IV Akustik von Stimme und Sprache

IV.1 Spracherzeugung

Der menschliche Sprachapparat ist ein äußerst komplexes biologisches System, an dem mehrere Organe und Muskeln beteiligt sind (z. B. Zwerchfell, Lunge, Brustkorbmuskulatur, Kehlkopf, Zunge, Schlund- und Mundmuskulatur) und der einer äußerst diffizilen neuronalen Steuerung und Regelung unterliegt. Mit unserer Stimme sind wir in der Lage, einen breiten Frequenz- und Pegelbereich zu überstreichen und andererseits auch Stimmungslagen und Emotionen auszudrücken, so daß dieses biologische System als ein Wunderwerk betrachtet werden kann, dessen genaue Funktionsweise im Detail noch längst nicht bekannt ist. Aus physikalisch-akustischer Sicht ist die Schwingungserzeugung und anschließende akustische Filterung und Schallabstrahlung von Interesse, sowie Methoden zur Charakterisierung und Analyse von Sprache, Sprachlauten und der akustischen Vorgänge bei Stimmstörungen.

Durch das Zwerchfell und die Brustkorb-Muskulatur wird die in der Lunge gespeicherte Luft unter Druck gesetzt, so daß eine Luftströmung durch die Luftröhre, den Kehlkopf und den Nasen- und Rachentrakt entsteht. Einen Überblick über die Anatomie des Kehlkopfes gibt die unten stehende Abbildung. Bei der normalen Atmung wird diese Luftströmung nicht unterbrochen, bei der Phonation (d. h. der Erzeugung von Stimmlauten) wird dagegen dieser Luftstrom entweder im Kehlkopf bei den **Stimmlippen** unterbrochen (dort wird eine periodische Schwingung erzeugt) oder an anderer Stelle zur Erzeugung von aperiodischen, rauschförmigen Schwingungen (z. B. an den Schneidezähnen zur Erzeugung eines scharfen „s“).

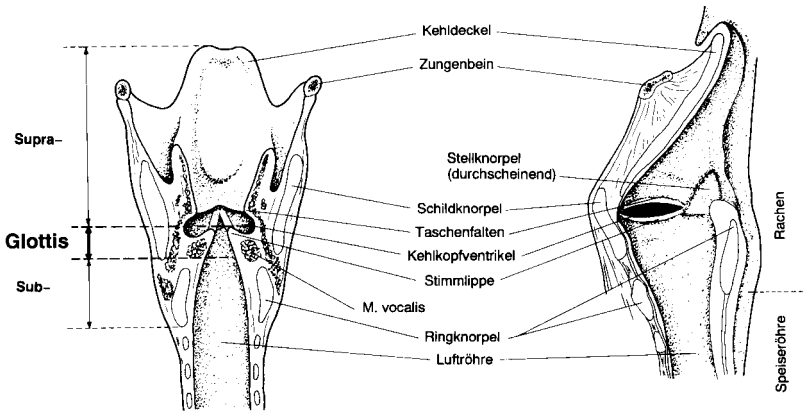

Abbildung 4.1: Kehlkopfinneres; Frontalschnitt mit Blick nach vorne und Sagittalschnitt mit Seitenansicht.

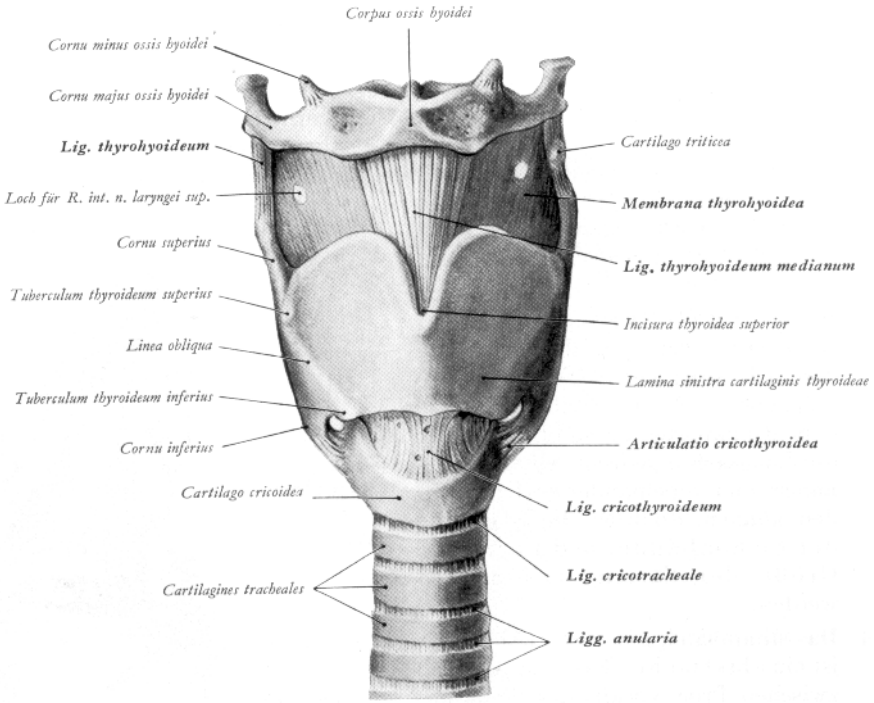

Abbildung 4.2: Der Bandapparat des Kehlkopfes von vorne gesehen.

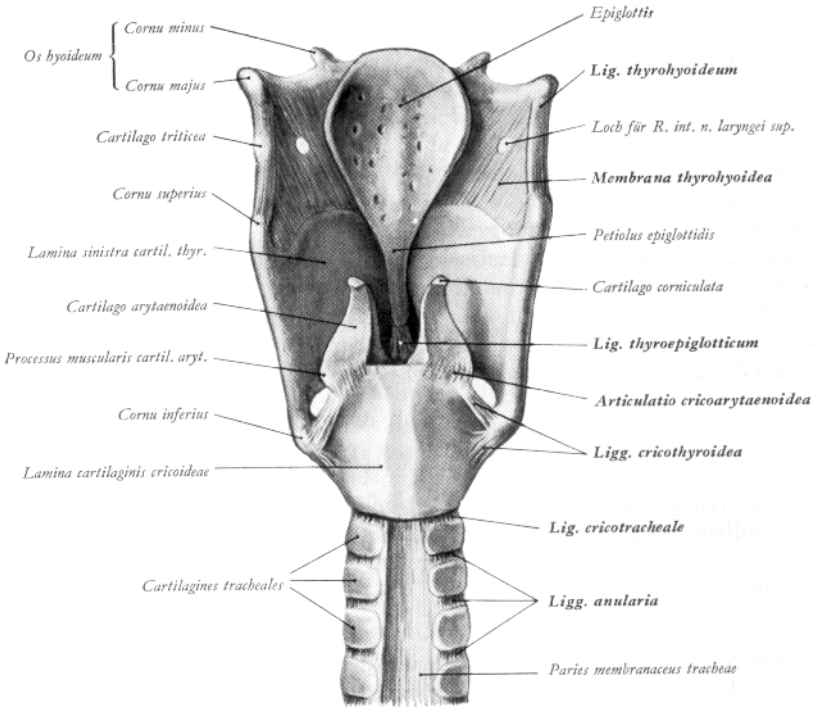

Abbildung 4.3: Der Bandapparat des Kehlkopfes von hinten gesehen.

Der Kehlkopf besteht dabei aus dem größeren Schildknorpel (vor dem sich die Schilddrüse befindet) und dem darunter befindlichen Ringknorpel auf dessen Hinterseite zwei spiegelsymmetrisch angeordnete kleine Ary-Knorpel sind (Cartilaginae Arytaenoideae). Zwischen der Vorderkante des Schildknorpels und den vorderen Ansätzen der Ary-Knorpel sind die Stimmlippen als eine Falte in der auskleidenden Schleimhaut angeordnet,

die man sich wie eine gelatineartige Masse vorstellen kann, die außen mit einer dünnen Haut umgeben ist und in deren Kern eine etwas festere, sehnige Struktur zu finden ist, die sich zwischen Vorderseite des Schildknorpel und Ansatz der Ary-Knorpel spannt. Durch die Einwirkung entsprechender Muskel (Einzelheiten im Anatomie- bzw. Physiologiebuch) können die Ary-Knorpel nun eine Translations- und Rotationsbewegung durchführen, bei der die Stimmlippen in ihrem hinteren Ende einander in der Mitte angenähert und angespannt werden können. In dieser Phonationsstellung besitzt der Kehlkopf dann folgenden schematischen Querschnitt:

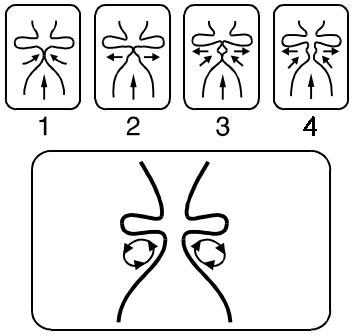

Abbildung 4.4: Stimmlippenschwingungen

Durch die Druckdifferenz zwischen den Lungen (bzw. der Trachea, d. h. Luftröhre) und dem Vokaltrakt wird eine **Kippschwingung** der Stimmritze (Glottis) erzeugt, die mit einer periodischen Öffnung und Schließung der Stimmritze verbunden ist. Der schematische Bewegungsablauf ist in Abb. 4.4 dargestellt. Zunächst wird der Glottis-Verschluß durch den Überdruck der Lungen „gesprengt“. Anschließend verschließt sich die Glottis wieder, einerseits, weil es sich um ein in der Resonanz betriebenes, schwingungsfähiges System handelt und andererseits, weil es durch den Bernoulli’schen Unterdruck zu einer Sogwirkung zwischen den Stimmlippen kommt. Dieser Bernoulli-Effekt beschreibt den Druckabfall des statischen Drucks p, wenn eine Strömung mit der Geschwindigkeit v in einem Medium der Dichte ρ auftritt:

$$p + \frac{1}{2}\rho v^2 = \text{const.} = p_0 \qquad \text{(IV.1)}$$

Ist die Strömungsgeschwindigkeit v = 0, wird der statische Luftdruck p_0 angenommen, während mit zunehmender Strömungsgeschwindigkeit der statische Druck p stark absinkt. Dieser Effekt wird auch bei den Tragflächen von Flugzeugen ausgenutzt, um eine Auftriebskraft zu erzeugen. Seine Bedeutung für den Stimmlippenverschluß ist zwar von theoretischer Bedeutung, sein Einfluß wird jedoch überschätzt, weil eine Rückstellkraft

zum Zusammenschluß der Glottis auch aufgrund der Elastizität der Glottis bewirkt wird, die letztlich durch Muskelanspannung verändert werden kann und zu einer Veränderung der Stimmfrequenz führt.

Die einfachste Beschreibung der Stimmlippen-Schwingung wird daher mit einem **Ein-Massenmodell** erreicht, das im Prinzip wie folgt aufgebaut ist:

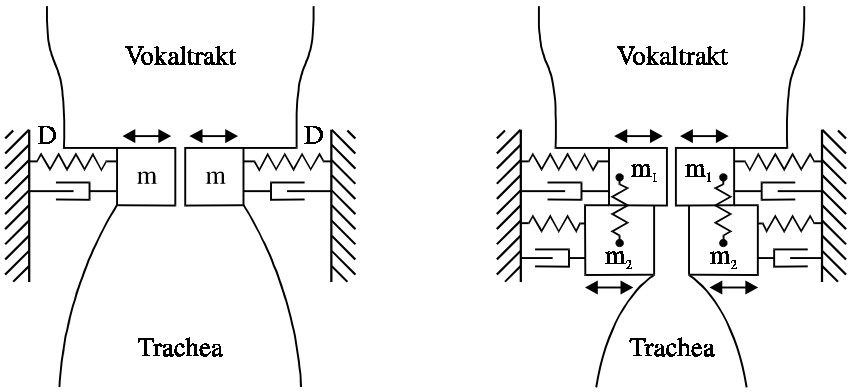

Abbildung 4.5: Ein-Massenmodell (links) und Zwei-Massenmodell (rechts)

Sämtliche schwingende Masse wird in die symmetrisch angeordnete Masse m verlagert, während die Elastizität durch die Federkonstante D und die Dämpfung durch einen zusätzlichen Parameter modellmäßig angesetzt wird. Ein derartiges Ein-Massen-Modell der Glottis liefert eine mathematische Beschreibung der Schwingung mit der Grundfrequenz:

$$f_0 = \frac{1}{2\pi} \cdot \sqrt{\frac{D}{m}} \qquad \text{(IV.2)}$$

Eine realistischere Beschreibung der tatsächlichen Form der Glottis und der Phasenverschiebung zwischen der Ober- und Unterseite der Stimmritze liefert ein Zwei-Massenmodell, was in der Abbildung 4.5 schematisch rechts dargestellt ist. Dabei sind die Massen m_1 und m_2 durch eine weitere Feder miteinander verbunden und können zu einer phasenverschobenen Schließung des Vokaltraktes führen.

Neben dieser periodischen Schwingungsanregung, die bei Vokalen eine entscheidene Rolle spielt, wird bei Verschluß-Konsonanten (Plosive, Frikative) der Vokaltrakt nicht im Bereich der Stimmritze, sondern in seinem weiteren Verlauf im Rachen-Mundraum am weitesten eingeengt. Bei einer starken Einengung des Vokaltraktes bewirkt die starke Querschnittserniedrigung bei gleichzeitig anhaltendem Luft-Volumen-Strom einen Umschlag von einer laminaren Luft-Strömung in eine turbulente Strömung.

Dieser Umschlag wird dadurch charakterisiert, daß die Reynolds-Zahl Re einen gewissen kritischen Wert überschreitet:

$$\mathrm{Re} = \frac{\rho \cdot \mathrm{v} \cdot \mathrm{d}}{\eta} > 1700 \qquad \text{(IV.3)}$$

Dabei bezeichnet ρ die Dichte der Luft, v die Strömungsgeschwindigkeit, d den Durchmesser und η die Viskosität der Luft. Die Reynoldszahl beschreibt damit das Verhältnis zwischen Trägheits-Kräften und Reibungs-Kräften in der Flüssigkeit. Falls die Reibungs-Kräfte im Verhältnis zu stark abnehmen, steigt die Reynolds-Zahl an und eine turbulente Strömung resultiert, die zu einer rauschartigen Schwingungsanregung im Vokaltrakt führt. Durch diese Rausch-Anregung können beispielsweise Frikative (z. B. das „f“ (wie Faß) und das „ ∫ “ (wie Scheibe)) artikuliert werden.

Die bisherigen Vorgänge beschäftigen sich mit der Schwingungsanregung im Vokaltrakt, also der akustischen **Quelle**. Diese von der Quelle erzeugte akustische Energie wird weiter im Vokaltrakt fortgeleitet und durch die verschiedenen Einengungen und Erweiterungen des Vokaltraktes in unterschiedlicher Weise akustisch **gefiltert**. Die dabei auftretende Filterwirkung des Vokaltraktes kann am ehesten durch die folgende Analogie zwischen dem Vokaltrakt und einem Schalldämpfer-Röhrensystem verdeutlicht werden:

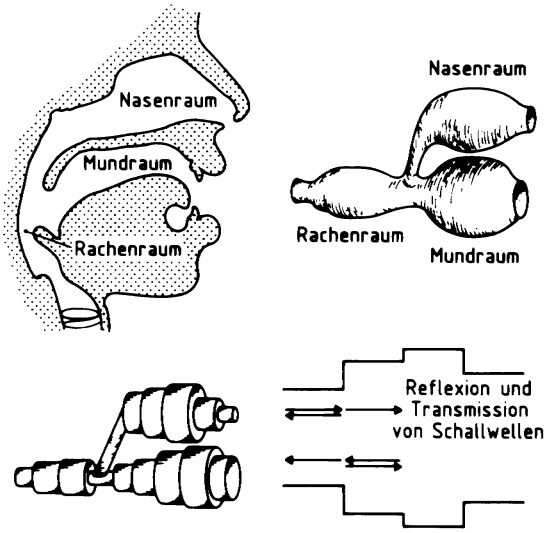

Abbildung 4.6: Analogie zwischen Vokaltrakt und Schalldämpfer-Röhrensystem

Wie in Kapitel 2 vorgestellt, führt jede Querschnittsänderung in einer Röhre zu einer Reflektion und teilweisen Transmission der einfallenden Schallwelle. Bei einer Hintereinanderschaltung von Röhrensegmenten mit

unterschiedlichem Durchmesser wird so eine Filterstruktur geschaffen, die formal einem digitalen Filter sehr ähnelt (Zeitverzögerung und Multiplikation mit einer Konstanten sowie anschließende Aufaddition). Dadurch wird deutlich, daß man durch unterschiedliche Wahl der Röhrendurchmesser (d. h. unterschiedliche Stellung des Artikulationstraktes) eine unterschiedliche akustische Filterwirkung erzeugt, die letztlich zu der Klangverfärbung führt, die für jeden Sprachlaut (insbesondere für jeden Vokal) von Bedeutung ist. Die Erzeugung von Sprache kann daher in erster Näherung beschrieben werden als die zeitvariante, akustische Filterung des Quellensignals (Glottis-Schwingung oder aperiodische Anregung) durch den Vokaltrakt. Die Aufgabe des Empfängers ist es nun, anhand dieser zeitvarianten Klangänderung auf die zugrundliegende Artikulationsstellung zurückzuschließen, aus der dann wiederum auf den Sprachlaut geschlossen werden kann, den der Sprecher artikuliert hat.

IV.2 Akustische Phonetik

Die akustische Phonetik beschäftigt sich mit der Beziehung zwischen der (abstrakten) Sprache und der akustischen Realisation dieser Sprache. So ist das kleinste bedeutungstragende Element der Sprache das **Phonem** d. h. durch Veränderung oder Weglassen eines Phonems wird der Sinn eines Wortes verändert. Ein derartiges Phonem kann (muß aber nicht) einem geschriebenem Buchstaben entsprechen. Beispielsweise hat das Wort „Sinn“ die drei Phoneme /z/, /ɪ/ und /n/. Wenn an Stelle des Phonems /ɪ/ ein anderes Phonem (z. B. der Vokal /ʌ/) steht, ändert sich der Sinn. Dasselbe Phonem /ɪ/ kann allerdings von sehr unterschiedlichen Sprechern ausgesprochen werden, die zu einer deutlich anderen akustischen Realisation ein und desselben Phonems führt. Außerdem wird die akustische Äußerung desselben Phonems /ɪ/ bei ein und derselben Person jedesmal ein anderes akustisches Signal erzeugen. Zu jedem Phonem gibt es daher eine (unendlich) große Zahl möglicher akustischer Realisierungen (**Phone**).

Um zu einer geeigneten akustischen Analyse gesprochener Sprache zu gelangen, anhand derer sich die Phone unterscheiden und klassifizieren lassen, bedient man sich des **Spektrogramms**, d. h. einer Kurzzeit-Spektralanalyse als Funktion der Zeit, die wir bereits im vorigen Kapitel kennengelernt haben. Wenn s(t) das Sprachsignal bezeichnet, wird das (diskrete) Spektrogramm (Sonagramm) bei einem bestimmten Zeitpunkt t_m und einer Frequenz f_n definiert als:

$$S(t_m, f_n) = \sum_{n=-(N-1)/2}^{(N-1)/2} s(t_m + n) \cdot w(n) \cdot e^{-2\pi i \frac{k \cdot a}{N}} \qquad \text{(IV.4)}$$

Man erhält also eine zwei-dimensionale Abbildung, bei der auf der Abszisse die Zeit und auf der Ordinate die Frequenz aufgetragen ist. Die Schwärzung bezeichnet dabei den Betrag des Kurzzeit-Leistungspektrum zu diesem Zeitpunkt t und zur betreffenden Frequenz f. Diese Darstellung entspricht ungefähr der Kurzzeit-Spektralanalyse, die auch im menschlichen Gehör durchgeführt wird.

Als Beispiel ist im folgenden die Zeit-Funktion und das Spektrogramm des Wortes „Phoniatrie“ aufgezeichnet. Während man im oberen Teilbild nur grob aus dem Verlauf der **Zeitfunktion** ersehen kann, daß erst ein Konsonant, dann drei unterschiedliche, halbwegs stationäre Vokale und dann wieder ein Konsonant mit anschließenden ausklingenden Vokal erfolgt, liefert das untere Teilbild (Spektrogramm) wesentlich mehr Informationen. Insbesondere kann man bei den Vokalen die Struktur eines harmonischen Tonkomplexes entdecken, d. h. eine Periodizität im Spektrum, die durch eine Grundfrequenz (Glottis-Frequenz) mit ihren harmonischen Obertönen erzeugt wird. Charakteristisch sind nun für diese Vokale die Lage der Maxima im Spektrum, die sogenannten **Formanten**, die gerade Resonanz-Frequenzen des Vokaltraktes entsprechen. Sie können als Bereiche besonderer Schwärzung in den Vokalen identifiziert werden. Die Konsonanten können dagegen durch ihren Zeitverlauf, das Auftreten von Pausen und den überdeckten Spektralbereich ansatzweise klassifiziert werden. Obwohl es nicht immer eindeutig möglich ist, anhand von Spektrogrammen auf die zugrundeliegenden Sprachelemente zu schließen (das Ohr ist auf diese Aufgabe wesentlich besser spezialisiert und Sprachwissenschaftler benötigen einige Übung, bis sie Spektrogramme „lesen“ können), soll im folgenden aufgezeigt werden, welche charakteristischen akustischen Merkmale von Sprachlauten auftreten und wie man so zu einer Identifizierung von Sprachlauten anhand akustischer Merkmale gelangen kann.

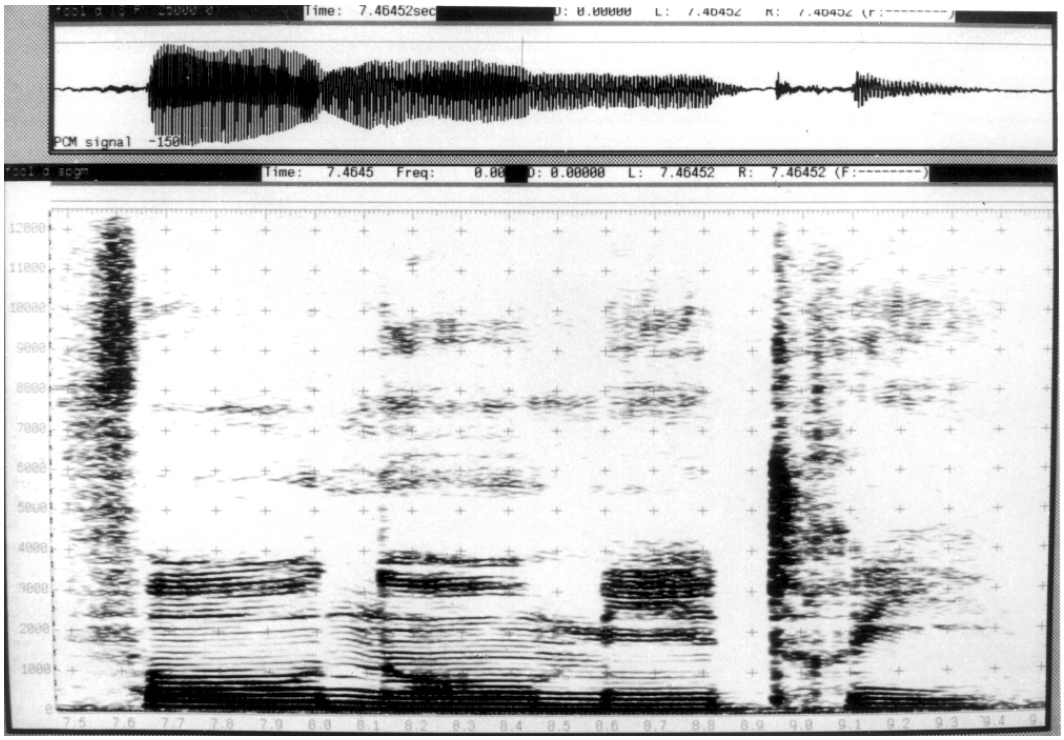

Abbildung 4.7: Zeitfunktion und Spektrogramm des Wortes „Phoniatrie“

Eine wichtige akustische Information sind die bereits erwähnten **Formanten,** die sich als (zeitlich veränderliche) Maxima der Vokaltrakt-Übertragungsfunktion bemerkbar machen.

Grundfrequenz F_0	Männer: 100-400 Hz, Frauen: 200-800 Hz
1. Formant F_1	300-1000 Hz
2. Formant F_2	600-2500 Hz
3. Formant F_3	1500-2500 Hz

Für die einzelnen Vokale gibt es relativ charakteristische Lagen dieser Formanten, die auch gut mit der Artikulationsweise der jeweiligen Vokale zusammenhängen: Beispielsweise weist das „u“ einen niedrigen ersten und niedrigen zweiten Formanten auf, während das „i“ einen hohen zweiten und einen relativ niedrigen ersten Formanten besitzt (s. Abbildung 4.8). In der Mitte von diesem als **Vokaldreieck** bezeichneten Graphen liegt der sogenannte „Reduktionsvokal“ „ə“, der im Englischen auch als „schwa“-Laut bezeichnet wird und im Deutschen als Zeichen heftigen Nachdenkens interpretiert wird (z. B. die Äußerung „...äh“).

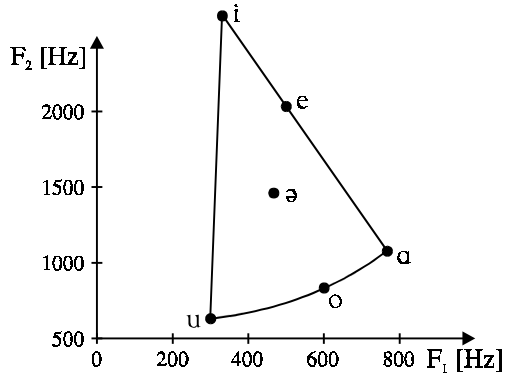

Abbildung 4.8: Vokaldreieck

Das in der F_2/F_1-Ebene aufgespannte Vokal-Dreieck kann auch als **artikulatorisches Vokaldreieck** interpretiert werden: Die Lage des zweiten Formanten korreliert relativ gut mit der Unterkiefer-Stellung (beim „i“ ist der Mund relativ weit geschlossen, während er beim „u“ und „a“ relativ weit geöffnet ist). Der erste Formant korreliert relativ gut mit der Vorne-Hinten-Artikulation, wobei das „a“ relativ weit vorne mit offenem Mund artikuliert wird, während das „u“ weit hinten mit gerundetem Mund artikuliert wird (d. h. 1. Formant: Zungenhöhe, 2. Formant: Mundöffnung).

Eine Liste der im Deutschen vorkommenden Vokale (als Phoneme) ist im folgenden aufgeführt. Man unterscheidet dabei zwischen Monophthongen und Diphthongen (d. h. Übergänge zwischen zwei Vokalen) und zwischen langen und kurzen Monophthongen.

	Monophthonge	Diphthonge	Reduktionsvokal
lang	/aː/ /eː/ /iː/ /oː/ /uː/ /yː/ /ɛː/	/ɔɪ̯/ /əɪ̯/ /əʊ̯/	/ə/
kurz	/a/ /ɛ/ /ɪ/ /ɔ/ /ʊ/ /ʏ/ /œ/		

Während die Vokale aufgrund ihrer Formantlage (die während der Dauer des Vokals relativ stabil ist) und ihrer Länge auch von der akustischen Analyse her relativ eindeutig zu klassifizieren sind, ist es bei den Konsonanten wesentlich schwieriger, zu einer Klassifikation zu gelangen. Eine Unterteilung aufgrund artkulatorischer Merkmale (d. h. aufgrund der Art, wie sie artikuliert werden und dem jeweiligen Artikulationsort) ist dagegen relativ einfach und führt zu der im folgenden aufgeführten Tabelle für die im Deutschen gebräuchlichen Konsonanten:

	Frikative	Plosive	Nasale	Approxi-manten	lateraler Approximant
stimmhaft	/v/ /z/	/b/ /d/ /g/	/m/ /n/ /ŋ/	/j/ /ʁ/	/l/
stimmlos	/f/ /s/ /ʃ/ /ç/ /x/ /h/	/p/ /t/ /k/			

Ein erster Schritt in die Richtung, eine Klassifikation von Sprachlauten aufgrund von rein akustischen Merkmalen zu erhalten, wird im System der „Distinktiven Sprachmerkmale“ versucht. Diese Sprachmerkmale sind binär (d. h. ein „+“ zeigt das Vorliegen des jeweiligen Sprachmerkmals an und ein „-“ bedeutet, daß das Merkmal nicht vorliegt). Jeder Konsonant ist gemäß nachstehender Tabelle (mit Beispielen) eindeutig durch eine Kombination von Merkmalen bestimmt. Während einige Sprachmerkmale artikulatorischer Natur sind (z. B. nasal, vokalisch bzw. konsonantisch) weisen andere Merkmale eher akustischer Natur auf (z. B. scharf, dunkel, abrupt).

	vokalisch	konsonantisch	kompakt	dunkel	nasal	abrupt	gespannt	stimmhaft	scharf
b (**B**ad)	-	+	-	+	-	+	-	+	-
d (**d**u)	-	+	-	-	-	+	-	+	-
f (**F**ee)	-	+	-	+	-	-	+	-	+
g (**g**ut)	-	+	+	+	-	+	-	+	-
h (**H**aar)	-	-	+	+	-	-	+	-	-
k (**K**ai)	-	+	+	+	-	+	+	-	-
l (**l**ag)	+	+	-	-	-	-	-	+	-
m (**M**al)	-	+	-	+	+	+	-	+	-
n (**n**un)	-	+	-	-	+	+	-	+	-
p (**P**ein)	-	+	-	+	-	+	+	-	-
r (**r**aus)	+	+	-	-	-	+	-	+	-
s (da**s**)	-	+	-	-	-	-	+	-	+
ʃ (**Sch**eu)	-	+	+	-	-	-	+	-	+
t (**T**al)	-	+	-	+	-	-	-	+	-
v (**V**ase)	-	+	-	+	-	-	-	+	-
x (Da**ch**)	-	+	+	+	-	-	-	-	-
z (**S**inn)	-	+	-	-	-	-	-	+	+
j (**J**od)	-	+	+	-	-	-	-	+	-

Eine vollständige akustische Klassifikation von Phonemen ist jedoch schwierig, weil es keine eineindeutige Beziehung zwischen Phonemen und ihrer akustischen Realisation gibt, d. h. für jedes Phonem gibt es eine fast unendlich große Vielzahl von akustischen Realisationsmöglichkeiten.

Das Zeitsignal von sprachlichen Äußerungen ist zu dieser akustischen Klassifikation relativ ungeeignet, da nur sehr wenige Spracheigenschaften direkt abgelesen werden können (z. B. Einhüllenden-Verlauf, Sprachpausen, „silent interval“ bei stimmlosen Plosiven). Eine bessere Visualisierung von sprachlichen Äußerungen bietet dagegen das Spektrogramm (s.o.).

Im Spektrogramm lassen sich die Formanten und die Formant-Übergänge als Merkmale ablesen. Unterschiedliche Vokale lassen sich daher durch die Lage der Formanten relativ gut charakterisieren. Stimmhafte bzw. stimmlose Konsonanten (z. B. „ba“-„pa“ oder „pa“-„fa“) lassen sich durch die sogenannte voice-onset-time von etwa 20 ms unterscheiden, die sich bei stimmlosen Konsonanten zwischen dem initialen Burst (z. B. Sprengung der Lippenöffnung bei „pa“) und dem Einsetzen der Stimmlippenschwingung beobachten läßt. Ein unterschiedlicher Artikulations-Ort (z. B. zur Unterscheidung von „ba“ und „ga“) läßt sich durch den Zeitverlauf der Formanten unterscheiden. Dieser unterschiedliche Zeitverlauf ist durch den Übergang des Vokaltraktes von der Artikulationsstellung des jeweiligen Konsonanten zum darauffolgenden Vokal bestimmt. Dabei sind die vom Gehör ausgewerteten Formanttransitionen relativ kurz, so daß es bei der Auswertung von Spektrogrammen selbst dem erfahrenen Sprachwissenschaftler schwerfällt, die einzelnen Phoneme voneinander zu unterscheiden. Nasale (z. B. „ma“) sind im Spektrogramm durch ihre Eigenschaft als Halbvokale gekennzeichnet (d. h. relativ stationäre Abschnitte im Zeitverlauf mit geringerer Energie als bei reinen Vokalen). Bei ihnen treten Nullstellen im Spektrum aufgrund der Interferenz zwischen der Abstrahlung durch die Nase und der Abstrahlung durch den geschlossenen Mund auf. Für weitere Details im Bereich der akustischen Phonetik sei auf die einschlägige Literatur (z. B. Kohler, K. Akustische Phonetik) verwiesen.

IV.3 Sprachübertragung und Sprachsynthese

Die Übertragung von Sprachsignalen mit niedrigen Bit-Raten ist insbesondere für das Telefonieren von Interesse, bei der einerseits möglichst viele Telefongespräche gleichzeitig auf einem Übertragungsweg (z. B. Übersee-Kabel, Satelliten-Übertragung oder Funk-Relais-Station für Funktelefone) übertragen werden müssen und andererseits keine wesentlichen Sprachverständlichkeits- und Qualitätsverluste hingenommen werden sollen. Verwandt mit diesem Problem ist das ebenfalls hochaktuelle Problem der Sprachsynthese, mit der z. B. in automatischen Auskunftssystemen Informationen über akustische Kommunikation weitervermittelt werden sollen oder eine akustische Mensch-Maschine-Kommunikation ermöglicht werden soll. Das den meisten Sprach-Übertragungsverfahren zugrundliegende Prinzip ist im folgenden Bild angegeben:

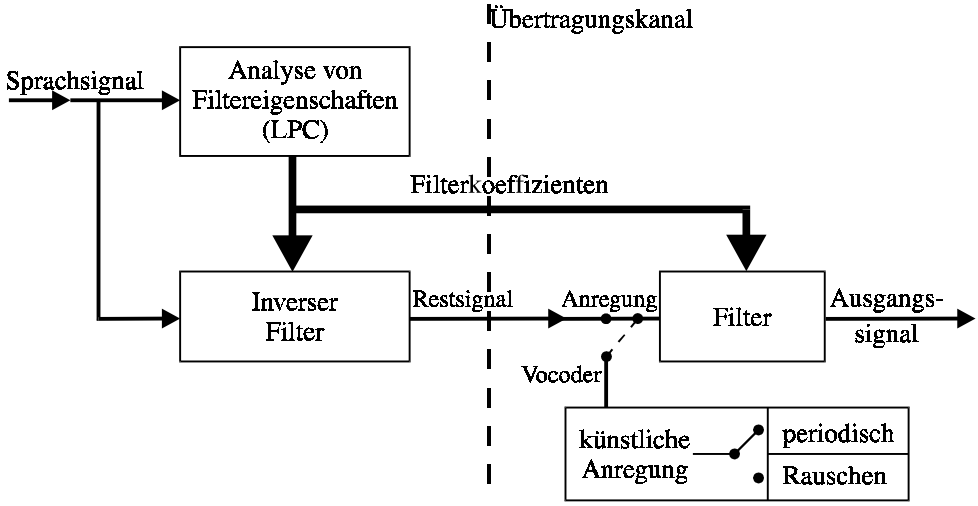

Abbildung 4.9: Prinzip der LPC-basierten Sprachübertragung

Das Ziel ist dabei, die hohe Redundanz im Sprachsignal auszunutzen, um nicht das gesamte Audio-Signal mit hoher Abtastrate übertragen zu müssen (z. B. 16 Bit pro Abtastwert bei einer Abtastrate von 20 kHz, die für die naturgetreue Sprachwiedergabe notwendig ist). Statt dessen soll die wirklich notwendige Information mit weniger Bits übertragen werden. Die grundsätzliche Idee ist dabei, daß das einkommende Sprachsignal zerlegt wird in eine langsam veränderliche (Vokaltrakt-)Filterfunktion und eine schnell veränderliche Anregungs-Funktion des Vokaltraktes mit einem flachen Spektrum, die entweder ein periodisches Signal (bei stimmhaften Konsonanten oder Vokalen) oder ein Rauschen darstellt (bei stimmlosen Vokalen bzw. Frikativen). Während die erste Größe in Form von Filterkoeffizienten (z. B. LPC-Filterkoeffizienten (s. u.) bzw. als hypothetischer Vokaltrakt-Längsschnitt) mit sehr niedriger Abtastrate übertragen werden kann, muß das zweite Signal mit relativ hoher Abtastrate übertragen werden. Allerdings enthält dieses sogenannte **„Restsignal“** oder Anregungssignal nur noch wenig Sprachinformation, so daß es mit einer sehr groben Quantisierung (d. h. mit hoher Ungenauigkeit) übertragen werden kann. Alternativ kann dieses Restsignal auch überhaupt nicht übertragen werden bzw. nur die Information übertragen werden, ob es sich um ein stimmhaftes oder ein stimmloses Signal handelt. In diesem Fall wird auf der Empfangsseite nicht mehr die Original-Stimme zur rekonstruieren versucht, sondern es wird eine Sprachsynthese nach dem Prinzip des **Vocoders** betrieben.

Beim Übergang von der vollständigen Übertragung beider Signale ohne Quantisierungsfehler und der Übertragung der Sprache mit sehr niedriger Datenrate (und entsprechend höherem Quantisierungsfehler) nimmt die Qualität der übertragenen Sprache ab, so daß sie im Extremfall des Vocoders zwar noch verständlich ist, aber sehr unnatürlich klingt. Als weitere

Literatur für Einzelheiten des Aufbaus von Sprachübertragungssystemen und Vocodern sei auf das Buch von Ince, Automatic Speech Processing, Kluver 1992 verwiesen.

Als gängigste Standardmethode zur Extraktion der „effektiven“ Vokaltrakt-Übertragungsfunktion, die in jedem modernen Funktelefon integriert ist, wird das Verfahren des **Linear Predictive Coding (LPC)** verwendet. Dieses Verfahren wurde unabhängig sowohl für die Sprachkodierung als auch für Anwendungen in der Geophysik entwickelt und ist formal äquivalent mit der minimalen Entropie-Spektralanalyse (vgl. Schroeder, M. R., 1989). Dabei wird das n-te Sample des Zeitsignals x(n) aus M vorhergehenden Samples x(n-k) wie folgt geschätzt:

$$x(n) = \sum_{k=1}^{M} a_k \cdot x(n-k) + e(n), \qquad \text{mit } e(n)\text{: Prädiktionssignal} \tag{IV.5}$$

Die Koeffizienten a_k stellen die sogenannten **Prädiktions-Koeffizienten** dar, die so optimiert werden müssen, daß das sogenannte Prädiktions-Fehlersignal e(n) minimal ist, d. h. daß das Signal x(n) optimal aus den vorhergehenden Sampeln vorhergesagt werden kann. Als Bedingung dafür folgt für den mittleren Fehler E_M:

$$\begin{aligned} E_M &= \overline{e^2(n)} \\ &= \overline{x^2(n)} - 2 \cdot \sum_{k=1}^{M} a_k \cdot \underbrace{\overline{x(n) \cdot x(n-k)}}_{=\phi_{k0}} + \sum_{k=1}^{M} \sum_{l=1}^{M} a_k \cdot a_l \cdot \underbrace{\overline{x(n-k) \cdot x(n-l)}}_{=\phi_{kl}} \\ &\overset{!}{=} \min \end{aligned} \tag{IV.6}$$

$$\Rightarrow \qquad 0 = \frac{\partial}{\partial a_k} E_M = -2\phi_{k0} + \sum_{l=1}^{M} a_l \phi_{kl} \tag{IV.7}$$

Diese Gleichung kann man für sämtliche ϕ_{kl} auch als Matrix schreiben, so daß gilt:

$$\underline{\varphi} = \underline{\underline{\phi}} \cdot \underline{a}, \quad \text{wobei } \underline{a} = (a_1, \ , a_M)^T,$$

$$\underline{\varphi} = (\phi_{10}, \phi_{20}, \ldots, \phi_{M0})^T = (R(1), \ldots, R(M))^T,$$

$$\underline{\underline{\phi}} = \begin{pmatrix} \phi_{11} & \cdots & \phi_{1M} \\ \vdots & & \vdots \\ \phi_{M1} & \cdots & \phi_{MM} \end{pmatrix} = \begin{pmatrix} R(0) & R(1) & \cdots & R(M-1) \\ R(1) & R(0) & \ddots & \\ \vdots & \ddots & \ddots & \\ R(M-1) & \cdots & & R(0) \end{pmatrix}, \qquad \text{(IV.8)}$$

$$\text{da } \phi_{kl} = \overline{x(n-k) \cdot x(n-l)} = R(k-l) = R(l-k)$$

(Autokorrelationsfunktion)

Die Elemente der Matrix $\underline{\underline{\phi}}$ sind dabei die Autokorrelations-Funktionswerte für die Verschiebungen um 0, 1 bzw. M – 1 Samples. Da diese Autokorrelationsfunktionen symmetrisch bzgl. ihres Arguments sind, ist auch die Matrix $\underline{\underline{\phi}}$ symmetrisch und hat zudem eine Diagonal-Streifen-Form, die auch als **Toeplitz-Form** bezeichnet wird (positive Diagonal-Streifen mit Maximum in der Hauptdiagonalen). Für dieses einfache lineare Gleichungssystem, das auch als Jule-Walker-Gleichungen bezeichnet wird, gibt es einen effizienten Algorithmus, der als **Levinson-Robinson-Durban** Algorithmus bezeichnet wird. Durch die Existenz dieses Algorithmus wird die LPC-Analyse mit einem ähnlich geringen Rechenaufwand praktisch berechenbar wie eine FFT (vgl. Markel, J. D, Gray, A. H.: Linear Prediction of Speech, Springer Verlag, Berlin 1976).

Um die Auswirkung der LPC-Analyse für den spektralen Gehalt des Analyse-Signals besser zu verstehen, betrachten wir die Darstellung der vorhergehenden Schritte im Frequenzbereich: Wenn das Zeitsignal x(n) ein Spektrum X(f) aufweist, denn gilt für das zeitverzögerte Zeitsignal x(n)

$$\begin{aligned} &x(n) \bullet\!\!-\!\!\circ X(f) \\ &x(n-k) \bullet\!\!-\!\!\circ X(f) \cdot e^{-2 \cdot \pi \cdot i \cdot f \cdot T \cdot K}, \ \text{mit k: Zeitverzögerung um } k \cdot T \end{aligned} \qquad \text{(IV.9)}$$

Für die Darstellung der Prädiktions-Gleichung ergibt sich dann im Frequenzbereich:

$$x(n) = \sum_{k=1}^{M} a_k \cdot x(n-k) + e(n) \qquad \text{(IV.10)}$$

$$X(f) = \underbrace{\sum_{k=1}^{M} a_k \cdot e^{-2 \cdot \pi \cdot i \cdot f \cdot T \cdot k}}_{A(F)} \cdot X(f) + E(f) \qquad \text{(IV.11)}$$

$$\Rightarrow\ X(f)\cdot\left(1-A(f)\right)=E(f),\qquad X(f)=\frac{E(f)}{1-A(f)} \tag{IV.12}$$

Durch das Berechnen der Prädiktions-Koeffizienten a_k wird daher ein Filter A(f) konstruiert, mit dem das Signal x durch Filterung des Fehlersignals e erzeugt wird. Die dabei auftretende Übertragungsfunktion $\frac{1}{1-A(f)}$ besitzt dabei nur Nullstellen im Nenner, d. h. es ist eine **Nur-Pole-Übertragungsfunktion**. Das zugehörige Modell, daß eine spektrale Schätzung aufgrund der Filterung mit einer Nur-Pole-Übertragungsfunktion vorsieht, wird als **AR-Modell** (Auto-Regressive Model) bezeichnet. Analog dazu gibt es auch ein nicht-rekursives Modell, das als **Moving Average (MA)** bezeichnet wird, sowie eine Kombination, die als **ARMA** bezeichnet wird.

Die hier vorgenommene Beschreibung des Sprachsignals durch die Filterung des Anregungssignals mit einer Nur-Pole-Übertragungs-Funktion entspricht daher gut der Vorstellung einer Resonanz-Filterung im Vokaltrakt, bei der die Formant-Frequenzen aufgrund von Helmholtz-Resonatoren erzeugt werden, die wiederum durch Pole in ihrer Laplace-Transformierten gekennzeichnet sind. Diese Analogie zur Resonanz-Filterung im Vokaltrakt kann sogar noch etwas weiter verfolgt werden, wenn anstelle der Prädiktor-Koeffizienten a_k, die daraus eineindeutig berechenbaren **ParCor-Koeffezienten r_k** berechnet werden, die nur einen sehr eingeschränkten Wertebereich zwischen -1 und 1 annehmen können. Diese ParCor-Koeffizienten können als Reflexions-Koeffizienten in Röhren-Segmenten des Röhren-Modells vom Vokaltrakt aufgefaßt werden. Diese Interpretation kann auch in der strukturellen Nachbildung der Wellenausbreitung im Vokaltrakt durch die sogenannte **Lattice-Struktur** erreicht werden, bei der jeweils eine hin- und rücklaufende Welle an aufeinanderfolgenden Grenzschichten reflektiert oder weitergeleitet werden können und die ParCor-Koeffizienten den jeweiligen Reflexionsgrad bestimmen.

IV.4 Spracherkennung

Um eine vollständige akustische Mensch-Maschine-Kommunikation mit gesprochener Sprache zu realisieren, ist neben der oben behandelten Sprachsynthese auch die Spracherkennung von gesprochener Sprache notwendig. Obwohl es in der Vergangenheit viele Anstrengungen in diese Richtung gegeben hat und auch einige Fortschritte erzielt worden sind, sind selbst die leistungsfähigsten Rechner und Algorithmen heute noch immer nicht in der Lage, eine ähnliche Leistungsfähigkeit bei der Spra-

cherkennung auch unter ungünstigen akustischen Situationen zu erreichen, wie der Mensch. Heutige Spracherkennungsalgorithmen erreichen bei **sprecherabhängiger Erkennung** (d. h. Sprecher ist bekannt) eine Erkennungsrate von etwa 95 % und bei **sprecherunabhängiger Erkennung** etwa eine Rate von 90 %. Dieser Wert ist natürlich stark abhängig vom verwendeten Sprachmaterial und dem Wortschatz, sowie einer Reihe weiterer Parameter wie den akustischen Aufnahmebedingungen.

Der grundlegende **Aufbau von Spracherkennungssystemen** ist in unten stehender Abbildung 4.10 skizziert: Das akustische Sprachsignal wird zunächst in einer Vorverarbeitungsstufe in eine spektrogrammähnliche Darstellung transformiert. Dazu wird zumeist eine Filterbank für die Frequenz-Analyse benutzt und der Zeitverlauf wird als Folge aufeinanderfolgender Analyse-Frames dargestellt. Um unabhängig von der jeweiligen Gesamt-Energie des Sprachsignals zu sein, ist eine Energie-Normierung zudem notwendig, die z. B. durch Logarithmierung der Zeit-Frequenz-Darstellung oder durch andere Adaptions-Algorithmen (z. B. Division durch die Gesamt-Energie des Eingangssignals) erreicht werden kann. In diesem Zusammenhang ist auch die **Präemphase** zu erwähnen, d. h. die Anhebung der hohen Frequenzen im Spektrum mit etwa 6 dB pro Oktave (Differenzierung des Zeitsignals), damit das analysierte Spektrum in erster Näherung die gleiche Energie bei hohen und tiefen Frequenzen aufweist. Auf dieser normierten Zeit-Frequenz-Darstellung können bereits einige Sprachmerkmale markiert oder besonders hervorgehoben werden (z. B. die Formanten, die sich insbesondere bei der LPC-Spektralanalyse als scharfe Maxima im Spektrum deutlich abzeichnen).

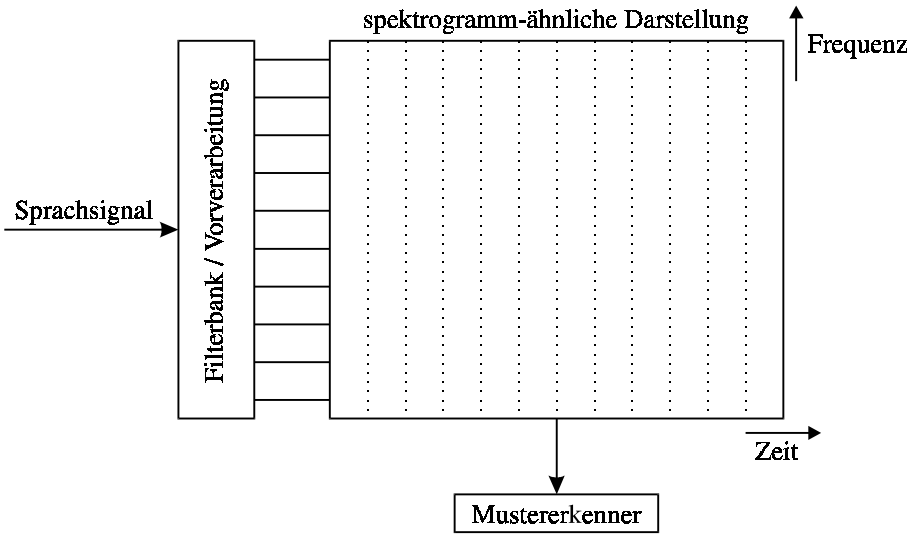

Abbildung 4.10: Prinzipieller Arbeitsweise eines Spracherkennungssystemes

Auf diesem vorverarbeiteten, zweidimensionalen Signal operiert in einem weiteren Schritt ein Mustererkennungs-Algorithmus, der das Muster des jeweils zu erkennenden Wortes aufgrund seiner Ähnlichkeit mit „gelernten" Mustern aus dem Trainings-Wortschatz erkennt. Für diesen Mustererkenner gibt es genau wie für die Vorverarbeitung eine große Vielzahl von Variations-Möglichkeiten, von denen die drei gängigsten im folgenden vorgestellt werden sollen:

IV.4.1 Dynamic-Time-Warping (DTW)-Algorithmus

Diesem Algorithmus liegt die Annahme zugrunde, daß das zu erkennende Wort durch lokale Zeit-Stauchungen und -Dehnungen aus einem der gespeicherten Referenzwörter entsteht. Dies kann z. B. bei der Äußerung desselben Wortes vom selben Sprecher, aber zu unterschiedlichen Zeiten bedingt sein. Der DTW-Algorithmus versucht nun, den Zeitverlauf der Einhüllenden des gesuchten Wortes mit der entsprechenden Einhüllenden des jeweiligen Referenz-Wortes möglichst gut in Einklang zu bringen, indem ein **optimaler Pfad** in der Ebene gesucht wird, die von dem Zeitverlauf des einen Wortes auf der X-Achse und dem Zeitverlauf des anderen Wortes an der Y-Achse aufgespannt wird (als Zeitfunktion kann auch jeweils eine Bandpaß-gefilterte Version des jeweiligen Wortes verwendet werden. Dasjenige Wort aus dem Trainingswortschatz wird erkannt, bei dem nach Optimierung des Angleichungs-Pfades der geringste Abstand resultiert. Der Vorteil dieses Verfahrens ist der sehr **geringe Trainingswortschatz** (für jedes zu erkennende Wort des Wortschatzes braucht im Prinzip nur eine akustische Realisation aufgenommen werden) und der relativ geringe Rechenaufwand beim Training, der allerdings einem etwas höheren Rechenaufwand in der Erkennungsphase gegenübersteht. Der Nachteil des DTW-Algorithmus ist seine Sprecherabhängigkeit (d. h. bei Verwendung eines anderen Sprechers als beim Referenz-Wortschatz sinkt die Erkennungsrate deutlich) und die insgesamt relativ hohe Fehlerrate des Algorithmus. Dieser Algorithmus wird daher in jüngerer Zeit relativ selten eingesetzt.

IV.4.2 Hidden-Markov-Modelle (HMM)

Diesem Muster-Erkennungs-Algorithmus liegt die Annahme zugrund, daß das gesprochene Wort durch eine **Abfolge von Zuständen** generiert wird, die mit einer gewissen Wahrscheinlichkeit aufeinander folgen und eine jeweils zufällige akustische Realisation bewirken. Jeder dieser Zustände kann beispielsweise ein Phonem (oder ein Teil eines Phonemes oder eine Phonemenkette) repräsentieren, das im nächsten Zeitschritt entweder noch immer vorliegt oder durch ein nachfolgendes Phonem abgelöst wird. Aufgrund der Übergangswahrscheinlichkeiten von einem Zustand zum nächsten Zustand kann dieser Vorgang als eine Markov-Kette

beschrieben werden. Die einzelnen Zustände dieser Markov-Kette sind jedoch nicht direkt beobachtbar, weil man zwar die akustische Realisation, nicht aber das hier zugrundeliegende Phonem kennt. Aus diesem Grunde redet man von „versteckten“ Zuständen der Markov-Kette (Zustände 1, 2, 3, 4, ... in unten stehender Abbildung). Für jeden Zustand j tritt bei seinem Vorliegen eine (zufällige) Auswahl von akustischen Realisationen dieses Zustandes S^i_k statt.

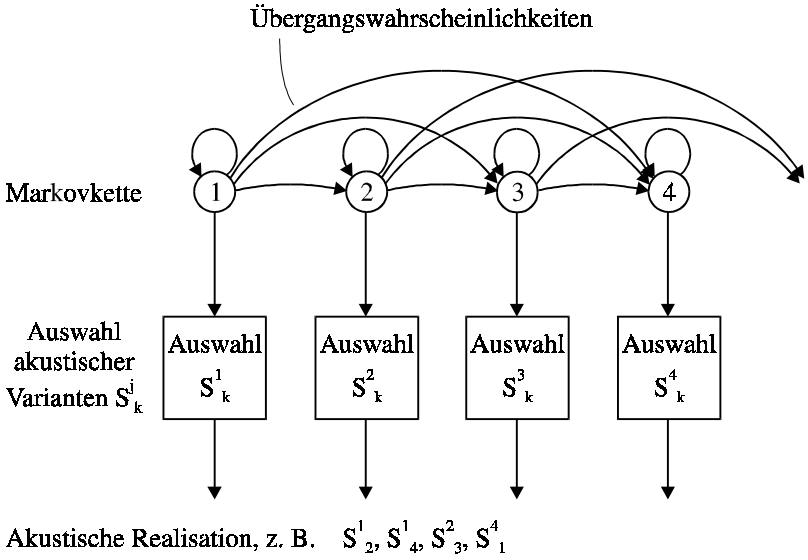

Abbildung 4.11: Struktur des Hidden-Markov-Modells

In der Trainingsphase muß nun für jedes Wort ein eigenes „Modell“ gelernt werden, das durch die Anzahl der „versteckten“ Zustände, ihre Übergangswahrscheinlichkeiten und ihre akustischen Realisationen mit jeweiliger Auftretenshäufigkeit charakterisiert wird. Diese Größen können nur durch eine Statistik über eine Vielzahl von Realisationen ein und desselben Wortes gewonnen werden, so daß ein HMM-Spracherkenner einen sehr großen Trainingswortschatz benötigt. Der Vorteil eines derartigen Algorithmus ist die **relativ hohe Trefferquote**, die auch sprecherunabhängig erreicht werden kann (solange der Trainingswortschatz auch von mehreren Sprechern aufgenommen wurde). Zu den Nachteilen des HMM-Algorithmus gehört der extrem hohe benötigte **Trainingswortschatz**, ohne den das Modell nicht erfolgreich funktionieren kann. Ein weiterer Nachteil ist der relativ **hohe Rechenzeitaufwand**, weil zu einem vorgegebenen Wort die Wahrscheinlichkeit für jedes Referenz-Modell errechnet werden muß, daß dieses jeweilige Modell die beobachtete Sequenz von akustischen Realisationen erzeugt hat. Aufgrund seiner hohen Erkennungsrate und sonstiger Vorteile ist der HMM-Erkenner derzeit die am häufigsten eingesetzte Form der künstlichen Spracherkennung.

IV.4.3 Neuronales Netz

In den letzten Jahren wurden zunehmend neuronale Netze für klassische Aufgaben der Mustererkennung eingesetzt aufgrund ihrer Eigenschaft, sich selbst zu organisieren und bei entsprechender Wahl der Parameter auch die einmal „gelernten“ Klassifikationsregeln generalisieren zu können. Das am häufigesten dabei angewandte Netz ist das **Multi-Layer-Perceptron**. Bei ihm tritt ein klarer Signalfluß auf, der bei einer Eingangsschicht von Neuronen startet, die Verbindungen zu einer (oder mehreren) Zwischenschichten aufweisen. Diese weisen wiederum nur Verbindungen zu den darauffolgenden Schichten (bzw. der darauffolgenden Schicht) auf. Auf die Eingangsschicht wird dabei die zweidimensionale Zeit-Frequenz-Darstellung des Sprachsignals gegeben. In der Ausgangsschicht sollte die „Zelle“ die maximale Aktivität entfalten, die zu dem zu erkennenden Wort gehört. Um das Netz zu **trainieren** werden mehrere Realisationen desselben Wortes benötigt, für die jeweils das Eingangsmuster und das gewünschte Ausgangsmuster dem Netz vorgegeben wird. In dem Lehrvorgang werden nun die Gewichte, mit denen die in der vorausgehenden Schicht liegenden Neuronen-Aktivitäten (für jedes Neuron in der nachfolgenden Schicht) verrechnet werden, gemäß einer Lernvorschrift adaptiert. Für das häufigst genutzte Beispiel des Multi-Layer-Perceptrons gibt es einen effizienten Adaptionsalgorithmus der Gewichte, den sogenannten Backpropergation-Algorithmus (vgl. Rummelhard McClellan., s. Literaturliste). Die zugrundeliegende Idee ist nun, daß das Wortmuster vom Netz in Form eines Eins-aus-n-Kodierers klassifiziert wird. Falls die Anzahl der Neuronen im Netz günstig gewählt ist, kann das Netz „generalisieren“, d. h. es kann auf wesentliche Eigenschaften der Eingangssignale ansprechen und für die Klassifikation unwichtige Eigenschaften der Zeit-Frequenz-Darstellung des Wortes ingnorieren. Der Vorteil von einem neuronalen Netz-Erkenner ist die sehr einfache und schnelle Berechnung des erkannten Wortes, sobald das Netz erst einmal trainiert ist. Der Nachteil des neuronalen Netz-Spracherkenners liegt in dem relativ großen Trainingswortschatz und den leider sehr begrenzten Erkennungsraten. Obwohl das Konzept des neuronalen Netzes sich eng an der Struktur von biologischen Nervensystemen anlehnt und auch einige interessante Eigenschaften (z. B. selbst organisiertes Lernen) damit erreicht werden können, ist dieses Konzept den konventionellen Verfahren zur Mustererkennung und Musterklassifikation (z. B. HMM-Algorithmus für die Spracherkennung) nicht überlegen (vgl. Behme, H., Dissertation, Universität Göttingen).

IV.5 Stimmpathologie

Entsprechend den bereits unter IV.1 beschriebenen Einzelheiten zur Stimmerzeugung und zur Physiologie der Stimme kann es eine Reihe von sehr unterschiedlichen Störungen in der Stimm-Mechanik geben, die zu pathologischen Stimmveränderungen führen. Neben Regulations- und Steuerungsstörungen durch das zentrale Nervensystem (vgl. Phoniatrie-Lehrbuch, z. B. Wendler et al., 1996) wird oft die Mechanik der Stimmlippen-Schwingung beeinträchtigt, auf die allerdings nur kurz eingegangen werden soll. Aus akustischer Sicht sind dabei die drei Grundtypen von Stimmpathologien zu unterscheiden: Hauchigkeit, Rauhigkeit und Diplophonie.

Eine „Hauchigkeit“ der Stimme tritt bei einem unvollständigen Glottis-Schluß auf, der beispielsweise durch eine einseitige Glottis-Lähmung hervorgerufen werden kann. Für die Phonation wird daher ein sehr großer Luftstrom benötigt, so daß zusätzlich zu der periodischen Schwingung starke Atemgeräusche hörbar werden und der akustische Eindruck einer verhauchten, nicht rein tönenden Stimme entsteht.

Für die „Rauhigkeit“ von Stimmlippenschwingungen sind Verdickungen bzw. Strukturunregelmäßigkeiten der Stimmlippen verantwortlich (z. B. Polypen, Schleim oder eine Entzündung). Sie führt zu veränderten Schwingungseigenschaften (z.B. zu Amplituden- und Frequenzmodulationen) und zu einem veränderten Spektralgehalt der Glottis-Schwingung, was sich subjektiv als „rauhe“ Stimme bemerkbar macht. Diese Rauhigkeit ist dabei nicht mit der psychoakustischen Rauhigkeit zu verwechseln.

Für die Diplophonie (d. h. Phonation mit zwei unterschiedlichen Frequenzen) ist eine Asymmetrie der Stimmlippen verantwortlich, die beispielsweise durch eine einseitige Lähmung oder durch einen asymmetrischen Befall mit Polypen bedingt ist. Dabei können die beiden Stimmlippen mit unterschiedlicher Frequenz schwingen, so daß der Eindruck einer nichtstabilen Grundfrequenz auftritt.

Als wichtigste Untersuchungstechnik in der Stimmdiagnostik wird die **Stimmgrundfrequenz-gesteuerte Stroboskopie** verwendet. Bei ihr wird die Glottis mit Hilfe eines starren oder flexiblen Endoskops beobachtet und zugleich mit einem Blitzlicht beleuchtet, das eine Blitzfolge-Frequenz aufweist, die leicht gegenüber der Stimmfrequenz verschoben ist. Durch die dabei auftretende Schwebung lassen sich sämtliche Phasen der periodischen Stimmlippenschwingung im zeitverlangsamten Ablauf beobachten, so daß man einen Überblick über die dynamische Bewegung der

Stimmlippen erhalten kann. Da diese Technik nur mit lang angehaltenen Vokalen mit relativ stabiler Grundfrequenz angewandt werden kann, stößt sie in der Praxis bei Patienten auf Schwierigkeiten, die nicht in der Lage sind, einen Ton vorgegebener Tonhöhe eine gewisse Zeit lang unter Endoskopbetrachtung auszuhalten. Als Alternative bietet sich daher die in jüngerer Zeit eingeführte Hochgeschwindigkeitsglottografie an, bei der die Stimmlippen-Schwingung mit mehreren Belichtungen bei jeder Periode der Stimmlippenschwingung abgebildet wird. Auf diese Weise können auch kurze Zeitabschnitte von Stimmlippen-Schwingungen beurteilt werden.