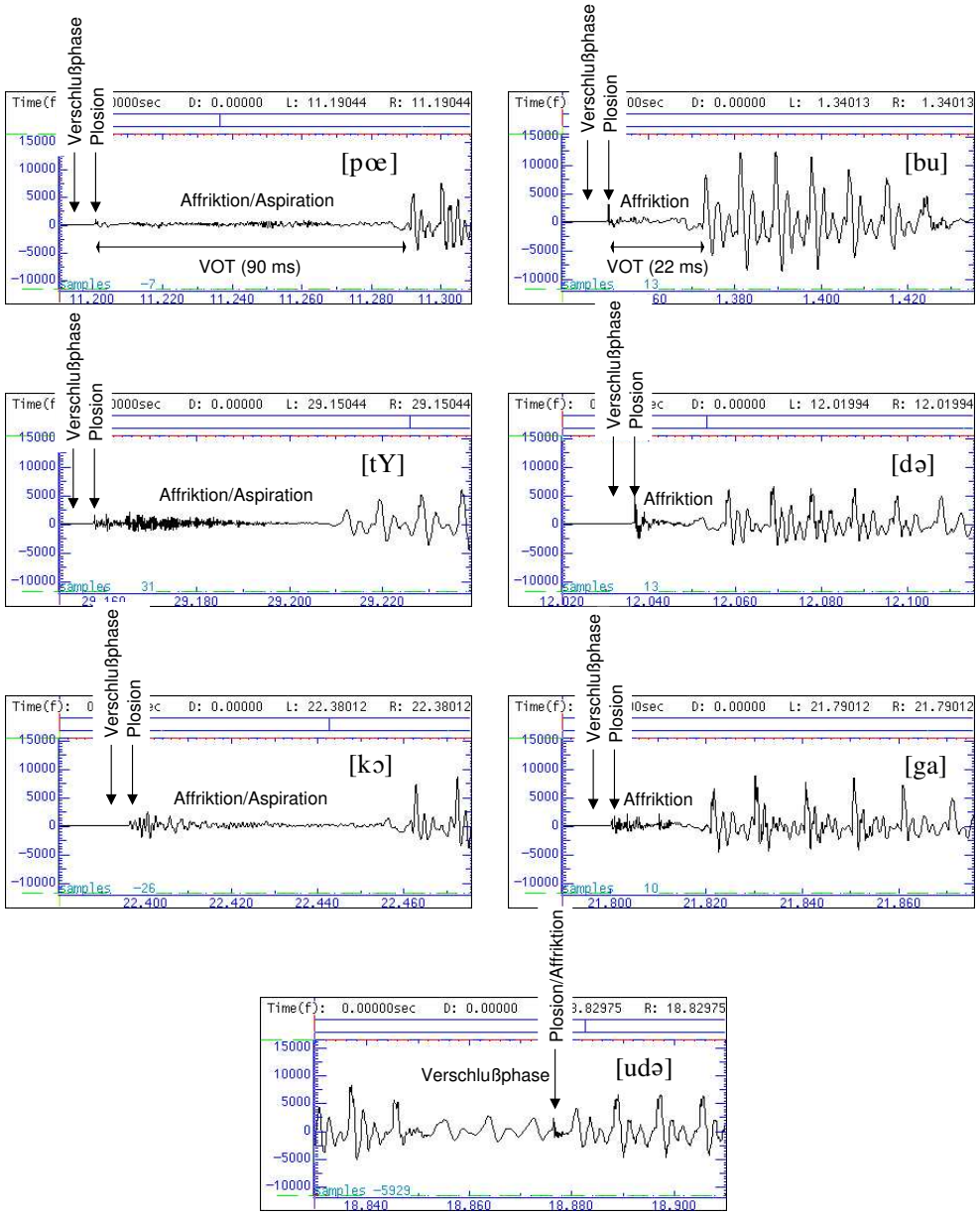


Abbildung 4.14: *Oszillogramme von Verschlusslauten; wortinitial (obere 3 Reihen) und wortmedial (ganz unten).*

Literaturverzeichnis

[1] o. A. (1999). *Handbook of the International Phonetic Association. A Guide to the Use of the International Phonetic Alphabet.* Cambridge University Press.
Das offizielle Handbuch zu allen Fragen rund um das IPA und die Transkription; enthält auch ExtIPA, die Erweiterungen zur Beschreibung gestörter Sprache.

[2] M. J. Ball, J. Rahilly & P. Tench (1996). *The Phonetic Transcription of Disordered Speech.* San Diego, London: Singular Publishing Group Inc.
Die Anwendung von ExtIPA zur Beschreibung gestörter Sprache.

[3] J. Clark & C. Yallop (1995). *An introduction to phonetics and phonology.* Blackwell Publishers.
Sehr gutes und materialreiches Lehrbuch sowohl für Phonetik als auch für Phonologie.

[4] H. Fastl & E. Zwicker (2007). *Psychoacoustics. Facts and Models.* Berlin, Heidelberg, New York: Springer.
Umfassendes, sehr detailliertes Standardwerk zur Psychoakustik mit Audiobeispielen auf CD.

[5] W. J. Hardcastle & J. Laver (1997). *The Handbook of Phonetic Sciences.* Blackwell Publishers.
Weiterführende Überblickskapitel zu vielen relevanten Themen der Phonetik; für Fortgeschrittene.

[6] K. Johnson (2. Aufl., 2003). *Acoustic and Auditory Phonetics.* Blackwell Publishers.
Sehr gute, eher knappe Einführung speziell in die akustische und auditorische Phonetik; mit Übungsaufgaben.

[7] K. J. Kohler (2. Aufl., 1995). *Einführung in die Phonetik des Deutschen.* Erich Schmidt Verlag.
Phonetische Beschreibung des Deutschen; mit Grundlagen der deutschen Phonologie.

[8] P. Ladefoged (3. Aufl., 1993). *A course in phonetics.* Harcourt Brace College Publishers.
Ein Klassiker; besonders zugeschnitten auf die Phonetik des Amerikanischen Englisch.

[9] N. J. Lass (Ed., 1996). *Principles of Experimental Phonetics.* Mosby.
Überblickskapitel zu allen möglichen Themen der Experimentalphonetik und gesonderte Abschnitte zu phonetischen Instrumenten und zur Methodologie.

[10] P. Lieberman & S.E. Blumstein (1988). *Speech physiology, speech perception, and acoustic phonetics.* Cambridge University Press.
Noch ein Klassiker; Grundlagen der 'naturwissenschaftlich orientierten' Phonetik.

[11] J. Neppert (1999). *Elemente einer akustischen Phonetik.* Hamburg: Buske.
Sehr gutes und ausführliches Lehrbuch speziell zur akustischen Phonetik.

[12] M. Pétursson & J. Neppert (1996). *Elementarbuch der Phonetik.* Hamburg: Buske.
Eine knappe, aber gute Einführung in die Phonetik.

[13] B. Pompino-Marschall (1995). *Einführung in die Phonetik.* Berlin/New York: Walter de Gruyter.
Gut verständliche Einführung in die Kernbereiche der Phonetik (Artikulatorische, akustische, auditorische und systematische Phonetik).

[14] K. N. Stevens (1998). *Acoustic Phonetics.* MIT Press.
Ein sehr umfassendes, exzellentes Standardwerk zur akustischen Phonetik; für Fortgeschrittene.

[15] I. R. Titze (1994). *Principles of Voice Production.* Englewood Cliffs: Prentice–Hall, Inc.
Sehr fundierte Darstellung sämtlicher Aspekte der Stimmproduktion.

Nachschlagewerke

[16] H. Bußmann (1990) *Lexikon der Sprachwissenschaft.* A. Kröner Verlag.
Umfassendes Fachwörterbuch mit Begriffsdefinitionen und weiterführender Literatur.

[17] D. Crystal (3. Aufl., 1991). *A Dictionary of Linguistics and Phonetics.* Blackwell Publishers.

[18] Duden, Band 6: *Das Aussprachewörterbuch.*
Phonetische und phonologische Grundlagen des Deutschen; sehr umfangreicher Lexikonteil mit der Standardaussprache sowie den wichtigsten Varianten.

Index