Entwicklung einer Klassifikationsmethode zur akustischen Analyse fortlaufender Sprache unterschiedlicher Stimmgüte mittels Neuronaler Netze und deren Anwendung

Dissertation
zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultäten
der Georg-August-Universität zu Göttingen

vorgelegt von
Jan Lessing
aus
Göttingen

Göttingen 2007

D 7
Referent: Prof. Dr. M. R. Schroeder
Koreferent: Prof. Dr. E. Kruse
Tag der mündlichen Prüfung: 17.07.2007

„Die edelste Beschäftigung des Menschen
ist der Mensch“

Gotthold Ephraim Lessing
(1729 – 1781)

Inhaltsverzeichnis

Abkürzungsverzeichnis

α	Momentum (Trägheitsterm) beim Training der NN
AKF	Autokorrelationsfunktion
CEP	Cepstrum
DPF	Directional Perturbation Factor
EPQ	Energy Perturbation Quotient
FFD	Fundamental Frequency Distribution
GHD	Göttinger Heiserkeits-Diagramm
GHDT	Göttinger Heiserkeits-Diagramm Textanalyse
GNE	Glottal to noise excitation ratio
HNR	Harmonics to noise ratio
HPS	Harmonisches Produktspektrum
IPA	International Phonetic Alphabet
KKK	Kreuzkorrelationskoeffizient
LPC	Linear Predictive Coding
LTAS	Long Term Average Spectrum
MLP	Multi-Layer-Perzeptron
MWMC	Mean Waveform Matching Coefficient
NDR	Nulldurchgangsrate
NN	Neuronales Netz(-werk)
η	Lernrate bei Training der NN
PA	Pitch Amplitude
PF	Perturbation Factor
PPQ	Pitch Perturbation Quotient
PQ	Perturbation Quotient
RMS	Root-Mean-Square
SAMPA	Speech Assessment Methods Phonetic Alphabet
SFR	Spectral Flatness Ratio
SFM	Spectral Flatness Measure
SNR	Signal to Noise Ratio
T_0	Grundperiodenlänge
WMC	Waveform Matching Coefficient

Vorwort

Diese Dissertation ist am Dritten Physikalischen Institut in Göttingen in der Arbeitsgruppe *Sprache und Neuronale Netze* angefertigt worden. Unter der Betreuung von Prof. Manfred R. Schroeder ist die vorliegende Arbeit in enger Zusammenarbeit mit der Abteilung Phoniatrie & Pädaudiologie des Universitätsklinikums in Göttingen unter der Leitung von Prof. Eberhard Kruse entstanden. Ziel dieser Arbeit war die Entwicklung einer automatischen Klassifikationsmethode für die akustische Analyse fortlaufender Sprachsignale jeglicher Stimmgüte auf der Grundlage von Kurzzeitsegmenten, die nach Ihrer Stimmanregung in unterschiedliche Klassen (stimmhaft/stimmlos) eingeteilt, und anschließend in zusammenhängenden Bereichen der gleichen Klasse analysiert werden.

Die akustische Analyse von Sprachsignalen normaler und gestörter Stimmfunktion basiert in der Praxis zumeist auf gehaltener Phonation aus Mangel an automatisierten Verfahren zur Analyse fortlaufender Sprache. Dieser Ansatz liefert nicht notwendigerweise repräsentative Ergebnisse auch für fortlaufend gesprochene Sprache, die den eigentlichen Anwendungsbereich von Sprache darstellt. Lediglich bei einzelnen Stimmstörungsbildern kann eine alleinige Beurteilung gehaltener Phonation ausreichend sein, jedoch keinesfalls für den Großteil der Stimmanalysen [AH86]. Eine alleinige Beschreibung der Stimmgüte anhand von akustischen Stimmgütemaßen reicht weder aus gehaltener Phonation noch aus fortlaufender Sprache für eine objektive Darstellung aus. Die gehaltene Phonation als Analysegrundlage spiegelt zudem eher die Eigenschaften der Singstimme wider [K90].

Um Aussagen über das Schwingungsverhalten der Stimmlippen anhand des akustischen Signals treffen zu können, ist eine Selektion von Teilsegmenten stimmhafter Anregung (periodische Stimmlippenschwingung) aus dem fortlaufenden Sprachsignal notwendig. Bei der Analyse gehaltener Phonation kann das gesamte Signal – abgesehen von Ein- und Ausschwingvorgängen – als stimmhaft betrachtet werden. Aus Mangel an zuverlässigen automatischen Klassifikationsmethoden zur Bestimmung von Teilsegmenten stimmhafter und stimmloser Phonation aus fortlaufender Sprache – gerade für hochgradig gestörte Stimmen – findet eine Analyse fortlau-

fender Sprachäußerungen zumeist anhand von perzeptiven Beurteilungen durch ein Gutachterkollektiv statt. Die Möglichkeit einer automatisierten, unüberwachten akustischen Analyse fortlaufender Sprachsignale ist apparativ kostengünstiger, objektiver und zuverlässiger reproduzierbar als eine perzeptive Beurteilung.

Die Verwendung akustischer Maße zur Quantifizierung der Stimmgüte auf Basis einer automatischen Klassifikationsmethode ermöglicht eine Beurteilung auch aus fortlaufender Sprache und liefert zusammen mit den aus gehaltener Phonation bestimmten Ergebnissen eine umfassende Beschreibung der Stimmgüte. Die Entwicklung dieser Methode stellt den Kern der vorliegenden Arbeit dar und findet direkte Anwendung in der Bestimmung und Validierung ausgewählter akustischer Stimmgütemaße.

Stand der Wissenschaft

Für einen als natürlich empfundenen Klang der Stimme sind geringe Variationen der Schwingungsamplitude und -dauer einzelner aufeinander folgender Schwingungszyklen der Stimmlippen verantwortlich [H63] und als physiologisch bekannt [L61]. Erreichen diese Perturbationen größere Ausmaße, führen sie zu einem akustisch wahrgenommenen *rauen* Stimmklang [C71]. Diese Irregularitäten eines normalen Schwingungsverhaltens der Stimmlippen resultieren zumeist aus pathologischen Störungen des laryngealen Systems [HK71]. Der Einsatz akustischer, aus dem Stimmsignal bestimmter Maße zur automatisierten Detektion und differenzierten Beurteilung einer Stimmstörung stellt den Ausgangspunkt diverser Arbeiten auf diesem Forschungsgebiet dar und liefert hohe Korrelationen zwischen laryngealen Pathologien und abgeleiteten akustischen Maßen [D81], [B87], [ECH90], [KFM98], [BO00].

Eine wichtige Basis für die Bestimmung wesentlicher akustischer Maße (bspw. Jitter und Shimmer) stellt die exakte Detektion der einzelnen Grundperioden im Signalverlauf dar. Die Berechnung relativer Differenzen der Periodenlängen aufeinander folgender Schwingungszyklen liefert bei Gubrynowicz, Mikiel und Zarnecki unter Einbeziehung eines Fuzzy Algorithmus mit diagnostischem Modell eine Klassifikationsgüte von 72 % bei der Unterscheidung normaler von gestörter Stimmfunktion [GMZ80]. Die Entwicklung immer robusterer Algorithmen zur Grundfrequenzbestimmung – insbesondere solcher, die unabhängig von Periodizitätskriterien arbeiten – hat die Entwicklung der akustischen Analyse maßgeblich unterstützt [M87], [MYC91], [DPH93], [TL93], [H95], [BKGD96], [RLM97], [PJ99].

Bereits 1971 haben Hecker und Kreul [HK71] erste Arbeiten zur akustischen Analyse fortlaufender Sprache von je 5 Sprechern mit und ohne Stimmstörung publiziert. Eine Analyse der Grundfrequenz und daraus abgeleiteter Perturbationsmaße des zweiten Satzes der Regenbogen-Passage[1] erfolgte anhand von Filmaufnahmen eines 2-Strahl-Oszillographenbildes des gefilterten Sprachsignals. Eine manuelle Identifikation einzelner Glottispulse auf dem entwickelten Filmmaterial bildete die Grundlage für die Selektion stimmhafter und stimmloser Teilsegmente. Die Beurteilung der Zuverlässigkeit einer Differenzierung zwischen gestörter und normaler Stimmfunktion anhand dieser berechneten Maße stand dabei im Vordergrund.

Kasuya und Wakita [KW79] nutzen aus der Linearen Prädiktion digitalisierter Sprachsignale abgeleitete akustische Maße (*back-to-total cavity volume ration (BTR)* und *RMS* der Signalenergie) zur sprecherunabhängigen stimmhaft/stimmlos-Klassifikation und erreichen bei ausgewählten Sprechern normaler Stimmfunktion Klassifikationsgüten von bis zu 93 %. Siegel und Bessey verwenden die Nulldurchgangsrate des Sprachsignals als Kriterium für Stimmhaftigkeit. Unterschreitet diese einen bestimmten Schwellwert, so handelt es sich um ein stimmhaftes Segment [SB82]. Probleme stellt diese Vorgehensweise allerdings bei hochgradigen Stimmstörungen dar. Die Beschreibung eines adaptiven Algorithmus zur automatischen Segmentierung fortlaufender Sprache von Stotterern in Bereiche stimmhafter und stimmloser Phonation sowie Sprechpausen stellen Lucas und Hudson 1994 vor [LH94]. Auf der Basis einer Filterbankanalyse sind verschiedene Formantfrequenzen berechnet und als Grundlage für die Klassifikation herangezogen worden. Auch bei Parsa et al. sind Klassifikationsuntersuchungen publiziert [PJ00].

Bettens, Grenez und Schoentgen stellen eine Analysemethode – ähnlich der von Qi, Hillman und Milstein [QHM99] basierend auf der Linearen Prädiktion – vor und vergleichen die Analyseergebnisse aus gehaltener Phonation und fortlaufender Sprache [BGS05]. Sie erhalten hohe Korrelationen der *signal-to-disperiodicity ratio* für gehaltene Vokale und fortlaufende Sprache und auch innerhalb der fortlaufenden Sprache zwischen Teilstücken mit wenigen und solchen mit häufigen *voiceonset* und *-offset* Elementen. Vergleichende Untersuchungen des Signal-Rausch-Verhältnisses aus gehaltener Phonation und fortlaufender Sprache von Klingholz [K90] unterstreichen die Unvollständigkeit einer alleinigen Analyse gehaltener Phonation. Die Ergebnisse der Bestimmung des SNR aus fortlaufender Sprache weisen in dieser Studie höhere Korrelationen mit den laryngealen Bedindungen auf als die aus gehaltener Phonation und stellen somit eine zusätzliche Information dar.

[1] Die *rainbow passage* [F60] dient als englischer Standardtext für Sprachdatenbanken (siehe Anhang 7.3).

Hammarberg et al. sind zu dem Ergebnis gekommen, dass eine alleinige Beurteilung der Stimmgüte aus gehaltener Phonation unzureichend ist, da sie die dynamischen Aspekte der fortlaufenden Sprache nicht erfasst [HFGS80]. De Krom unterstreicht diese Aussage, da gehaltene Phonation nicht repräsentativ für den alltäglichen Gebrauch der Stimme in der Kommunikation ist [K95]. Auch andere Arbeitsgruppen gelangen in ihren Untersuchungen zu dem gleichen Ergebnis [AH86], [TK75], [QH97], [S89], [LFMS99], [WSDHEL06]. Die gehaltene Phonation, wie auch die fortlaufende Sprache weisen individuelle Vor- und Nachteile in der Beurteilung der Stimmgüte auf und stellen nur gemeinsam eine umfassende Grundlage dar [PJ01].

Die Anpassung bereits etablierter akustischer Maße aus der Beurteilung gehaltener Phonation an fortlaufende Sprache und auch deren Weiterentwicklung stellen die Grundlage verschiedener Publikationen dar, wie bei Klingholz [K90]. So stellen bspw. Qi und Hillman 1997 robustere Verfahren zur Bestimmung des HNR im Zeit- und Frequenz-Bereich vor. Ebenfalls von Qi, Hillman und Milstein stammt eine verfeinerte Methode zur Bestimmung des SNR aus fortlaufender Sprache [QHM99]. Die Auswahl und Zusammensetzung repräsentativer Maße zur umfassenden Beschreibung der Stimmqualität spielen dabei eine wichtige Rolle [PMWH87], [MS95], [FSK97], [PJ01].

Vergleichsuntersuchungen perzeptiver Ergebnisse mit denen akustischer Analysemethoden fortlaufender Sprache sind unter anderem bei Hammarberg et al. [HFGS80] beschrieben. Anhand von Energieschwellwerten in den niedrigen Frequenzbändern einer 51-kanaligen Filterbank zur Bestimmung stimmloser Teilsegmente ist eine differenzierte Beurteilung von stimmhafter Phonation im Vergleich zum gesamten Sprachsignal vorgenommen worden. Askenfeld und Hammarberg untersuchten weiterhin die Aussagekraft von sieben verschiedenen Perturbationsmaßen im Vergleich zu parallel erhobenen perzeptiven Beurteilungskriterien und haben festgestellt, dass Variationen in Tonhöhe und Lautstärke in fortlaufender Sprache wichtige Indikatoren für eine gestörte Stimmfunktion darstellen [AH86]. Auch bei Eadie und Doyle sind Vergleichsuntersuchungen anhand verschiedener akustischer Maße zu perzeptiven Beurteilungen beschrieben [ED05].

Umapathy et al. beschreiben in 2005 einen Zeit-Frequenz-Ansatz zur Klassifikation pathologischer Stimmen ohne die sonst notwendige stimmhaft/stimmlos-Segmentierung mittels eines adaptiven Zeit-Frequenz-Transformationsalgorithmus und daraus abgeleiteter Maße, wie *Oktav-Maße, energy-ratio, length-ratio, frequency-ratio* und andere. Die eigentliche Klassifikation erfolgt unter Verwendung eines statistischen Muster-Klassifikators [UKPJ05].

Der Einsatz trainierter künstlicher Neuronaler Netze zur Klassifikation stimmhafter und stimmloser Segmente aus dem Sprachsignal – wie in dieser Arbeit – ist auch in anderen Arbeiten verfolgt worden, wie bei Qi und Hunt [QH93], [LFMS99b], [LMFK01]. Godino-Llorente und Gomez-Vilda benutzen ebenfalls ein Multi-Layer Perceptron zur Klassifikation [GG04]. Callan et al. differenzieren mittels selbstorganisierender Merksmalskarten unterschiedliche Stimmfunktionen [CKRT99]. Das in dieser Arbeit entwickelte Klassifikationsverfahren basiert auf der Beurteilung der Stimmhaftigkeit einzelner Kurzzeitsegmente mittels trainierter Neuronaler Netze und ermöglicht somit eine nichtlineare Klassifikation der Kurzzeitsegmente entsprechend dem artikulierten Phonem.

In der Literatur sind allerdings auch vereinzelt Untersuchungen publiziert, die eine Analyse fortlaufender Sprache zur Beschreibung der Stimmgüte aus Mangel an zuverlässigen Verfahren nicht empfehlen, da der Einfluss des phonetischen Kontextes, der Anspannung des Sprechers oder Variationen der Intonation die Analyseergebnisse u. U. verfälschen könnten [HMD73], [IL70], [MD80]. Diese Studien unterstreichen den Bedarf an zuverlässigen Klassifikationsalgorithmen zur Analyse fortlaufender Sprache. An dieser Stelle versucht die vorliegende Arbeit eine Lücke zu schließen und stellt ein automatisiertes Verfahren zur akustischen Analyse fortlaufender Sprache jeglicher Stimmgüte vor.

Aufbau

Nach einem kurzen Umriss der Thematik und einer einleitenden Motivation in Kapitel 1 werden in Kapitel 2 akustische Maße zur Beschreibung fortlaufender Sprachsignale und deren stimmstörungsbedingter Perturbationen bestimmter Stimmeigenschaften vorgestellt. Die Entwicklung einer Methode zur Selektion stimmhafter Phoneme aus dem fortlaufenden Sprachsignal unter Verwendung von Modellen künstlicher Neuronaler Netze ist in Kapitel 3 dargelegt und stellt eine wesentliche Grundvoraussetzung für die Analyse dar. Daran anschließend werden in Kapitel 4 der Ablauf und die apparativen Voraussetzungen für die Datenakquise sowie das analysierte Sprachdatenmaterial beschrieben. Die Darstellung der Ergebnisse zur Entwicklung der Klassifikationsmethode und der unter Verwendung dieser Methode berechneten akustischen Stimmgütemaße erfolgt in Kapitel 5. Eine abschließende Zusammenfassung mit Diskussion und Ausblick in Kapitel 6 soll diese Arbeit im Umfeld anderer Methoden zur akustischen Analyse fortlaufender Sprache beleuchten. Den Abschluss dieser Arbeit bilden der Anhang, das Literaturverzeichnis sowie ein Index.

1 Einleitung

Das wohl wesentlichste Mittel der menschlichen Kommunikation stellt die Sprache dar. Über einen Zeitraum von mehreren 10.000 Jahren hat sich die menschliche Lautsprache, eng verbunden mit der gesamten Entwicklung des Menschen, ausgebildet. Die heutige Anatomie des menschlichen Stimmtraktes hat sich in rund 200.000 Jahren bis vor ungefähr 35.000 Jahren – dem Auftreten des *Cro-Magnon*-Menschen[2] – entwickelt [Z86]. Die Evolution hat in dieser Zeitspanne eine optimale Anpassung des menschlichen Apparates zur Erzeugung der Sprache – mit dem daraus resultierenden akustischen Signal – und der Mechanismen der Sprachaufnahme im Gehör und Gehirn aneinander hervorgebracht.

Im Gegensatz zu anderen Säugetieren hat sich beim Menschen ein Lauterzeugungsmechanismus ausgebildet, der durch einen extrem tief sitzenden Kehlkopf, eine rund in den Hals abfallende Zunge und einen besonders hoch gewölbten harten Gaumen gekennzeichnet ist. Diese anatomischen Merkmale führen zu einem vergrößerten, fein modulierbaren Resonanzraum oberhalb der Stimmlippen und erlauben eine sehr differenzierte und umfangreiche Lautbildung, die Voraussetzung für eine komplexe Lautsprache – wie die menschliche Sprache – ist [Z86].

1.1 Sprachproduktion

Der menschliche Sprechapparat ermöglicht die Bildung unterschiedlicher Laute, die bei Verkettung mit einer speziellen Grammatik – die sich je nach Sprache unterscheiden kann – eine Sprachäußerung darstellen. Auf diesem Wege haben sich weltweit die unterschiedlichsten Sprachen und Dialekte zur Kommunikation entwickelt. Der Mechanismus der Sprachproduktion ist dabei allen Sprachen gleich und lässt sich in verschiedene Komponenten aufteilen, die sich getrennt beschreiben und unabhängig voneinander modellieren lassen: die Stimmanregung, die Artikulation und die Abstrahlung.

[2] *Homo sapiens sapiens*: Von Handwerkern 1868 bei Cro-Magnon in Frankreich entdeckt.

Die Stimmanregung erfolgt auf glottaler Ebene durch Modulation eines Luftstroms und stellt die *Quelle* des Sprachsignals dar. Die Artikulation findet im supralaryngealen Bereich statt und kodiert die zu übermittelnde Information in unterschiedlichen Lauten durch *Filterung* des Anregungssignals. Diese werden schließlich an den Lippen abgestrahlt und pflanzen sich als Schalldruckwelle in der Luft fort.

Bei der Modellierung dieses Zusammenhangs mit digitalen Filtern ergibt sich das Ausgabesignal $x(n)$ als Faltungsprodukt von Eingangssignal $e(n)$ und Impulsantwort $h(n)$. Da eine Faltung im Zeitbereich einer Multiplikation im Frequenzbereich entspricht, kann man obiges Modell in Notation der z-Transformierten [J64] darstellen. Das Sprachsignal $X(z)$ lässt sich in ein Modell der glottalen Anregungsfunktion $E(z)$, ein Glottismodell $G(z)$, ein Vokaltraktmodell $V(z)$ und eine Modellvorstellung für die Abstrahlung an den Lippen $L(z)$ entsprechend Gleichung 1.1 zerlegen.

$$X(z) = E(z)G(z)V(z)L(z) \tag{1.1}$$

Die glottale Anregung $E(z)$ stellt bei stimmgesunden Sprechern in stimmhafter Phonation eine periodische Folge von Glottispulsen dar. Bei der Artikulation stimmloser Phoneme steht die Glottis offen und der Vokaltrakt wird durch Rauschen mit flachem Spektrum angeregt. An Verengungen im glottalen Bereich können dabei Turbulenzen entstehen. Bei gestörten Stimmen setzt sich $E(z)$ in stimmhafter Phonation infolge irregulärer Stimmlippenschwingungen – und einem daraus resultierenden inkompletten Glottisschluss – aus einer Überlagerung dieser beiden Anregungsformen zusammen. Diese Überlagerungen treten auch bei stimmgesunder Artikulation bestimmter Laute, wie z.B. der stimmhaften Frikative, auf.

Dieser lineare Modellansatz der Sprachproduktion wird als *Quelle-Filter*-Modell bezeichnet [MG76], [F81]. Obwohl dieser, in Abbildung 1.1 skizzierte Modellansatz

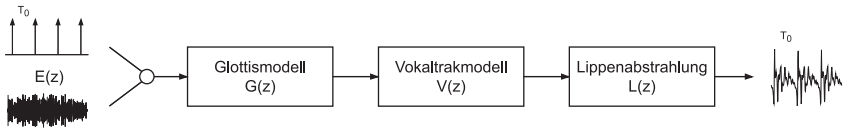


Abb. 1.1: *Lineares Modell der Spracherzeugung.*

eine Vereinfachung der physikalischen Wirklichkeit darstellt, erlaubt er doch eine hinreichend gute Approximation der für die Wahrnehmung wichtigen Sprachmerkmale. Nennenswerte Nichtlinearitäten ergeben sich bspw. durch Ankopplung eines selbstschwingenden oder aerodynamischen Glottismodells.

Der physiologische Aufbau des menschlichen Sprechapparates ist dem Mediosagittalschnitt in Abbilundung 1.2 zu entnehmen.

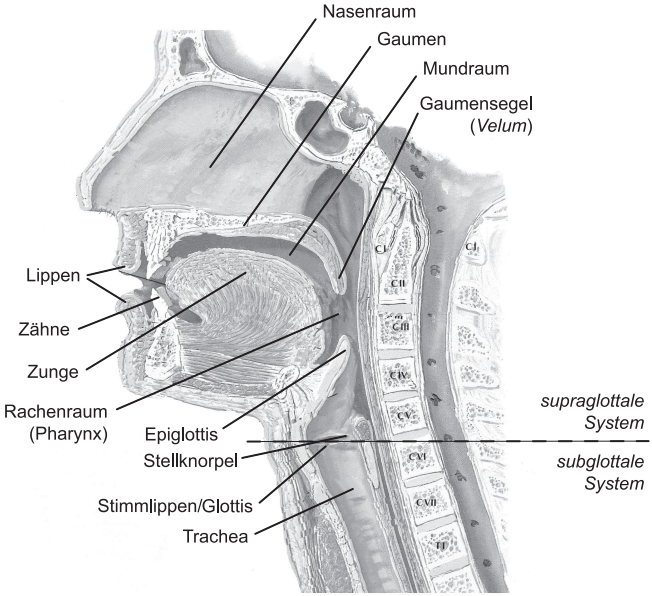


Abb. 1.2: *Mediosagittalschnitt des Sprachproduktionsapparates des Menschen.*

Der aus den Lungen beim Ausatmen entweichende Luftstrom trifft nach Durchlaufen der Luftröhre (*Trachea*) auf die Stimmritze (*Glottis*). Durch Positionsveränderungen der Stellknorpel (*Aryknorpel, Cartilagines arytenoideae*) im Kehlkopf (*Larynx*) werden die Stimmlippen, deren elastische Eigenschaften sich durch Variation der Muskelspannung im *Vokalismuskel*[3] verändern lassen, in Phonationsstellung gebracht. Der durch die somit verschlossene Glottis gestaute Luftstrom aus den

[3] Der *Vokalismuskel* stellt den Hauptbestandteil des Stimmlippengewebes dar.

Lungen führt zu einem Anstieg des subglottalen Drucks, der beim Überschreiten eines bestimmten Druckschwellwertes ein Auseinanderdrücken der Stimmlippen – in Abhängigkeit vom Stimmlippentonus – in lateraler Richtung bewirkt. Die Glottis öffnet sich und der gestaute Luftstrom entweicht schlagartig.

Der von der Luft durchströmte Querschnitt ist auf Stimmlippenebene deutlich geringer als ober- und unterhalb des Larynx, so dass die Luft diese Engstelle mit höherer Geschwindigkeit durchströmt. Hierdurch entsteht im Bereich der Stimmlippen ein Unterdruck, der quer zur Strömungsrichtung wirkt und als *Bernoulli*-Effekt bezeichnet wird [B58]. Die *myoelastische Rückstellkraft* der gespannten Muskeln und die Bernoullikraft übersteigen zusammen die durch den subglottalen Druck wirkende Kraft deutlich, so dass sich die Stimmlippen wieder aufeinander zu bewegen. Die lateral zurückschwingenden Stimmlippen verschließen beim Aufeinandertreffen abrupt die Glottis wieder. Mit Unterbrechung des Luftstroms fällt auch die Bernoullikraft weg und der subglottale Druck steigt daraufhin erneut an und der Vorgang wiederholt sich.

Um eine selbsterregte Schwingung aufrechtzuerhalten, ist ein ununterbrochener Energietransfer vom Luftstrom auf das schwingende Stimmlippengewebe erforderlich, der im theoretischen Modell durch einen negativen Dämpfungsterm dargestellt wird. Die Bernoulli-Kräfte allein reichen nicht aus, um diesen Energieübertrag zu gewährleisten und die Schwingung würde langsam ausklingen. Videostroboskopische Untersuchungen haben gezeigt, dass die glottale Kante des Stimmlippengewebes in einer Art Wellenbewegung (*mucosal wave*) in lateraler Richtung schwingt, bei der die *kaudale* (untere) Stimmlippenkante der *kranialen* (oberen) etwas vorauseilt. Beim Zusammenschwingen der Stimmlippen ist deren unterer Teil dadurch bereits dichter zusammen als deren oberer Teil und der Luftstrom divergiert. Im Gegensatz dazu konvergiert der glottale Luftstrom während der Phase des Auseinanderschwingens. Aus einem geringfügig unterschiedlichen mittleren Luftdruck in der Glottis während dieser beiden Phasen resultiert die notwendige Druckasymmetrie, die die Schwingung aufrecht erhält [T94].

Die aus der Stimmlippenbewegung resultierende Modulation des Luftstroms beschreibt bei stimmgesunden Sprechern eine periodische Pulsfolge (*Glottispulse*), deren Frequenz von der Vokalismuskelspannung, dem subglottalen Druck und der schwingenden Masse der Stimmlippen abhängt und als Grundfrequenz der Phonation bezeichnet wird. Die Frequenz der Stimmlippenschwingung wird vom Hörer als Tonhöhe (*pitch*) wahrgenommen. Dieser Zusammenhang ist in Abbildung 1.3 dargestellt.

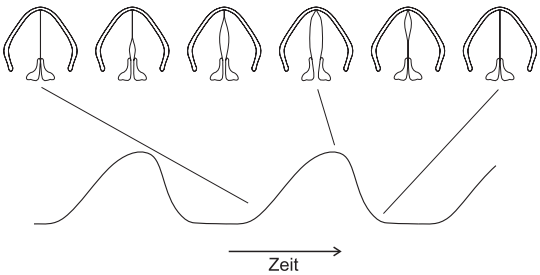


Abb. 1.3: *Schnittbild der sich öffnenden und wieder schließenden Glottis (oben) und resultierende Abschnitte des Glottispulses (unten) während der Stimmlippenschwingung. Deutliche zu erkennen sind die Unterschiede in der ansteigenden und abfallenden Flanke.*

An die Glottis schließt sich der supraglottale Bereich an, bestehend aus Vokaltrakt – Rachen- (*Epiglottis, Pharynx*) und Mundraum – sowie dem Nasaltrakt. Die Geometrie des Vokaltrakts ist bei jedem Menschen verschieden. Diese Unterschiede resultieren in einem individuellen Klang der Stimme. Wesentlich für die Bildung unterschiedlicher Laute ist die variable Stellung der Artikulatoren, zu denen u. a. das Gaumensegel, Zungenrücken, Unterkiefer, Zungenspitze, Zähne und Lippen zählen. In Abhängigkeit von der Position dieser Artikulatoren im Vokaltrakt auftretende Resonanzen (*Formanten*) bewirken Verstärkungen bestimmter Spektralbereiche, die sich in einer unterschiedlichen akustischen Färbung des primären, glottalen Anregungssignals äußern. Über diese Variation findet die Kodierung der zu übermittelnden Sprache mit einem akustischen Alphabet statt.

Der so auf glottaler Ebene erzeugte und durch den Vokaltrakt modulierte Luftstrom wird schließlich an den Lippen als Schalldruckwelle abgestrahlt.

1.2 Aufbau von Sprache

Bei der Beschreibung des Aufbaus von Sprache muss man zwischen zwei Betrachtungsweisen differenzieren. In der *Phonetik*, die den Klang auf Signalebene beschreibt, bezeichnet das *Phon* (Laut, Sprachlaut) – als atomarer Baustein – die kleinste, durch Segmentierung gewonnene Einheit einer konkreten sprachlichen Äußerung. Phone können anhand ihrer artikulatorischen und akustischen Eigenschaften

identifiziert und beschrieben werden. Aus der Betrachtungsweise der *Phonologie* – der linguistischen Ebene – dagegen werden Laute durch Systematisierung und Verallgemeinerung klassifiziert und in *Phoneme* eingeteilt. Das einzelne Phonem kann durchaus auf verschiedene Weisen akustisch realisiert werden in Form unterschiedlicher Phone, deren Menge innerhalb der Phonologie als Allophone des Phonems bezeichnet wird. Nach Swadesh [S34] lautet eine treffende Definition des Phonems: *„Wenn Phoneme für einheimische Sprecher verständliche Einheiten der Muttersprache sind, so werden sie doch nicht isoliert als verständliche Einheiten erfahren. ... Phoneme sind verständliche Einheiten in dem Sinne, dass der Einheimische solche Wörter als verschieden erkennen kann, die sich in einem der Phonembestandteile unterscheiden.“* [S83]

Abhängig von Sprache und Dialekt kommen dabei unterschiedlich viele verschiedene Phoneme (zwischen 10 und 65) zum Einsatz. In der deutschen Sprache unterscheidet man, je nach Phonemlexikon, ungefähr 60 Phoneme (SAMPA, *Speech Assessment Methods Phonetic Alphabet*, siehe Anhang 7.4). Eine weitere international gebräuchliche Konvention beschreibt das IPA (*International Phonetic Alphabet*). Es beinhaltet als phonetisches Alphabet eine Sammlung phonetischer Zeichen, mit deren Hilfe die Laute aller menschlichen Sprachen genau beschrieben werden können. SAMPA stellt dabei kein echtes eigenständiges phonetisches Alphabet dar, sondern liefert eine maschinenlesbare und tastaturfreundliche ASCII-Kodierung einer Teilmenge des IPA.

Erst die Verkettungen der inhaltlich bedeutungslosen Phoneme zu *Halbsilben* und *Silben* tragen Bedeutung, die zu *Wörtern* und *Sätzen* zusammengefasst die *fortlaufend gesprochene Sprache* (im folgenden *fortlaufende Sprache*, engl. *continuous speech, running speech, connected speech*) ergeben. Im Gegensatz dazu spricht man bei lang anhaltenden Lautäußerungen eines einzelnen, isoliert gesprochenen Phonems (bspw. lang gehaltene Vokale) von *gehaltener Phonation* (engl. *sustained* oder *isolated vowels*). Diese gehaltene Phonation spiegelt allerdings eher die Eigenschaften der Singstimme wider und die fortlaufende Sprache die des täglichen Einsatzes zur Kommunikation.

Da sich auf Grund von Bewegungsträgheit der Artikulatoren in fortlaufender Sprache die idealisierten Artikulationspositionen der einzelnen Phoneme nicht immer exakt einstellen können, kommt es zu *Koartikulationseffekten* aufeinander folgender Phoneme, die dabei akustische Mischlaute bilden. Die Formanteinstellungen für artikulierte Vokale werden in diesem Fall nicht immer erreicht und der Vokal wird in seiner Qualität reduziert. Unterschiedliche Vokale werden akustisch ähnlicher. Dieses Phänomen wird als *Vokal-Reduktion* bezeichnet.

Eine Gruppierung der verschiedenen Phoneme kann auf der Grundlage ihrer Anregung im menschlichen Sprechapparat erfolgen. Diese kann bei einem gesunden Stimmapparat aus einer periodischen Pulsfolge oder Rauschen bestehen. Entsprechend diesen unterschiedlichen Anregungsmöglichkeiten lassen sich die Phoneme in zwei Klassen einteilen:

- *stimmhafte* und
- *stimmlose* Phoneme.

Als stimmhaft (*voiced*) werden Phoneme mit periodischer glottaler Anregung bezeichnet und als stimmlos (*unvoiced*) solche mit ausschließlicher Rauschanregung. Die der Artikulation eines stimmhaften Phonems zugrunde liegende periodische Schwingung der Stimmlippen wird im Wesentlichen durch eine Wechselwirkung myoelastischer und aerodynamischer Kräfte hervorgerufen [B58].

Im Gegensatz zur stimmhaften Anregung steht bei der stimmlosen Anregung die Glottis offen, und der Luftstrom kann ungehindert in den supraglottalen Bereich entweichen. Ohne Gewebe auf der glottalen Ebene zu Schwingungen mit großer Amplitude anzuregen, führt eine Verengung des Luftweges an der Glottis oder im supralaryngealen Bereich allerdings zu Turbulenzen. In Anhang 7.4 ist eine Liste der Phoneme nach SAMPA-Konvention und ihre in dieser Arbeit vorgenommene Klassifizierung in stimmhaft/stimmlos zu finden.

Bei der menschlichen Kommunikation spielen neben der Informationsübermittlung durch Sprache in Form von Lauterzeugung und -verkettung ebenso Emotionen beim Sprechen und die gesamte Körpersprache eine wichtige Rolle. Das emotionale Befinden fließt zumeist unbewusst in den Klang der Sprachäußerung mit ein und äußert sich unter anderem in der Sprechgeschwindigkeit, der Lautstärke, dem Sprechfluss oder der Höhe der Grundfrequenz. Insbesondere ausgebildete Sprecher, Schauspieler und Sänger verfügen über ein sehr feines Vermögen, durch geringfügige Modifikation der an der Spracherzeugung beteiligten Elemente der Stimme einen differenzierten Ausdruck zu verleihen.

Dient die zwischenmenschliche Kommunikation primär der Informationsübermittlung, spielt die Qualität der Stimmgebung eine untergeordnete Rolle und der Zuhörer beschränkt seine Aufmerksamkeit auf die Aufnahme der inhaltlichen Information.

1.3 Stimmstörung

Sind Teile des menschlichen Sprechapparates durch Krankheit oder Überlastung in ihrer Funktion eingeschränkt, gestört oder sogar außer Funktion, so ist die betroffene Person wesentlich in ihren Kommunikationsmöglichkeiten beeinträchtigt. Der Phoniater unterscheidet zwischen verschiedenen Störungsursachen. Das Spektrum reicht dabei von neurologischen Störungen über Veränderungen der Stimmlippen durch Knötchen, Zysten, Papillome und Ödeme sowie Teilresektionen und Lähmungen bis hin zu kompletter Aphonie. Je nach Ausprägung und Dauer der Störung ist der Betroffene in seiner Lebensqualität deutlich beschränkt. Bei Berufsgruppen wie Lehrern oder Schauspielern z.B., die auf eine uneingeschränkte Funktion ihrer Stimme angewiesen sind, kann eine lang anhaltende Stimmstörung sogar eine Berufsunfähigkeit bedingen. Das rechtzeitige Erkennen und Behandeln solcher Störungen durch einen Phoniater, bzw. richtige Therapieren durch Logopäden und Sprecherzieher, spielt dabei für die Erhaltung dieses für den Menschen ungemein wichtigen Kommunikationsorgans eine wesentliche Rolle. Treten diese Störungen permanent auf, so spricht man von einer *pathologischen* Stimme (in Abgrenzung zur *Normalstimme*).

Durch Überlastungen oder krankheitsbedingte Funktionsstörungen des Sprechapparates kann es beim Sprechen zu Unregelmäßigkeiten in der glottalen Schwingung sowie einem inkompletten Glottisschluss kommen. Der perzeptive Höreindruck variiert deutlich zwischen unterschiedlichen Pathologien, und auch der Grad der Störung spielt für den Klang der Stimme eine wichtige Rolle. Ein unvollständiger glottaler Schluss führt bspw. zu einem erhöhten Anteil turbulenten Rauschens im Stimmschallsignal. Die Stimme wird dadurch *behauchter*. Eine irreguläre Schwingung der Stimmlippen aufgrund von Variationen der Schwingungsfrequenz oder -amplitude äußert sich dagegen in einer *rauen* Stimme. Die Periodizität der Schwingung der Stimmlippen spielt dabei eine wichtige Rolle zur Beurteilung der Stimmqualität – die auch als *Stimmgüte* bezeichnet wird – insbesondere pathologischer Stimmen.

Die Stimmgüte lässt sich quantitativ durch verschiedene Verfahren bestimmen, die sich in zwei Gruppen gliedern lassen:

- perzeptive Beurteilungen,
- Bewertungen nach akustischen Maßen.

Zu den perzeptiven Beurteilungen zählen z. B. die in Deutschland verbreitete RBH-Skala (*Rauigkeit, Behauchung, Heiserkeit*) [WRK86] oder die von der japanischen Gesellschaft für Logopädie vorgeschlagene GRBAS-Skala (*Grade, Rough, Breathy, Asthenic, Strained*) [H81]. Die einzelnen Parameter werden auf einer ganzzahligen Skala von *nicht vorhanden* bis *hochgradig* anhand von Hörproben durch ein Gutachterkollektiv beurteilt [NAW94]. Die perzeptiven Beurteilungsmethoden sind folglich sehr zeit- und personalintensiv und zudem nur bedingt objektiv und vergleichbar. Die Zusammensetzung des Gutachterkollektivs kann dabei entscheidenden Einfluss auf die Ergebnisse haben [PSIJN06]. Da eine Automatisierung nicht möglich ist, finden in dieser Arbeit perzeptive Beurteilungsmethoden keine Anwendung und weitere Beschreibung.

Die zweite Gruppe umfasst die Beschreibung der Stimmgüte unter Verwendung akustischer Maße. Die akustische Analyse ist gegenüber der perzeptiven Beurteilung objektiver, zuverlässiger reproduzierbar, zeiteffizienter und in der Regel auch apparativ kostengünstiger. Es muss dabei natürlich sichergestellt sein, dass die berechneten akustischen Maße zuverlässig sind und sich die Verhältnismäßigkeit unterschiedlicher Pathologien auch in den bestimmten Werten wiederfinden lässt. Die ersten Publikationen zur akustischen Analyse pathologischer Stimmen gehen auf Liebermann aus dem Jahre 1961 zurück [L61], der in seiner Arbeit Änderungen der Grundperiodenlänge in aufwendiger Handarbeit quantifiziert hat. Seitdem sind viele verschiedene Methoden und Maße entwickelt worden, deren einer Teil der Beschreibung additiven (turbulenten) Rauschens im Stimmsignal dient, ein anderer Teil der Beschreibung frequenz- und amplitudenmodulierenden Rauschens durch Schwankungen der Periodizität des Signals.

Neben Methoden zur Beurteilung des akustischen Stimmsignals findet im Rahmen der phoniatrischen Diagnostik auch eine *phonoskopische* Beurteilung der Stimmgebung während der Phonation statt. Erste indirekte Methoden zur Beobachtung des glottalen Sprachproduktionsapparates, wie die transcutane Durchleuchtung des Kehlkopfes (*Transillumination*) [S60], elektrische Impedanzmessungen durch die Glottis [F57], Ultraschallspiegelungen [MKH68] oder auch die Erfassung von Röntgenbildsequenzen [H65] während fortlaufender Sprache sind bereits seit Ende der 50er Jahre durchgeführt worden [CSAL71]. Zur Beurteilung der Stimmfunktion ist heutzutage eine visuelle Betrachtung des Stimmlippenschwingungsverhaltens während gehaltener Phonation unter Verwendung eines starren Laryngoskops durch den geöffneten Mund des Sprechers üblich. Ein Phonoskop mit flexibler Optik, das durch die Nase eingeführt wird und einen Blick aus dem Rachenraum auf

die Stimmlippen erlaubt, ermöglicht die Beurteilung des Sprechapparates während fortlaufender Sprache. Die Zeitauflösung des menschlichen Auges (und nachgeschalteten Gehirns) ist allerdings nicht ausreichend, um einzelne Schwingungsperioden in einem üblichen Grundfrequenzbereich der menschlichen Sprache aufzulösen. Die direkte Beurteilung von geringen Irregularitäten ist demzufolge nicht möglich. Eine digitale Aufzeichnung des Schwingungssignals ermöglicht eine anschließende Beurteilung einzelner Schwingungsphasen auf anderen Zeitskalen.

Die Bildraten klassischer digitaler Videosysteme, die bisher unter dauerhafter Beleuchtung (*Videolaryngoskopie*) bzw. unter stroboskopischer Beleuchtung (*Videostroboskopie*) zur Beurteilung des Schwingungsverhaltens verwendet werden, reichen nicht aus, um einzelne Schwingungszyklen getrennt beobachten zu können. Zur Umgehung dieses limitierenden Faktors ist die Analyse einzelner Bildzeilen (*Videokymogramm* [SS96]) über den Zeitverlauf entwickelt worden. Sie erlaubt zwar höhere Zeitauflösungen aufgrund geringerer Datenmengen und die Beurteilung von Schwingungsirregularitäten, insbesondere Seitenunterschiede im Schwingungsverhalten beider Stimmlippen, einen wesentlichen Durchbruch im Bereich der bildgebenden Verfahren hat allerdings erst die *Hochgeschwindigkeitsglottografie* gebracht. Fortschritte in der Hochgeschwindigkeitsbildgebung erlauben inzwischen auch in der medizinischen Anwendung eine digitale Aufzeichnung der Stimmlippenschwingungen mit Bildraten bis zu 4000 Bildern/s bei parallel archiviertem Tonsignal[4]. Die Entwicklung von Bilderkennungsalgorithmen zur Identifikation bspw. der Glottis oder der Stimmlippen in diesen Bildsequenzen steht noch am Anfang und ist hochaktuelles Forschungsgebiet auch in der Arbeitsgruppe Sprache und Neuronale Netze des Dritten Physikalischen Instituts in Göttingen, in der auch diese Arbeit angefertigt wurde.

1.4 Akustische Analyse von Sprache

Als akustische Analyse von Sprachäußerungen lässt sich die Bestimmung unterschiedlicher Stimmgütemaße zusammenfassen, die auf der digitalen Verarbeitung des aufgezeichneten akustischen Signals basiert. Da diese Vorgehensweise nicht invasiv ist, bietet sie Vorteile gegenüber der direkten Betrachtung des Schwingungsverlaufs mittels optischer Methoden. Bei der Stimmanalyse von Sprechern ohne bekannte Stimmstörungen mittels optischer und akustischer Methoden stellt sich

[4] Bei einer Grundfrequenz von bspw. 200 Hz steht eine Sequenz von 20 Bildern pro Schwingungszyklus für die Analyse zur Verfügung.

heraus, das auch bei diesen vermeintlich „normalen“ Stimmen durchaus geringgradige Stimmstörungen vorliegen können, die bei einer rein perzeptiven Beurteilung nicht dokumentiert worden sind. Die akustische Analyse bietet folglich auch Vorteile gegenüber der perzeptiven Beurteilung.

Da sich in fortlaufender Sprache – im Gegensatz zur gehaltenen Phonation – unterschiedliche Phonationsstellungen der Artikulatoren und verschiedene glottale Anregungsmechanismen auf sehr kurzen Zeitskalen abwechseln, ist eine Einteilung des Sprachsignals in kürzere Zeitabschnitte für eine akustische Analyse notwendig. Das akustische Signal gehaltener Vokale kann dagegen während der gesamten Phonationsdauer – abgesehen von Ein- und Ausschwingvorgängen – annähernd als stationär (*fast-stationär*) betrachtet werden und eignet sich deshalb sehr gut für akustische Analysen. Gehaltene Vokale spiegeln aber eher die akustischen Eigenschaften der Singstimme wider und stellen deshalb nicht notwendigerweise auch eine verlässliche Repräsentation der Stimmfunktion während fortlaufender Sprache dar [K90], [PMWH87], [QHM99], [PJ01].

Im Unterschied zur gehaltenen Phonation treten in fortlaufender Sprache zusätzliche Eigenschaften der Stimmfunktion, wie plötzlicher Stimmeinsatz und -ende, gehaltene oder schnelle Positionsänderungen des Phonationsmechanismus, Variationen der Grundfrequenz und Amplitude durch Satzmelodie sowie Stimmabbrüche (*voice breaks*) auf, die für eine gesamtheitliche Beurteilung der Stimmgüte relevant sind. Speziell für die Beurteilung pathologischer Stimmen sind akustische Analysen sowohl während gehaltener Phonation als auch während fortlaufender Sprache notwendig, um zuverlässige Aussagen über die Stimmgüte treffen zu können [AH86], [TK75], [LFMS99], [PJ01].

Akustische Verfahren zur Analyse fortlaufender Sprache, die im Folgenden auch als *Textanalyse*[5] bezeichnet wird, sind im Vergleich zu Methoden der Vokalanalyse bisher erst in sehr geringer Zahl entwickelt und publiziert worden [KW79], [SB82], [LH94], [LSK98], [PJ00], [SMLA03]. Es existieren Veröffentlichungen im Zusammenhang mit perzeptiven Beurteilungen von gesprochenem Text [HFGS80], [AH86] über die Wahl akustischer Maße zur Beschreibung der Stimmgüte aus fortlaufender Sprache [FSK97] sowie Korrelationsuntersuchungen mit perzeptiven Ergebnissen, um die Aussagekraft akustischer Maße bei der Beurteilung pathologischer Stimmen beurteilen zu können [K90].

Hecker und Kreul haben bereits 1971 [HK71] Schwankungen der Grundfrequenz

[5] Nicht zu verwechseln mit der Analyse von geschriebenem Text im Sinne einer Text- oder Schrifterkennung.

aus fortlaufender Sprache unterschiedlicher Stimmgüte untersucht. Die aus dem ersten Satz der Regenbogen-Passage bestimmten absoluten Jitterwerte ließen allerdings keine Unterscheidung zwischen pathologischen und normalen Stimmen zu. Askenfeld und Hammarberg [AH86] haben Schwankungen verschiedener akustischer Maße mit perzeptiven Beurteilungsergebnissen anhand prä- und postoperativer Aufnahmen von Stimmpatienten korreliert und für einzelne akustische Maße hohe Korrelationskoeffizienten (für *Directional Perturbation Factor (DPF)* Korrelationswert von $r = 0{,}86$) berechnet. Eine Untersuchung von Schoentgen [S89] zur Differenzierung von Stimmen unterschiedlicher Stimmgüte anhand von Jitterwerten, die sowohl aus fortlaufender Sprache als auch gehaltener Phonation bestimmt wurden, zeigte keine eindeutigen Vorteile für gehaltene oder fortlaufende Phonation. Parsa und Jamieson haben 2001 [PJ01] eine Arbeit publiziert, in der sie mehrere akustische Maße auf ihre Aussagekraft zur Trennung pathologischer von Normalstimmen hin untersucht haben.

Für die Aufnahme der zu analysierenden Sprachsignale werden den Sprechern phonetisch ausgewogene Standardtexte (bspw. der *Nordwind und Sonne*-Text, die *Buttergeschichte* oder die *Regenbogen-Passage*, siehe Anhang) zum Vorlesen vorgelegt, von denen Übersetzungen in verschiedene Sprachen existieren [IPA49].

Alle Verfahren zur Bestimmung akustischer Maße aus fortlaufender Sprache unterliegen speziellen Eigenschaften der fortlaufenden Sprache, die bei der Analyse gehaltener Phonation nicht ins Gewicht fallen:

1. Während der Ein- und Ausschwingphasen (*voice-onset* und *voice-offset*) fortlaufender Sprache treten hohe Grundperiodenlängenschwankungen auf, bis das glottale Schwingungssystem einen quasistationären Zustand erreicht hat. Lediglich eine fundierte Auswahl der der Analyse zugrunde liegenden Sprachsegmente garantiert aussagekräftige Ergebnisse. Auch der Einfluss der Sprachmelodie und anderer antrainierter oder zufälliger Eigenschaftsänderungen im Sprachproduktionsapparat können die Ergebnisse verfälschen.

2. Änderungen in der Dynamik des Sprachsignals durch Vokaltraktmodulationen oder Schwingungsmodulationen auf glottaler Ebene haben direkten Einfluss auf akustische Maße zur Beschreibung von Amplitudenirregularitäten. Auch in diesem Bereich ist eine zuverlässige Klassifikation bestimmter Sprachsegmente notwendig.

Aus Mangel an robusten Verfahren zur Klassifikation relevanter Sprachsegmente

sind häufig akustische Methoden zur Beschreibung pathologischer Stimmen herangezogen worden, die auf längeren Zeitfenstern als einzelnen Phonemen arbeiten. Hammarberg et al. [HFGS80] haben bspw. bei der Berechnung von Langzeitspektren (*Long-Term-Average-Spectra, LTAS*) und der Verteilung der Grundfrequenz (*Fundamental Frequency Distribution, FFD*) hohe Korrelationen zu perzeptiven Bewertungen pathologischer Stimmen erhalten. Klingholz [K90], Qi et al. [QHM99] und weitere Autoren haben Methoden zur akustische Beschreibung glottalen Rauschens wie HNR (*Harmonics-to-Noise-Ratio*) oder SNR (*Signal-to-Noise-Ratio*) aus gehaltener Phonation auf die Bestimmung aus fortlaufender Sprache erweitert und neu entwickelt.

Die meisten der Publikationen zur akustischen Analyse fortlaufender Sprache zielen auf eine Differenzierung normaler von pathologischer Stimmfunktion ab. Die individuelle Berechnung einzelner Stimmgütemaße in periodischer Stimmanregung erfordert eine aufwendige Selektion dieser stimmhaften Sprachsegmente aus dem fortlaufenden Signal. Da sich diese Klassifikation insbesondere bei sehr ausgeprägten Stimmstörungen als problematisch erweist, versagt ein Großteil der publizierten Methoden in diesem Bereich. Die Entwicklung einer zuverlässigen Methode zur Klassifikation des Sprachsignals in Bereiche unterschiedlicher Phonation (stimmhaft und stimmlos) sowie der Exklusion von Sprechpausen stellt einen wesentlichen Bestandteil dieser Arbeit dar. Auf Basis dieser Klassifikation ist eine akustische Analyse nach unterschiedlichen Maßen für Normalstimmen und pathologische Stimmen des gesamten Stimmgütebereichs möglich und durchgeführt worden.

2 Akustische Maße zur Stimmgütebeschreibung

Im Folgenden soll ein Überblick über akustische Maße zur Stimmgütebeschreibung gegeben werden, die aus dem akustischen Zeitsignal der fortlaufenden Sprache (Analysen im *Zeitbereich*) oder deren Transformierten (Analysen im *Frequenzbereich*) abgeleitet werden können. Es handelt sich dabei zumeist um akustische Maße, die bereits in der Analyse gehaltener Phonation Anwendung finden, deren Berechnungsalgorithmen allerdings auf die Verwendung in fortlaufender Sprache angepasst worden sind. Die Auswahl der akustischen Maße stellt dabei keinerlei Anspruch an Vollständigkeit, sondern beschreibt die in dieser Arbeit zum Teil verwendeten Maße, die sich zumeist in vielzähligen Untersuchungen der verschiedensten Autorengruppen gegenüber anderen akustischen Maßen als „aussagekräftiger“ und im Kollektiv als jeweils weitgehend unkorreliert herausgestellt haben.

Eine der Basisgrößen zur Beschreibung von Schwingungen stellt neben der Schwingungsamplitude die Schwingungsfrequenz dar. Sie spielt in der Stimmanalyse eine wesentliche Rolle und einige der im Folgenden vorgestellten Größen basieren auf der *Grundperiode* der Stimmlippenschwingung. Die exakte Bestimmung der Grundperiodenlänge (bzw. von deren Kehrwert, der *Grundfrequenz*) ist Thema diverser Publikationen und auch immer noch aktuelles Forschungsgebiet [S68], [S81], [H83], [DMKM89], [ECH90], [TL93], [DPH93], [BKGD96], [PJ99]. Die unterschiedlichen Methoden zur Grundfrequenzbestimmung (*Pitch Determination Algorithm, PDA*) beruhen zum einen auf Methoden im Zeitbereich und zum anderen auf solchen im Frequenzbereich. Bei einer Bestimmung im Zeitbereich treten bei den meisten Methoden Probleme bei stark irregulären Signalen oder einem großen Grundfrequenzbereich im Signal auf [H83].

2.1 Bestimmung der Periodenlänge

Die Periodenlänge T (Grundperiode) einer periodischen Schwingung beschreibt die kleinste Zeitdauer bis zum Wiedereintreten des gleichen Schwingungszustandes des Systems. Unter der Voraussetzung exakt periodischer Vorgänge ist eine Bestimmung von T zu jedem beliebigen Zeitpunkt der Schwingung $s(t)$ als die kleinste zeitliche Differenz des Wiedereintretens eines bestimmten Schwingungszustandes möglich, unter der Vorgabe:

$$T: \qquad s(t) = s(t+T); \qquad T > 0$$

Da die Vorgabe der exakten Periodizität im menschlichen Sprechapparat allerdings nicht gegeben ist, ist die Bedingung $s(t) = s(t+T)$ fast nie erfüllt. Durch Änderungen der physikalischen Parameter, die die Schwingung beeinflussen – wie des Drucks, des aus den Lungen kommenden Luftstroms oder Schwankungen der Muskelspannung im Vokalismuskel bspw. – entstehen Variationen der Schwingungsfrequenz und -amplitude der Stimmlippen, die sich im akustischen Zeitsignal wiederfinden lassen.

Bei der Berechnung der Periodenlänge als Grundlage zur Bestimmung von Periodenlängenschwankungen oder zur Ableitung akustischer Maße auf Basis der Periodenlänge werden zwei Vorgehensweisen unterschieden:

1. **Fensterweise Mittelung:** Zum einen kann in einem Signalteilstück fester Länge (*Fenster*) ein mittlerer Periodenlängenwert T für alle Perioden, die dieses Signalfenster überdeckt, angegeben werden. Die Fensterlänge kann je nach Anwendung so gewählt werden, dass entsprechend viele Perioden erfasst werden. Aufeinander folgende Fenster können sich gegebenenfalls überlappen. Es existieren sowohl Methoden, die auf dem Zeitsignal arbeiten, als auch solche, die zur Analyse im Frequenzbereich eingesetzt werden.

2. **Ereignisbasierte Methoden:** Zum anderen kann, ausgehend von einem detektierten oder empirisch bestimmten Startzeitpunkt $t = t_0$, jede einzelne folgende Periodenlänge T_i (mit $i \in \mathbf{Z}^+$) durch Beurteilung ausgezeichneter Schwingungszustände im Signalverlauf (maximal positive oder negative Signalamplitude oder steilster Nulldurchgang bspw.) ermittelt werden. Der Periodenstartzeitpunkt der folgenden Schwingung $t = t_{i+1}$ ergibt sich iterativ aus dem Periodenstartzeitpunkt t_i und der ermittelten Periodenlänge T_i für die aktuelle Periode.

Häufig werden Methoden zur fensterweisen Mittelung mehrerer Perioden als initiale Grobabschätzung der Periodenlänge benutzt, um in einem zweiten Schritt in einer Umgebung dieser abgeschätzten Periodenlänge eine feinere Bestimmung der exakten Periodenlänge vornehmen zu können.

2.1.1 Fensterweise Mittelung der Periodenlänge

Die unterschiedlichen, hier beschriebenen Verfahren, von denen die Autokorrelationsfunktion in dieser Arbeit als Vorverarbeitungsstufe Verwendung findet, weisen dabei individuelle Vor- und Nachteile auf [PJ99].

Autokorrelationsfunktion

Die Kurzzeit-*Autokorrelationsfunktion* (AKF) eines diskreten Signals $s(t); t = 1 \ldots N$ ist definiert als das Skalarprodukt eines Signalbereichs endlicher Länge mit einem um τ zeitlich verschobenen Signalbereich gleicher Länge desselben Signals.

$$R(\tau) = \sum_{t=1}^{N-\tau} s(t)s(t+\tau) \tag{2.1}$$

Im Fall periodischer Signale weist die AKF ein relatives Maximum bei Verschiebung um die Länge der Grundperiode ($\tau = T_0$) auf. Die AKF weist neben dem absoluten Maximum bei keiner Verschiebung ($\tau = 0$) insbesondere bei Verschiebungen um Vielfache der Grundperiode ($\tau = 2T_0, 3T_0, \ldots$) – den Sub-Harmonischen – weitere Maxima auf. Zusätzliche dominante Maxima können bei den Formantfrequenzen auftreten, die gerade bei schnellen Phonemübergängen häufig das Maximum der Grundfrequenz überragen.

Zur Bestimmung der Grundperiodenlänge T_0 ist lediglich in einem gegebenen Periodenlängenbereich ($[\tau_{\min}, \tau_{\max}]$) das Maximum der AKF zu suchen [RS78].

$$T_0 = \underset{\tau_{\min} \leq \tau \leq \tau_{\max}}{\operatorname{argmax}} \; R(\tau) \tag{2.2}$$

Wird die AKF mit der Signalenergie $R(0)$ normiert, so spricht man von der *normierten Autokorrelationsfunktion*. Somit ist ein Vergleich absoluter AKF-Werte aus unterschiedlichen Signalfenstern möglich.

Der zu untersuchende Periodenlängenbereich ist für gehaltene Phonation empirisch bekannt und erstreckt sich bei normaler, erwachsener Sprechstimmlage typischerweise von $\tau_{\min} = 3\,\text{ms}$ (hohe Frauenstimme) bis $\tau_{\max} = 14\,\text{ms}$ (tiefe Männerstimme). Bei Kinderstimmen muss dieser Analysebereich eventuell zu kürzeren Periodenlängen erweitert werden. Eine mögliche Fehlerquelle besteht in der Anfälligkeit der AKF für Oktavfehler, da der angegebene Periodenlängenbereich mehr als eine Oktave umfasst. Signaleigenschaften wie Periodenverdopplung oder geringfügige Instationaritäten können dazu führen, dass die AKF bei der halben oder bei der doppelten (wahrgenommenen) Grundfrequenz ihr Maximum aufweist.

Die angesprochenen Nebenmaxima und Formanteinflüsse des Vokaltraktes in der AKF können die Bestimmung der Grundperiode deutlich erschweren. Eine Möglichkeit der Optimierung besteht darin, das Spektrum im Bereich der höheren Frequenzen abzuflachen (*spectral flattening*). Als besonders effektiv hat sich dabei eine nichtlineare Verzerrung durch das *center clipping* [S68a] erwiesen. Eine entscheidende Rolle spielt bei dieser Vorgehensweise natürlich die Wahl des Schwellwertes.

Eine wesentlich effizientere Methode, um die bei der AKF störenden Formanteinflüsse und Harmonischen bei der Grundperiodensuche zu eliminieren, stellt das Cepstrum dar, das sich diese zunutze macht.

Cepstrum

Als *Cepstrum* (CEP) (von J. W. Tukey abgeleitetes Anagramm von „Spectrum“) bezeichnet man die Fourierrücktransformierte des logarithmierten Leistungsspektrums des Signals $x(t)$:

$$CEP_r(q) = \mathcal{F}^{-1}\left\{\log\left|\mathcal{F}\left\{x(t)\right\}\right|^2\right\} \tag{2.3}$$

Die periodische Struktur der Grundfrequenz und ihrer Harmonischen im logarithmierten Leistungsspektrum bildet die Grundlage dieser Methode[6]. Zwischen dem Cepstrum und der AKF besteht eine enge Verwandtschaft. Die Autokorrelationsfunktion AKF$(x(t))$ eines Signals $x(t)$ steht in Bezug zu dessen Fouriertransformierten $\mathcal{F}(x(t))$ über das *Wiener-Khinchin*-Theorem[7]:

[6] Die Entwicklung des Cepstrums diente eigentlich der Unterscheidung unterirdischer nuklearer Explosionen von Erdbeben.

[7] Das Wiener-Khinchin-Theorem stellt einen Spezialfall des Kreuzkorrelations-Theorems dar. Eine Faltung zweier Signale $f(x)$ und $g(x)$ im Zeitbereich entspricht einer Multiplikation im Frequenzbereich: $f(x) * g(x) \leftrightarrow F(y) \cdot G(y)$; Spezialfall Wiener-Khinchin für $f(x) = g(x)$.

$$\mathrm{AKF}(x(t)) \longleftrightarrow |\mathcal{F}\{x(t)\}|^2 \tag{2.4}$$

Die Autokorrelation eines Signals in einem Bereich (Zeit- oder Frequenz-) entspricht der Berechnung des Leistungsspektrums in dem jeweils anderen Bereich.

Basierend auf dem Quelle-Filter-Ansatz aus Gleichung 1.1 kann man ein stimmhaftes Sprachsignal $x(t)$ vereinfacht als Faltungsprodukt aus periodischen Anregungspulsen $e(t)$ und Vokaltraktantwort $v(t)$ (inklusive Glottispuls und Abstrahlung der Lippen) betrachten:

$$x(t) = e(t) * v(t) \tag{2.5}$$

Eine Fouriertransformation $\mathcal{F}$ (auch $\widehat{\ }$) ermöglicht die Trennung dieses Faltungsprodukts in eine einfache Multiplikation:

$$\mathcal{F}(x(t)) = \mathcal{F}\{e(t) * v(t)\} = \mathcal{F}\{e(t)\} \cdot \mathcal{F}\{v(t)\} \tag{2.6}$$

$$\widehat{x}(\omega) = \widehat{e}(\omega) \cdot \widehat{v}(\omega) \tag{2.7}$$

Durch eine Logarithmierung lässt sich das Produkt in eine Summe transformieren:

$$\log \widehat{x}(\omega) = \log \widehat{e}(\omega) + \log \widehat{v}(\omega) \tag{2.8}$$

Eine erneute Fouriertransformation erlaubt die Rücktransformation in den Zeitbereich, in dem der Einfluss des Vokaltraktes und der der Grundfrequenz sich nun infolge der Addition separieren lassen:

$$CEP_c(q) = \widehat{\log \widehat{x}(q)} = \widehat{\log \widehat{e}(q)} + \widehat{\log \widehat{v}(q)} \tag{2.9}$$

Die neue Variable q wird als *Quefrenz*[8] bezeichnet und stellt das Argument des komplexen Cepstrums CEP_c dar. In einem vorgegebenen Quefrenzintervall findet daraufhin die Bestimmung eines relativen Maximums, das die Grundperiodenlänge charakterisiert, statt [NS64], [N64], [N67]. Wichtig ist dabei die Wahl der unteren Grenze des Intervalls, da die niedrigen Quefrenzen (0 – 3 ms) zum Großteil durch den Einfluss des Vokaltraktes bestimmt sind. Da es sich bei diesen Spektren um nichtnegative, reellwertige Funktionen handelt, ersetzt der reelle Logarithmus den komplexen und zur Grundfrequenzbestimmung wird in der Regel folglich das reellwertige Leistungs-Cepstrum CEP_r aus Gl. 2.3 anstelle des komplexen Cepstrums CEP_c aus Gl. 2.8 verwendet [H83].

[8] *Quefrenz* ist ebenfalls ein Anagramm von J. W. Tukey aus „Frequenz“ als Argument des Cepstrums in der Dimension *Zeit*.

Harmonisches Produktspektrum

Die Analyse des *Harmonischen Produktspektrums (HPS)* geht auf Schroeder [S68] zurück und basiert auf dem Auftreten deutlicher Maxima bei Vielfachen der Grundfrequenz im Amplitudenspektrum, der sog. Harmonischen. Das Harmonische Produktspektrum wird als Produkt von R Replikas dieses Amplitudenspektrums, die jeweils auf der linearen Frequenzachse um einen ganzzahligen Faktor r gegenüber dem Originalspektrum gestaucht sind, berechnet.

$$HPS(n) = \sqrt[R]{\prod_{r=1}^{R} |X(e^{in\,\Delta\omega\,r})|} \tag{2.10}$$

Höhere Harmonische fallen durch die Stauchung mit der Grundfrequenz selbst zusammen und liefern bei periodischen Signalen ein deutliches Maximum bei der Grundfrequenz, die durch eine Extremwertsuche aus dem HPS bestimmt wird. Betrachtet man bspw. ein um Faktor 2 auf der linearen Frequenzachse gestauchtes Spektrum, so fällt der Peak der zweiten Harmonischen mit dem der Grundfrequenz im ungestauchten Spektrum zusammen und verstärkt sich durch die Multiplikation. Gleiches gilt für die dritte Harmonische im um Faktor 3 gestauchten Spektrum und die höheren ganzzahligen.

Gerade bei Signalen mit niedrigem Signal-Rausch-Verhältnis (SNR, *signal to noise ratio*) oder einem generell hohen Rauschanteil weist diese Methode Vorteile gegenüber den bisher dargestellten auf. Der Einfluss der unkorrelierten Rauschkomponenten nimmt durch die Multiplikation der gestauchten Spektren ab und es resultiert ein deutlicher Peak bei der Grundfrequenz.

Ein wesentlicher Nachteil der bisher dargestellten Methoden besteht allerdings darin, dass auch bei kurzen Signalabschnitten keine Aussage über den exakten Start- bzw. Endzeitpunkt einzelner Schwingungsperioden möglich ist. Es wird lediglich eine lokal gemittelte Periodenlänge für das analysierte Segment berechnet.

Über die genannten Methoden hinaus existieren weitere Verfahren zur Grundfrequenzbestimmung auf Kurzzeitsegmenten, wie die Analyse des Modulationsspektrums [S00], [SQ02] bspw., die hier allerdings nicht weiter ausgeführt werden.

2.1.2 Ereignisbasierte Methoden

Im Gegensatz zu den im vorigen Abschnitt beschriebenen fensterbasierten Methoden werden bei den ereignisbasierten die individuellen Periodenlängen der einzelnen Schwingungszyklen direkt berechnet. Gängige ereignisbasierte Verfahren stellen die *peak picking*-Methode oder die Bestimmung der *Nulldurchgangsrate (zero crossing rate, ZCR)* dar. Automatisch detektierte Extremwerte in zu analysierenden Signalabschnitten markieren beim *peak picking* die gesuchten Periodengrenzen, wie in Abbildung 2.1 schematisch dargestellt ist.

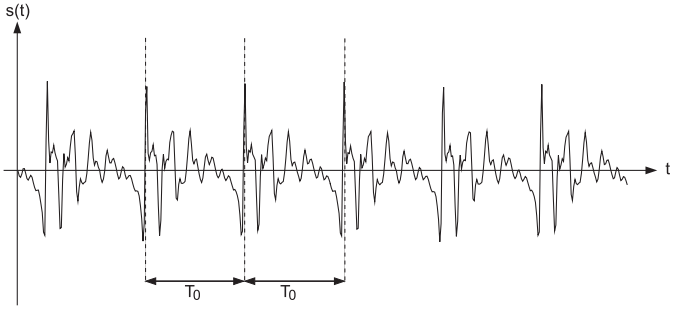


Abb. 2.1: *Periodenlängenbestimmung mittels peak picking-Verfahren. Der Abstand zwischen zwei Extremwerten innerhalb eines vorgegebenen Intervalls definiert die jeweilige Periodenlänge T_0.*

Diese Verfahren funktionieren zufriedenstellend bei normalen Stimmen und nicht zu hoher zeitlicher Auflösung. Sie werden häufig in kommerziellen Systemen zur Stimmanalyse eingesetzt [K92], da sie einfach zu implementieren sind und bis auf hochgradig gestörte Stimmen zumeist zuverlässig funktionieren.

Solch Einzelereignis-basierte (*event based*) Methoden liefern allerdings bei pathologischen, insbesondere stark behauchten Stimmen nur unzureichende Ergebnisse. Ein vorhandener additiver Rauschanteil dieser Stimmen kann zu Verschiebungen der relevanten, zu beurteilenden Ereignisse einer Periode um wenige Abtastwerte führen und das Ergebnis somit verfälschen. Bei der Analyse pathologischer Stimmen sind solche Verfahren folglich nur eingeschränkt anzuwenden.

Robuster gegen additives Rauschen sind integrale Verfahren, wie der im folgenden Abschnitt vorgestellte Waveform Matching Algorithmus.

Waveform Matching Algorithmus

Beim *Waveform Matching* Algorithmus [TL93], [MYC91], [M87], [PJ99] handelt es sich um ein integrales Verfahren, das die gesamte Information des Zeitsignals zweier aufeinander folgender Perioden für die Grundperiodenlängenbestimmung nutzt. In einem vorgegebenen Intervall ($T_{\min} \leq T_\tau \leq T_{\max}$) – voreingeschränkt unter Verwendung einer fensterbasierten Periodenlängenbestimmungsmethode aus Abschnitt 2.1.1 – werden dazu Kurzzeit-Kreuzkorrelationskoeffizienten (KKK_τ) paarweise aufeinander folgender Signalabschnitte $x(t)$ und $y(t)$ der Länge τ bei festgehaltenem Startpunkt t_0 bestimmt. Die lokale Grundperiodenlänge T_0 ist über die entsprechende Segmentlänge τ an der Stelle des maximalen Korrelationskoeffizienten KKK_τ gegeben, wie in Abbildung 2.2 dargestellt ist.

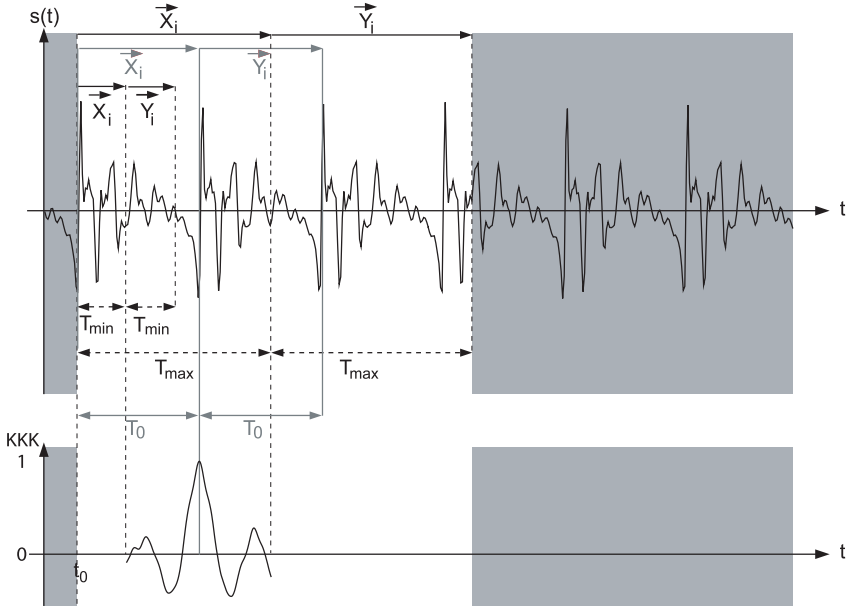


Abb. 2.2: *Schematische Darstellung des Waveform Matching Algorithmus mit entsprechendem Korrelationssignal. Das Maximum der KKK im Intervall* $[T_{\min}; T_{\max}]$ *bestimmt die Periodenlänge.*

Die Signalteilstücke $x(t)$ und daran anschließend $y(t)$ werden als τ-dimensionale Vektoren aufgefasst. Die Abtastwerte des ersten Segments $(x_{t_0}, \ldots, x_{t_0+\tau-1})$ werden zum Vektor $\vec{X_\tau}$ der Dimension τ zusammengefasst, zum Vektor $\vec{Y_\tau}$ die des zweiten Segments $(x_{t_\tau}, \ldots, x_{t_{2\tau-1}})$. Der KKK_τ errechnet sich als normiertes Skalarprodukt der beiden Vektoren $\vec{X_\tau}$ und $\vec{Y_\tau}$. Das Maximum der KKK_τ wird als Periodenkorrelationskoeffizient oder *Waveform Matching Coefficient* (WMC) bezeichnet.

$$\mathrm{WMC} = \underset{T_{\min} \leq T_\tau \leq T_{\max}}{\operatorname{argmax}} \mathrm{KKK}_\tau = \frac{\vec{X_\tau} \cdot \vec{Y_\tau}}{|\vec{X_\tau}| \cdot |\vec{Y_\tau}|} \tag{2.11}$$

Die berechnete Periodenlänge T wird bezogen auf den Startpunkt der Suche t_0 als Ausgangspunkt der folgenden Iteration des Verfahrens benutzt.

Die Wahl der zulässigen Periodenlängengrenzen $T_{\min}$ und $T_{\max}$ ist wichtig für eine vollautomatische Implementierung dieses Verfahrens. Einen initialen Anhaltspunkt bietet das erste Maximum der Autokorrelationsfunktion T_{AKF} für Fenster der Länge 200-500 ms. Die beiden Periodenlängengrenzwerte werden darauf basierend angesetzt, als [MFS98], [FMSK00]:

$$T_{\min} = 0{,}5 \cdot T_{\mathrm{AKF}} \quad \text{und} \quad T_{\max} = 1{,}5 \cdot T_{\mathrm{AKF}}$$

Die Bestimmung der Periodenlänge ist bei dieser Vorgehensweise in der Auflösung von ganzzahligen Abtastperioden möglich, die allerdings für bestimmte Anwendungen nicht ausreichend ist. Durch eine anschließende parabolische Interpolation des maximalen Kreuzkorrelationskoeffizienten $\mathrm{KKK}_{\mathrm{WMK}}$ und seiner beiden Nachbarwerte ($\mathrm{KKK}_{\mathrm{WMK-1}}$ und $\mathrm{KKK}_{\mathrm{WMK+1}}$) kann eine Verfeinerung der Periodenlängenauflösung weit über die Zeiteinheit einer Abtastperiode hinaus erreicht werden [MYC91].[9]

Die Bestimmung des Kreuzkorrelationskoeffizienten gestattet darüber hinaus eine Aussage über die Ähnlichkeit zweier aufeinander folgender Perioden. Ein hoher Korrelationswert $[0{,}1]$ nahe 1 beschreibt eine sehr ähnliche Signalform der beiden Perioden. Mittelt man die berechneten Korrelationskoeffizienten über ein zu analysierendes Zeitfenster, so spricht man vom *Mittleren Periodenkorrelationskoeffizienten* oder auch *Mean Waveform Matching Coefficient* (MWMC). Dieser zusätzlich gewonnene Parameter stellt ein weiteres Stimmgütemaß dar.

[9] Die parabolische Interpolation ist notwendig, da bspw. bei einem mit $f_s = 48$ kHz abgetasteten Signal einer Grundfrequenz von 200 Hz eine Grundperiode lediglich 240 Abtastwerte umfasst. Bei einer Zeitauflösung von einem Abtastwert würde eine Abweichung von einem Abtastwert bereits zu einem Periodenlängenfehler von $0{,}41\bar{6}\%$ führen.

2.2 Stimmgütemaße zur Beschreibung von Schwingungsirregularitäten

Da es sich beim menschlichen Sprechapparat um ein komplexes biomechanisches System handelt, bei dem verschiedenste Teile mit einwirken, die über den Schwingungsverlauf nicht notwendigerweise konstante Eigenschaften aufweisen (siehe Abschnitt 1.1), sind geringe Schwankungen in der Grundfrequenz und Amplitude auch bei stimmgesunden Sprechern vorzufinden. Gerade die Möglichkeit der sehr feinen Steuerung und die geringen Schwankungen wesentlicher Stimmproduktionseigenschaften bewirken einen *natürlichen* Stimmklang und erlauben eine breite Vielfalt an sprachlichen Ausdrucksformen, im Gegensatz zu synthetisierten Stimmen, bei denen eine exakte Periodizität und Konstanz der Schwingungsamplitude teilweise durch das implementierte System vorgegeben ist [L61], [H63]. Diese natürlichen Schwankungen sind bei stimmgesunden Sprechern nur sehr klein und ihre Erfassung durch akustische Messgrößen geht direkt auf die Periodenlängenbestimmung zurück. Schwingungsirregularitäten treten nach Untersuchungen von Orlikoff et al. allerdings bereits allein durch den Pulsschlag auf [OB89]. Insbesondere der *musculus thyroarytaenoideus (vocalis)*, der zwischen Schild- und Stellknorpeln verläuft, ist davon betroffen und verursacht Irregularitäten.

Die Grundlage dieser Schwingungsirregularitätsmaße stellt die Periodenlänge dar. Der Terminus *Jitter* wird dabei als ein Maß für die Periodenlängenschwankung verwendet. Zur Beschreibung der Schwankungen der Schwingungsamplitude bzw. Energie eines Schwingungszyklus wird der Terminus *Shimmer* verwendet. Bei hochgradig gestörten Stimmen ist die Aussagekraft dieser Schwingungsirregularitätsmaße nur bedingt gegeben, da u. U. keinerlei annähernd periodische Schwingung glottalen Gewebes mehr vorliegt [TL92]. Aus diesem Grund ist die Verwendung eines Periodenlängenbestimmungsalgorithmus – wie des Waveform Matching Algorithmus – notwendig, der a priori keinerlei Periodizität des Signals voraussetzt [FMSK97].

2.2.1 Jitter

Erstmalig publiziert wurde die Bestimmung des Jitters als Maß für die Schwankung der Grundperiodenlänge aus fortlaufender Sprache von Liebermann [L61]. In einer späteren Arbeit von Liebermann sind Jittermessungen zur Beurteilung der Stimmqualität pathologischer Stimmen durchgeführt worden [L63]. Trotz diverser

Publikationen, in denen der Jitter als akustisches Stimmgütemaß Verwendung findet, existiert keine exakte Definition und Vorschrift zur Bestimmung [L61], [L63], [S89], [PT90], [SG91], [KES93], [TL93], [VFMD93], [SG95]. Der berechnete Wert des Jitters hängt deshalb direkt von der Wahl der Grundperiodenbestimmungsmethode ab [TL93]. Ebenso uneinheitlich definiert in der Literatur ist die Angabe des Jitters a) als prozentuale Abweichung der Periodenlänge oder b) als absolutes Maß.

Die Anwendung unterschiedlicher Grundperiodenbestimmungsmethoden resultiert zwar in verschieden großen Werten des berechneten Jitters, die Relationen zwischen den Jitterwerten verschiedener Stimmen bleiben allerdings gleich, wie Titze [TL92] für gehaltene Phonation zeigt. Deshalb ist bei der Angabe von Jitterwerten auch immer eine Angabe der angewandten Berechnungsmethode (sowie der Grundperiodenbestimmungsmethode) notwendig, um diese Werte in Bezug zu anderen Studien beurteilen zu können.

Parallel zur Entwicklung unterschiedlicher Jitter-Bestimmungsmethoden haben sich verschiedene Bezeichnungen zur Quantifizierung der Periodenlängenschwankungen, die auch als *Periodenperturbationen* bezeichnet werden, etabliert. Um diese unterschiedliche Nomenklatur in der Literatur zu vereinheitlichen haben Pinto und Titze [PT90], [T94a] diese auf mathematische Begriffe – sog. Perturbationsmaße – zurückgeführt.

Perturbationsmaße

Bei der Verwendung von Perturbationsmaßen wird die individuelle Abweichung einer Messgröße in Bezug zu einer gemittelten Abweichung in einer lokalen Umgebung betrachtet und über die Gesamtstatistik gemittelt. Der *Perturbation Faktor* (PF) stellt solch ein Perturbationsmaß dar, bei dessen Berechnung über die lokal normierte Abweichung von zwei aufeinander folgenden Einheiten gemittelt wird.

$$\mathrm{PF} = \frac{100\%}{N-1} \sum_{n=1}^{N-1} \left| \frac{u(n) - u(n-1)}{u(n)} \right| \tag{2.12}$$

$u(n)$ beschreibt die Folge des betrachteten Parameters, dessen Perturbation bestimmt werden soll und N ist durch die Gesamtlänge des Signals in Abtastwerten gegeben. Dient der PF der Beschreibung der Grundperiodenlänge, so spricht man vom *Pitch Perturbation Factor (PPF)*. Der PPF ist allerdings sensitiv gegenüber

Grundperiodenlängenänderungen, die auf größeren Zeitskalen als dem zeitlichen Umfang der lokalen Mittelung erfolgen. Der Einfluss der Satzmelodie in fortlaufender Sprache kann demzufolge die Irregularitäten der Stimmlippenschwingung bei der Berechnung mittels PF verfälschen.

Einen wesentlichen Vorteil bei der akustischen Analyse fortlaufender Sprache bietet die Mittelung der Abweichungen in einer lokalen Umgebung im Gegensatz zur lokalen Periodenlänge. Der *Perturbation Quotient* (PQ) stellt solch ein Perturbationsmaß dar und geht auf Koike [K71] zurück. Er ist für eine Folge $u(n)$ definiert, als [KES93]:

$$\mathrm{PQ} = \frac{100\%}{N-K} \sum_{n=\frac{K-1}{2}}^{N-\frac{K-1}{2}-1} \left| \frac{u(n) - \frac{1}{K}\sum_{k=-\frac{K-1}{2}}^{\frac{K-1}{2}} u(n+k)}{\frac{1}{K}\sum_{k=-\frac{K-1}{2}}^{\frac{K-1}{2}} u(n+k)} \right| \tag{2.13}$$

$u(n)$ beschreibt die Folge des betrachteten Parameters, N ist durch die Gesamtlänge des Signals in Abtastwerten gegeben und K beschreibt die Breite des lokalen Mittelungsbereichs[10]. Entsprechend Gleichung 2.13 erhält man als Perturbationsmaß eine Prozentangabe.

Für die Bestimmung des PQ der Periodenlänge, des *Period Perturbation Quotient* (PPQ), wird innerhalb des zu analysierenden Signalbereichs der Länge N jeder Periodenlängenwert T_n in seiner Umgebung der K Nachbarperiodenwerte $T_{(n-\frac{K-1}{2})},\ldots,T_n,\ldots,T_{(n+\frac{K-1}{2})}$ bewertet [K71][KES93]. In einer Umgebung von $K = 3$ mit $u(n) := T(n)$ lautet der PPQ:

$$\mathrm{PPQ} = \frac{100\%}{N-3} \sum_{n=1}^{N-2} \left| \frac{T_n - \frac{1}{3}(T_{n-1} + T_n + T_{n+1})}{\frac{1}{3}(T_{n-1} + T_n + T_{n+1})} \right| \tag{2.14}$$
$$\text{mit} \quad T_n = T(n) \quad \text{und} \quad n = 0,\ldots,N-1.$$

Für die Jitterberechnung aus fortlaufender Sprache wird in dieser Arbeit $K = 3$ gewählt, um insbesondere den Einfluss der Satzmelodie und der Phonemübergänge auf die lokale Periodenlänge weitestgehend ausschließen zu können. Der Jitter wird in Prozent angegeben und liegt für stimmgesunde Sprecher im Bereich von $0{,}1 - 1\%$ für gehaltene Vokale. Für fortlaufende Sprache liegen diese Werte etwas höher [SG95].

[10] K muss ungerade sein, damit ein zentraler Wert der Umgebung existiert.

2.2.2 Shimmer

Als Shimmer werden die Schwankungen der Amplitude der einzelnen Grundperioden bezeichnet. Die davon abgeleitete Betrachtung der Schwankungen der Energie im Gegensatz zur Amplitude gewährleistet bei starken individuellen Schwankungen der Amplitude infolge von Rauschen eine größere Unabhängigkeit. Eine Quantifizierung kann ebenfalls mit dem Perturbation Quotient erfolgen und wird entsprechend als *Energy Perturbation Quotient* (EPQ) mit $u(n) := E(n)$ und $K = 3$ nach Gleichung 2.15 berechnet:

$$\text{EPQ} = \frac{100\%}{N-3} \sum_{n=1}^{N-2} \left| \frac{E_n - \frac{1}{3}(E_{n-1} + E_n + E_{n+1})}{\frac{1}{3}(E_{n-1} + E_n + E_{n+1})} \right| \tag{2.15}$$

$$\text{mit} \quad E_n = E(n) \quad \text{und} \quad n = 0, \ldots, N-1.$$

Shimmerwerte für stimmgesunde Sprecher liegen für gehaltene Phonation im Bereich von $1-10\%$ und werden in dieser Arbeit ebenfalls in einer lokalen Umgebung von $K = 3$ für fortlaufende Sprache berechnet.

2.2.3 Periodenkorrelationskoeffizient

Ein Maß für die Ähnlichkeit der Signalform aufeinander folgender Perioden ist deren Kurzzeit-Kreuzkorrelationskoeffizient. Er wird bei der Grundperiodenlängenbestimmung nach dem Waveform Matching Algorithmus berechnet (siehe Abschnitt 2.1.2). Der Mittelwert dieser Korrelationskoeffizienten paarweise aufeinander folgender Perioden in einem Signalabschnitt wird auch als mittlerer Periodenkorrelationskoeffizient (*Mean Waveform Matching Coefficient*, MWMC) bezeichnet.

$$\text{WMC} = \underset{T_{\min} \leq T_\tau \leq T_{\max}}{\text{argmax}} \; \text{KKK}_\tau = \frac{\vec{X_\tau} \cdot \vec{Y_\tau}}{|\vec{X_\tau}| \cdot |\vec{Y_\tau}|} \tag{2.16}$$

Er liegt im Intervall $[0,1]$ und erreicht sein Maximum bei Signalen mit identischer Periodenform (exakt periodische Signale). Mit zunehmenden Unregelmäßigkeiten in Länge und Form der Perioden fällt der MWMC im Wert ab.

2.2.4 Directional Perturbation Factor

Der *Directional Perturbation Factor (DPF)* geht auf eine Arbeit von Hecker und Kreul [HK71] zurück und ist definiert als der prozentuale Anteil der Grundperiodenlängendifferenzen aufeinander folgender Schwingungsperioden unterschiedlichen Vorzeichens von der Gesamtzahl der Periodenlängenänderungen.

$$DPF = \frac{\Delta_{\pm}}{\sum \Delta_T} \tag{2.17}$$

In einer Untersuchung von Askenfeld et al [AH86] an fortlaufenden Sprachäußerungen von stimmgestörten Sprechern vor und nach Stimmtherapie zeigt der DPF eine hohe Korrelation mit perzeptiven Beurteilungen der Stimmgüte.

2.2.5 Fundamental Frequency Distribution

Hammarberg et al. [HFGS80] haben die Verteilung der Grundfrequenz (*Fundamental Frequency Distribution, FFD)* über einem Signalabschnitt untersucht und hohe Korrelationen mit perzeptiven Beurteilungen der Stimmgüte erhalten. Der Kehrwert der lokal bestimmten Periodenlänge – die Grundfrequenz jeder einzelnen Periode – wird dabei in einem Frequenzhistogramm mit einer Auflösung von 1 Hz pro Bin aufgetragen. Bei Hammarberg et al. wird zur Bestimmung der Periodenlänge das Signal eines Kontaktmikrophons am Hals des Sprechers auf Höhe des Kehlkopfes analysiert. Die Analyse des akustischen Sprachsignals sollte bei einer ausreichend hohen zeitlichen Auflösung allerdings auch entsprechende Ergebnisse liefern.

Eine Beurteilung dieser Frequenzverteilung erfolgt anhand der Steigung zweier Geraden, die an die beiden Flanken beidseits des Maximums des Histogramms angepasst werden.

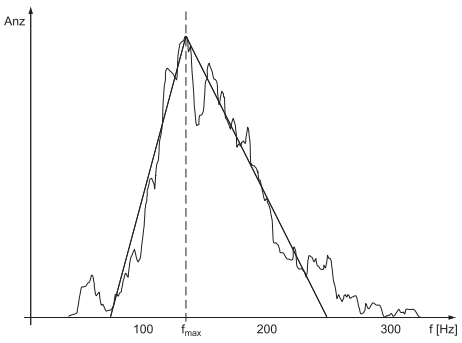


Abb. 2.3: *FFD: Exemplarisches Grundfrequenz-Histogramm mit angepassten Geraden bei 1Hz Auflösung pro Bin.*

2.3 Untersuchungen des Residualsignals

Entsprechend der Theorie der *Linearer Prädiktion* treten im Prädiktionsfehlersignal eines stimmgesunden Sprechers – insbesondere zu Beginn eines Schwingungszyklus – deutlich erkennbare Prädiktionsfehlermaxima auf [MG76]. Eine Bestimmung der Grundperiodenlänge ist demzufolge anhand dieses Fehlersignals – des Residualsignals – relativ zuverlässig möglich [CL91]. Bei pathologischen Sprechern sind – bedingt durch einen zumeist inkompletten Glottisschluss – diese Maxima nicht so deutlich zu erkennen [KM75], [PMWH87]. Änderungen im Stimmlippenschwingungsverhalten aufgrund von Stimmstörungen sollten sich folglich in den Eigenschaften des Residualsignals (*residue features*) dokumentieren lassen.

Die Betrachtung der Eigenschaften des Residualsignals gründet sich auf die Annahme, das dieses hochgradig mit der Schwingung der Stimmlippen korreliert ist. Das Residualsignal wird durch Inversfilterung des Sprachsignals bestimmt – entsprechend dem Quelle-Filter-Modell der Sprachproduktion nach Gleichung 1.1. Die Koeffizienten der Inversfilterung werden mittels Linearer Prädiktion berechnet [PMWH87]. Da die Lineare Prädiktion den Einfluss des Vokaltraktes auf das Sprachsignal parametrisiert, ist das Fehlersignal der Linearen Prädiktion mit dem glottalen Anregungssignal korreliert. Bei stimmgesunder Phonation eines stimmhaften Phonems sollte das glottale Anregungssignal aus einer Pulsfolge bestehen und das Residualsignal entsprechend ein flaches Spektrum aufweisen, was bei hinreichend hoher Prädiktorordnung allerdings immer gegeben ist [D76], [PMWH87].

Davis hat bereits 1976 [D76] eine erste „automatische“ akustische Analysemethode publiziert, die eine Trennung pathologischer Stimmen von Normalstimmen anhand von Eigenschaften des Residualsignals ermöglicht. Bestimmte Eigenschaften stellen auch nach Prosek et al. [PMWH87] zuverlässige Indikatoren für eine Differenzierung zwischen stimmgesunden und pathologischen Stimmen dar. Eskenazi et al. [ECH90] und Parsa et al. [PJ01] dokumentieren die Möglichkeit, anhand von Residualparametern wie *SFR* (siehe Abschnitt 2.3.2) und *PA* (siehe Abschnitt 2.3.3) eine Unterscheidung verschiedener Pathologien durchführen zu können.

2.3.1 Lineare Prädiktion

Die erste Beschreibung der Terminologie der *Linearen Prädiktion* (*LP*) geht auf Wiener [W66] zurück. Im Gebiet der Sprachverarbeitung werden Analyse- und Kodierungsanwendung mittels LP (*Linear Predictive Coding, LPC-Analyse*) erstmals von

Schroeder und Atal [AS70], [AH71] entwickelt. Obwohl die Hauptanwendung der LPC-Analyse sicherlich im Bereich der Sprachcodierung[11] liegt, bildet sie doch die Grundlage für einige wichtige Stimmgütemaße und spielt ebenso bei der stimmhaft/stimmlos-Klassifizierung (siehe Abschnitt 3) eine wesentliche Rolle.

Der Ansatz der LPC-Analyse besteht darin, die Abtastwerte des zeitlich diskretisierten Sprachsignals $x(n)$ innerhalb eines kurzen Signalbereichs durch eine Linearkombination der vorangegangenen k Abtastwerte zu approximieren.

$$\text{Prädiziertes Signal:} \quad \hat{x}(n) = \sum_{k=1}^{K} a_k \, x(n-k) \tag{2.18}$$

Die LP beschreibt folglich die Modellierung eines rein rekursiven Digitalfilters, das sich der Einhüllenden des Kurzzeitspektrums anpasst. Diese Vorgehensweise nutzt die Tatsache, dass sich das Sprachsignal im Vergleich zur Abtastrate nur sehr langsam ändert und erlaubt eine Beschreibung linear redundanter Eigenschaften (Grobstruktur des Spektrums) des Signals durch wenige Parameter. Die Koeffizienten a_k in der Linearkombination werden dabei als Prädiktorkoeffizienten bezeichnet und für jeden analysierten Kurzzeitsignalbereich getrennt berechnet.

Die zu Beginn der Analyse unbekannten LPC-Koeffizienten a_k werden angepasst, indem das Fehlersignal $e(n)$, das sich aus der Differenz des Originalsignals $x(n)$ und des mit den Koeffizienten prädizierten Signals $\hat{x}(n)$ ergibt, im Sinne des kleinsten Fehlerquadrats minimiert wird.

$$\begin{aligned} \text{Fehlersignal:} \qquad e(n) &= x(n) - \hat{x}(n) \\ &= x(n) - \sum_{k=1}^{K} a_k x(n-k) \end{aligned} \tag{2.19}$$

$$\begin{aligned} \text{Fehlerquadrat:} \quad E = \sum_n e^2(n) &= \sum_n \left[x(n) - \sum_{k=1}^{K} a_k x(n-k) \right]^2 \\ &\stackrel{!}{=} \text{Minimum} \end{aligned} \tag{2.20}$$

[11] In der heutigen Mobilfunktechnik erfolgt die Sprachcodierung auf Basis einer Erweiterung der LPC-Analyse, der sog. CELP (*Codebook excited linear predicitve coding*) [SA85], [S04].

Zur Minimierung der Fehlerquadratsumme in Gleichung 2.20 müssen die partiellen Ableitungen nach den Prädiktorkoeffizienten gleich Null gesetzt werden:

$$\frac{\partial E}{\partial a_k} = 2 \sum_n e(n) \cdot \frac{\partial e}{\partial a_k} = 0 \tag{2.21}$$

$$\sum_n \sum_{k=1}^{K} \sum_{l=1}^{K} a_k x(n-k)x(n-l) = \sum_n \sum_{l=1}^{K} x(n-l)x(n). \tag{2.22}$$

Zur Lösung dieses Gleichungssystems 2.22 existieren zwei unterschiedliche Ansätze für die Wahl der Summationsgrenzen über n: die *Kovarianzmethode* und die *Autokorrelationsmethode* [MW72].

Kovarianz- und Autokorrelationsmethode

Die beiden Berechnungsmethoden unterscheiden sich im Intervall, in dem der zu minimierende Prädiktionsfehler berechnet wird. Bei der *Kovarianzmethode*[12] wird die Fehlerbestimmung lediglich im Intervall $[K,N-1]$ des N Abtastwerte umfassenden Signalabschnitts durchgeführt. Zur Berechnung der Kovarianzmatrixelemente werden allerdings alle N Abtastwerte verwendet.

Für die Bestimmung der Prädiktorkoeffizienten a_k ergibt sich damit folgendes Gleichungssystem für die Kovarianzmethode:

$$\sum_{k=1}^{K} a_k \Phi(l,k) = \Phi(l,0)\,, \qquad l = 1,\ldots,K \tag{2.23}$$

$$\text{mit} \quad \Phi(l,k) = \sum_{n=K}^{N-1} x(n-k)x(n-l). \tag{2.24}$$

$\Phi(l,k)$ bestimmt dabei die Kreuzkorrelation zwischen den beiden bei $(K-l)$ und $(K-k)$ beginnenden Signalsegmenten der Länge $N-K$.

Bei Verwendung der Autokorrelationsmethode zur Lösung des Gleichungssystems 2.22 werden die Intervallgrenzen als $\pm\infty$ angenommen und die Abtastwerte außer-

[12] Der Name „Kovarianzmethode“ stammt aus der Ähnlichkeit der sich aus der Kreuzkorrelation ergebenden Koeffizientenmatrix mit einer Kovarianzmatrix.

halb des betrachteten Signalfensters als $x(n) = 0$ für $n < 0$ und $n \geq N$ angenommen. Das Gleichungssystem vereinfacht sich dadurch mit $\Phi(l,k) = R(|l-k|)$ zu:

$$\sum_{k=1}^{K} a_k R(|l-k|) = R(l)\,, \qquad l = 1,\ldots,K \tag{2.25}$$

$$\text{mit} \quad R(l) = \sum_{n=0}^{N-1-l} x(n)x(n+l) \quad \text{und} \quad l \geq 0 \tag{2.26}$$

Das Analysefenster muss bei der Autokorrelationsmethode in stimmhafter Phonation mehrere Grundperioden umfassen. Die Kovarianzmethode kann im Gegensatz dazu auch auf kürzeren Signalsegmenten im Bereich der Grundperiodenlänge oder darunter angewandt werden und ist insbesondere für grundperiodensynchrone Untersuchungen geeignet. Werden mehrere Grundperioden analysiert, so liefert die Anwendung beider Methoden ähnliche Ergebnisse, da bei hinreichend vielen Abtastwerten die Koeffizienten der Kovarianzmatrix sich den Autokorrelationskoeffizienten angleichen.

Der Vorteil der Autokorrelationsmethode liegt zum einen in einem geringeren Rechenaufwand zur Koeffizientenbestimmung, da mit dem Levinson-Durbin-Algorithmus [M75] bspw. eine effiziente iterative Lösung des Gleichungssystems 2.25 existiert. Zum anderen ist bei der Autokorrelationsmethode ein stabiles Filter in Gl. 2.29 (s. u.) garantiert [MG76].

Prädiktorordnung

Die Anzahl K der berechneten Prädiktorkoeffizienten a_k wird als Prädiktorordnung bezeichnet. Die Wahl dieser Ordnung spielt eine wichtige Rolle für die Modellierung der Vokaltraktresonanzen. Um den Vokaltrakteinfluss auf die Grobstruktur des Spektrums (spektrale Einhüllende) unter idealisierten Bedingungen adäquat parametrisieren zu können, sollte die Übertragungsfunktion $A(z)$ mindestens eine Anzahl von Filterkoeffizienten aufweisen, die die doppelte Länge der Schalllaufzeit von der Glottis bis zu den Lippen umfasst. Diese beträgt $2\,L/c$ mit der Länge des Vokaltraktes L und der Schallgeschwindigkeit c. Unter der Annahme einer Schallgeschwindigkeit in Luft von $c \approx 34\,\text{cm/ms}$ und einer Vokaltraktlänge von $L \approx 17\,\text{cm}$ entspricht dies einer Laufzeit von $1\,\text{ms}$, die die Übertragungsfunktion mit einschließen sollte. Bei einer für die LPC-Analyse gebräuchlichen Abtastfrequenz von $12\,\text{kHz}$ ergibt sich somit eine minimale Filterordnung von $K = 12$.

Da der Einfluss der glottalen Anregung und der Abstrahlung an den Lippen in dem obigen Modell bisher nicht enthalten ist, kann diese Prädiktorordnung nur als unterer Grenzwert betrachtet werden. Markel [M71] schlägt als sinnvolle Angabe der Prädiktorordnung einen Wert vor, der sich aus Abtastfrequenz in kHz (f_s) zuzüglich 4 oder 5 weiterer Koeffizienten ergibt [MG76].

Präemphase

Bei der Berechnung der Prädiktorkoeffizienten können Probleme auftreten, wenn die Matrix des Gleichungssystems 2.22 singulär ist. Im Frequenzbereich spiegelt sich dies in einer Instabilität des Filters (Gl. 2.29) wider. Mit einer Höhenanhebung des Sprachsignals durch Differenzierung lassen sich diese Stabilitätsprobleme in der Regel beheben [MG76]. Dies geschieht durch Filterung des Sprachsignals mit einem 1-Nullstellen-Filter mit der Übertragungsfunktion:

$$H(z) = 1 - \mu \cdot z^{-1} \tag{2.27}$$

Dieses Digitalfilter 1. Ordnung entspricht einem Hochpass. Der Höhenanhebungs- oder Präemphasenfaktor μ (*preemphasis*) sollte für Sprachsignale innerhalb des Intervalls $[0{,}9; 1{,}0]$ liegen [MG76]. Ein gebräuchlicher Wert ist $\mu = 0{,}9375$ [RS78].[13]

Interpretation im Frequenzbereich

Über den linearen Ansatz der Sprachproduktion ist eine Interpretation der LPC-Analyse im Frequenzbereich möglich. Die LPC-Analyse kann dabei als eine parametrische Schätzung des Leistungsdichtespektrums aufgefasst werden. Nach Anwendung der z-Transformierten auf die Gleichung 2.19 für den Prädiktionsfehler ergibt sich:

$$E(z) = \left[1 - \sum_{k=1}^{K} a_k z^{-k}\right] \cdot X(z) \quad = \frac{1}{G(z)} \cdot X(z) \tag{2.28}$$

[13] 0,9375 entspricht dem rationalen Verhältnis $15/16$.

$E(z)$ und $X(z)$ beschreiben die z-Transformierten des Fehlersignals $e(n)$ und des Zeitsignals $x(n)$. Das *Nur-Pole*-Filter $G(z)$ lässt sich folglich durch die Prädiktorkoeffizienten a_k beschreiben:

$$G(z) = \frac{1}{1 - \sum_{k=1}^{K} a_k z^{-k}} \tag{2.29}$$

Über die Berechnung der LPC-Koeffizienten aus dem Zeitsignal ist somit eine Beschreibung der spektralen Grobstruktur möglich, die als *LPC-Formant-Spektrum* oder auch *LPC-Spektrum* bezeichnet wird.

2.3.2 Maße der Spektralen Flachheit

Laryngeale Pathologien können – wie schon angesprochen – zu Veränderungen im Schwingungsverhalten der Stimmlippen und zu einem Anstieg des Rauschanteils im Stimmsignal führen. Diese Veränderungen haben Einfluss auf die Flachheit des Spektrums (*spectral flatness*). Ein höherer Rauschanteil in einem Sprachsignal bedingt ein höheres Maß der Spektralen Flachheit (*Spectral Flatness Measures, SFM*) des Energiespektrums in dB, das als Verhältnis des geometrischen zum arithmetischen Mittel der spektralen Energieverteilung eines Signalfensters j definiert ist [MG76]. Mit x_i^j als i-ter Amplitudenwert des j-ten Spektrums:

$$SFM_j = \frac{\text{Geometrisches Mittel}}{\text{Arithmetisches Mittel}} = \frac{\sqrt[N]{\prod |x_i^j|^2}}{\frac{1}{N}\sum_{i=0}^{N-1} |x_i^j|^2} \tag{2.30}$$

N ist dabei die Fensterbreite (Punktezahl der DFT). Der Wertebereich des SFM erstreckt sich von $[0; 1]$ und entspricht dem Wert 1 für ein exakt flaches Spektrum.

Es ist üblich, das Maß für die Spektrale Flachheit als Logarithmus dieses Verhältnisses in dB entsprechend Gleichung 2.31 anzugeben [ECH90].

$$SFM_j = 10 \lg \left(\frac{\sqrt[N]{\prod |x_i^j|^2}}{\frac{1}{N}\sum_{i=0}^{N-1} |x_i^j|^2} \right) = \frac{1}{N} \sum_{i=0}^{N-1} 10 \lg(|x_i^j|^2) - 10 \lg(\frac{1}{N} \sum_{i=0}^{N-1} (|x_i^j|^2)) \tag{2.31}$$

Der Wertebereich ändert sich entsprechend zu $[-\infty; 0]$, wobei ein großer negativer Wert eine gesunde Stimme (eher stimmhaft) beschreibt und ein kleiner negativer Wert nahe 0 (eher stimmlos) entsprechend eine gestörte Stimme.

Spektrale Flachheit des Residualsignals

Die spektrale Flachheit des Residualsignals (SFR) kann als ein Maß für die Maskierung der Harmonischen der Grundfrequenz durch Rauschen betrachtet werden [D76], [PMWH87]. Der Wert des SFR erlaubt nach Parsa et al. [PJ01] bereits eine nahezu fehlerfreie Klassifikation zwischen normaler und pathologischer Stimmfunktion auf der Basis von gehaltener Phonation. Ein hoher negativer Wert des SFR stimmt mit einer gesunden Stimmfunktion überein, ein kleiner negativer Wert nahe Null mit einer gestörten Stimmfunktion.

Ein weiteres Maß der spektralen Flachheit stellt das des Inversfilters (üblicherweise der Nenner von Gl. 2.29) dar. Nach Yanagihara stellt die spektrale Flachheit des Inversfilters (SFF) ein Maß für die Maskierung der Formanden im Spektrum durch Rauschen dar [Y67].

2.3.3 Pitch Amplitude

Die *Pitch Amplitude (PA)* ist definiert als Amplitudenwert des ersten Nebenmaximums der normierten Autokorrelationsfunktion des Residualsignals [RS78] und beschreibt die Periodizität des Stimmlippenschwingungsverhaltens. Der Wertebereich erstreckt sich von $[0; 1]$, wobei stark behauchte Stimmen einen geringen Wert des PA aufweisen und stimmgesunde Normalstimmen einen hohen Wert. Davis zeigte 1976 bereits [D76], dass sowohl heisere als auch behauchte Stimmen kleinere PA-Werte als Normalstimmen aufweisen.

Basierend auf dem linearen Modell der Sprachproduktion weist das Prädiktionsfehlersignal zu Beginn eines Schwingungszyklus hohe Fehlerwerte bei stimmgesunden Sprechern auf. Bei hochgradig gestörten Stimmen mit großer Glottisöffnungsfläche während der Phonation sind die Näherungen des linearen Sprachmodells nicht mehr erfüllt, da der Resonanzraum Vokaltrakt am glottalen Ende nicht abgeschlossen ist. Eine zunehmende Vermischung von pulsartiger und rauschhafter Anregung äußert sich im Residualsignal in undeutlicheren Fehlermaxima zu Beginn eines Schwingungszyklus. Der Wert der Pitch Amplitude weist entsprechend geringere Werte bei pathologischen Stimmen auf.

Plant et al. [PHW97] haben in Untersuchungen an gehaltener Phonation hohe Korrelationen ($r = 0{,}85$) des berechneten PA-Wertes mit perzeptiven Beurteilungskriterien der Gesamtstimmqualität gefunden.

2.4 Stimmgütemaße zur Beschreibung additiven Rauschens

Bei der Bestimmung einiger glottaler Rauschmaße wird im Gegensatz zur Vorgehensweise bei Schwingungsirregularitätsmaßen der Ansatz verfolgt, das Sprachsignal aus einer Signal- und einer Rauschkomponente zusammengesetzt zu betrachten [YGB82], [KOME86], [K87], [MBWMF88], [K93]. Das relative Verhältnis (*ratio*) der Energien dieser beiden Anteile stellt dabei ein grundlegendes Maß zur Beschreibung des glottalen Rauschanteils dar; durch den Wert des *Harmonics-to-Noise-Ratio* oder des *Signal-to-Noise-Ratio* beispielsweise. Die Bestimmung der beiden Komponenten aus dem akustischen Signal kann sowohl auf Basis des Zeitsignals [YGB82], als auch im Frequenzbereich erfolgen [KOME86], [K95].

2.4.1 Harmonics-to-Noise-Ratio

Ein Maß für die Quantifizierung der perzeptiv wahrgenommenen Heiserkeit einer Stimme geht auf Yumoto et al. [YGB82] zurück und wird als *Harmonics-to-Noise-Ratio* (HNR) bezeichnet. Der Wert des HNR ist durch das relative Verhältnis der harmonischen Signalenergie zur Energie des Rauschanteils bestimmt. In der Literatur sind dazu inzwischen verschiedene Ansätze zur Bestimmung der beiden Anteile neben der grundlegenden Idee von Yumoto et al. zu finden, bei der die Energie des harmonischen Signalanteils auf Basis einer gemittelten Periodenlänge aus 50 Perioden bestimmt wird. Die Verwendung solch langer Zeitfenster erlaubt die Bestimmung des HNR lediglich aus gehaltener Phonation und lässt keine aussagekräftigen Ergebnisse in fortlaufender Sprache zu. Ein weiterer Ansatz basiert auf der Analyse eines gemittelten Amplitudenverlaufs mehrerer aufeinander folgender Signalperioden [YGB82] oder schließlich eine Methode zur Schätzung des Rauschspektrums [KOME86].

Eine von Qi und Hillman [QH97] publizierte Methode zur Bestimmung des HNR aus fortlaufender Sprache operiert im Frequenzbereich (*frequency-based HNR, FHNR*) und basiert auf dem Cepstrum eines 200 ms langen Signalsegments. Nach Bestimmung des Cepstrums werden in diesem die Maxima identifiziert, die der Grundfrequenz und ihren Harmonischen entsprechen. Eine cepstrale Fensterfunktion wird weiterhin benutzt, um den Teil der hohen *Quefrenzen* aus dem Cepstrum zu *liftern*[14]. Dieses gelifterte Cepstrum wird zurück in den Frequenzbereich trans-

[14] *Quefrenzen* und *liftern* sind Wortschöpfungen von Tukey und als Anagramm von den Bezeichnungen „Frequenz“, bzw. „filtern“ abgeleitet.

formiert, sodass das resultierende Spektrum den geglätteten Rauschanteil darstellt.

Den eigentlichen Wert des FHNR bestimmt schließlich das relative Verhältnis der spektralen Energie des harmonischen und des Rauschanteils des jeweiligen Signalfensters.

2.4.2 Signal-to-Noise-Ratio

Das Signal-(zu)-Rausch-Verhältnis (*Signal-to-Noise-Ratio, SNR*) ist ein Maß für das relative Verhältnis zwischen Signal- und Rauschanteilen eines Sprachsignals. Für die Analyse fortlaufender Sprache existieren Ansätze von Klingholz [K87], [K90] oder auch von Qi et al. [QHM99]. Untersuchungen von Klingholz zur Folge liefert die Bestimmung des SNR aus fortlaufender Sprache eine um 5,6% bessere Klassifikationsleistung bei der Beschreibung pathologischer Stimmen als aus gehaltener Phonation [K90].

Eine von Qi et al. [QHM99] publizierte Methode zur Bestimmung des SNR aus fortlaufender Sprache basiert auf dem Modell der Zusammensetzung eines Sprachsignals aus regelmäßigen und somit prädizierbaren Anteilen – dem Signal – und unregelmäßigen, nicht vorhersagbaren Anteilen – dem Rauschen. Das relative Verhältnis beider Signalanteile zueinander wird nach Qi et al. als Signal-Rausch-Verhältnis bezeichnet. Die Bestimmung des prädizierbaren Anteils erfolgt dabei schrittweise durch systematisches Entfernen der Regelmäßigkeiten aus dem Sprachsignal, bis das Restsignal nahezu gaußverteilt ist [SA85].[15]

Entsprechend der akustischen Theorie der Sprachproduktion sind in Sprachsignalen sowohl Korrelationen auf kurzen als auch solche auf langen Zeitskalen vorhanden. Kurzzeitkorrelationen basieren dabei auf einer Prädizierbarkeit anhand direkt aufeinander folgender Abtastwerte und sind primär durch Vokaltraktresonanzen bestimmt. Die Betrachtung von Langzeitkorrelationen erfolgt im Gegensatz dazu nicht anhand direkt aufeinander folgender Abtastwerte, sondern ist vielmehr durch die quasiperiodische Signalstruktur der Sprache bestimmt. Die charakteristischen Signaleigenschaften zu Beginn eines jeden Stimmlippenschwingungszyklus lassen sich zu einem gewissen Teil aus den vorherigen Zyklen ableiten [RK89].

Zur Bestimmung der prädizierbaren Anteile der Sprache, sowohl der Kurz- als auch der Langzeitkorrelationen, werden Methoden der Linearen Prädiktion, die in

[15] Die Aufteilung des Sprachsignals in Kurzzeit- sowie Langzeit-Korrelationen und gausssches Rauschen finden auch im Bereich der Sprachcodierung in der Telekommunikation Anwendung, CELP [SA85].

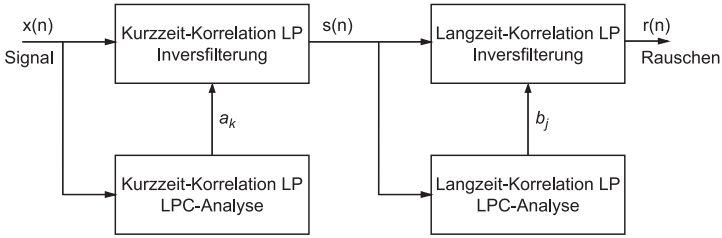


Abb. 2.4: *SNR aus fortlaufender Sprache.*

Abschnitt 2.3.1 bereits dargestellt worden sind, benutzt. Die Bestimmung von Kurzzeitkorrelationen erfolgt durch inverse Filterung des Sprachsignals $x(n)$ mit den aus Linearer Prädiktion bestimmten Filterkoeffizienten a_k. Typischerweise werden in diesem Schritt Analysefensterlängen von 20 ms und eine Prädiktorordnung von 14 verwendet [QHM99]. Das Residualsignal $s(n)$ der Inversfilterung stellt das von Kurzzeitkorrelationen bereinigte Signal dar, das nachfolgend auf Langzeitkorrelationen analysiert wird.

Zur Bestimmung der Langzeitkorrelationen wird ein Intervall von 2,5 ms verwendet. Diese Intervalllänge reicht aus, um die pulshafte Struktur des Residualsignals in periodischer Stimmlippenschwingung mit einzuschließen. Unter Verwendung einer Prädiktorordnung von 3 wird in einem Zeitintervall von 1,25 – 17,5 ms nach dem Koeffizientensatz b_j gesucht, der einen minimalen Prädiktionsfehler liefert. Mit diesem Filter wird das im ersten Schritt von Kurzzeitkorrelationen befreite Signal nun von Langzeitkorrelationen durch Inversfilterung bereinigt. Das inversgefilterte Signal $r(n)$ wird schließlich als Rauschanteil des ursprünglichen Sprachsignals betrachtet.

Der Wert des SNR wird als Verhältnis aus gemittelter RMS-Amplitude des Originalsignals und des von Kurz- und Langzeitkorrelationen bereinigten Signals bestimmt. Da das Originalsignal die beiden Anteile Signal und Rauschen umfasst, wird das berechnete Verhältnis noch durch Subtraktion von 1 bereinigt und anschließend logarithmiert [QHM99].

2.4.3 Glottal-to-Noise-Excitation-Ratio

Der *Glottal-to-Noise-Excitation-Ratio (GNE)* nach Michaelis et al. [MGS97], [MSZK94] stellt ein akustisches Maß zur Beschreibung relativen Rauschens in der Glottispulsfolge dar. Er ist ebenfalls am Dritten Physikalischen Institut in Göttingen entwickelt worden und seine Verwendung bei der akustischen Analyse pathologischer Stimmen weist Vorteile gegenüber anderen Rauschmaßen, wie dem *Normalized-Noise-Energy* Parameter (NNE) [KOME86] oder dem *Cepstrum-based-Harmonics-to-Noise-Ratio* (CHNR) [K95] auf, da der GNE, auf Grund seiner Bestimmungsmethode, nicht mit Schwingungsirregularitätsmaßen, wie Jitter und Shimmer korreliert [MGS97], [MFS98]. Der GNE berechnet sich als maximaler Korrelationskoeffizient zwischen Hilberteinhüllenden [SH94] des invers- und bandpassgefilterten Sprachsignals in verschiedenen Frequenzbändern und stellt somit ein Maß für die Qualität des glottalen Schlusses dar.

Die möglichen Rauschanteile im Sprachsignal setzen sich zum einen aus Anteilen in der Glottispulsfolge, die beim Aneinanderschlagen der beiden Stimmlippen entstehen, und zum anderen aus turbulentem Rauschen, das von unvollständigem Glottisschluss herrühren kann, zusammen. Bei vollständigem Glottisschluss wird der gesamte Frequenzbereich gleichmäßig von der Pulsfolge, die sich durch Inversfilterung bestimmen lässt, angeregt [FMLSK01]. Die Hilberteinhüllende verschiedener Frequenzbänder[16] zeigt dann eine annähernd gleiche Form. Die Form des anregenden Glottispulses ist dabei unabhängig von vorangegangenen oder nachfolgenden Pulsen und die berechneten Korrelationskoeffizienten weisen in diesem Fall hohe Werte nahe 1 auf.

Bei pathologischen Stimmen mit inkomplettem Glottisschluss entstehen bei der Stimmanregung Turbulenzen auf glottaler Ebene, die sich in einem breitbandigen Rauschanteil im Signal widerspiegeln. Unterschiedliche Rauschanteile in den verschiedenen Frequenzbändern führen zu verschiedenen Formen der Hilberteinhüllenden und entsprechend zu niedrigeren Korrelationskoeffizienten. Der Wert des GNE $[0; 1]$ liegt somit deutlich unter 1 und sinkt mit zunehmendem Rauschanteil in der Stimme bis auf Werte im Bereich von 0 ab. Für Normalstimmen weist der GNE Werte nahe bei 1 auf [MGS97].

[16] Die Bandpassfilterung erfolgt mit 3 kHz Frequenzbreite bei unterschiedlichen Mittenfrequenzen. Für den Korrelationskoeffizienten finden nur Bänder mit hinreichend großem Abstand der Mittenfrequenzen ($\Delta >$ 3 kHz) Verwendung.

2.5 Göttinger Heiserkeits-Diagramm

Das *Göttinger Heiserkeits-Diagramm (GHD)* geht auf eine Arbeit von Michaelis, Fröhlich und Strube am Dritten Physikalischen Institut in Göttingen zurück und stellt eine zweidimensionale, quantitative grafische Darstellung zur Beschreibung der Stimmgüte gehaltener Phonation dar [MFS98]. Die Verwendung des GHD ermöglicht sowohl die Beurteilung stimmgesunder Normalsprecher, als auch hochgradig gestörter pathologischer Stimmen [FMSK97], [FMK98], [FMSK98], [FMSK00]. In die Ergebnisse der akustischen Analyse nach den Parametern des GHD fließt eine Beurteilung der Schwingungsirregularitäten durch die *Irregularitätskomponente* und additiven Rauschens im Sprachsignal durch die *Rauschkomponente* ein.

Die Irregularitätskomponente I wird aus drei unterschiedlichen Schwingungsirregularitätsmaßen berechnet und stellt somit ein Maß für die Unregelmäßigkeit der Stimmlippenschwingung dar. Der Logarithmus des Jitters, des Shimmers und des mittleren Periodenkorrelationskoeffizienten fließen normiert und zu gleichen Teilen in die Berechnung der Irregularitätskomponente I nach Gleichung 2.32 ein. Die Verwendung des Logarithmus garantiert dabei, einen großen Parameterbereich im GHD darstellen zu können und trotzdem eine gute Auflösung zur Differenzierung zwischen einzelnen Stimmen ähnlicher Stimmgüte zu gewährleisten.

$$I = 5 + \frac{1}{\sqrt{3}} \left(\frac{\log(1 - wmc) + 1{,}614}{0{,}574} + \frac{\log(j3) + 0{,}374}{0{,}645} + \frac{\log(s15) - 0{,}757}{0{,}368} \right) \tag{2.32}$$

Die Bestimmung des Jitters und Shimmers erfolgt anhand der mittels des Waveform Matching Algorithmus bestimmten Periodenlängen und in Form des Perturbation Quotient (siehe Abschnitt 2.2.1) über einen Analysebereich von $K = 3$ beim Jitter ($j3$) und $K = 15$ beim Shimmer ($s15$). Die Berechnung des Periodenkorrelationskoeffizienten (wmc) erfolgt für paarweise aufeinander folgende Perioden.

Die Rauschkomponente R stellt ein Maß für den Rauschanteil in der Sprache und damit für den glottalen Schluss bei der Phonation dar. Ihr Wert wird nur durch den im vorigen Abschnitt 2.4.3 vorgestellten GNE bestimmt und nach Gleichung 2.33 berechnet.

$$R = 1{,}5 + \frac{(0{,}695 - gne3)}{0{,}242} \tag{2.33}$$

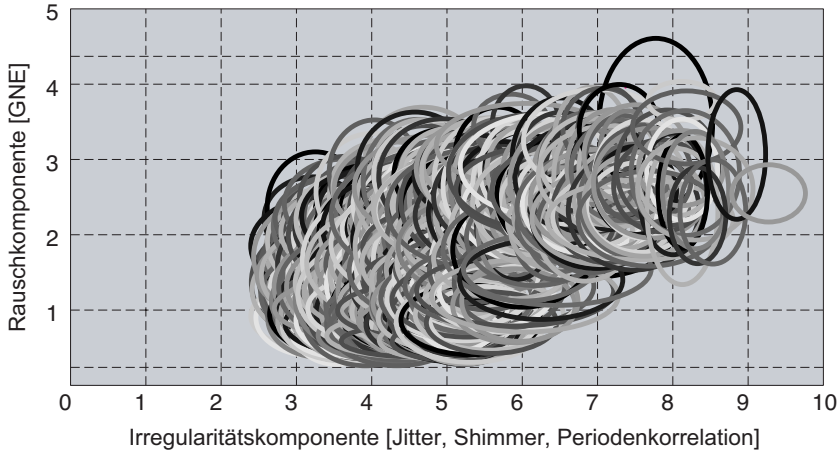


Abb. 2.5: *Darstellung der Stimmgüte für 500 unterschiedliche Sprecher im GHD. Der Mittelpunkt jeder Ellipse beschreibt dabei den Mittelwert der Irregularitäts- und Rauschkomponente des akustischen Signals, die beiden Halbachsen deren Standardabweichungen.*

Die berechneten Werte der Irregularitäts- und Rauschkomponente werden als Ellipse – mit dem Fehlermaß der beiden Komponenten als Halbachsen – grafisch im Heiserkeits-Diagramm dargestellt. Stimmgesunde Normalstimmen haben niedrige Werte der Irregularitäts- und Rauschkomponente, hochgradig gestörte pathologische Stimmen weisen hohe Werte in beiden Komponenten auf. Die Lage der Ellipse im GHD stellt dabei ein Maß für die Stimmgüte dar, wie für die Einzelergebnisse mehrerer hundert Sprecher unterschiedlichster Stimmgüte in Abbildung 2.5 dargestellt ist.

2.5.1 Erweiterung für fortlaufende Sprache

Die Skalierung des Göttinger Heiserkeits-Diagramms ist auf der Grundlage von 447 Vokalen [MFS98] durchgeführt worden und stellt demzufolge nicht zwangsläufig eine gute Repräsentation für fortlaufende Sprache dar. Auch die Bestimmung des Shimmers über ein Analysefenster von 15 aufeinander folgenden Perioden ist

in fortlaufender Sprache aufgrund kürzerer quasistationärer Signalabschnitte nicht immer gewährleistet.

Aus diesem Grund ist im Rahmen dieser Arbeit eine Neuskalierung des GHD für fortlaufende Sprache durchgeführt worden. Die direkte Vergleichbarkeit der berechneten Komponentenwerte zwischen gehaltener Phonation und fortlaufender Sprache geht dabei allerdings auf Kosten einer feineren Differenzierung unterschiedlicher Stimmgüten verloren [LFMS99], [LFMSK00], [KL03].

2.6 Langzeitspektren

Im Gegensatz zur Analyse der Signalspektren kurzer Zeitsegmente (Kurzzeitanalyse), die die zeitlich lokalen Eigenschaften des Sprachsignals beschreiben, enthalten Langzeitspektren zusätzliche Informationen über den Verlauf des Sprachsignals. Hammarberg et al. [HFGS80] haben dazu fortlaufende Sprachsignale von Sprechern mit Stimmstörungen mit einer 51-kanaligen Filterbank äquidistanter Filterbreite von je 250 Hz gefiltert und die Intensität jeder dieser Kanäle über die Zeit gemittelt (*Long Term Average Spectra, LTAS*).

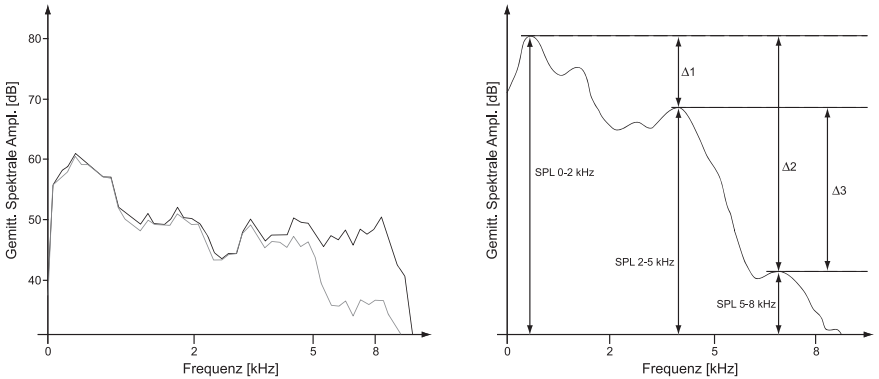


Abb. 2.6: *LTAS eines 40 s Sprachsegments aus einer 51-kanaligen Filterbank mit 250 Hz Filterbreite; heller Graph ist von stimmlosen Segmenten befreit (links) und Illustration der abgeleiteten Maße in drei Hauptfrequenzbändern 0-2, 2-5 und 5-8 kHz (nach [HFGS80]).*

Eine Bewertung dieser LTAS ist zum einen über eine Berechnung des spektralen Abfalls in dB pro Oktave dieser gemittelten Spektren möglich. Zum anderen ist es bei Vergleichen mit perzeptiven Bewertungen häufig üblich, die Frequenzachse dieser Langzeitspektren (f_s = 16 kHz) in drei Bereiche (0–2; 2–5; 5–8 kHz) zu unterteilen. Der maximale Pegel dieser drei Bereiche wird bei Hammarberg zu Vergleichen mit perzeptiven Bewertungen eines Gutachterkollektivs aus Logopäden und Phoniatern herangezogen und liefert hohe Korrelationen.

Neben einer Betrachtung des Gesamtsignals für die Bestimmung und Beurteilung der LTAS ist zusätzlich eine getrennte Analyse lediglich der stimmhaften Sprachanteile sinnvoll. Die Eliminierung der stimmlosen Bereiche erfolgte bei Hammarberg et al. über einen Energieschwellwertvergleich der unteren Frequenzbänder.

3 Klassifikation fortlaufender Sprache

Im Gegensatz zur gehaltenen Phonation isolierter Vokale stellt die fortlaufende Sprache eine Mischung aus stimmhaften und stimmlosen Bereichen sowie Sprechpausen dar, wie in Abbildung 3.1 exemplarisch verdeutlicht ist.

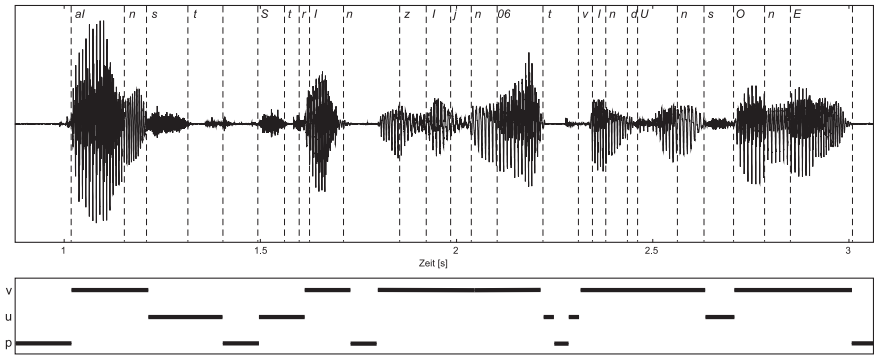


Abb. 3.1: *Sprachsignal: „Einst stritten sich Nordwind und Sonne“ eines stimmgesunden Sprechers und Klassifizierung stimmhafter (v) und stimmloser (u) Phonation sowie Sprechpausen (p).*

Ein Teil der in Kapitel 2 dargestellten akustischen Maße beschreibt Irregularitäten im Stimmlippenschwingungsverhalten in stimmhafter Phonation. Für die Bestimmung dieser Maße über das gesamte fortlaufende Sprachsignal müssen die stimmhaften Segmente (*v*) getrennt von stimmlosen (*u*) und Pausensegmenten (*p*) analysiert werden. Andere akustische Maße werden sowohl in stimmhafter als auch stimmloser Phonation berechnet (SNR, LTAS). Auch die Bestimmung von Sprechpausen und deren Länge kann zur Beschreibung der Stimmgüte beitragen.

Eine zuverlässige Methode zur automatischen Selektion der stimmhaften Segmente aus Sprachaufnahmen jeglicher Stimmgüte stellt eine wesentliche Grundvoraus-

setzung zur akustischen Analyse fortlaufender Sprache dar. Sämtliche bisher publizierten Methoden erlauben allerdings keine robuste Segmentierung bei hochgradig gestörten Stimmen. In der Entwicklung einer zuverlässigen, automatischen Klassifikationsmethode zur Selektion stimmhafter Segmente im fortlaufenden Sprachsignal bestand deshalb einer der wesentlichen Schwerpunkte dieser Arbeit.

Ausgangspunkt für die Entwicklung einer solchen Klassifikationsmethode war die Suche nach akustischen Gemeinsamkeiten von gesunden und gestörten Stimmen, die sich aus dem Sprachsignal ableiten lassen. Für unterschiedliche Sprecher – unabhängig von deren Stimmgüte – sollte der der Segmentierung zugrunde liegende, gesuchte „Parametersatz" bei der Artikulation desselben Phonems ähnliche Werte liefern. Die Selektion der stimmhaften Segmente erfolgt dabei nicht anhand einer Beurteilung der glottalen Anregung, sondern basiert auf der *Stimmhaftigkeit* des artikulierten Phonems[17]. Dadurch soll auch bei aphonen Sprechern die Selektion z. B. eines gesprochenen Vokals – also eines stimmhaften Phonems – möglich sein, obwohl bei dieser extremen Form der Stimmstörung keinerlei kontrollierte periodische Schwingung der Stimmlippen zu beobachten ist. Die vorgenommene Gruppierung der einzelnen Phoneme in stimmhafte und stimmlose Anregung ist dem Anhang 7.4 zu entnehmen.

Für die Segmentierung bieten sich zwei unterschiedliche Vorgehensweisen an:

1. Detektion, Isolation und Erkennung der einzelnen Phoneme im Sprachsignal und getrennte Analyse, je nach vorgegebener Zuordnung in stimmhafte/ stimmlose Phonation.

2. Analyse äquidistanter Kurzzeitsegmente fester Länge; Klassifizierung jedes einzelnen Kurzzeitsegments und anschließende Analyse zusammenhängender Signalbereiche der gleichen Phonationsklasse.

Während bei der ersten Methode eine Phonemerkennung unter Verwendung von Spracherkennungsalgorithmen durchgeführt werden muss, ist bei der zweiten Herangehensweise für jedes zu analysierende Kurzzeitsegment fester Länge eine Eingruppierung in stimmhaft oder stimmlos notwendig.

Da die Segmentgrenzen im zweiten Fall in der Regel nicht mit den Phonemgrenzen zusammenfallen, spielt die Länge des Analysefensters eine wichtige Rolle. Zu kurze Fensterlängen erlauben keine Bestimmung einer Grundperiodenlänge, zu lange

[17] In der Literatur wird Stimmhaftigkeit von Lauten in der Regel im Zusammenhang mit einer periodischen, glottalen Anregung verwendet [S83].

Segmente schließen mitunter mehrere Phoneme ein, sodass keine eindeutige Zuordnung des Gesamtsegments mehr vorgenommen werden kann. In der Sprachverarbeitung haben sich Analysefensterlängen im Bereich von 20 – 50 ms als sinnvoll erwiesen, da in diesem Zeitbereich das fortlaufende Sprachsignal als annähernd stationär betrachtet werden kann.

Eine sprecherunabhängige Phonemerkennung aus fortlaufender Sprache nicht bekannten Inhalts stellt schon bei stimmgesunden Normalsprechern eine komplexe Aufgabenstellung dar. Viele Phonemerkenner basieren auf aus dem Sprachsignal extrahierten Merkmalen, bei denen eine Beurteilung des Stimmlippenschwingungsverhaltens mit einfließt. Gerade die glottale Anregung der Phonation, die bei gestörten Stimmen deutlich vom Verhalten bei Stimmgesunden abweichen kann, soll bei der Klassifizierung nicht betrachtet werden, damit auch gesprochene stimmhafte Phoneme von Sprechern mit gestörter Stimmfunktion als solche analysiert werden und nicht aufgrund ungenügender Periodizität im Schwingungsverhalten als stimmlos klassifiziert werden.

Die Verwendung äquidistanter Kurzzeitfenster scheint für die Analyse pathologischer Stimmen besser geeignet. Nichtsdestotrotz ist auch bei dieser Methode eine zuverlässige Klassifizierung notwendig.

3.1 Pausendetektion

Die Detektion der Sprechpausen erfolgt auf Basis der Signalenergie, die in diesen Bereichen wesentlich kleinere Werte als während der Phonation aufweist. Die Bestimmung der Energie E_i gefensterter Kurzzeitsegmente der Länge M erlaubt über den Vergleich mit einem Schwellwert E_{thresh} die Klassifizierung von Sprechpausen. Segmente mit Werten unterhalb dieses Schwellwertes werden als Pause identifiziert und fließen in die weitere akustische Analyse nicht mit ein.

$$\text{RMS}_i = \sqrt{\frac{E_i}{M}} = \sqrt{\frac{\sum_{m=0}^{M-1} x(n+m)^2}{M}} \qquad \begin{cases} \leq E_{\text{thresh}} & : \text{ Pause} \\ > E_{\text{thresh}} & : \text{ Analyse} \end{cases} \tag{3.1}$$

$$\text{mit} \quad \text{Signal } x(n) \quad \text{und} \quad n = i \cdot M/4 \quad \text{bei} \quad i = 0,1,2,\ldots,4(N-M)/M.$$

Die Festlegung des Schwellwertes E_{thresh} erfolgt individuell für jede Sprachaufnahme. Für sich überlappende Kurzeitfenster (Fensterlänge/Fenstervorschub = 4/1) wird der *Root-Mean-Square*-Wert (RMS) berechnet. Unterschreitet dieser RMS-Wert

eines Kurzzeitsegments den Schwellwert von 3 % des Mittelwertes der drei größten RMS-Wertes aller Segmente, so wird das Segment als Sprechpause klassifiziert.[18]

$$E_{\text{thresh}} = 0{,}03 \cdot (1/3) \sum_{i=1}^{3} \text{argmax}_i(\text{RMS}) \tag{3.2}$$

Die relative Gesamtlänge der Sprechpausen im Verhältnis zur Gesamtphonationslänge kann ein weiterer Indikator für Stimmstörungen sein. Bei stimmgesunden Sprechern mit normalem Sprachfluss ist dieses Verhältnis kleiner als bei pathologischen Sprechern, die aufgrund eines erhöhten Anstrengungsgrades beim Sprechen u. U. häufiger Pausen zum Luftholen einlegen müssen.

Die Klassifikation stimmhafter Segmente erfolgt in einem zweiten Schritt auf den von Sprechpausen befreiten, zusammenhängenden Signalteilen.

3.2 Lineare stimmhaft/stimmlos-Klassifikation

Ein erster Ansatz basiert auf dem Vergleich charakteristischer akustischer Eigenschaften des Signals mit einem empirisch bestimmten Schwellwert dieser Eigenschaft, der zur Trennung der stimmhaften und stimmlosen Segmente geeignet gewählt sein muss [LSK98].

3.2.1 Nulldurchgangsrate

Die Nulldurchgangsrate (NDR, *zero crossing rate*) des Sprachsignals in einem Zeitfenster ist ein Anhaltspunkt für die Segmentierung [SB82]. Ein stimmhafter Bereich weist bei einem stimmgesunden Sprecher im Unterschied zu einem stimmlosen in der Regel eine NDR von unter 3 kHz auf [RS78].

NDR $\leq$ 3kHz	stimmhaftes Segment,
NDR $>$ 3kHz	stimmloses Segment.

Die Anwendung dieses Verfahrens versagt allerdings bei pathologischen Stimmen. Aufgrund des zumeist höheren Rauschanteils ist auch die Nulldurchgangsrate deutlich erhöht und eine zuverlässige Klassifizierung ist nicht möglich.

[18] Dieser 3 %-Wert ist aus empirischen Untersuchungen an der Sprachdatenbank ermittelt worden. Bei der Wahl aller empirisch gewählten Kriterien spielen die Bandbreite und Präemphase eine große Rolle.

3.2.2 Energie und AKF-Maximum

Auch über die Signalenergie eines Kurzzeitsegments ist eine Abschätzung stimmhafter bzw. stimmloser Phonation möglich. Parsa und Jamieson [PJ01] klassifizieren ein Kurzzeitsegment als stimmhaft, sobald folgende Bedingungen gleichzeitig erfüllt sind:

1. NDR < 1,5 kHz
2. Normierte Signalenergie $E > 30\,\%$ der Gesamtsignalenergie
3. Wert des 1. Nebenmaximums der normalisierten AKF > 0,3

Die Analyse mehrerer charakteristischer Eigenschaften liefert zwar bessere Ergebnisse als die Verwendung lediglich eines dieser Kriterien, sie erlaubt aber trotzdem nur eine grobe Unterscheidung zwischen pathologischen und nichtpathologischen Stimmen und versagt bei hochgradig gestörten Stimmen.

3.3 Nichtlineare stimmhaft/stimmlos-Klassifikation

Eine rein lineare Betrachtung bestimmter Signaleigenschaften liefert für die gesamte Bandbreite an Stimmqualität folglich keine zuverlässigen Ergebnisse. Ein Schwellwertvergleich stellt zudem immer einen Kompromiss dar: bei stark gestörten Stimmen trotzdem die stimmhaften Sprachanteile zu detektieren, bei stimmgesunden Sprechern aber dennoch keine stimmlosen Segmente fälschlicherweise als stimmhaft zu klassifizieren. Generell ist eine höhere *Falsch-negativ*-Quote (stimmhaft nicht erkannt) besser als eine entsprechende *Falsch-positiv*-Quote (stimmlos als stimmhaft klassifiziert). Einige wenige, als stimmhaft klassifizierte stimmlose Segmente können die Ergebnisse einzelner akustischer Maße nachhaltig verfälschen, während vereinzelte nicht analysierte stimmhafte Segmente bei der Vielzahl an Segmenten in fortlaufender Sprache keinen so gravierenden Einfluss haben.

Die Verwendung geeigneter Modelle Neuronaler Netze erlaubt die Lösung auch nichtlinear separierbarer Probleme, wie des XOR-Problems beispielsweise. In einer vorausgehenden Trainingsphase wird das Modell auf die Klassifizierung bestimmter Eigenschaften anhand von definiertem Material trainiert. Das so ausgebildete Neuronale Netz erlaubt bei erfolgreicher Generalisierung eine Klassifizierung auch nicht bekannten Materials.

3.3.1 Parametrisierung des Vokaltraktes

Der in dieser Arbeit entwickelte Ansatz basiert auf einer Parametrisierung des Vokaltrakteinflusses und nicht von Eigenschaften der glottalen Anregung. Die Stellung der Artikulatoren und die sich daraus unterschiedlich ausbildenden Resonanzen im Vokaltrakt (*Formanten*) haben wesentlichen Einfluss auf die Grobstruktur des Sprachspektrums. Diese sollte bei der Phonation desselben Phonems unabhängig vom Grad der Stimmstörung bei allen Sprechern ähnlich sein.

Grundlage dieser Überlegungen ist das bereits in Abschnitt 1.1 vorgestellte lineare Quelle-Filter-Modell der Sprachproduktion. Die einzelnen Komponenten des Modells sind dafür auf ihre Unabhängigkeit von der Stimmgüte und ihre Eigenschaften zur Klassifikation der stimmhaften Bereiche hin untersucht worden.

$$X(z) = E(z)G(z)V(z)L(z) \qquad \text{(Notation durch z-Transformierte).} \tag{3.3}$$

Die Verwendung eines festen Modells für die Form der Glottispulse $G(z)$, die Resonanzeigenschaften des Vokaltraktes $V(z)$ und die Lippenabstrahlung $L(z)$ haben keinen Einfluss auf die Stimmgüte. Die Modelldarstellung aus Gleichung 3.3 kann deshalb reduziert werden, zu:

$$X(z) = E(z)H(z) \qquad \text{mit} \qquad H(z) = G(z)V(z)L(z). \tag{3.4}$$

Das Modell des Filters $H(z)$ ist näherungsweise ein *Nur-Pole*-Filter, das die Grobstruktur des Signalspektrums beschreibt. Das *Nur-Nullstellen*-Filter $1/H(z)$ wird auch als *inverses* Filter bezeichnet und gibt bei Filterung des Sprachsignals mit diesem inversen Filter Auskunft über die Anregung $E(z)$. Geht man von einem zeitinvarianten festen Glottismodell und einer definierten, zeitinvarianten Abstrahlung der Lippen aus, so bietet sich für die gesuchte Differenzierung zwischen stimmhaften und stimmlosen Phonemen eine Beschreibung des Einflusses des Vokaltraktes an.

Um diesen Ansatz zu verdeutlichen, ist in Abbildung 3.2 exemplarisch ein Kurzzeitspektrum aus dem stimmhaften Bereich der Äußerung „einst“ für einen stimmgesunden und einen pathologischen Sprecher aufgetragen. Es ist eine gute Übereinstimmung der spektralen Grobstruktur im Frequenzbereich von 0-4 kHz trotz der unterschiedlichen Stimmqualitäten zu erkennen.

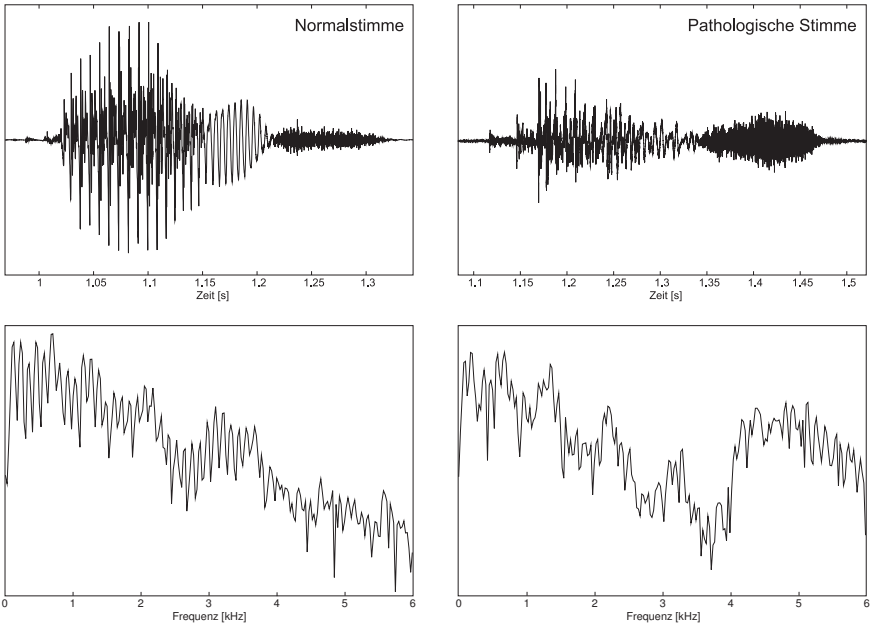


Abb. 3.2: *Äußerung „einst" und ein* 40 *ms Kurzzeitspektrum aus dem stimmhaften Bereich eines Sprechers mit gesunder (links) und gestörter Stimmfunktion (rechts).*

Zur Klassifizierung der parametrisierten spektralen Grobstruktur sind in einer früheren Arbeit [L98] verschiedene Parametersätze untersucht worden, die sich aus dem Zeitsignal durch Lineare Prädiktion berechnen lassen: LPC-Koeffizienten [M75], PARCOR-Koeffizienten [IS69], Reflexionskoeffizienten und Log-Area-Koeffizienten [W73]. Die besten Klassifikationsergebnisse lassen sich mit transformierten LPC-Spektren aus den Prädiktorkoeffizienten erzielen.

Eine Transformation ist zur Reduktion der Datendimensionalität notwendig, da bei der Verwendung mehrschichtiger Neuronaler Netze – wie in dieser Arbeit – jedes Element des Eingangsdatensatzes auf eine Eingangszelle abgebildet wird. Ist die Netzdimensionalität im Verhältnis zur Anzahl der Trainingsdatensätze zu groß, generalisiert das NN nicht, sondern lernt die Trainingsdatensätze auswendig. Eine zuverlässige Klassifikation neuer Datensätze ist in diesem Fall nicht gegeben.

Aus dem akustischen Zeitsignal berechnete LPC-Spektren eignen sich demzufolge nicht direkt für die Klassifikation mit einem Neuronalen Netz. Sie werden deshalb *Bark*-transformiert und durch eine Trapezfensterung auf der Frequenzachse in Frequenzbändern (Barkkanälen) zusammengefasst.

3.3.2 Barkskalierung

Eine nichtlineare Skalierung der Frequenzachse der LPC-Spektren nach der Barkskala ermöglicht eine effektive Reduktion der Datendimensionalität unter Beibehaltung der für die gesuchte Klassifikation wesentlichen Information durch ein Zusammenfassen von Frequenzbereichen in Frequenzbändern. Die Vorschrift für diese Skalierung ist von der Verarbeitung von Schallsignalen auf der Basilarmembran im menschlichen Ohr abgeleitet. Der repräsentierte Frequenzumfang eines gleich langen Abschnitts auf der Membran nimmt zu höheren Frequenzen hin zu. Eine äquidistante Skala auf der Basilarmembran beschreibt die *Tonheit* mit ihrer Maßeinheit *Bark*[19]. In Gleichung 3.5 ist ein analytischer Zusammenhang zwischen Frequenz f und Tonheit z näherungsweise dargestellt [TL87]:

$$z[\text{Bark}] = \frac{26{,}81 \cdot f}{1960\text{Hz} + f} - 0{,}53 \tag{3.5}$$

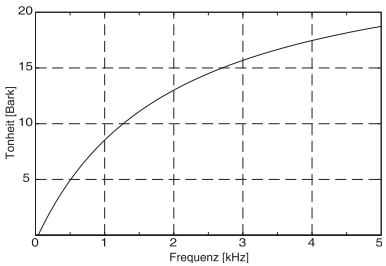


Abb. 3.3: *Qualitativer Verlauf der nichtlinearen Barkskalierung.*

Anhand dieser Beziehung kann jeder Frequenz f eine Tonheit z zugeordnet werden. Frequenzbänder, die einem Tonheitsintervall von 1 Bark Breite entsprechen,

[19] Nach Heinrich Georg Barkhausen, deutscher Physiker (1881-1956).

werden in sog. Barkkanälen zusammengefasst. In der Psychoakustik spricht man in diesem Zusammenhang auch von *Frequenzgruppen*. Diese Frequenzintegration wird durch sich im Frequenzbereich mit 0,5 Bark überlappende Trapezfenster von 1,5 Bark Breite realisiert. Die obere Grundlinienbreite der Trapezfenster beträgt 0,5 Bark, wie in Abbildung 3.4 dargestellt ist.

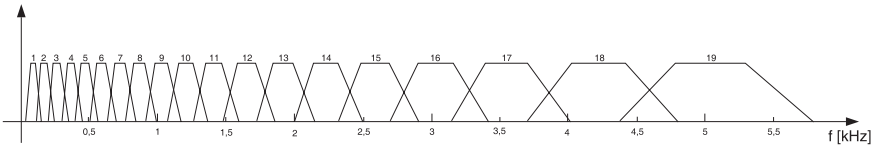


Abb. 3.4: *Schaubild der Trapezfensterung auf einer linearen Frequenzachse zur Berechnung der Barkkanäle, um die zu höheren Frequenzen anwachsende Barkkanalbreite zu veranschaulichen.*

Jeder dieser Frequenzgruppen entspricht auf der Basilarmembran die gleiche räumliche Ausdehnung. Für die unteren 5 Frequenzgruppen besteht ein annähernd linearer Zusammenhang zur Frequenz in Hz, wobei ein Bark ungefähr 100 Hz entspricht. Oberhalb von 5 Bark wachsen die Intervalle auf einer logarithmischen Skala an. Die dadurch gegebene feinere Auflösung tieffrequenter Bereiche spielt eine wesentliche Rolle, da diese für die Klassifikation wichtigere Informationen als die hochfrequenten enthalten. Speziell die Lage der niedrigen Formanten F1 und F2 bleibt unter dieser nichtlinearen Transformation erhalten. Der gesamte für den Menschen hörbare Frequenzbereich erstreckt sich über 24 Barkkanäle.

3.3.3 Dynamikkompression

Zwicker hat festgestellt, dass der sehr komplexe Zusammenhang zwischen physikalischem Schalldruck und der Empfindung am Ohr unter bestimmten Bedingungen durch ein Potenzgesetz nach Gleichung 3.6 angenähert werden kann [Z82]:

$$N \propto P^d \tag{3.6}$$

Ein Exponent von $d = 0{,}23$ spiegelt diesen Zusammenhang für einen breiten Hörbereich annähernd wider. Durch diese Dynamikkompression ist eine Steigerung der Klassifikationsleistung erreicht worden.

3.3.4 Normierung der Spektren

Um bessere Klassifikationsergebnisse erzielen zu können, werden die LPC-Spektren nach der Barkskalierung und der Dynamikkompression normiert. Es wird dazu eine zweidimensionale Maximum-Normierung durchgeführt, bei der sowohl in allen Barkkanälen z_i als auch über den gesamten Zeitverlauf T sämtlicher Barkspektren die maximale Intensität eines Barkkanals gesucht wird. Durch Multiplikation des gesamten Barkspektrums mit dem Kehrwert dieser Maximalamplitude wird die Normierung nach Gleichung 3.7 durchgeführt:

$$N^*_{tz} = \frac{N_{tz}}{\max\limits_{\substack{1<\tau<T \\ 1\leq\zeta\leq Z}} (\{N_{\tau\zeta}\})} \tag{3.7}$$

$$\text{für} \quad 0 \leq t < T \quad \text{und} \quad 1 \leq z \leq Z.$$

Um den Einfluss der Signalenergie – die nach dem *Parseval-Theorem* auch im Spektrum wiederzufinden ist – auf die Klassifikation auszuschließen ist eine solche Normierung notwendig und spielt insbesondere bei der Verwendung des *Error-Backpropagation*-Trainingsalgorithmus (siehe Abschnitt 3.4.6) für Neuronale Netze mit sigmoider Aktivierungskennlinie der Neuronen eine wichtige Rolle.

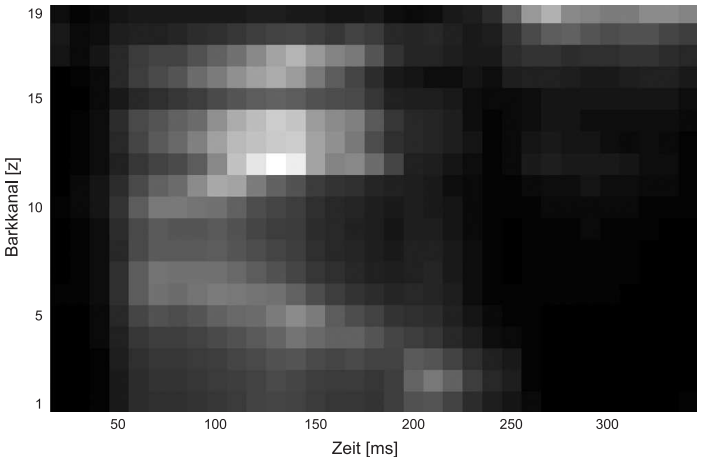


Abb. 3.5: *19-Kanal Barkspektrogramm der Äußerung „einst“ eines Normalsprechers.*

3.4 Neuronale Netzwerke

Das Ziel bei der Entwicklung künstlicher *Neuronaler Netzwerke* (NN, *artifical neural networks*) bestand darin, mathematische Modelle zu entwickeln, die die enorme kognitive Leistungsfähigkeit des menschlichen Gehirns nachempfinden. In Analogie zu realen, biologischen Neuronalen Netzen basiert diese auf dem hohen Parallelisierungsgrad ihrer Verarbeitung. Die einzelnen Elemente, die *Neuronen*, stellen dabei relativ einfache Verarbeitungseinheiten dar, deren Funktionsweise heutzutage gut verstanden ist. Erst durch die Vernetzung in Schichten über *Synapsen* werden sie in die Lage versetzt, komplexere Aufgabenstellungen effizient zu lösen [L87]. Die Stärke künstlicher Neuronaler Netze liegt dabei in der Lösung von Aufgaben im kognitiven Bereich, wie Lernen, Generalisierung, Musterklassifikation oder Optimierung. Für klassische, algebraische Berechnungen wie der Addition oder Multiplikation sind sie hingegen wenig geeignet.

Neuronale Netze finden Anwendung in den unterschiedlichsten Bereichen der Sprachsignalverarbeitung. Neben der in dieser Arbeit entwickelten stimmhaft/stimmlos-Klassifikation werden sie zur Spracherkennung, Formantextraktion, Sprachsynthese oder auch zur Erkennung nonverbaler Anteile der Sprache – wie Emotionen – verwendet. Es hat sich dabei ein breites Spektrum an unterschiedlichen Netztypen entwickelt, die jeweils individuelle Vorteile in Effizienz, Konvergenzverhalten, Robustheit oder Geschwindigkeit aufweisen. Für den Erfolgszug des *Neurocomputing* sind zum einem die Generalisierungsfähigkeit der Neuronalen Modelle – die ein echtes „Lernen und Optimieren“ ermöglicht – und zum anderen eine ausgeprägte Fehlertoleranz bei gestörten Eingangssignalen verantwortlich.

3.4.1 McCulloch und Pitts Neuron

Erste Modelle (künstlicher) Neuronaler Netzwerke tauchen bereits in den fünfziger Jahren bei den Mathematikern McCulloch und Pitts auf [MP43]. McCulloch und Pitts beschreiben ein binäres Entscheidungselement als ein *logisches Schwellwertelement* mit zwei möglichen Zuständen und bezeichnen es als Neuron, wenngleich es wesentlich einfacher als sein reales, biologisches Vorbild aufgebaut ist. Jedes dieser Elemente kann lediglich die Zustände $y = 0{,}1$ annehmen, wobei $y = 1$ einer Aktivierung und $y = 0$ einem Ruhezustand entspricht[20]. Über n Eingangsleitungen

[20] Die Aktivitäten 1 und 0 werden in Anlehnung an das Feuern bzw. Nichtfeuern eines Aktionspotentials biologischer Neuronen verwendet.

x_i mit $i = 1,\ldots,n$ („afferente Axone“) wird dem Neuron ein Informationsmuster angeboten, aus dessen Verarbeitung das Ergebnis an der einen Ausgangsleitung y („efferentes Axon“) berechnet wird.

Durch die Aktivitätszustände der Eingangsleitungen x_i wird eine Information kodiert. Die Aktivität der Ausgangsleitung y ergibt sich als Funktion A des Eingangs x_i, des Gewichtsvektors ω_i und des Schwellwertes s, der durch ω_0 mit $x_0 = 1$ bestimmt ist.

$$y = A(\sum_{i=1}^{n} \omega_i \cdot x_i - s) \tag{3.8}$$

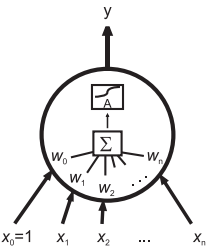


Abb. 3.6: *Neuron.*

Der Gewichtsvektor symbolisiert dabei die individuelle, synaptische Verbindungsstärke. Die Aktivierungsfunktion A bestimmt den Ausgabezustand als Funktion der Differenz aus dieser Summe und dem Schwellwert. Ursprünglich wurden dafür Stufenfunktionen (liefert 0 für negative Werte und 1 für 0 und positive Werte) verwendet, heute sind aber auch differenzierbare sigmoide Funktionen üblich.

Eine entscheidende Rolle bei der Lösung einer Klassifizierungsaufgabe spielt folglich der Gewichtsvektor ω. Dessen Wahl und Zusammensetzung bestimmen die Aktivierung des Neurons für ein bestimmtes Eingangsmuster und somit die Verbindungsstärke (Synapse) der Netzausgabe einer Zelle als Eingabe einer anderen. Ein iterativer Lernprozess nach vorgegebenem Algorithmus passt die Gewichte schrittweise an, um eine bestmöglich Klassifizierung zu erreichen. Diesen Vorgang bezeichnet man als *Lernphase* oder auch *Training* des Neuronalen Netzes.

3.4.2 Modellierung des Lernens

Das einfache McCulloch und Pitts Neuron führt die Abbildung mit einem fest definierten, vorgegebenen Satz an Gewichten durch. Das eigentliche Potential der Neuronalen Netze steckt allerdings in der Fähigkeit der Adaption des Netzwerkes, um eine Abbildungsfunktion möglichst gut zu approximieren. Zur Findung der optimalen Konfiguration dient eine vorherige Trainingsphase des NN. In dieser *erlernt* das Neuronale Netz anhand von Trainingsdatensätzen die gesuchte Klassifikation, bis eine gewünschte Abbildungsgenauigkeit erreicht ist.

Während dieser *Lern-/Trainingsphase* des Neuronalen Netzes werden unterschiedliche Vorgehensweisen des Lernens unterschieden:

1. **Überwachtes Lernen** (*supervised learning*): Bei dieser Trainingsform werden dem NN gleichzeitig ein Pärchen aus Eingangs- und zugehörigem Soll-Ausgabesignal präsentiert. Über eine Gewichtsanpassungsfunktion wird iterativ versucht, eine fehlerminimale Abbildung des Ausgabevektors als Funktion des Eingangsvektors zu erreichen. Ist die iterative Adaption für das gesamte Trainingsdatenmaterial in ein Gesamtfehlerminimum konvergiert, ist das Training abgeschlossen und die Gewichte werden fixiert. Diese Trainingsform setzt allerdings das Vorhandensein von Trainingsdatensätzen voraus, die zum jeweiligen Eingangsdatensatz auch die passende Ausgabe beinhalten. Typisches Beispiel ist der Error-Backpropagation-Algorithmus.

2. **Überwachtes, bestärkendes Lernen** (*graded, reinforcement learning*): Im Gegensatz zum überwachten Training wird bei dieser Form nicht die Sollausgabe mit angeboten. Es erfolgt vielmehr eine Beurteilung über die Güte (richtig/falsch, ggf. detailliertere Graduierung) der Netzausgabe, die die Grundlage für die Adaption der Gewichtsmatrix darstellt. Der exakte Wert der Abweichung wird nicht einbezogen.

3. **Unüberwachtes Lernen, Selbstorganisation** (*unsupervised learning, selforganization*): Bei der am weitesten autonomen Trainingsmethode werden dem Neuronalen Netz lediglich Eingangsdaten präsentiert. Das Netz entwickelt eigenständig Topologie und Gewichtsadaption in Ähnlichkeitsklassen, um eine Klassifizierung bestmöglich vorzunehmen; bspw. Cluster-Bildung.

Ziel während der Lernphase ist die selbständige Minimierung eines Fehlermaßes, welches die Güte der Abbildung von Eingangssignal des Neuronalen Netzes auf das Ausgabesignal quantifiziert, nach einem vorgegebenen Lernalgorithmus. Als Fehlermaß $E(\omega)$ findet zumeist der gemittelte quadratische Gesamtfehler (*mean squared error*) Verwendung.

Da es in der Regel keine exakte Lösung gibt, erfolgt – von einer initialen Konfiguration startend – eine iterative Adaption des Gewichtsvektors. Geht man unter verallgemeinerten Umständen davon aus, dass $E(\omega)$ differenzierbar nach ω ist, so erreicht man das Fehlerminimum durch iterativen Abstieg in Richtung des negativen Gradienten $-\nabla_\omega E(\omega)$ auf der Fehlerlandschaftsoberfläche, die durch die Parameter aufgespannt wird.

$$\omega_{\text{neu}} = \omega_{\text{alt}} - \eta \nabla_{\omega} E(\omega) \tag{3.9}$$

η wird dabei als Lernrate/Schrittweite bezeichnet und bestimmt als Faktor die Stärke der Änderung des Gewichtsvektors. Sein Wert ist typischerweise eine kleine positive Zahl. Dieser iterative Prozess wird solange fortgeführt, bis der Abbildungsfehler gegen ein Minimum konvergiert.

Maßgeblich für ein erfolgreiches Training mittels Gradientenabstiegsverfahren ist die Wahl dieser Lernrate η. Eine zu groß gewählte Lernrate bewirkt starke Sprünge in der Fehlerlandschaft. Ist die Lernrate dagegen zu klein, nimmt die Trainingsdauer rapide zu und beim Erreichen von flachen Plateaus kann die Adaption sogar zum Erliegen kommen, wie in Abb. 3.7 visualisiert ist. Komplexe Daten werden in der Regel besser erkannt und generalisiert, wenn die Lernrate klein gehalten wird.

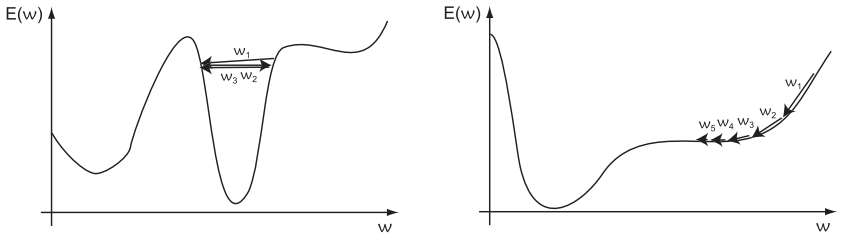


Abb. 3.7: *Die Wahl der Lernrate η hat entscheidenden Einfluss auf das Konvergenzverhalten. Links eine in dieser Anwendung zu groß gewählte Lernrate, die in einem Hin- und Herspringen in der Fehlerlandschaft resultiert. Rechts ein zu klein gewähltes η, sodass die Adaption auf flachen Plateaus zum Erliegen kommt.*

Die Gefahr bei dieser „einfachen“ Adaption besteht allerdings darin, in einem lokalen Fehlerminimum zu enden und das absolute/globale Minimum der Fehlerlandschaft nicht zu erreichen. Auf die Implementierung von Stabilisierungsparametern zur Vermeidung dieses Problems während der Lernphase wird im späteren Verlauf dieses Kapitels noch eingegangen.

Auf der Basis des einfachen McCulloch und Pitts-Modells und der Entwicklung von Lernregeln hat Rosenblatt [R59] ein Modell eines künstlichen Neuronales Netzes mit der Bezeichnung *Lineares Perzeptron* entwickelt. Dieses Modell ist von Minsky und Papert mathematisch analysiert und in mehrere Klassen von Perzeptrons eingeteilt worden [MP69].

3.4.3 Lineares Perzeptron

Das Modell des linearen Perzeptrons[21] beschreibt ein Neuronales Netz, bestehend aus einer festen, unbegrenzten Anzahl von Elementen, denen über N Eingangsleitungen N-dimensionale Merkmalsmuster zugeführt werden. Durch eine unterschiedliche Gewichtung der Eingangswerte und Vergleich der Summe mit einem Schwellwert wird ein Wert für die Ausgabezelle des Neuronalen Netzes berechnet.

Mit einem reellwertigen Perzeptron ist es unter Verwendung eines iterativen Gradientenverfahrens (*Delta*-Regel) möglich, ein gegebenes Klassifikationsproblem dergestalt zu lösen, dass in einem Merkmalsraum der Dimension d lineare (d-1)–dimensionale Separationshyperebenen konstruiert werden. Ein iteratives Gradientenverfahren führt nach einem Konvergenztheorem von Minsky und Papert in endlich vielen Schritten genau dann zum Ziel, wenn das zugrunde liegende Klassifikationsproblem *linear separabel* ist [MP69]. Für nicht linear separable Klassifikationsprobleme – wie das XOR bspw. – sind mehrschichtige Perzeptrons mit nichtlinearen Elementen erforderlich [MR90].

Lineare Netze können allerdings auch nur lineare Abbildungsfunktionen approximieren. Um nichtlineare Fragestellungen lösen zu können, sind nichtlineare Aktivierungsfunktionen notwendig, die in Form von nichtlinearen Kennlinien der Aktivierungsfunktionen implementiert sind. Typischerweise werden Funktionen mit sigmoider Kennlinie verwendet, die allesamt die Treppenfunktion approximieren und differenzierbar sind (vgl. Abb. 3.8). Beliebige funktionale Beziehungen zwischen den Merkmalsvektoren (Eingabedaten) und der Ausgabemenge können mit mehrschichtigen Neuronalen Netzen mit nichtlinearen Aktivierungsfunktionen der Neuronen abgebildet werden.

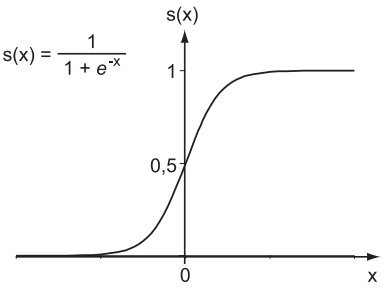


Abb. 3.8: *Sigmoide Fermi-Funktion als Aktivierungsfunktion des Neurons zur Lösung nichtlinearer Abbildungen.*

[21] Perzeptron leitet sich aus dem englischen *perceptron* als Verarbeitungseinheit einer sensorischen Empfindung *perception* ab.

3.4.4 Multi-Layer-Perzeptron

Mehrschichtige Perzeptrons (*Multi-Layer-Perceptron*, MLP) sind aus hintereinander geschalteten Schichten jeweils mehrerer Neuronen aufgebaut und entsprechen von ihrem Aufbau damit wesentlich eher realen, biologischen Neuronalen Netzen. Eine Eingangs-Schicht dient der Informationsaufnahme und leitet diese durch das Neuronale Netz an nachgelagerte Schichten zu einer Ausgabe-Schicht weiter. Bestehen dabei lediglich Verbindungen des Neuronenausgangs einer Schicht zu Eingangselementen von Neuronen der nächsten Schicht und keine rückwärts gerichteten oder Intraschichtverbindungen, so spricht man von *vorwärts gekoppelten* Neuronalen Netzen (*feed forward*). Ist jedes einzelne Netzelement darüber hinaus mit jedem Element der nächsten Schicht verknüpft, so handelt es sich um ein von *vollständig verbundenes* oder *vollvermaschtes* Netz. Unter bestimmten Voraussetzungen sind auch Rückkopplungen zu vorgelagerten Schichten im Neuronalen Netz sinnvoll. Zu beachten ist dann auch eine zeitliche Verarbeitungsabfolge innerhalb des NN, da im Gegensatz zu den statischen, rein vorwärtsgerichteten Netzen das Ergebnis der Abbildung nicht direkt berechnet werden kann. In dieser Anwendung kommen allerdings nur vollständig verbundene, vorwärts gekoppelte Multi-Layer-Perceptrons zum Einsatz.

Die einfachste Form eines mehrschichtigen Feed Forward Netzes besteht aus einer Eingangsschicht, einer versteckten „inneren“ Zwischenschicht und einer Ausgabeschicht. Eine solche Netztopologie wird auch als *dreischichtiges* Neuronales Netz bezeichnet und ist zur Veranschaulichung mit seinen Neuronen, Verbindungen und Gewichtsfaktoren exemplarisch in Abbildung 3.9 dargestellt.

Die einzelnen Neuronen i und h der Eingangs- und Zwischenschicht (bzw. h und o der Zwischen- und Ausgabeschicht) sind über Gewichtsfaktoren ω_{ih} (bzw. ω_{ho}) miteinander verbunden (*Gewichtsmatrix*), mit der die Ausgabe einer Zelle als Eingabe für die in der nächsten Schicht liegende Zelle multipliziert wird. Für die Berechnung der Ausgabe des mehrschichtigen Netzes als Funktion der gesamten Gewichtsmatrix und Schwellwertmatrix wird die in der jeweiligen Schicht gewichtete Eingabe vorwärts durch das Netz propagiert. Jedes Neuron erhält von allen Neuronen der vorgelagerten Schicht dessen Aktivität als Input.

Mehrschichtige Netze besitzen im Gegensatz zum zweischichtigen Perzeptron durch die zusätzlichen Neuronen in den versteckten Schichten (*hidden layer*) die Möglichkeit der Umkodierung der Daten in den Zwischenschichten. Den Vorteil der erweiterten Anwendungsmöglichkeiten einer mehrschichtigen Architektur er-

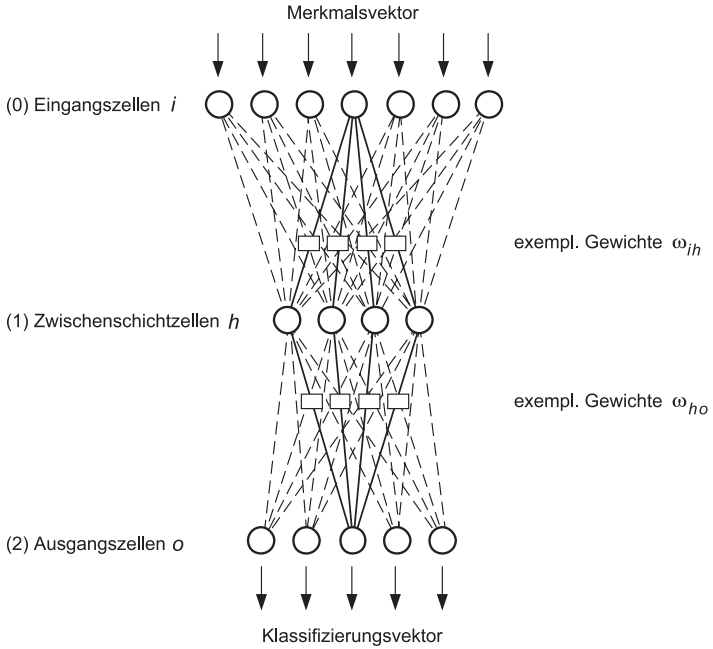


Abb. 3.9: *Schematische Darstellung eines dreischichtigen vollständig verbundenen, vorwärts gekoppelten Multi-Layer-Perzpetrons mit Verarbeitung der Information von oben (Eingangszellen) durch die versteckte Schicht (Zwischenschicht) nach unten (Ausgangszellen).*

kauft man sich allerdings – neben einem erheblich größeren Rechenaufwand während der Trainingsphase – durch ein ungewisses Konvergenzverhalten bei der Adaption. Die zugrunde liegende Fehlerlandschaft weist im Allgemeinen mehrere lokale Minima auf, was das Auffinden des gesuchten globalen Minimums erschweren kann. Die Verarbeitung in den versteckten Neuronen bleibt nach außen hin verborgen und kann nicht direkt beeinflusst werden.

Der Wahl der „richtigen“ Dimensionalität des Neuronalen Netzes kommt folglich eine wichtige Bedeutung zu, die in der Praxis durch wiederholtes Training mit variierender Netzkonfiguration verifiziert wird.

3.4.5 Netztopologie – Dimension

Die Auswahl der Art und Topologie des Neuronales Netzes erfolgt nach heuristischen Verfahren, da analytisch im Voraus die Eignung einer Konfiguration nicht bestimmbar ist. Die Anzahl der Eingabeneuronen ist durch die Dimensionalität des Merkmalsraums bestimmt. Die Dimension der Ausgabeschicht ist ebenfalls durch eine geeignete Kodierung der Anzahl der zu separierenden Klassen gegeben. Für die Bestimmung der Anzahl der Zwischenschichten und deren Dimensionalität existieren zwar Berechnungsansätze, allerdings keine analytische Vorschrift. Die besten Klassifizierungsergebnisse lassen sich generell mit einem möglichst kleinen – aber ausreichend großen – Netz erzielen. Eine Beschränkung der Dimensionalität nach oben existiert für MLP dabei insofern, als dass für ein Training mindestens so viele unterschiedliche Trainingsdatensätze wie Elemente der Gewichtsmatrix des Neuronalen Netzes existieren müssen. Anderenfalls würde das NN die Trainingsdatenmenge auswendig lernen und keine Generalisierung vollziehen.

An einem dreischichtigen Netz, wie es auch in dieser Arbeit zur Klassifikation benutzt wird, soll im folgenden Abschnitt 3.4.6 ein modifiziertes Gradientenverfahren zur Adaption der Gewichtsmatrix vorgestellt werden. Um beim Training des Netzes mit diesem Algorithmus eine Überanpassung (*overfitting*) [H91] und keine Generalisierung der Datensätze zu vermeiden, sollte die Anzahl der Zellen in den versteckten Schichten so klein wie möglich sein. In der Regel sind eine, maximal zwei versteckte Schichten ausreichend, um die gesuchte Klassifikation mit dem Neuronalen Netz abzubilden.

3.4.6 Backpropagation-Algorithmus

Der *(Error)-Backpropagation-Algorithmus* beschreibt eine wesentliche Trainingsmethode für Multi-Layer-Perceptrons und geht auf Werbos [W74] zurück. Der Mangel an Adaptionsverfahren für die versteckten Schichten mehrschichtiger Netze hat das Potential der MLP lange brach liegen lassen. Dieser mehrstufige Algorithmus basiert auf einer iterativen Adaption der Gewichtsmatrix mit dem Ziel der Minimierung der Differenz zwischen Sollausgabe und mit derzeitigen Gewichten berechneter Netzausgabe. Da dies in der Regel analytisch nicht möglich ist, werden die Gewichte – ausgehend von einer Startkonfiguration (zufällige oder feste Werte) – nach einem Gradientenabstiegsverfahren iterativ so geändert, dass sich der berechnete Fehler am Netzausgang in der hochdimensionalen Fehlerlandschaft im Mittel in Richtung eines lokalen Minimums ändert [RHW86].

Der Ablauf des Backpropagation-Algorithmus gliedert sich dabei in drei Stufen. In der ersten Stufe (*forward pass*) wird die Netzeingabe eines Trainingsmusters Schicht für Schicht durch das Netz geleitet und die Netzausgabe berechnet. In einem zweiten Schritt – der Fehlerbestimmung – wird die Abweichung dieser Netzausgabe von der bekannten Sollausgabe berechnet. Überschreitet der Fehler einen vorgegebenen Schwellwert, liefert das Neuronale Netz noch nicht die erforderliche Abbildungsgüte und eine Adaption der Gewichtsmatrix ist notwendig. In der dritten Stufe (*backward pass*) wird dazu dieses Fehlermaß in entgegengesetzter Richtung Schicht für Schicht durch das Netz *zurück*-geleitet (*backpropagation*) und eine Adaption der Gewichte nach der Lernregel vorgenommen.

Jedes Neuron besitzt in dieser speziellen Anwendung eine kontinuierliche Ausgangsaktivität s im Bereich $[0;1]$. Die Aktivität s_h eines Neurons h der inneren Schicht (1) und die Aktivität s_o eines Neurons o der Ausgabeschicht (2) (vgl. Abbildung 3.9) berechnet sich entsprechend Gleichungen 3.10 und 3.11.

$$s_h = \sigma(h_h) = \sigma(\sum_{i=1}^{N_i} \omega_{ih} \cdot s_i + \vartheta_h) \tag{3.10}$$

$$s_o = \sigma(h_o) = \sigma(\sum_{h=1}^{N_h} \omega_{ho} \cdot s_h + \vartheta_o) \tag{3.11}$$

$$= \sigma(\sum_{h=1}^{N_h} \omega_{ho} \cdot \sigma(\sum_{i=1}^{N_i} \omega_{ih} \cdot s_i + \vartheta_h) + \vartheta_o) \tag{3.12}$$

Der Index h kennzeichnet dabei die N_h versteckten Zellen der Zwischenschicht ($1 \leq h \leq N_h$) und o die N_o Zellen der Ausgangsschicht ($1 \leq o \leq N_o$). Die Funktion $\sigma(x)$ beschreibt die Antwort eines Neurons x auf die gesamte synaptische Eingabe aller verbundenen Zellen der vorherigen Schicht und wird als Aktivierungsfunktion des Neurons bezeichnet. Für die Wahl der nichtlinearen Aktivierungsfunktion $\sigma(x)$ ist eine sigmoide Funktion, wie die *Fermifunktion*[22] oder der *Tangens hyperbolicus* (tanh), gebräuchlich. Die s_i bestimmen die Aktivitäten der N_i Neuronen ($1 \leq i \leq N_i$) in der Eingangsschicht (0) und entsprechen den Elementen des N_i-dimensionalen Merkmalsvektors. ϑ_h und ϑ_o sind Schwellwerte der Neuronen, mit denen die Gesamtsumme der Eingabe verglichen wird.

[22] Fermifunktion $\sigma(x) = 1/(1+e^{-x})$. vgl. Abb. 3.8. Die Begrenzung des Wertebereichs solcher Funktionen liefert Vorteile bei der technischen Implementierung

Durch die Gleichungen 3.10 und 3.11 wird jedem der N_p Eingabemuster $\vec{x}^p$ aus der Mustermenge[23] $(\vec{x},\vec{y})_p = (\vec{x}^p,\vec{y}^p)$ ein Ausgabemuster $\vec{y}^p$ nach folgender Beziehung zugeordnet:

$$y_o^p = s_o(\vec{x}^p) \tag{3.13}$$

In der Trainingsphase werden die Gewichte ω_{ih} und ω_{ho} sowie die Schwellwerte ϑ_h und ϑ_o schrittweise so adaptiert, dass sie die N_p Eingabemuster $\vec{x}^p$ im Sinne des kleinsten quadratischen Abbildungsfehlers $E(\omega_{ih},\vartheta_h,\omega_{ho},\vartheta_o)$ nach Gleichung 3.14 auf die Sollausgabe y_o^{p*} abbilden:

$$E(\omega_{ih},\vartheta_h,\omega_{ho},\vartheta_o) = \frac{1}{2}\sum_{p}^{N_p}\sum_{o}^{N_o}[s_o^p - y_o^p]^2 \tag{3.14}$$

Die Gewichts- und Schwellwertmatrix ist für die Abbildung eines festen Mustersatzes $(\vec{x},\vec{y})_p$ dann optimal gewählt, wenn der Fehler E minimal ist.

Die Bestimmung der optimalen Abbildungsmatrizen läuft folglich auf ein Minimierungsproblem des Abbildungsfehlers hinaus, für das sich als einfache Möglichkeit ein modifiziertes Gradientenabstiegsverfahren anbietet. Zu diesem Zwecke bestimmt man den Gradienten von E durch partielles Ableiten nach den unterschiedlichen Gewichten und Schwellwerten.

$$\begin{aligned}
\frac{\partial E}{\partial \omega_{ho}} &= \frac{\partial E}{\partial s_o}\frac{\partial s_o}{\partial \omega_{ho}} && (3.15)\\
&= (s_o - y_o)\cdot\sigma'(h_o)\cdot\frac{\partial h_o}{\partial \omega_{ho}} && (3.16)\\
&= (s_o - y_o)\cdot s_o(1-s_o)\cdot s_h && (3.17)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial E}{\partial \vartheta_o} &= \frac{\partial E}{\partial s_o}\frac{\partial s_o}{\partial \vartheta_o} && (3.18)\\
&= (s_o - y_o)\cdot\sigma'(h_o)\cdot\frac{\partial h_o}{\partial \vartheta_o} && (3.19)\\
&= (s_o - y_o)\cdot s_o(1-s_o) && (3.20)
\end{aligned}$$

[23] Der Index p (für *pattern*) bezeichnet das p-te Musterpaar aus der Menge der N_p Trainingsmusterpaare.

mit

$$\frac{\partial E}{\partial s_o} = (s_o - y_o) \quad , \quad \frac{\partial s_o}{\partial \omega_{ho}} = s_o(1 - s_o)s_h \quad , \quad \frac{\partial s_o}{\partial \vartheta_o} = s_o(1 - s_o) \tag{3.21}$$

und für die sigmoide Aktivierungsfunktion gilt:

$$\sigma'(x) = \frac{\partial}{\partial x}\sigma(x) = \sigma(x)(1 - \sigma(x)) \tag{3.22}$$

und den partiellen Ableitungen von E nach ω_{ih} und ϑ_h:

$$\frac{\partial E}{\partial \omega_{ih}} = \sum_{o=1}^{N_o} \frac{\partial E}{\partial s_o}\frac{\partial s_o}{\partial s_h}\frac{\partial s_h}{\partial \omega_{ih}} \tag{3.23}$$

$$= \sum_{o=1}^{N_o} (s_o - y_o) \cdot \sigma'(h_o)\frac{\partial h_o}{\partial s_h} \cdot \frac{\partial s_h}{\partial \omega_{ih}} \tag{3.24}$$

$$= \sum_{o=1}^{N_o} (s_o - y_o) \cdot \sigma'(h_o)\omega_{ho} \cdot \frac{\partial s_h}{\partial \omega_{ih}} \tag{3.25}$$

$$= \sum_{o=1}^{N_o} (s_o - y_o) \cdot \sigma'(h_o)\omega_{ho} \cdot \sigma'(h_h)\frac{\partial h_h}{\partial \omega_{ih}} \tag{3.26}$$

$$= \sum_{o=1}^{N_o} (s_o - y_o) \cdot s_o(1 - s_o)\omega_{ho} \cdot s_h(1 - s_h)s_i \tag{3.27}$$

$$\frac{\partial E}{\partial \vartheta_h} = \sum_{o=1}^{N_o} \frac{\partial E}{\partial s_o}\frac{\partial s_o}{\partial s_h}\frac{\partial s_h}{\partial \vartheta_h} \tag{3.28}$$

$$= \sum_{o=1}^{N_o} (s_o - y_o) \cdot \sigma'(h_o)\frac{\partial h_o}{\partial s_h} \cdot \frac{\partial s_h}{\partial \vartheta_h} \tag{3.29}$$

$$= \sum_{o=1}^{N_o} (s_o - y_o) \cdot \sigma'(h_o)\omega_{ho} \cdot \frac{\partial s_h}{\partial \vartheta_h} \tag{3.30}$$

$$= \sum_{o=1}^{N_o} (s_o - y_o) \cdot \sigma'(h_o)\omega_{ho} \cdot \sigma'(h_h)\frac{\partial h_h}{\partial \vartheta_h} \tag{3.31}$$

$$= \sum_{o=1}^{N_o} (s_o - y_o) \cdot s_o(1 - s_o)\omega_{ho} \cdot s_h(1 - s_h) \tag{3.32}$$

mit

$$\frac{\partial E}{\partial s_o} = (s_o - y_o) \quad , \quad \frac{\partial s_o}{\partial s_h} = s_o(1 - s_o)\omega_{ho} \quad , \quad \frac{\partial s_h}{\partial \omega_{ih}} = s_h(1 - s_h)s_i \tag{3.33}$$

Die Änderung der Gewichtsmatrizen $\Delta\omega$ und Schwellwerte $\Delta\vartheta$ für den aktuellen Lernschritt S erfolgt für ein hinreichend kleines positives η entlang der Richtung des steilsten Abfalls von E. Der Faktor η beschreibt die Lernrate (oder *Schrittweite*).

$$\Delta\omega_{ho}(S) = -\eta \cdot \frac{\partial E}{\partial \omega_{ho}} = -\eta \cdot (s_o - y_o) \cdot s_o(1 - s_o) \cdot s_h \tag{3.34}$$

$$\Delta\vartheta_o(S) = -\eta \cdot \frac{\partial E}{\partial \vartheta_o} = -\eta \cdot (s_o - y_o) \cdot s_o(1 - s_o) \tag{3.35}$$

$$\Delta\omega_{ih}(S) = -\eta \cdot \frac{\partial E}{\partial \omega_{ih}} = -\eta \cdot \sum_{o=1}^{N_o} (s_o - y_o) \cdot s_o(1 - s_o)\omega_{ho} \cdot s_h(1 - s_h)s_i \tag{3.36}$$

$$\Delta\vartheta_h(S) = -\eta \cdot \frac{\partial E}{\partial \vartheta_h} = -\eta \cdot \sum_{o=1}^{N_o} (s_o - y_o) \cdot s_o(1 - s_o)\omega_{ho} \cdot s_h(1 - s_h) \tag{3.37}$$

Bei Betrachtung der Gleichungen 3.34 und 3.36 für die Änderung der Gewichtsmatrizen fällt auf, dass der Faktor $(s_o - y_o) \cdot s_o(1 - s_o) \cdot s_h$ in beiden Gleichungen auftaucht. Die Änderung an der Ausgabeschicht $\Delta\omega_{ho}$ wird somit *rückwärts* – entgegen der Richtung der synaptischen Verbindungen – an die vorherige Schicht durch das Netzwerk zurückpropagiert und entspricht dem *backward pass* des Backpropagation-Algorithmus. Bei der Betrachtung der Schwellwertänderungen nach Gleichung 3.35 und 3.37 gilt entsprechendes.

Es ist allerdings zweifelhaft, ob diese „Rückpropagation" bei biologischen Neuronen ebenfalls möglich ist. Für die Simulation und letztendlich den Erfolg Neuronaler Netze auf Computern hat dieses einfach zu implementierende, schnelle und sehr robuste Lernverfahren einen wesentlichen Grundstein gelegt.[24]

[24] Seit ca. 1986 hat sich das Gebiet geradezu explosiv entwickelt. Es gibt eine Vielzahl von wissenschaftlichen Zeitschriften zum Hauptthema Neuronaler Netze (Neural Networks, Neural Computation, Neurocomputing, IEEE Trans. on Neural Networks, etc.), große anerkannte wissenschaftliche Gesellschaften wie die INNS (International Neural Network Society), die ENNS European Neural Network Society), eine große IEEE Fachgruppe über neuronale Netze und Fachgruppen nationaler Informatik-Gesellschaften wie die GI (Gesellschaft für Informatik).

Trägheitsparameter

Ein wichtiger Parameter zur Stabilisierung des Trainings mittels des Backpropagation-Algorithmus ist der Trägheitsparameter α (*Momentum*). Er beschreibt eine Richtungsträgheit beim Gradientenverfahren zur Minimierung des Netzfehlers E. Diese Modifikation des Verfahrens ist in allgemeinerer Form auch unter dem Namen *konjugierter Gradientenabstieg (conjugate gradient descent [PFTV86])* bekannt [MR90]. Der Trägheitsparameter spielt insbesondere in stark zerklüfteten Regionen und auf flachen Plateaus der Fehlerlandschaft eine wichtige Rolle.

Bei der Berechnung der Adaption der Gewichtsmatrix für den aktuellen Lernschritts S fließt eine Bewertung der Änderungen des vorherigen Lernschritts $S-1$ mit dem Trägheitsparameter α nach folgendem Zusammenhang mit ein:

$$\Delta\omega_{ho}(S) = -\eta \sum_{p=1}^{N_p} \left(\frac{\partial E}{\partial \omega_{ho}} \right) + \alpha \Delta\omega_{ho}(S-1) \tag{3.38}$$

$$\Delta\vartheta_{o}(S) = -\eta \sum_{p=1}^{N_p} \left(\frac{\partial E}{\partial \vartheta_{o}} \right) + \alpha \Delta\vartheta_{o}(S-1) \tag{3.39}$$

$$\Delta\omega_{ih}(S) = -\eta \sum_{p=1}^{N_p} \left(\frac{\partial E}{\partial \omega_{ih}} \right) + \alpha \Delta\omega_{ih}(S-1) \tag{3.40}$$

$$\Delta\vartheta_{h}(S) = -\eta \sum_{p=1}^{N_p} \left(\frac{\partial E}{\partial \vartheta_{h}} \right) + \alpha \Delta\vartheta_{h}(S-1) \tag{3.41}$$

Bei geeigneter Wahl von $\alpha \in [0;1]$ verläuft das Training des Neuronalen Netzes stabiler, da so ein „Hin- und Herspringen“ (Oszillation) in der Fehlerlandschaft bei stark wechselnden Gradienten unterbunden wird. Für $\alpha = 0$ ergibt sich die schon beschriebene verallgemeinerte δ-Regel aus Gleichung 3.34 und folgende.

Um einer immer weiter fortschreitenden Adaption der Gewichtsmatrix während des Trainings Rechnung zu tragen, werden der Trägheitsparameter und die Lernrate zumeist adaptiv implementiert. Mit fortschreitendem Trainingsverlauf werden die beiden Parameter abgesenkt, um ein schnelles Konvergenzverhalten zu erreichen.

3.4.7 Modifizierter Backpropagation-Algorithmus

Bei dem von Rumelhart vorgestellten Backpropagation-Algorithmus können Probleme bei extremen Werten $[0;1]$ an den Ausgabezellen s_o des Neuronalen Netzes auftreten. Entsprechend Gleichung 3.21 liefert der Term $s_o(1-s_o)$ für $s_o \approx 1$ oder $s_o \approx 0$ sehr kleine Fehlerwerte. Die unter Einbeziehung dieses Fehlermaßes berechneten Änderungen der Matrizen nach Gleichung 3.34, 3.35, 3.36 und 3.37 kann dann trotz eines möglichen maximal falschen Netzausgabewert (1 anstatt 0 oder 0 anstatt 1) keine entsprechend notwendige, große Änderung bewirken. Dies kann zu einer massiven Verzögerung der Adaption führen.

Dieses Problem bringt die Ableitung der Aktivierungsfunktion $\sigma(x)$ nach Gleichung 3.22 mit sich.

$$\sigma'(x) = \sigma(x)(1-\sigma(x))$$

Im Grenzwert der beiden Sättigungspunkte der sigmoiden Funktion verschwindet deren Ableitung.

Durch eine geringfügige Modifikation der Fehlerfunktion $E(\omega_{ih},\vartheta_h,\omega_{ho},\vartheta_o)$ des Backpropagation-Algorithmus nach van Ooyen et al. [ON92] lässt sich dieses Problem beheben. Anstatt den quadratischen Fehler der Differenz zwischen Sollausgabe y_o^p und tatsächlicher Ausgabe s_o^p für das aktuelle Musterpaar p des Netzes zu minimieren, wird eine modifizierte Fehlerfunktion verwendet:

$$E = -\sum_{p=1}^{N_p}\sum_{o=1}^{N_o}[y_o \ln s_o + (1-y_o)\ln(1-s_o)] \tag{3.42}$$

Die Änderung der partiellen Ableitungen von E nach den Gewichten und Schwellwerten lautet dann exemplarisch für $\partial E/\partial\omega_{ho}$:

$$\frac{\partial E}{\partial \omega_{ho}} = (s_o - y_o)\cdot s_h \tag{3.43}$$

Der zurückpropagierte Fehler ist somit direkt proportional zur Differenz von Netzwerk- und Sollausgabe $(s_o - y_o)$.

3.4.8 Training des Neuronalen Netzes

Bevor ein entsprechend dimensioniertes Multi-Layer-Perzeptron zur Klassifikation einer Problemstellung verwendet werden kann, muss es in einer separaten Trainingsphase unter Verwendung eines Lernalgorithmus – wie des Backpropagation-Algorithmus bspw. – zuerst iterativ auf die zu leistende Abbildung trainiert werden. Es existieren bisher keine Algorithmen, die direkt aus den Trainingsdatensätzen die Gewichtsmatrizen für eine gewählte Topologie berechnen können. Während der Trainingsphase findet vielmehr eine gerichtete Minimumsuche auf der hochdimensionalen Fehleroberfläche statt, die durch die Parameter der einzelnen Neuronen aufgespannt wird.

Die nach Abschluss des iterativen Trainings fixierten Gewichts- und Schwellwertmatrizen des trainierten Neuronalen Netzes werden in der *Erkennungs-* oder *Klassifikationsphase* zur Klassifikation unbekannter Eingabedaten des gleichen Datentyps verwendet. Eine Abbildung der zu erkennenden Daten unter Verwendung dieser adaptierten Matrizen liefert entsprechend der Trainingsvorgaben eine Netzausgabe und somit Klassifikation. Ist das Neuronale Netz in der Lage, auch nicht zum Trainingsdatensatz gehörige Datensätze korrekt zu klassifizieren, so spricht man von einer erlangten *Generalisierung*, die notwendig für die Klassifikation unbekannten Datenmaterials ist.

Die Verwendung von Trainingsmethoden, wie dem Gradientenabstiegsverfahren, mit der Vorgabe der Minimumsuche eines Fehlermaßes, liefern – vorausgesetzt eine adäquate Wahl der Trainingsparameter – eine stetige Minimierung des absoluten Netzausgabefehlers mit fortschreitenden Trainingsiterationen. Die daraus resultierende Adaption der Gewichtsmatrizen hat im späteren Trainingsverlauf nicht notwendigerweise auch eine Steigerung der Klassifikationsleistung für Nichttrainingsdaten zur Folge. Das Neuronale Netz verliert u. U. bei fortschreitendem Training seine zwischenzeitlich erlangte Generalisierungsfähigkeit und versucht, irrelevante Details der Trainingsdatensätze abzubilden, unter der Maßgabe, den absoluten Netzausgabefehler weiter zu minimieren. Dieses häufig zu beobachtende Phänomen der Überanpassung (*overfitting*) ist in Abbildung 3.10 visualisiert. Nach ca. 230 Lernschritten weist das Neuronale Netz einen minimalen Klassifikationsfehler beim Test mit Nichttrainingsdaten auf und es besteht zu diesem Zeitpunkt eine maximale Generalisierungsfähigkeit. Mit weiteren Trainingsschritten nimmt der Trainingsfehler zwar weiter ab, die Klassifikationsleistung wird allerdings auch wieder schlechter.

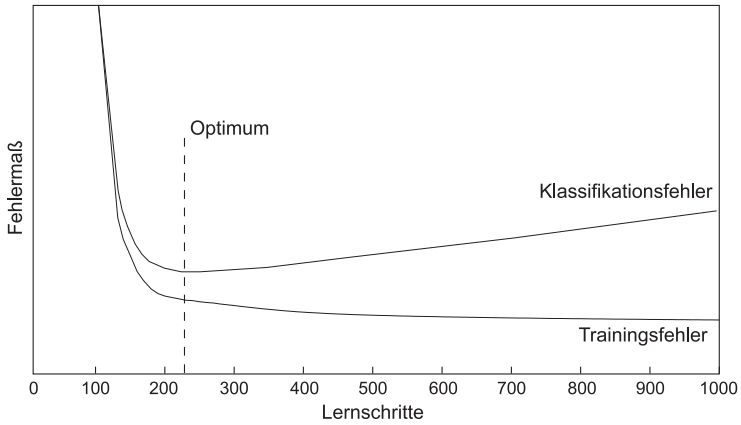


Abb. 3.10: *Verlauf des absoluten Netzfehlers und des Klassifikationsfehlers für Nichttrainingsdaten (Generalisierungsfähigkeit) während der Trainingsiterationen.*

Dieses Problem umgeht man durch das sog. *leave one out*-Training[25], bei dem nicht mit dem gesamten Trainingsdatensatz das Netz trainiert wird, sondern einige Musterpaare aus Eingabe und Sollausgabe für die Beurteilung der Klassifikationsleistung zurückgehalten werden. In definierten Abschnitten der Trainingsphase wird das derzeitige Netz zur Abbildung dieser nicht zur Trainingsmenge gehörenden Datensätze in einer Klassifikationsphase verwendet. Der Klassifikationsfehler sollte solange fallen, bis die maximale Generalisierungsfähigkeit erreicht ist. Danach beginnt die Phase der Überanpassung und die Abbildungsleistung geht zurück. Der absolute Netzfehler (Trainingsfehler) kann durchaus in nachfolgenden Trainingsiterationen weiter fallen, das NN beginnt allerdings damit, für die Gesamtmenge der Datensätze nicht repräsentative Details der Auswahl der Trainingsdaten abzubilden. Das Training sollte an dieser Stelle abgebrochen werden, sofern natürlich die geleistete Klassifikation den gewünschten Anforderungen genügt. Ansonsten ist die Topologie zu überdenken oder mit einer modifizierten Startkonfiguration der Matrizen erneut zu trainieren.

Die statistischen Eigenschaften des Trainingsmaterials sollten den gesamten rele-

[25] In der Literatur sind unterschiedliche Anwendungen des *leave one out-Trainings* zu finden, wobei sowohl das Auslassen eines einzelnen Datensatzes als auch das Auslassen einiger weniger Datensätze – wie in dieser Arbeit – Verwendung findet.

vanten Datenbereich des späteren Einsatzmaterials in der Erkennungsphase abdecken und uniform verteilt sein, denn diese Zusammensetzung spielt für die spätere Erkennungsleistung eine wichtige Rolle. Durch „einseitiges“ Training mit einem speziellen Datentyp lernt das Netz nicht notwendigerweise, sich auf globale Merkmale zu konzentrieren. Stattdessen lernt es u. U. Spezifikationen dieses speziellen Datentyps darzustellen und liefert bei der Klassifikation anderer Daten dann keine guten Erkennungsergebnisse.[26] Durch mehrmalige Wiederholung mit wechselnder Zusammensetzung aus Trainings- und Klassifikationsdaten gewinnt man Informationen über die Repräsentativität der ausgewählten Trainingsmuster für den Gesamtdatenbereich.

[26] Für die Selektion der stimmhaften Anteile anhand von Bark-LPC-Spektren würde ein einseitiges Training z. B. eines mit ausschließlich Normalstimmen bedeuten. Ein so trainiertes Netz hätte möglicherweise Schwierigkeiten mit der Klassifikation pathologischer Stimmen.

4 Datenmaterial

In der Abteilung Phoniatrie und Pädaudiologie des Universitätsklinikums Göttingen werden im Rahmen der klinischen Diagnostik sowohl phonoskopische Methoden wie die Laryngoskopie, Laryngo-Stroboskopie oder Hochgeschwindigkeitsglottografie zur Beurteilung der Stimme und des Schwingungsverhaltens der Stimmlippen angewandt als auch akustische Stimmaufnahmen und deren akustische Analyse durchgeführt. Sowohl die optischen als auch die akustischen Methoden unterstützen den Phoniater bei seiner Diagnostik und ermöglichen eine objektive Beurteilung der Stimmfunktion und ggf. eines Therapieerfolges.

In enger Zusammenarbeit mit den Mitarbeitern dieser Abteilung sind – unter der Leitung von Prof. Dr. E. Kruse – seit Ende 1995 bereits über 70.000 akustische Aufnahmen von Sprechern (Patienten) unterschiedlichster Stimmgüte unter definierten und weitgehend identischen Aufnahmebedingungen durchgeführt, analysiert und archiviert worden. Diese Aufnahmen umfassen zusätzlich zum gesamten Spektrum an Stimmstörungen auch rund 8% an *Normalstimmen*. Als Normalstimmen werden dabei Aufnahmen von Sprechern[27] ohne bekannte oder erkennbare Stimmprobleme bezeichnet. Aufnahmen von Sprechern mit jeglicher Form und Ausprägung einer Stimmstörung werden als Gruppe der *pathologischen Stimmen* zusammengefasst.

Das Spektrum an Stimmstörungen reicht dabei von neurologischen Störungen über Veränderungen der Stimmlippen durch Knötchen oder Zysten sowie Teilresektionen und Lähmungen bis hin zu kompletter Aphonie, einer hochgradigen Stimmstörung, bei der keinerlei periodische glottale Schwingung während der Phonation zu beobachten ist.[28] Zur Klassifizierung der unterschiedlichen Stimmstörungen wurde ein 87-teiliger Diagnoseschlüssel mit 23 unterschiedlichen Diagnosezusätzen entwickelt (siehe Anhang 7.6, 7.7). Über diese Verschlüsselung ist auch eine Patientengruppenbildung nach unterschiedlichen Phonationsmechanismen möglich.

[27] „Sprecher“ bzw. „Patient“ wird synonym für weibliche und männliche Personen benutzt.
[28] Äußerungen aphoner Sprecher klingen wie Flüsterstimmen.

4.1 Aufnahme-System

In einem speziell für diese akustischen Aufnahmen vorgesehenen, schallisolierten und reflexionsarmen Raum der Abteilung Phoniatrie und Pädaudiologie werden die Sprachaufnahmen unter identischen Aufnahmebedingungen durchgeführt. Dieser Raum ist dazu von sämtlichen Störgeräuschquellen wie Computern, Lüftern oder Ähnlichem befreit.

Zur Aufnahme des Sprachsignals wird ein Kopfmikrofon des Typs BEYERDYNAMIC HEM 191.15 mit Kugelcharakteristik verwendet, dessen Signal über einen ALPHA-RECORDS Mikrofon-Vorverstärker MIC/DAT 2 direkt in eine Computer-Soundkarte geleitet wird. Die Verwendung eines Kopfmikrofons garantiert einen gleichbleibenden Abstand des Aufnahmemikrofons vom Mund des Sprechers. Dieser Abstand wird vor Beginn der Aufnahme vom eingewiesenen Aufnahmeleiter auf ca. 10 cm eingestellt und die Mikrofonkapsel selbst etwa auf Kinnhöhe unterhalb des Sprechluftstroms platziert, um Störeinflüsse durch Atemgeräusche zu vermeiden.

Bei der Soundkarte handelt sich um das Modell SOUNDBLASTER PCI 128 der Firma CREATIVE LABS. Diese Soundkarte ist in einen PC eingebaut, der aus Gründen der Störgeräuschvermeidung in einem benachbarten Raum platziert ist. Die analogen Signale des Mikrofonvorverstärkers werden von den AD-Wandlern auf der Soundkarte mit 48 kHz digitalisiert. Die 48 kHz werden aus Gründen der Konsistenz zu älteren DAT-basierten akustischen Stimmaufnahmen verwendet. Auf Basis des Betriebsystems Linux und einer speziell für diese Anwendungen in der Abteilung entwickelten C++-Softwarebibliothek (*leaf*) [MFLK01] wird das akustische Signal über eine grafische Benutzeroberfläche dem Aufnahmeleiter visualisiert. Die Steuerung der Aufnahme erfolgt am Computermonitor im Aufnahmeraum über die Benutzeroberfläche (Monitor, Tastatur und Maussignal werden zum entfernten Computer verlängert) und das akustische Signal wird abschließend auf Festplatte im WAV-Format gespeichert.

Im direkten Anschluss an die akustische Aufnahme besteht über diese Programmoberfläche die Möglichkeit des Nachbearbeitens des aufgenommen Sprachsignals, um bspw. Sprechpausen am Aufnahmeanfang oder -ende abzuschneiden, einleitende Instruktionen des Aufnahmeleiters zu entfernen oder um bestimmte Bereiche im Signal zu markieren.

Die nachbearbeiteten Stimmaufnahmen werden automatisch in einer für diesen Zweck programmierten, SQL-basierten MySQL-Datenbank auf einem mit dem

Aufnahme-PC vernetzten Server strukturiert archiviert. Zusammen mit diesen akustischen Aufnahmen werden weitere sprecherbezogenen Daten gespeichert. Neben den Stammdaten wie Name, Geburtsdatum und Geschlecht werden das Datum der akustischen Aufnahme und insbesondere die phoniatrischen Diagnosecodes aus der vorausgegangenen phoniatrischen Untersuchung mit dem Datensatz zusammen abgelegt. Zu jeder akustischen Aufnahme liegt demzufolge die phoniatrische Diagnose vom selben Tag und gegebenenfalls eine phonoskopische Videoaufnahme in der Datenbank vor.

Da von den Patienten prä- bzw. postoperativ, während des Therapieverlaufs und auch nach Therapieabschluss zu Kontrollzwecken Aufnahmen zu unterschiedlichen Zeitpunkten gemacht werden, können auch Therapieverläufe mit den Analyseergebnissen dokumentiert werden.

4.2 Aufnahmeprotokoll

Während jeder Aufnahmesitzung werden nach der Aufzeichnung von Spontansprache des Sprechers Stimmaufnahmen von gehaltenen, isoliert gesprochenen Vokalen in unterschiedlichen Tonlagen sowie von einem vorgelesenen Standardtext erstellt:

- Spontansprache (freies Erzählen des Sprechers),
- Vokalset [ɛ: a: e: i: o: u: ɛ:] in normaler Stimmlage,
- Vokalset [ɛ: a: e: i: o: u: ɛ:] in tiefer Stimmlage,
- Vokalset [ɛ: a: e: i: o: u: ɛ:] in hoher Stimmlage,
- Standardtext „Nordwind und Sonne“ (vgl. Anhang 7.1),
- Vokalset [ɛ: a: e: i: o: u: ɛ:] in normaler Stimmlage (belastet[29]).

Als Standardtext, der die Grundlage für das akustische Datenmaterial dieser Arbeit bildet, wird der phonetisch ausgewogene „Nordwind und Sonne“ -Text verwendet, der im Anhang 7.1 aufgeführt ist und von dem Übersetzungen in über 50 verschiedene Sprachen existieren [IPA49]. Die Aufnahmen gehaltener Phonation dienen in dieser Arbeit dem Vergleich der Ergebnisse der entwickelten akustischen Analysemethoden und der Validierung der erhaltenen Ergebnisse.

[29] „Belastet“ steht hier für die Belastung durch die vorangegangen Aufnahmen.

4.3 Fortlaufende Sprache

Grundlage der in Abschnitt 5 berechneten Ergebnisse ist eine Auswahl von 480 akustischen Stimmaufnahmen fortlaufender Sprache aus der Datenbank der Abteilung Phoniatrie und Pädaudiologie. Diese Auswahl umfasst 50 Normalstimmen (*Gruppe der Normalstimmen*) sowie eine für den gesamten Datenraum repräsentative Auswahl von 430 Aufnahmen jeglicher Stimmstörungen (*Gruppe der pathologischen Stimmen*). Alle 480 Aufnahmen stammen dabei von verschiedenen Sprechern, sodass es in den Ergebnissen zu keiner Ungleichgewichtung zu Gunsten eines einzelnen Sprechers kommen kann.

Bestimmte Untersuchungen werden zu Vergleichszwecken getrennt für die beiden Gruppen der stimmgesunden Normalstimmen und der stimmgestörten pathologischen Stimmen durchgeführt. Die Zusammensetzung (Vergabe von Mehrfachdiagnosen ist möglich) der unterschiedlichen Pathologien in der Auswahlgruppe und der gesamten Sprechergruppe ist der Tabelle 4.1 zu entnehmen.

Stimmstörung	**Anzahl Auswahl**	**Anzahl Gesamt**
Normalstimme	50	157
Dysphonien	99	338
Lähmungen	114	399
Neubildungen (gutartig)	83	330
Neubildungen (bösartig) und Vorstadien	59	315
Resektionen und Ersatzphonationen	113	558
Laryngitis	12	54
Mutationen und Mutationsstörungen	14	63
Aphonien	3	23
Sonstiges	29	122

Tabelle 4.1: *Zusammensetzung der Sprechergruppe.*

Der Altersbereich der Sprecher in der Gruppe der Normalstimmen bewegt sich zwischen 14 und 66 Jahren; 63% der Aufnahmen stammen dabei von Frauen und 37% von Männern, wobei in den weiteren Betrachtungen zwischen männlichen und weiblichen Sprechern nicht unterschieden wird. In der Gruppe der pathologischen Stimmen stammen 42% der Aufnahmen von Frauen und 58% von Männern, bei einem Altersbereich zwischen 9 und 88 Jahren.

4.4 PHONDAT-Sprachaufnahmen

Da es sich bei den 480 Aufnahmen in der Sprachdatenbank um ungelabeltes Datenmaterial handelt, kann dieses nicht direkt für ein Training des Neuronalen Netzes eingesetzt werden, für das ein stimmhaft/stimmlos-Klassifikationssollausgabesignal benötigt wird. Aus diesem Grund ist für ein grundlegendes Training des Neuronalen Netzes die in der Literatur bekannte PHONDAT I-Sprachdatenbank [P93] aus Normalsprechern eingesetzt worden, für die eine Phonemklassifikation vorliegt. Für das Training und die Beurteilung der Klassifikationsleistung des Neuronalen Netzes sind 32 Sprachaufnahmen aus dem Korpus der PHONDAT I-Sprachdatenbank verwendet worden (siehe Anhang 7.5).[30] Es handelt sich dabei um nach einem leicht modifizierten SAMPA-Phonemlexikon (siehe Anhang 7.4) gelabelte Aufnahmen fortlaufender Sprache. 17 der 32 Aufnahmen stammen von Frauen und 15 von Männern. Von den Sprechern dieser 32 Aufnahmen sind keine Stimmstörungen in der Datenbank dokumentiert und auch Hörproben der Aufnahmen deuten auf eine Zuordnung in die Gruppe der Normalstimmen hin.

Der gesprochene Text entspricht in 16 Aufnahmen dem auch in der Abteilung Phoniatrie und Pädaudiologie in Göttingen verwendeten „Nordwind und Sonne“ -Text (siehe Anhang 7.1). Die anderen 16 Aufnahmen behandeln den Text der „Buttergeschichte“ (siehe Anhang 7.2).

Die Aufnahmen liegen als Dateien im WAV-Format auf CD vor und sind mit 16 kHz Abtastfrequenz und 16 Bit quantisiert. In einer separaten Textdatei sind die Phonemgrenzen zu jeder Aufnahme als Zeitmarken gespeichert. Unter Verwendung der *leaf*-Softwarebibliothek ist es möglich, aus dem akustischen Signal und den Phonemgrenzmarken eine Klassifikation der einzelnen Phoneme vorzunehmen, um das Datenmaterial für das Training der Neuronalen Netze mit Sollausgabesignal generieren zu können [MFLK01].

[30] Der gesamte PHONDAT-Sprachdatenkorpus ist so konzipiert, dass er alle im Deutschen möglichen 1308 Phonemverbindungen umfasst.

5 Ergebnisse

Der Ergebnisteil dieser Arbeit gliedert sich in drei wesentliche Bereiche. Im ersten Teil (5.1) werden die Ergebnisse der Entwicklung eines zuverlässigen Klassifikationssystems – auf Basis einzelner Kurzzeit-Signalabschnitte des akustischen Signals – in Bereiche stimmhafter und stimmloser Phonation sowie Sprechpausen dargestellt. Auf dieser Klassifikation aufbauend werden im zweiten Abschnitt (5.2) die Resultate der akustischen Analyse fortlaufender Sprachsignale nach definierten Stimmgütemaßen beschrieben und eine Erweiterung des Göttinger Heiserkeits-Diagramms für fortlaufende Sprache eingeführt. Der dritte Abschnitt (5.3) beschreibt eine Anwendung zur automatisierten Vokalanalyse in Anlehnung an die Klassifikationsverarbeitung zur stimmhaft/-los-Klassifikation.

Die Beschreibung akustischer Stimmgütemaße und deren Bestimmungsmethoden in Kapitel 2 erfordert eine Klassifikation des fortlaufenden Sprachsignals in Bereiche stimmhafter und stimmloser Phonation sowie Sprechpausen. Für diese komplexe Klassifikation bieten sich Neuronale Netze, in diesem Fall Multi-Layer-Perzeptrons, an. Die Parametrisierung der zu klassifizierenden Kurzzeitsegmente erfolgt anhand von aus dem Sprachsignal bestimmten Vokaltraktparametern, die weitestgehend unabhängig von der Qualität der Stimmlippenschwingung während der Sprachproduktion sind, wie in Kapitel 3 gezeigt worden ist. Diese Unabhängigkeit gewährleistet eine Klassifizierung sowohl von Normalstimmen als auch von Sprechern mit gestörter Stimmfunktion bis zur Aphonie.

Bisher in der Literatur publizierte Verfahren und auch das populäre kommerzielle Stimmanalysesystem CSL (*Computerized Speech Lab*) der Firma KAY ELEMETRICS [K92] versagen bei ausgeprägten Stimmstörungen, da eine Klassifikation von stimmhaften und stimmlosen Teilsegmenten zumeist auf Periodizitätskriterien oder einer Bewertung bestimmter Frequenzbänder beruht, die bei hochgradig gestörten Stimmen nicht notwendigerweise zielführend ist. Die hier vorgestellte Methode liefert sowohl für Normalstimmen als auch für jegliche Stimmstörungen zuverlässige Klassifikationsresultate und stellt somit eine deutliche Erweiterung des Analysespektrums gestörter Stimmfunktion fortlaufender Sprache dar.

5.1 Klassifikation fortlaufender Sprache

Ausgehend vom digitalisiert vorliegenden, mit einer Abtastfrequenz von $f_s = 48\,\text{kHz}$ bei 16 Bit linearer Amplitudenauflösung diskretisierten Sprachsignal, folgen die einzelnen Teilschritte der Klassifikation – entsprechend dem in Kapitel 3.3 motivierten Ansatz – dem nachfolgend skizzierten Ablauf.

1. Unterabtastung des Sprachsignals,
2. Fensterung des Signals,
3. Bestimmung der Signalenergie zum Ausschluss von Sprechpausen,
4. Bestimmung der Prädiktorkoeffizienten der Linearen Prädiktion bei Verwendung der Autokorrelationsmethode,
5. Berechnung der LPC-Spektren aus den Prädiktorkoeffizienten,
6. Transformation der LPC-Spektren in 19-kanalige Barkspektren,
7. Dynamikkompression der Barkspektren,
8. Amplitudennormierung der Barkspektren,
9. Abbildung der Spektren mit einem Neuronalen Netz,
10. Klassifikation über Schwellwertvergleich der Netzausgabe.

Dieser Ablauf, der zur Veranschaulichung mit seinen Teilschritten in Abbildung 5.1 visualisiert ist, lässt sich in drei wesentliche Bereiche gliedern:

a) die *Vorverarbeitung* (Schritte 1–3),

b) die *spektrale Transformation* (Schritte 4–8),

c) die eigentliche *Klassifikation* (Schritte 9 und 10).

Diese einzelnen Teilschritte werden im Kontext des Gesamtablaufs im folgenden detaillierter erläutert.

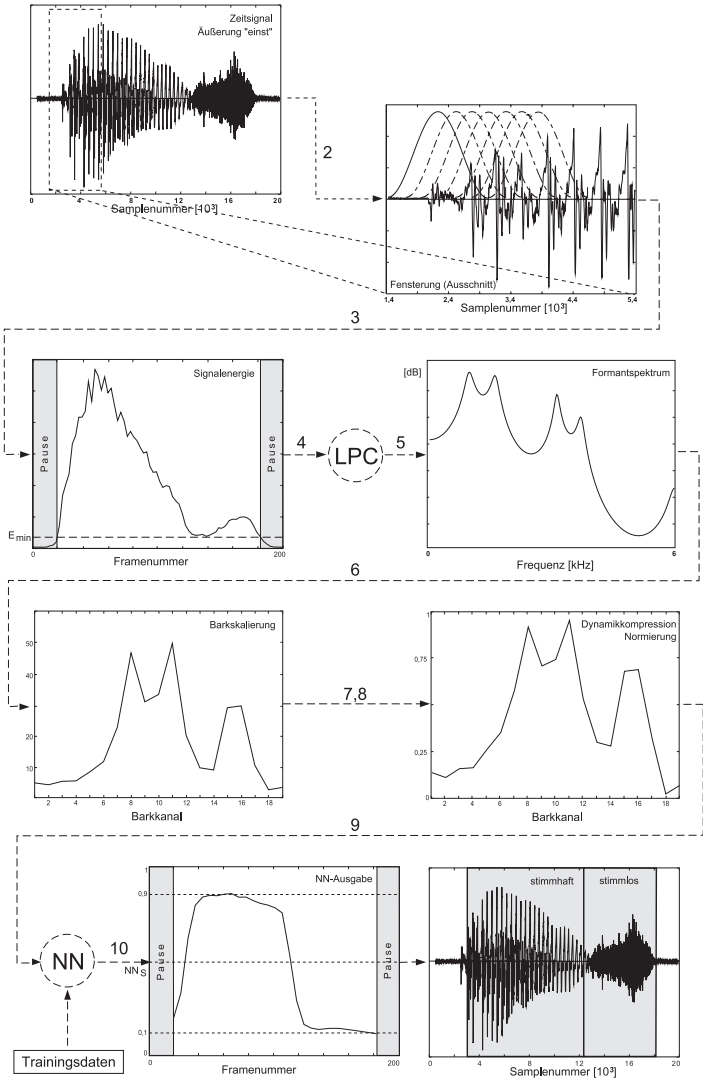


Abb. 5.1: *Ablauf der Klassifikation der fortlaufenden Sprache in Bereiche stimmhafter und stimmloser Phonation.*

5.1.1 Vorverarbeitung

Die Vorverarbeitung umfasst die Pausendetektion und notwendige Schritte der digitalen Signalverarbeitung von Sprachsignalen für den weiteren Ablauf.

Unterabtastung des Sprachsignals

Eine Unterabtastung des Sprachsignals ist für die Lineare Prädiktion sinnvoll, da die für die Klassifikation relevante Information der spektralen Einhüllenden insbesondere durch die Lage und Intensität der niedrigen Formanten des Vokaltraktes bestimmt wird. Die Abtastfrequenz des Sprachsignals wird deshalb unter Verwendung der *sinc*-Funktion[31] auf 12 kHz verringert.

Fensterung des Signals

Da die fortlaufende Sprache im Gegensatz zur gehaltenen Phonation durch eine Abfolge sich auf zumeist kurzen Zeitskalen ändernder Artikulatorstellungen – und dadurch variierender Resonanzeigenschaften – gekennzeichnet ist, ist eine Einteilung des Sprachsignals in kürzere, einzeln zu analysierende Segmente quasistationärer Phonation notwendig. Dies erfolgt durch Fensterung mit einem Hann-Fenster[32] von 40 ms Länge bei einem überlappenden Vorschub von 10 ms.

Die Form dieser Fensterfunktion gewährleistet eine scharfe Hauptkeule im Frequenzspektrum bei gleichzeitiger Unterdrückung der Nebenkeulen – im Gegensatz zu einer einfachen Rechteckfensterung mit zwar schärferer Hauptkeule, aber starken Nebenkeulen (*Leakage-Effekt*). Auch ähnliche Fensterfunktionen, wie das Hamming- oder Kaiserfenster können verwendet werden.

Energiebestimmung

Die Berechnung der Signalenergie jedes dieser Kurzzeitsegmente dient der Sprechpausenexklusion. Durch Vergleich des lokalen RMS-Wertes der Energie mit ei-

[31] $\mathrm{sinc}(x) = \sin(x)/x$ *Sinus cardinalis*. Die Fouriertransformierte der sinc-Funktion ist bis auf einen Faktor die Rechteckfunktion.

[32] Im englischsprachigen Raum ist in Anlehnung an das verwandte Hammingfenster auch die Bezeichnung *Hanningfenster* in der Literatur verbreitet. Die Namensgebung stammt von R. B. Blackmann und J. Tukey und ist nach Julius von Hann benannt. $w(n) = \frac{1}{2}\left[1 + \cos(\frac{2\pi n}{M})\right]$ mit $n = -\frac{M}{2}, \ldots, \frac{M}{2}$ und M der Fensterbreite

nem Prozentsatz eines globalen Energiewertes, der als Mittelwert der drei größten RMS-Werte aus dem gesamten fortlaufenden Sprachsignal $x(n)$ bestimmt wird, ist über ein Schwellwertkriterium eine Bestimmung der Nicht-Sprechpausensegmente möglich. Dieser Schwellwert liegt bei 3% und wird individuell für jede Sprachaufnahme berechnet (vgl. Abschnitt 3.1).

$$\mathrm{RMS}_i = \sqrt{\frac{E_i}{M}} = \sqrt{\frac{\sum_{m=0}^{M-1} x(n+m)^2}{M}} \qquad \begin{cases} \leq \mathrm{RMS}_{\min} & : \text{ Pause} \\ > \mathrm{RMS}_{\min} & : \text{ Analyse} \end{cases} \tag{5.1}$$

$$\text{mit} \quad \mathrm{RMS}_{\min} = 0{,}03 \cdot \frac{1}{3} \sum_{k=1}^{3} \mathrm{RMS}_{\max_k}$$

$$\text{und} \quad n = i \cdot M/4 \quad \text{bei} \quad i = 0{,}1{,}2, \ldots, 4(N-M)/M.$$

Lediglich die Bereiche der „Nicht-Pausen“ (siehe grau unterlegte Bereiche in Abb. 5.2) werden der weiteren akustischen Analyse zugeführt.

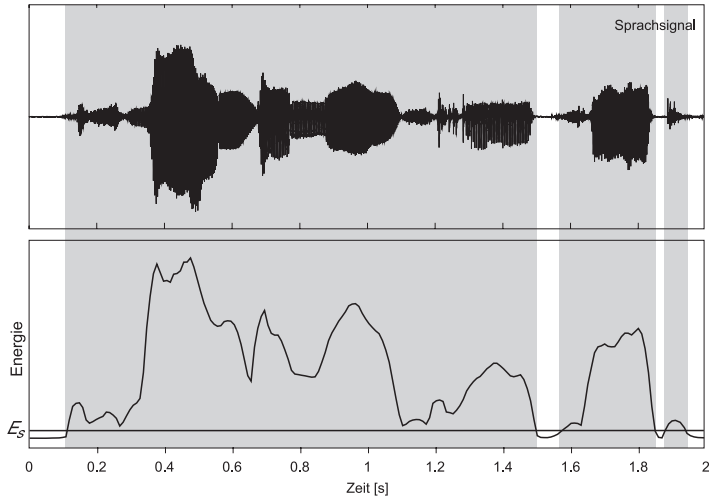


Abb. 5.2: *Sprachsignal* „Es war in Berlin zu einer Zeit“ *(oben), Verlauf der Energie der 40 ms Kurzzeitfenster (unten). Grau markiert sind die Signalbereiche oberhalb des Energieschwellwertes E_s, die der weiteren Analyse zugeführt werden.*

5.1.2 Spektrale Transformation

Die weitere Verarbeitung und Klassifikation der Kurzzeitsegmente erfolgt im Spektralbereich, da die Parametrisierung über Vokaltraktparameter erfolgen soll, die sich durch die Einhüllende des Kurzzeitsignalsspektrums beschreiben lassen. Für jedes dieser sich überlappenden 40 ms Kurzzeitsegmente werden die Prädiktorkoeffizienten der Linearen Prädiktion bestimmt, aus denen sich das Kurzzeitspektrum berechnen lässt.

Allen Untersuchungen zur optimalen Parameterwahl der Linearen Prädiktion und auch der späteren Klassifikationsanwendung des trainierten Neuronalen Netzes ist eine Höhenanhebung des Sprachsignals entsprechend Gleichung 5.2 mit einem Präemphasefaktor von $\mu = 0{,}9375$ vorausgegangen.

$$H(z) = 1 - \mu \cdot z^{-1} \tag{5.2}$$

Diese Höhenanhebung liefert einen besseren Prädiktor und führt somit zu einer Steigerung der Klassifikationsleistung.

LPC Prädiktorordnung

Auf der Grundlage einer modellbasierten Abschätzung der Prädiktorordnung (siehe Abschnitt 2.3.1) ist deren Einfluss auf die Klassifikationsleistung von jeweils drei trainierten Multi-Layer-Perzeptrons (NN1, NN2, NN3) gleicher Topologie bei Verwendung von 8, 10, 12, 14, 16, 18 und 20 Prädiktorkoeffizienten für die Berechnung der LPC-Spektren untersucht worden. Die Prädiktorordnungen 12, 14 und 16 entsprechen der von Markel [M71] angegebenen sinnvollen Größenordnung bei vorliegender Abtastfrequenz von $f_s = 12\,\text{kHz}$. Die übrigen Prädiktorordnungen sind hinzugenommen worden, um den Einfluss auf die Klassifikationsleistung umfassender beurteilen zu können. Auch wenn eine Prädiktorordnung von 8 nicht ausreichend für eine Parametrisierung der Einhüllenden des Kurzzeitspektrums der LPC-Analyse scheint, gilt trotzdem zu prüfen, ob sie u. U. zu besseren Klassifikationsleistungen führt. Des Weiteren sind eventuelle Klassifikationsunterschiede bei Verwendung der Autokorrelations- (acf) bzw. Kovarianzmethode (cov) zur Bestimmung des Koeffizientensatzes aus Gl. 2.22 untersucht worden, deren Ergebnisse der Tabelle 5.1 zu entnehmen sind.

Präd-Ord K	Methode	Fehlklass. NN 1	Fehlklass. NN 2	Fehlklass. NN 3	Mittelwert (SD)
8	acf	4,63%	4,71%	4,69%	4,67% (0,04)
10	acf	4,62%	4,59%	4,56%	4,59% (0,03)
12	acf	4,55%	4,56%	4,59%	4,57% (0,02)
14	acf	4,64%	4,58%	4,63%	4,62% (0,03)
16	acf	4,62%	4,62%	4,64%	4,62% (0,01)
18	acf	4,64%	4,68%	4,71%	4,68% (0,03)
20	acf	4,66%	4,65%	4,58%	4,63% (0,04)
8	cov	5,60%	5,63%	5,60%	5,61% (0,02)
10	cov	5,61%	5,63%	5,58%	5,61% (0,03)
12	cov	5,42%	5,47%	5,52%	5,47% (0,05)
14	cov	5,57%	5,64%	5,73%	5,64% (0,08)
16	cov	5,51%	5,47%	5,39%	5,45% (0,06)
18	cov	5,48%	5,45%	5,48%	5,47% (0,02)
20	cov	5,52%	5,46%	5,62%	5,53% (0,08)

Tabelle 5.1: *Vergleich der Fehlklassifikationsrate von drei Multi-Layer-Perceptrons (NN 1-3) bei Variation der Prädiktorordnung K (8, 10, 12, 14, 16, 18, 20) und LPC-Methode (Autokorrelations- (acf), Kovarianz- (cov)) bei einem Klassifikationsschwellwert von $N_s = 0{,}5$.*

Die niedrigste Fehlklassifikationsrate ist bei einer Prädiktorordnung von K=12 bei Verwendung der Autokorrelationsmethode (acf) und K=16 bei der Kovarianzmethode (cov) zu finden. Generell scheint die Autokorrelationsmethode bessere Klassifikationsergebnisse als die Kovarianzmethode zu liefern. Für die Berechnung der LPC-Spektren wird in dieser Arbeit die Autokorrelationsmethode bei einer Prädiktorordnung von 12 verwendet, da diese Kombination zur geringsten Fehlklassifikationsrate geführt hat, auch wenn eine Prädiktorordnung von K=16 nach Markel [M71] eher den Vorgaben von:

$$K \quad \approx \quad \textit{Abtastfrequenz in kHz zzgl. 4 oder 5 weiterer Koeffizienten}$$

entsprechen würde. Da die LPC-Analyse an dieser Stelle allerdings als Vorstufe einer Klassifizierung verstanden werden muss, steht allein die bestmögliche Klassifikationsleistung als Entscheidungskriterium im Vordergrund.

Barkskalierung, Dynamikkompression und Normierung

Die auf diese Weise berechneten LPC-Spektren werden in einem weiteren Schritt nach der Barkskala transformiert. Der Frequenzbereich des unterabgetasteten Signals überdeckt 19 Barkkanäle, deren Intensität durch sich überlappende Trapezfenster auf der Frequenzachse bestimmt wird. Eine anschließende Dynamikkompression der Barkspektren – in Anlehnung an Zwickers Modell für die Verarbeitung im menschlichen Ohr – gewährleistet eine weitere Steigerung der Klassifikationsleistung. Abschließend ist eine 2-dimensionale Normierung dieser zeitlichen Abfolge von Barkspektren notwendig, um einen Einfluss der Signalenergie auf die Klassifikationsleistung auszuschließen.

Für diese Normierung wird die Amplitude jedes Barkkanals aller Barkspektren mit der maximalen Amplitude aller Barkkanäle und -spektren normiert, vgl. Abschnitt 3.3.4.

$$N^*_{tz} = \frac{N_{tz}}{\max\limits_{\substack{1<\tau<T\\1\le\zeta\le Z}}(\{N_{\tau\zeta}\})} \qquad \text{für} \quad 0 \le t < T \quad \text{und} \quad 1 \le z \le Z$$

Diese Folge von 19-kanaligen Barkspektren – alle 10 ms – stellt das Eingangsdatenmaterial des Neuronalen Netzes für den eigentlichen Klassifikationsschritt dar.

Eine deutliche Ähnlichkeit der Barkspektrogramme unterschiedlicher Stimmfunktion ist in Abb. 5.3 b) zu erkennen. Insbesondere im Bereich der stimmhaften Phoneme der Äußerung *„einst“* in den ersten zwei Dritteln des Signalausschnitts ist ein ähnlicher Verlauf der Formanten während der Phonation zu erkennen. Im stimmlosen letzten Drittel des Signals ist bei der gestörten Stimmfunktion im Barkspektrogramm ein größerer Anteil hoher Frequenzen zu beobachten als bei der Normalstimme. Die Energieverteilung der Barkkanäle der einzelnen Barkspektren weist doch auch hier viel versprechende Ähnlichkeiten für eine Musterklassifikation auf (exemplarisch siehe Abb. 5.3 c) und d)).

Auf Grund der differierenden Anregungen des Vokaltrakts auf glottaler Ebene – entsprechend der Stimmfunktion – sind natürlich Unterschiede in den Barkspektrogrammen zu beobachten. Die Ähnlichkeit und damit die Aussagekraft der gewählten Parametrisierung zeigt allerdings eine gute Basis für eine erfolgreiche Klassifikation mit Neuronalen Netzen auf.

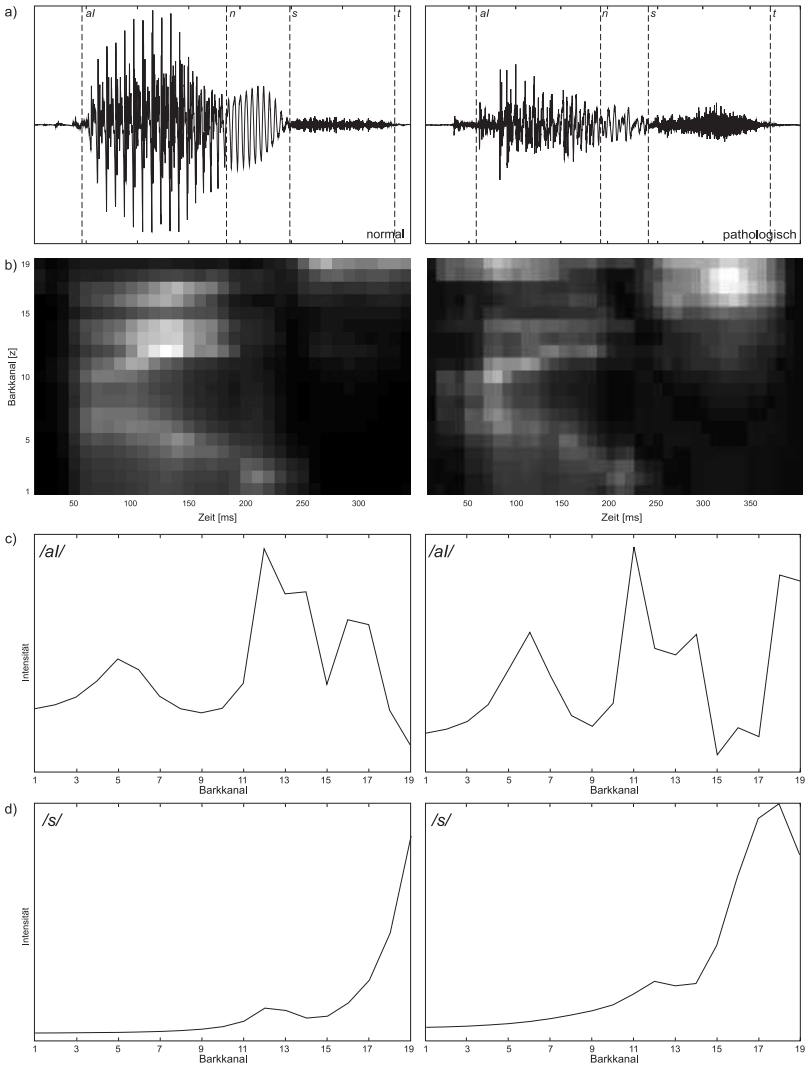


Abb. 5.3: *a) Äußerung „*einst*“ eines Sprechers mit gesunder (links) und gestörter (rechts) Stimmfunktion, b) zugehörige Folge von normierten Barkspektren alle 10 ms als Eingangsdaten des Neuronalen Netzes und einzelne Barkspektren aus c) stimmhaften und d) stimmlosen Signalabschnitten.*

5.1.3 Klassifikation mit Neuronalem Netz

Sowohl sämtliche Schritte der Vorverarbeitung und spektralen Transformation des Eingangsdatenmaterials als auch die Topologie und Trainingsparameter des Neuronalen Netzes sind in dieser Arbeit auf eine optimale Erkennung während der Trainingsphase angepasst worden, um abschließend ein Neuronales Netz mit hoher Generalisierungsfähigkeit zu erhalten. Die Bestimmung einer geeigneten Topologie für die vorliegende Klassifizierungsaufgabe spielt eine wesentliche Rolle für dessen Klassifikationsleistung.

5.1.4 Topologie des Neuronalen Netzes

Für die gesuchte Klassifikation der Kurzzeit-Sprachsegmente in Bereiche stimmhafter und stimmloser Phonation bietet sich die Verwendung mehrschichtiger Neuronaler Netze (*Multi-Layer-Perzeptron, MLP)* an, deren Aufbau in Abschnitt 3.4.4 beschrieben ist. Mit dem modifizierten Error-Backpropagation-Algorithmus (siehe Abschnitt 3.4.7) und der Einbeziehung von Momentum-Termen steht ein Lernverfahren zur Verfügung, das gute Konvergenzeigenschaften besitzt und in akzeptabler Rechenzeit in ein Minimum konvergiert. Da sich diese Kombination für Klassifikationsaufgaben der vorliegenden Art als geeignete Wahl erwiesen hat, sind auch keine weiteren Typen Neuronaler Netze untersucht worden.

Eine weiterer Vorteil dieser Architektur besteht in der einfachen Skalierbarkeit der Topologie. Die Dimensionalität der Eingangs- und Ausgangsschicht des neuronalen Klassifikators ist vorgegeben durch die Dimensionalität des zu analysierenden Datenmaterials (19 Eingangszellen, je eine pro Barkkanal) und die gewünschte Ausgabedimension (1 Ausgabezelle als bimodaler Entscheider). Voruntersuchungen haben gezeigt, das ein einschichtiges Perzeptron, also ein Neuronales Netz ohne versteckte Neurone in Zwischenschichten, für die gesuchte Klassifikation nicht ausreichend ist. Die Verwendung mehrschichtiger Netze liefert dagegen gute Klassifikationsergebnisse, wie in den folgenden Abschnitten zu sehen ist.

In der Regel lässt sich die optimale Topologie des Neuronalen Netzes für eine Klassifikationsaufgabe nicht analytisch bestimmen. Die Faustregel *„so klein wie möglich, so groß wie nötig“* liefert allerdings eine gute Basis. Je nach Klassifikationsproblem kann eine bestimmte Mindestanzahl versteckter Neuronen notwendig sein, um überhaupt eine Generalisierungsfähigkeit während des Trainings zu erreichen. Eine Steigerung der Anzahl versteckter Zellen (*hiddim* von *hidden dimension*) sollte

hiddim	Fehlklass NN 1	Fehlklass NN 2	Fehlklass NN 3	Mittelwert (SD)
4	4,84% (± 0,35)	4,83% (± 0,33)	4,87% (± 0,36)	4,85% (± 0,02)
6	4,83% (± 0,32)	4,89% (± 0,29)	4,81% (± 0,32)	4,84% (± 0,03)
8	4,82% (± 0,35)	4,80% (± 0,38)	4,81% (± 0,30)	4,81% (± 0,03)
10	4,80% (± 0,32)	4,84% (± 0,32)	4,80% (± 0,31)	4,81% (± 0,02)
12	4,76% (± 0,25)	4,83% (± 0,38)	4,82% (± 0,34)	4,80% (± 0,03)
20	4,84% (± 0,35)	4,83% (± 0,35)	4,80% (± 0,33)	4,82% (± 0,03)

Tabelle 5.2: *Über 3 Paare mit je 2 Testsprechern gemittelte Fehlklassifikationsrate und Standardabweichung jeweils dreier trainierter Neuronaler Netze (NN 1, NN 2, NN 3) bei Variation der Neuronenzahl (hiddim = hidden dimension) in der versteckten Zwischenschicht und einem Klassifikationsschwellwert von* $N_s = 0{,}5$.

die Klassifikationsleistung prinzipiell verbessern, da unter der Annahme von Null-Elementen in der Gewichtsmatrix für die zusätzlichen Zellen keine generelle Verschlechterung durch Hinzufügen weiterer Zellen eintreten kann. Die Gefahr bei zu großen Netzen besteht allerdings darin, keine Generalisierung zu erlernen. Dies kann daraus resultieren, das das Neuronale Netz die individuellen Trainingsdatensätze auswendig lernt, was bei Nichttrainingsdaten zu schlechten Klassifikationsergebnissen führen kann. Des Weiteren besteht bei zu großen Netzen die Gefahr der Überbewertung von Individualitäten einzelner Datensätze – dem Overfitting – mit dem Resultat, ebenfalls keine Generalisierung des Neuronalen Netzes zu erlernen.

Für die Wahl der Anzahl der versteckten Zellen in der Zwischenschicht (hiddim) sind Untersuchungen durchgeführt worden, deren Ergebnisse der Tabelle 5.2 zu entnehmen sind. Die Verwendung mehrerer versteckter Schichten ist in diesem Zusammenhang nicht näher untersucht worden, da bei detaillierter Betrachtung der Netzausgabe für eine versteckte Zwischenschicht das Auftreten einzelner Fehlklassifikationen eher auf die im folgenden Abschnitt aufgezeigten Fehlerquellen zurückzuführen ist und nicht auf ein generelles Abbildungsproblem, für das die gewählte Netztopologie unzureichend sein könnte. Exemplarisch ist zusätzlich zu den in Tabelle 5.2 dargestellten Ergebnissen eine Testreihe mit 40 Zellen in der versteckten Schicht durchgeführt worden, die das Modell des Overfitting durch sich verschlechternde Klassifikationsleistungen bestätigt hat.

Anhand der berechneten Fehlklassifikationsraten aus Tabelle 5.2 – die alle in einem ähnlichen Bereich liegen – liefert ein Neuronales Netz mit 12 Neuronen in der versteckten Zwischenschicht sowohl die absolut niedrigste Fehlklassifikationsrate (NN1=4,76%, hiddim=12), als auch den niedrigsten Mittelwert der drei mit jeweils identischen Trainingsparametern trainierten Neuronalen Netze NN1-3. Für die Klassifikation wird folglich ein vollständig verbundenes, vorwärtsgerichtetes Multi-Layer-Perzeptron mit 19 Eingangszellen, 12 versteckten Zellen in einer Zwischenschicht und einer Ausgangszelle trainiert. Diese resultierende Topologie des Neuronalen Netzes ist in Abbildung 5.4 illustriert.

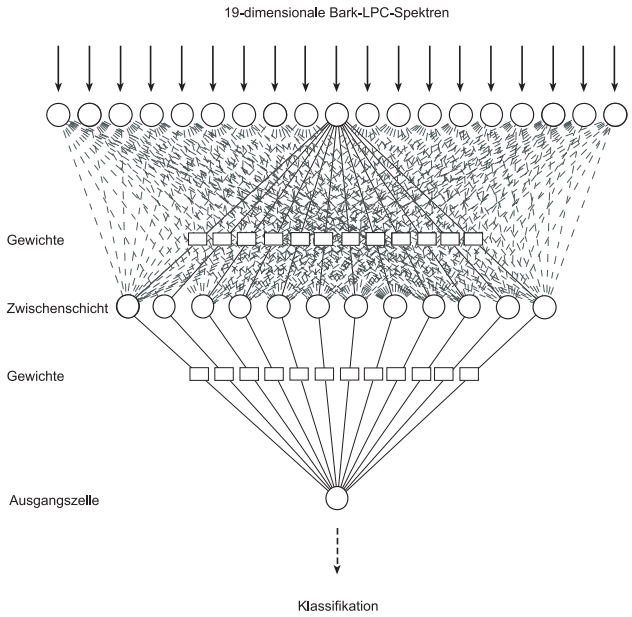


Abb. 5.4: *Optimale Topologie des Neuronalen Netzes zur Klassifikation stimmhafter/stimmloser 40 ms Kurzzeit-Sprachsegmente anhand normierter, dynamikkomprimierter, barkskalierter LPC-Spektren.*

Nachdem die Schritte der Vorverarbeitung und spektralen Transformation sowie die grundlegende Topologie des Neuronalen Netzes auf eine bestmögliche Klassifikation angepasst worden sind, muss solch ein generiertes NN trainiert werden.

5.1.5 Training des Neuronalen Netzes

Grundlage einer erfolgreichen Klassifikation ist das vorherige Training des NN, das mit einer Auswahl vorgegebener Trainingsdatensätze durchgeführt wird. Es handelt sich in diesem Fall in einer ersten grundlegenden Trainingsphase um akustische Aufnahmen fortlaufender Sprache des „Nordwind und Sonne"-Textes bzw. der „Buttergeschichte" von 32 Sprechern (siehe Anhang 7.5) mit normaler Stimmfunktion aus der PHONDAT-Sprachdatenbasis (16x „Nordwind und Sonne" (siehe Anhang 7.1) und 16x „Buttergeschichte" (siehe Anhang 7.2)). Für diese Sprachaufnahmen ist unter Verwendung der selbst entwickelten Softwarebibliothek `leaf` [MFLK01] das entsprechende Sollausgabesignal des Neuronalen Netzes auf der Basis von Phonemlabelmarken und deren Einordnung in stimmhafte bzw. stimmlose Phonation entsprechend Anhang 7.4 berechnet worden. In einer nachfolgenden Verfeinerungsphase soll das bis dahin auf die grundlegende stimmhaft/stimmlos-Klassifikation von Normalstimmen trainierte Neuronale Netz durch ein weiteres Training mit Sprachmaterial von Sprechern gestörter Stimmfunktion nachtrainiert werden. Da keine phonemgelabelte Sprachdatenbank gestörter Stimmfunktion zur Verfügung steht, ist entsprechendes Datenmaterial aus der umfangreichen Datenbank der Abteilung Phoniatrie und Pädaudiologie in Handarbeit erstellt und mit Sequenzen gehaltener Vokale ergänzt worden.

Als Trainingsdatenmaterial standen 154.550 Barkspektren von 40 ms langen, sprechpausenbereinigten Kurzzeitsegmenten des akustischen Signals aus den 32 Sprachaufnahmen normaler Stimmfunktion zur Verfügung. Das Training ist in verschiedenen Kombinationen von 30 dieser 32 Sprecher in jeweils maximal 5.000 Trainingsiterationen mit den phonemgelabelten Barkspektrogrammen und dem zugehörigen Sollausgabesignal der Netzausgangszelle durchgeführt worden, um den Einfluss der Zusammensetzung des Trainingsmaterials weitestgehend zu minimieren und ein hohes Maß an sprecherübergreifender Generalisierung zu erreichen. Wichtig ist dabei die Beurteilung der Klassifikationsleistung mit bekannten Datensätzen, die nicht zum Trainingsmaterial gehören (*leave one out*-Training).

Die Gewichtsmatrix wird zu Beginn des Trainings mit zufällig gewählten Werten aus einem gleichverteilten Wertebereich von $[-1; 1]$ initialisiert. Um das Verharren in einem lokalen Fehlerminimum auf Grund einer ungünstigen Wahl der zufälligen Startkonfiguration auszuschließen, sind jeweils 3 Neuronale Netze mit identischen Trainingsparametern trainiert worden, deren Mittelwert der Erkennungsrate zur Beurteilung herangezogen wird. Die Anpassung der Gewichtsmatrix während des Trainings erfolgt unter Verwendung des modifizierten Error-Backpropagation-

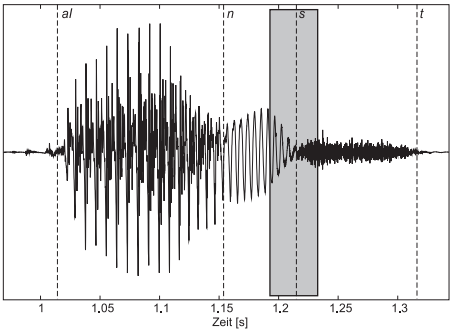


Abb. 5.5: *Grau unterlegt ist ein 40 ms Kurzzeit-Signalfenster am Phonemübergang /n/ - /s/. Eine minimale Verschiebung dieses Fensters kann bereits den 50 % Anteil stimmhaft ($s_o = 0{,}9$) zu stimmlos ($s_o = 0{,}1$) verschieben, ohne dass große Änderungen im Barkspektrum zu erwarten wären.*

Algorithmus aus Abschnitt 3.4.7 nach der *Train-by-Pattern* Methode, bei der nach jedem dargebotenen Musterpaar eine Adaption stattfindet. In Gegensatz dazu werden beim *Train-by-Epoch* Verfahren – das auch getestet wurde – die berechneten Gewichtsänderungen für alle Trainingsdatensätze summiert und am Ende einer Iteration zur Änderung der Gewichtsmatrix nach Gl. 5.3 und 5.4 herangezogen.

$$\Delta\omega_{ij}(s) = -\eta \sum_{p=1}^{N_p} \left(\frac{\partial E}{\partial \omega_{ij}} \right) + \alpha \Delta\omega_{ij}(s-1) \tag{5.3}$$

$$\Delta\vartheta_o(s) = -\eta \sum_{p=1}^{N_p} \left(\frac{\partial E}{\partial \vartheta_o} \right) + \alpha \Delta\vartheta_o(s-1) \tag{5.4}$$

Für den Wert des Trägheitsmoments α zur Stabilisierung des Trainings hat sich in Voruntersuchungen $\alpha = 0{,}8$ als sinnvolle Wahl abgezeichnet. Als Sollausgabe der sigmoiden Aktivierungsfunktion der Netzausgangszelle werden die Werte $s_o = 0{,}1$ für Kurzzeitsegmente stimmloser und $s_o = 0{,}9$ für Segmente stimmhafter Phonation verwendet. Diese Werte erlauben im Gegensatz zu 0 und 1 – den Grenzwerten der verwendeten sigmoiden Funktion – eine bessere Adaptionsmöglichkeit, da diese Extremwerte erst im Grenzwert $\pm\infty$ der Funktion erreicht werden.

An Phonemübergängen oder bei wieder einsetzender Phonation nach einer Sprechpause ist für jedes Analysefenster bei der Bestimmung des Sollausgabesignals zu entscheiden, ab welchem prozentualen Signalanteil eine Klassifikationsnetzausgabe entsprechend der Stimmhaftigkeit des Phonems erfolgen soll. In dieser Arbeit erhält ein 40 ms Signalfenster die Sollklassifikation, die der Stimmhaftigkeit von mindestens 50% der in diesem Segment enthaltenen Phoneme entspricht. Sobald folglich mindestens 20 ms eines Analysefensters ein als stimmhaft klassifiziertes Phonem überdecken, wird der gesamte Signalabschnitt auf die NN-Sollausgabe *stimmhaft* trainiert. Entsprechendes gilt bei stimmlosen Phonemen für eine Training auf *stimmlos*. Die Notwendigkeit, auch diese Phonemübergangsbereiche mit einer eindeutigen Klassifikation stimmhaft/stimmlos zu versehen, stellt im Extremfall natürlich ein Potential für anschließende Fehlklassifikationen dar, da eigentlich keine eindeutige Klassifikation vorgenommen werden kann. Die den Trainingsdatensätzen zugrunde liegenden Barkspektrogramme werden keine gravierenden Unterschiede aufweisen, ob ein Kurzzeitsegment eines Phonemübergangs ein stimmhaftes Phonem zu 49 % (Klassifikation: stimmlos) oder zu 51 % (Klassifikation: stimmhaft) überdeckt. Ein solcher Grenzfall ist in Abb. 5.5 illustriert.

In Abbildung 5.6 ist exemplarisch der absolute Netzausgabefehler im Verlauf der 5.000 Trainingsiterationen aufgetragen. Die Höhe des absoluten Werts des Netzausgabefehlers spielt allerdings keine entscheidende Rolle, sondern lediglich dessen Minimierung während des Trainings. Die Klassifikationsgüte und damit die Generalisierungsfähigkeit wird vielmehr im Re-Test mit Nichttrainingsmaterial bestimmt und als Entscheidungskriterium zur Netzauswahl herangezogen.

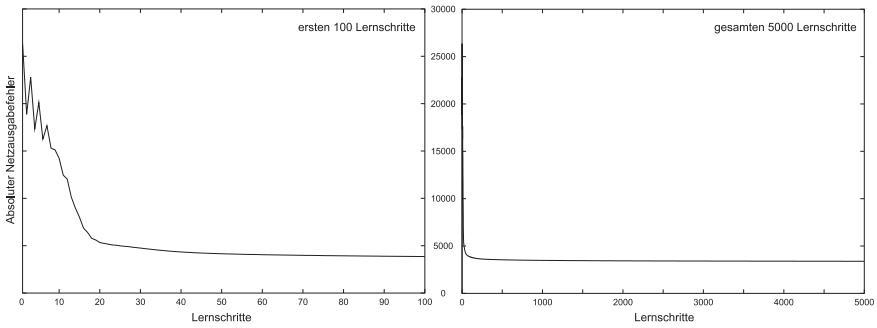


Abb. 5.6: *Verlauf des absoluten Netzausgabefehlers als Detail der ersten 100 Lernschritte (links) und im Gesamtverlauf der 5.000 Trainingsiterationen (rechts).*

5.1.6 Klassifikationsgüte

Zur Bestimmung der Klassifikationsgüte – und damit des Erfolgs des Trainings – wird das so trainierte Neuronale Netz für die Abbildung der jeweils restlichen Datensätze der 2 Nicht-Trainingssprecher (je nach Trainingszusammensetzung zwischen 10.867 und 12.447 Barkspektren) verwendet, von denen ebenfalls das NN-Sollausgabesignal vorliegt. Durch Vergleich dieser Sollausgabe mit der tatsächlichen Ausgabe des Neuronalen Netzes ist eine Bewertung der geleisteten Klassifikation für jedes Kurzzeitsignalfenster möglich. Der prozentuale Anteil an Fehlklassifikationen wird schließlich durch den Anteil der Abweichungen vom Sollsignal bestimmt.

Durch einen Schwellwertvergleich werden alle sprechpausenbereinigten Kurzzeit-Sprachsegmente mit Klassifikationswert oberhalb eines Netzausgabeschwellwertes NN_s als stimmhaft betrachtet, alle unterhalb als stimmlos. Die Wahl dieses Schwellwertes spielt eine entscheidende Rolle. Gerade bei stark gestörten Stimmen weist die Ausgabe des Neuronalen Netzes in stimmhafter Phonation teilweise kleinere Werte als bei stimmgesunder Phonation auf, da durch die Rauschanregung auf glottaler Ebene der Anteil der hohen Frequenzen stärker repräsentiert ist als bei normaler Stimmfunktion und die Erkennungsleistung des NN dadurch gemindert werden kann. Die geeignete Wahl des Schwellwertes garantiert, lediglich die „gesuchten“ stimmhaften Segmente von den stimmlosen zu trennen. Als Schwellwert wird $\mathrm{NN}_s = 0{,}5$ verwendet, sofern der höchste Netzausgabewert des gesamten klassifizierten Signals größer als $0{,}9$ ist. Sollte der höchste Wert im Intervall $[0{,}8; 0{,}9]$ liegen, so wird der Schwellwert auf $\mathrm{NN}_s = 0{,}45$ reduziert. Bei hochgradig gestörten Stimmen erreicht die Netzausgabe in Einzelfällen nur Maximalwerte von knapp unterhalb $0{,}8$. In diesem Fall wird der Klassifikationsschwellwert auf $\mathrm{NN}_s = 0{,}4$ abgesenkt, um die als stimmhaft erkannten Segmente zu detektieren. Die Absenkung des Schwellwertes stellt in diesem Fall kein Problem dar, da der niedrige Klassifikationswert nicht auf einen hohen Netzausgabewert eines stimmlosen Segmentes zurückzuführen ist, sondern vielmehr auf ein zugrunde liegendes stimmhaftes Segment stark gestörter Stimmfunktion.

Der absolute Wert an der Ausgangszelle des Neuronalen Netzes spielt – abgesehen vom Schwellwertvergleich mit NN_s – keine Rolle. Er kann natürlich dennoch für eine Abschätzung der Zuverlässigkeit der Klassifikation herangezogen werden. Extreme Netzausgabewerte bei $s_o \approx 0{,}1$ und $s_o \approx 0{,}9$ (den Sollausgaben während des Trainings) deuten auf eine zuverlässige Erkennung als stimmlos bzw. stimmhaft

hin, wohingegen Werte im Bereich von $s_o \approx 0{,}5$ auf eine eher ungewisse Klassifikation hinweisen und insbesondere in Phonemübergangsbereichen zwischen stimmhaft/stimmlos zu finden sind.

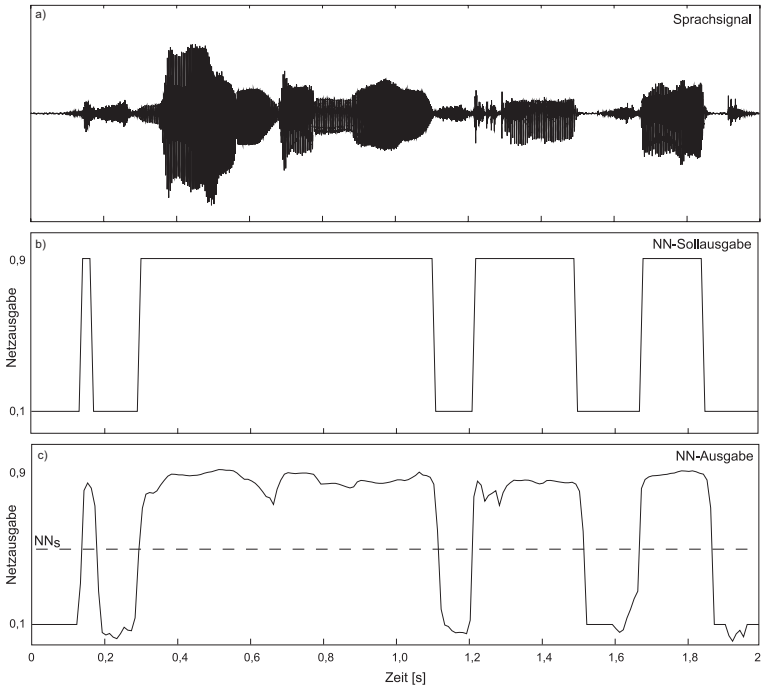


Abb. 5.7: *a) Sprachäußerung „*Es war in Berlin zu einer Zeit*“, b) Sollausgabe des trainierten NN und c) tatsächliche Erkennungsleistung des NN. $NN_s = 0{,}5$ gibt den Schwellwert als gestrichelte Linie an.*

In Abbildung 5.7 ist für die Sprachäußerung „*Es war in Berlin zu einer Zeit*“ eines stimmgesunden Sprechers aus der PHONDAT-Datenbank das Sprachsignal (oben), die Sollausgabe des Netzausgangs (mitte) und die Klassifikationsleistung eines entsprechend den Vorgaben trainierten Neuronalen Netzes (unten) grafisch aufgetragen. Sehr gut ist die hohe Übereinstimmung des Klassifikationsergebnisses mit der Sollausgabe zu erkennen. Das NN ist in der Lage, eine weitgehend eindeutige Klassifikation vorzunehmen, da keine Netzausgabewerte im Bereich des Schwellwertes

NN_s liegen, sondern vornehmlich in den beiden Extrema *stimmhaft* (0,9) und *stimmlos* (0,1). Ebenso ist die an den Phonemübergängen leicht absinkende Netzausgabe zu erkennen.

Bei Betrachtung dieser hohen Übereinstimmung und der Fehlklassifikationsraten aus den Tabellen 5.1 und 5.2 wird deutlich, das das nach den ermittelten Vorgaben trainierte Neuronale Netz bei knapp 96 % der Nichtpausensegmente der PHONDAT-Trainingsdatensätze bei normaler Stimmfunktion die richtige Sollausgabe klassifiziert. Diese hohe Klassifikationsrate unterstreicht die Wahl der in dieser Arbeit entwickelten Vorgehensweise zur Klassifikation.

5.1.7 Ursachen für Fehlklassifikationen

Die ideale Fehlklassifikationsrate von 0 % – also 100 % richtige Klassifikation – lässt sich bei Sprachdatenmaterial eines so großen Variationsbereichs an Stimmgüte, und auch generell bei Signalen mit natürlicher Streubreite, nicht erwarten. Umso bemerkenswerter ist die hohe Klassifikationsgüte von knapp 96 % des trainierten Neuronalen Netzes. Bei einer genaueren Analyse der einzelnen Fehlklassifikationen lässt sich ein Großteil dieser rund 4,5 % auf nachfolgend beschriebene Ursachen zurückführen.

In Abbildung 5.8 ist bei einer detaillierten Betrachtung der einzelnen Fehlklassifikationen deutlich zu erkennen, dass in diesem exemplarischen, aber dennoch repräsentativen Fall alle Abweichungen der Netzausgabe vom Sollsignal in Segmenten der Phonemübergänge – von Sprechpause zu Phonation, aus Phonation zu Sprechpause, von stimmhafter zu stimmloser Phonation oder umgekehrt – lokalisiert sind. Ein zu klassifizierendes Kurzzeitsegment mit >50 % Sprechpausenanteil und knapp unter 50 % stimmhaftem Phonemanteil wird vom Neuronalen Netz zumeist noch als stimmhaft erkannt, hätte auf Grund des Pausenanteils von größer 50 % allerdings als „Sollpause“ klassifiziert werden müssen. Ähnlich verhält es sich an Phonemübergängen von stimmhafter zu stimmloser Phonation, bei denen das erste Segment mit Stimmlos-Anteil von über 50 % durchaus noch gravierende Einflüsse des stimmhaften vorangegangenen Phonems im Barkspektrogramm aufweisen kann. Diese „vermeintlichen“ Fehlklassifikationen an Grenzbereichen stellen einen Großteil der Abweichungen der Ist-Klassifikation des trainierten NN von der Sollklassifikation dar und reduzieren die Klassifikationsleistung dieses Ansatzes teilweise ungerechtfertigterweise.

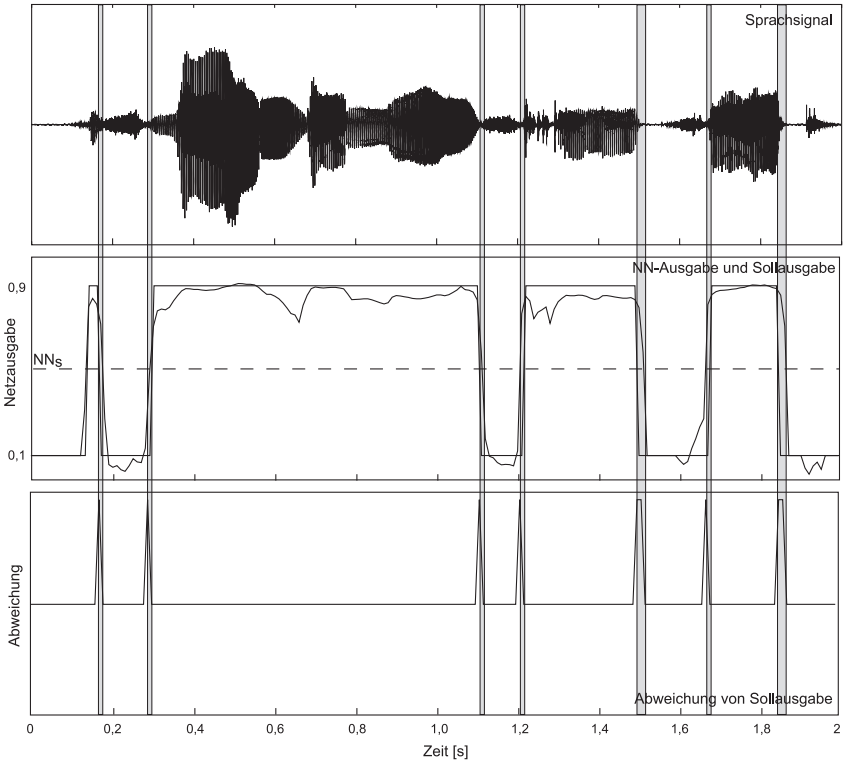


Abb. 5.8: *Sprachäußerung „*Es war in Berlin zu einer Zeit*"(oben), Netzausgabe und Sollausgabe des trainierten NN (Mitte) und Abweichung (grau unterlegt) der klassifizierten Netzausgabe von der Sollausgabe (unten).*

Die Wahl eines höheren Phonemanteils als Klassifikationsgrundlage würde das Auftreten solcher Fehlklassifikationen am Phonationsbeginn und -ende nur verstärken, die eines niedrigeren Anteils stellt die Aussagekraft in Frage, wenn ein Gesamtsegment nach einem Anteil von lediglich 25 – 30 % bewertet und der Inhalt der übrigen 70 – 75 % vernachlässigt wird. Auch bei Phonemübergängen von stimmhafter zu stimmloser Phonation oder umgekehrt innerhalb eines Segments kann dies zu uneindeutigen Klassifikationsergebnissen führen, die von der Soll-

ausgabe abweichen, wobei es sich dennoch nicht notwendigerweise um Fehlklassifikationen handeln muss.

Eine weitere mögliche Fehlerquelle liegt in der Position der von Hand gesetzten Phonemmarken in der PHONDAT-Sprachdatenbank, die die Grundlage für die Generierung des Sollausgabesignals darstellen. Mitunter entspricht die Ausgabe des trainierten Neuronalen Netzes der „richtigen" Klassifikation, die auf Grund einer leicht verschoben gesetzten Phonemgrenzmarke im Vergleich mit dem daraus generierten Sollausgabesignal als „falsch" klassifiziert wird und die Fehlklassifikationsrate ungerechtfertigterweise ansteigen lässt. Hier genügt mitunter eine Deplatzierung der Grenzmarke von wenigen ms, die die Verhältnisse stimmhaft/stimmlos bereits entscheidend verändern können.

Um den kritischen Einfluss dieser Phonemübergangsbereiche auf die nachfolgende akustische Analyse weitestgehend auszuschließen, werden für die Weiterverarbeitung zusammenhängender Segmente gleicher Phonationsklasse das jeweils erste und letzte Kurzzeit-Segment dieses Blocks – in denen typischerweise Phonemübergänge oder Phonationsbeginn/-ende lokalisiert sind – verworfen. Der Einfluss weniger dieser Grenzsegmente kann sensible akustische Maße bereits nachhaltig verfälschen.

Abweichungen der Netzausgabe von der Sollausgabe treten auch bei Plosiven, wie /b/, /d/ oder /g/ auf, die aus einer Verschluss-, einer Plosions- und einer Formantübergangsphase bestehen. Das Neuronale Netz erkennt die Formantübergangsphase der stimmhaften Plosive als stimmhaft, die ebenfalls in die Phonemmarkierung mit eingeschlossene Verschluss- und Plosionsphase allerdings auf Grund mangelnden akustischen Signals nicht, sodass es in diesen Übergangsbereichen zu Fehlklassifikationen kommen kann. Um den Einfluss dieser Mehrdeutigkeit auf die akustischen Analyseergebnisse weitestgehend auszuschließen, werden sämtliche Plosive in dieser Arbeit als stimmlos klassifiziert betrachtet (vergleiche Anhang 7.4).

Ein Anstieg der Fehlklassifikationsrate für stark gestörte Stimmen ist sicherlich zu erwarten, da die bisher präsentierten Ergebnisse auf Sprechern mit normaler Stimmfunktion basieren. Eine exakte Beurteilung der Klassifikationsleistung ist allerdings auch nur mit phonemgelabelten Sprachdaten möglich, und die liegen für gestörte Stimmfunktion nicht in diesem Umfang vor. Die generelle Leistungsfähigkeit dieses Ansatzes und des trainierten Neuronalen Netzes ist damit allerdings bestätigt.

Positiv fällt bei der kritischen Betrachtung der Fehlklassifikationen auf, dass der

Anteil „echter Fehlklassifikationen“ relativ gering ist und die Aussagekraft der anschließenden akustischen Analyse somit stützt. Die Fehlklassifikationsraten aus den Voruntersuchungen in den Tabellen 5.1 und 5.2 weisen durchweg sehr niedrige Werte im Bereich von knapp 5 % auf. Zieht man des Weiteren die obigen Überlegungen bzgl. der Ursachen für Fehlklassifikationen – insbesondere des hohen Anteils an vermeintlichen Fehlklassifikationen bei Phonemübergängen – hinzu, so weist die tatsächliche Klassifikationsleistung noch höhere und somit sehr gute Werte auf, wie in Abb. 5.9 dargestellt ist.

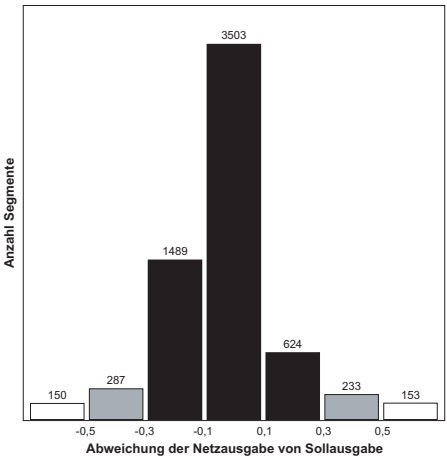


Abb. 5.9: *Histogramm über die Differenz zwischen abgebildeter Netzausgabe und Sollausgabe des Triggersignals für alle Kurzzeitsegmente eines Sprechers der Buttergeschichte. Die schwarz markierten Säulen des Histogramms bezeichnen die Segmente, bei denen die Netzausgabe und Sollausgabe übereinstimmen. In den grau markierten Säulen liegt die Abweichung bereits in einem kritischen Bereich um den Klassifikationsschwellwert. Die weiß markierten Säulen beschreiben Abweichungen über 0,5 (d.h. bei Sollausgabe 0,9 liegt der Netzausgabewert unter 0,4, entsprechend bei Sollausgabe 0,1 über 0,6).*

Diese gesammelten Informationen fließen abschließend in das Training des endgültigen NN ein, das in einer zweiten Trainingsphase auch auf die Klassifikation von Sprachdatenmaterial gestörter Stimmfunktion adaptiert wird.

5.1.8 Endgültige Netzkonfiguration und Klassifikationsleistung

Für die optimale Wahl der Netztopologie (Anzahl versteckter Zellen in der Zwischenschicht), der Trainingsparameter und der Zusammensetzung des Trainingsmaterials sind 6.750 unterschiedliche Neuronale Netze als Kombination folgender Parameter trainiert und getestet worden:

- 5 Topologien (Anzahl versteckter Zellen),
- 5 Lernraten,
- 3 Klassifikationsschwellwerte für die Ausgangszelle,
- 3 Zusammensetzungen des Trainingsmaterials,
- 10 Iterationsvarianten (Overfitting ausschließen),
- jeweils 3 Wiederholungen mit zufälliger Startkonfiguration der Gewichte.

Die jeweilige Klassifikationsleistung ist für all diese 6.750 trainierten Neuronalen Netze mit rund 12.000 Barkspektren und deren Sollausgabesignal getestet worden mit dem Ziel, das NN mit den besten Klassifikationseigenschaften und dem damit verbundenen höchsten Grad an Generalisierung zu bestimmen. Dieses Basisnetz ist in einer zweiten Phase mit 12.500 Barkspektren gestörter Stimmfunktion nachtrainiert worden, um eine umfassendere Repräsentation des gesamten Spektrums an Stimmgüte zu erlangen. Für dieses Nachtraining ist eine kleinere Lernrate gewählt worden, um die bereits antrainierte Basisklassifikation von Normalstimmen nicht wieder zu verlernen. Eine Beurteilung des Einflusses dieses Nachtrainings ist anhand von Vergleichen der Klassifikationsleistung für das bekannte PHONDAT-Sprachdatenmaterial möglich.

In Abbildung 5.10 ist exemplarisch die Ausgabe des endgültigen Neuronalen Netzes für 3 Stimmen unterschiedlicher Stimmgüte dargestellt, die die Klassifikationsfähigkeit dieses final trainierten NN dokumentieren soll. Während sich sowohl das akustische Signal als auch die Netzausgabe des stimmgesunden Sprechers (oben) und die des Sprechers mit glotto-ventrikulärer Ersatzphonation (Mitte) ähneln, sind im Vergleich zu den Signalen des aphonen Sprechers (unten) deutliche Unterschiede zu erkennen. Dennoch liefert das Neuronale Netz eine Klassifikation stimmhafter Phoneme, obwohl – entsprechend dem physiologischen Befund – keinerlei Schwingung auf glottaler Ebene bei aphoner Stimmfunktion vorliegt.

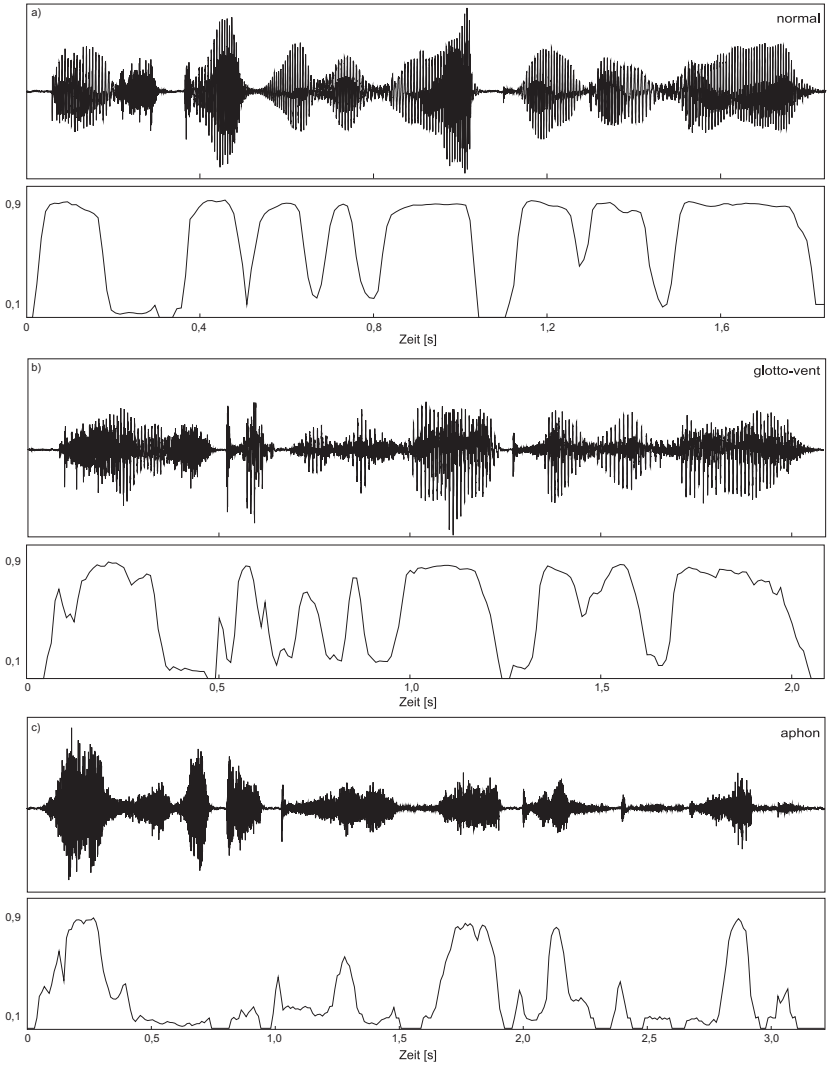


Abb. 5.10: *Ausgabesignal des Neuronalen Netzes zur Klassifikation des fortlaufenden Sprachsignals* „Einst stritten sich Nordwind und Sonne“ *von 3 Sprechern mit unterschiedlicher Stimmgüte: a) Normalstimme, b) glotto-ventrikuläre Ersatzphonation und c) Aphonie.*

PHONDAT-Sprecher	Fehlklassifikationsrate (Klassifikationsrate)	Anteil Phonemübergang (Klassifikationsrate ohne Phonemübergangsfehler)
nordwind01	3,78% (96,22%)	61,31% (98,54%)
nordwind02	4,30% (95,70%)	50,23% (97,86%)
nordwind03	3,61% (96,39%)	63,41% (98,68%)
nordwind04	4,28% (95,72%)	55,73% (98,10%)
nordwind05	4,82% (95,18%)	55,22% (97,84%)
nordwind06	4,86% (95,14%)	47,79% (97,46%)
nordwind07	4,47% (95,53%)	57,89% (98,12%)
nordwind08	5,26% (94,74%)	57,56% (97,77%)
nordwind09	3,59% (96,41%)	66,92% (98,81%)
nordwind10	3,05% (96,95%)	58,73% (98,74%)
nordwind11	4,30% (95,70%)	56,28% (98,12%)
nordwind12	3,85% (96,15%)	59,06% (98,42%)
nordwind13	4,46% (95,54%)	60,28% (98,23%)
nordwind14	5,35% (94,65%)	45,98% (97,11%)
nordwind15	4,98% (95,02%)	56,63% (97,84%)
nordwind16	4,86% (95,14%)	57,14% (97,92%)
berlin01	6,56% (93,44%)	55,04% (97,05%)
berlin02	4,73% (95,27%)	48,98% (97,58%)
berlin03	4,82% (95,18%)	57,98% (97,97%)
berlin04	5,18% (94,82%)	44,91% (97,14%)
berlin05	6,31% (93,69%)	53,23% (97,05%)
berlin06	6,26% (93,74%)	41,96% (96,37%)
berlin07	3,85% (96,15%)	56,24% (98,31%)
berlin08	4,22% (95,78%)	58,00% (98,23%)
berlin09	7,41% (92,59%)	48,49% (96,18%)
berlin10	6,31% (93,69%)	49,84% (96,83%)
berlin11	4,36% (95,63%)	48,11% (97,74%)
berlin12	5,86% (94,13%)	49,08% (97,01%)
berlin13	4,77% (95,23%)	54,73% (97,84%)
berlin14	6,17% (93,83%)	48,89% (96,85%)
berlin15	3,34% (96,66%)	39,70% (97,98%)
berlin16	5,14% (94,86%)	50,29% (97,45%)

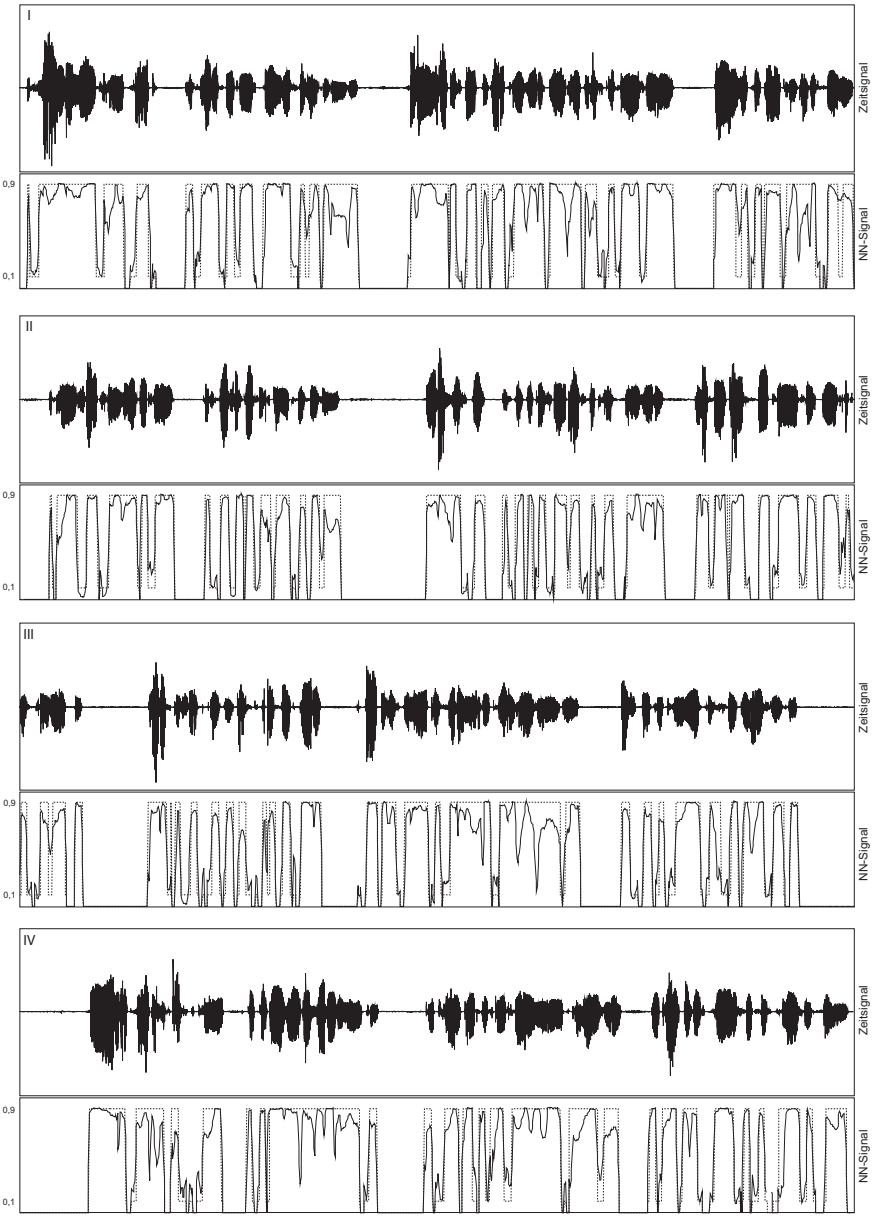
Tabelle 5.3: *Übersicht der individuellen Fehl-/Klassifikationsrate durch Vergleich mit Sollausgabe für die Trainingsdatensätze der* PHONDAT*-Sprecher; prozentualer Anteil an Fehlklassifikationen bei Phonemübergängen (stimmhaft/stimmlos oder Pause) und Klassifikationsrate ohne Phonemübergangsfehler, die bei der Analyse ausgeschlossen werden.*

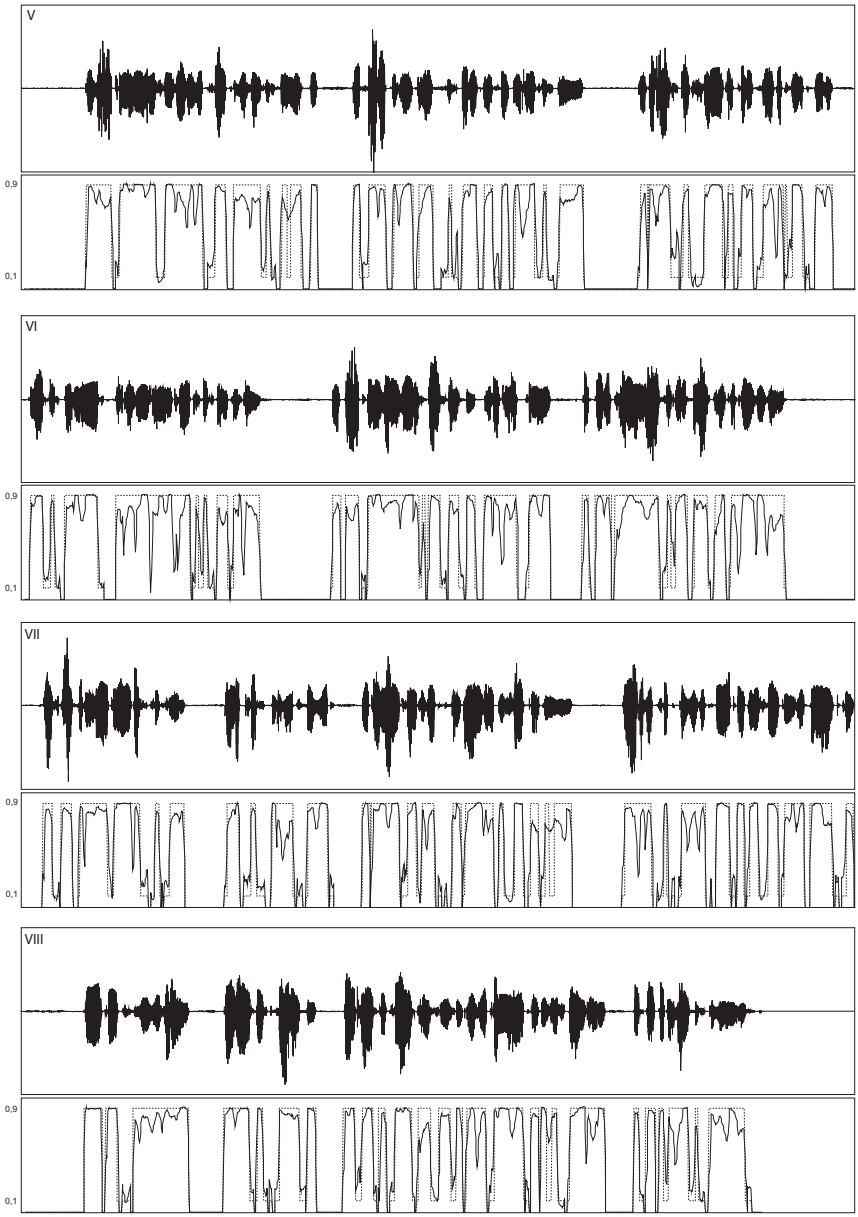
Deutlich zu erkennen sind in Abb. 5.10 allerdings auch die Unterschiede zwischen normaler und aphoner Stimmfunktion sowohl im Sprachsignal als auch im Klassifikationssignal. Beim Vergleich der dargestellten Sprachsignale fällt ein abweichender Sprechrhythmus des aphonen Sprechers im Vergleich zu den anderen beiden Äußerungen – neben einer natürlichen Individualität jedes Sprechers – auf. Eine deutlich kürzere relative Dauer der vokalischen Anteile in den Äußerungen ist auch im auditiven Vergleich festzustellen. Als Konsequenz daraus sind die Segmente stimmhafter Phonation im Vergleich zu den anderen beiden Sprechern kürzer, wobei bei einem rein visuellen Vergleich der Klassifikationssignale die unterschiedliche Zeitskalierung zu beachten ist.

Der Ausgabewert des NN für stimmhafte Phoneme aphoner Stimmfunktion erlangt teilweise Werte knapp unterhalb des Klassifikationsschwellwertes von $NN_s = 0{,}45$ (maximaler $NN_{out} \epsilon [0{,}8; 0{,}9]$, vergl. Abschnitt 5.1.6), was in einer weiteren Reduzierung des stimmhaften Anteils resultiert. Auf der anderen Seite sind die Extreme – stimmhaft und stimmlos mit Klassifikationswert 0,9 und 0,1 – deutlich ausgeprägt wiederzufinden. Insgesamt liefert der in dieser Arbeit entwickelte neuronale Klassifikator durchaus zuverlässige Klassifikationen zusammenhängender Segmente stimmhafter Phonation über das gesamte Spektrum der Stimmfunktion.

Die hohe Klassifikationsgüte zeigt sich bei der Analyse des PHONDAT-Trainingsmaterials, für das die geleistete Klassifikation direkt mit dem Sollausgabesignal verglichen werden kann. In Tabelle 5.3 sind die Fehlklassifikationsraten jedes einzelnen Trainingssprechers mit zusätzlicher Angabe zum Anteil der Fehlklassifikationen an Phonemübergängen – einer der wesentlichen Fehlerquellen – aufgelistet. Der durchweg hohe Fehleranteil an Phonemübergängen – und der dadurch entsprechend niedrigere eigentliche Fehleranteil – unterstreicht die Zuverlässigkeit der Klassifikation, da das jeweils erste und letzte Teilsegment für die akustische Analyse verworfen werden. Exemplarisch ist die vollständige Ausgabe für einen Normalsprecher der Buttergeschichte in Abb. 5.11 mit Signalverlauf, zugehörigem Klassifikationssignal sowie Sollklassifikation dargestellt.

Abb. 5.11: ***Siehe nächste Doppelseite:*** *Vollständiges Sprachsignal in 8 Teilabschnitten (I-VIII, jeweils obere Darstellung) und vom NN abgebildetes Klassifikationssignal (durchgezogene Linie untere Darstellung) sowie das Klassifikations-Sollsignal (gestrichelte Linie untere Darstellung) für einen Normalsprecher. Bereiche, in denen das Sollsignal auf 0 abfällt, sind als Pausensegmente bestimmt.*


I
Zeitsignal
0,9
0,1
NN-Signal
II
Zeitsignal
0,9
0,1
NN-Signal
III
Zeitsignal
0,9
0,1
NN-Signal
IV
Zeitsignal
0,9
0,1
NN-Signal


V
Zeitsignal
0,9
0,1
NN-Signal
VI
Zeitsignal
0,9
0,1
NN-Signal
VII
Zeitsignal
0,9
0,1
NN-Signal
VIII
Zeitsignal
0,9
0,1
NN-Signal

Basierend auf der entwickelten Klassifikation des Sprachsignals erfolgt die akustische Analyse anhand ausgewählter akustischer Maße, die zum einen lediglich Bereiche stimmhafter Phonation bewerten, zum anderen aber auch das gesamte Sprachsignal bis hin zum Pausenanteil als Grundlage heranziehen. Um die Gesamtanalyseergebnisse durch einzelne isolierte Signalsegmente nicht zu verfälschen, werden zusammenhängende Signalbereiche erst als stimmhaft/stimmlos gekennzeichnet und der weiteren akustischen Analyse zugeführt, sobald mindestens sieben aufeinander folgende Kurzzeitsegmente vom Neuronalen Netz in die gleiche Kategorie klassifiziert worden sind.

Als Ergebnis der stimmhaft/stimmlos-Klassifikation mit nachgeschalteter Grundperiodenanalyse mittels Waveform Matching Algorithmus in den stimmhaften zusammenhängenden Segmenten ergibt sich ein Ausgangsdatensatz entsprechend Abb. 5.12 für die Bestimmung akustischer Maße zur Beschreibung der Grundperioden, wie Jitter oder Shimmer beispielsweise.

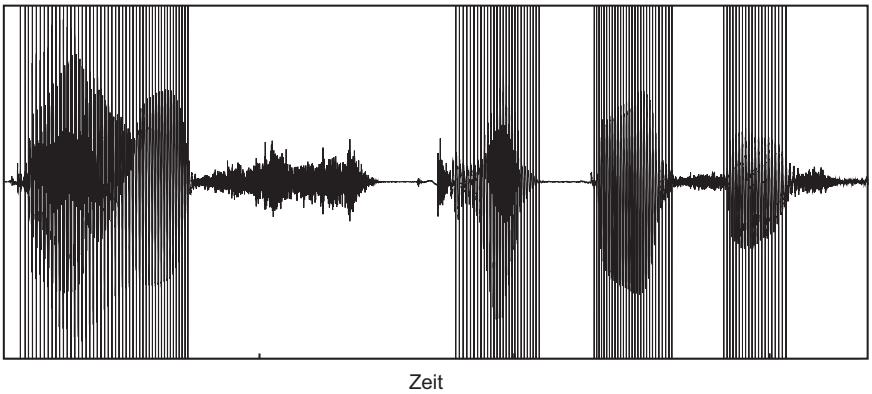


Abb. 5.12: *Sprachsignal „Einst stritten sich“ und vom Verarbeitungsalgorithmus automatisch positionierte Periodenmarken in stimmhaft klassifizierten Segmenten nach dem Waveform Matching Algorithmus als Grundlage zur Bestimmung weiterer akustischer Maße. Jede senkrechte Linie entspricht einer Grundperiodenmarke.*

Die bereits angesprochenen Fehlklassifikationen an Phonemübergängen aus/zu Sprechpausen und von stimmhafter zu stimmloser Phonation (oder umgekehrt)

können ebenfalls zu einer Verzerrung der Analyseergebnisse beitragen und werden deshalb ausgeschlossen, indem das jeweils erste und letzte Kurzzeitsegment eines zusammenhängenden Signalabschnitts gleicher Phonationsklasse aus dem zu analysierenden Segment eliminiert wird. Da in dieser Arbeit mit einem Analysefenstervorschub von 10 ms vorgegangen wird, werden dadurch lediglich 20 ms an Signal verworfen, die bei der Vielzahl von Segmenten keinen gravierenden Einfluss haben.

An dieser Stelle sei noch einmal darauf hingewiesen, dass der in der vorliegenden Arbeit verfolgte Ansatz nicht die Stimmhaftigkeit einer Lautäußerung anhand periodischer Schwingungen auf glottaler Ebene beurteilt, sondern die generelle Stimmhaftigkeit/Stimmlosigkeit eines Phonems als Klassifikationsgrundlage benutzt. Der in der Literatur zumeist verbreitete erstere Ansatz würde bei aphoner Phonation – bei der keinerlei Schwingung auf glottaler Ebene zu beobachten ist – keine Klassifikation stimmhafter Sprachanteile ermöglichen. Im Gegensatz dazu erlaubt die im Rahmen dieser Arbeit entwickelte Vorgehensweise sehr wohl eine Differenzierung zwischen stimmhaften und stimmlosen Phonemen, die eine wichtige Rolle für die anschließende Bestimmung ausgewählter akustischer Maße darstellt.

5.2 Akustische Maße

Zur quantitativen Beschreibung von Stimmstörungen und Therapiekontrollen finden unterschiedliche akustische Maße Anwendung, die aus dem digitalisierten Sprachsignal berechnet werden. Typischerweise werden diese Maße aus Aufnahmen gehaltener Phonation bestimmt und beschreiben Perturbationen der Grundfrequenz und Intensität auf kurzen und langen Zeitskalen sowie glottales Rauschen. Allerdings liefert die alleinige Analyse gehaltener Phonation kein umfassendes Bild der Stimmstörung, da sie die individuellen dynamischen Eigenschaften der fortlaufenden Sprache nicht mit erfasst. Die akustische Analyse fortlaufender Sprache stellt eine wesentliche Erweiterung zur Beschreibung von Stimmstörungen dar und bewertet insbesondere den alltäglichen Einsatzbereich der Stimme, da die Analyse gehaltener Phonation eher den Einsatz als Singstimme charakterisiert. Diese Notwendigkeit der umfassenden Analyse dokumentieren eine Vielzahl von Untersuchungen verschiedenster Arbeitsgruppen, die unter gezielter Betrachtung einzelner Maße – und teilweise Vergleichen mit perzeptiven Beurteilungen – signifikante Abweichungen der Ergebnisse aus gehaltener Phonation und fortlaufender Sprache erhalten haben [HFGS80], [AH86], [S89], [K90], [K95], [QH97], [PJ01].

Bei den bereits in Kapitel 2 vorgestellten akustischen Maßen handelt es sich zum Teil um Erweiterungen zur Analyse gehaltener Phonation, aber auch um eigenständige Maße, die spezielle Eigenschaften und Variationen der fortlaufenden Sprache quantifizieren. Die akustische Analyse mit bekannten Maßen aus gehaltener Phonation basiert auf der im vorausgegangenen Abschnitt 5.1 beschriebenen Methode zur Klassifikation des fortlaufenden Sprachsignals in zusammenhängende Bereiche stimmhafter bzw. stimmloser Phonation sowie Sprechpausen. Diese Klassifikation ist notwendig, da einige dieser Maße lediglich in stimmhaften Teilsegmenten der fortlaufenden Sprache zuverlässige Ergebnisse liefern. Direkt aufeinander folgende Analysefenster der gleichen Klassifikationsklasse des Neuronalen Netzes werden zusammenhängend analysiert, sofern sie eine Mindestlänge von 100 ms (7 aufeinander folgende Kurzzeitsegmente) aufweisen.[33] Für solch ein zusammenhängendes stimmhaftes Segment können dann – analog zur Analyse gehaltener Phonation – die akustischen Maße bestimmt werden und entweder deren Verlauf über das Gesamtsignal oder ihr Mittelwert interpretiert werden.

[33] Diese Mindestsegmentlänge ist notwendig, um auch bei tiefen Grundfrequenzen eine ausreichende Anzahl an Grundperioden zu erfassen, um eine Bestimmung von Maßen wie Jitter und Shimmer zuverlässig vornehmen zu können.

5.2.1 Auswahl der akustischen Maße

Ein Großteil der Literatur zur Analyse fortlaufender Sprache zielt auf eine durch akustische Maße gestützte Differenzierung pathologischer Stimmen von Normalstimmen ab. Anhand von individuellen Kombinationen ausgewählter Maße findet eine Beurteilung der Stimmgüte statt, die eine möglichst zuverlässige Klassifizierung erlauben soll [HFGS80], [AH86], [S89], [K90], [QHM99]. Die reine Differenzierung zwischen normaler und gestörter Stimmfunktion steht bei der vorliegenden Arbeit allerdings nicht im Vordergrund. Durch die Entwicklung der dargestellten Klassifikationsmethode liegt vielmehr die Basis für eine detaillierte akustische Analyse des fortlaufenden Sprachsignals vor.

Bei der Bestimmung der Stimmgüte – sowohl aus gehaltener Phonation als auch aus fortlaufender Sprache – reicht eine Beurteilung lediglich eines Aspektes zumeist nicht aus. Eine gezielte Auswahl repräsentativer akustischer Maße, die eine Beschreibung unterschiedlicher Eigenschaften der Stimmfunktion – und natürlich deren Störungen – erlauben, ist notwendig, um sämtliche Aspekte beurteilen zu können. Da die Anzahl und Vielfalt der inzwischen publizierten Stimmgütemaße unüberschaubar groß ist, muss eine entsprechende Vorauswahl getroffen werden. Um bei dieser Auswahl möglichst viel Einzelinformation zu kombinieren, ist eine Selektion entsprechend dem Informationsgehalt in der Kombination der akustischen Maße notwendig. Bei der Auswahl dieser Maße gilt es folglich die Frage zu beantworten, welche Kombination unabhängiger akustischer Maße die Schwingungsirregularitäten und additives Rauschen in pathologischen Stimmen bestmöglich beschreibt. Lediglich wenn diese beiden Eigenschaften unabhängig voneinander bestimmt werden können, sind Rückschlüsse und Interpretationen zu den perzeptiven Beurteilungskriterien, wie Rauigkeit und Behauchtheit, und damit verbundenen physiologischen Gegebenheiten auf glottaler Ebene möglich.

Durch die Auswahl der in Kapitel 2 vorgestellten akustischen Maße ist bereits eine bewusste Selektion vorgenommen worden. Diese Auswahl stellt ein untereinander weitestgehend unabhängiges, aber dennoch homogenes Ensemble dar und umfasst in der aktuellen Literatur bevorzugt verwendete Maße. Sie beinhaltet dabei zum einen aus der Analyse gehaltener Phonation bekannte Maße wie Jitter, Shimmer, Periodenkorrelationskoeffizienten, PA, SFR, GNE oder das Göttinger Heiserkeits-Diagramm, zum anderen aber auch spezielle Maße der Analyse fortlaufender Sprache, wie z.B. Langzeitspektren (LTAS) oder eine Beurteilung der Fundamental Frequency Distribution (FFD).

Das Göttinger Heiserkeits-Diagramm (vgl. Kapitel 2.5) nimmt unter den vorgestellten akustischen Stimmgütemaßen eine gewisse Sonderstellung ein, da es kein Einzelmaß ist, sondern – als Ergebnis einer Untersuchung zur Auswahl relevanter Maße – auf mehreren Maßen basiert. Das Konzept des Göttinger Heiserkeits-Diagramms hat sich in einer Vielzahl von Studien bewährt und soll in dieser Arbeit auf Basis der entwickelten Klassifikationsmethode für die Analyse fortlaufender Sprache erweitert werden [MFS98], [FMK98], [FMSK98], [FMSK00], [FFHSKK06].

5.2.2 Göttinger Heiserkeits-Diagramm

GHD für gehaltene Phonation

Grundlage der Entwicklung des Göttinger Heiserkeits-Diagramms (*Goettingen Hoarseness Diagram, GHD*) [MFS98] war ein informationstheoretischer Ansatz mit dem Ziel, aus der Vielzahl an publizierten Stimmgütemaßen für gehaltene Phonation eine niedrigdimensionale, repräsentative Teilmenge zu extrahieren, deren Komponenten weitgehend unkorreliert sein sollten, aber dennoch einen ausreichend großen Teil der Varianz der akustischen Analyse von Stimmen unterschiedlicher Stimmgüte abdeckt. Diese Vielzahl lässt sich entsprechend ihrer Beschreibung bestimmter Signalcharakteristika – neben anderen möglichen Klassifizierungen – in zwei Gruppen teilen: Schwingungsirregularitätsmaße (*aperiodicity features*) und Rauschmaße (*noise features*). Zu den ersteren zählen u. a. die populären Maße Jitter, Shimmer und MWMC, zu der zweiten Gruppe bspw. NNE, CHNR oder GNE. Die gesuchte Auswahl sollte eine möglichst unabhängige Beschreibung von Irregularitäten und additivem glottalen Rauschen ermöglichen, um auch die perzeptiven Eindrücke von Rauigkeit und Behauchtheit widerzuspiegeln.

Aus der Fülle von akustischen Maßen hat sich als Ergebnis der Entwicklung des Göttinger Heiserkeits-Diagramms – auf der Basis der Analyse von Korrelationen, Rang-Korrelationen, *Mutual Information*-Analyse und Hauptachsentransformation anhand von 447 Vokalen des gesamten Spektrums an Stimmgüte – eine Linearkombination aus Jitter, Shimmer und mittlerem Periodenkorrelationskoeffizienten in der einen Dimension und GNE in der orthogonalen Dimension herausgestellt. Grundlage der daraus abgeleiteten Irregularitäts- und Rauschkomponente des GHD waren normierte (mittelwertbefreite und durch die Standardabweichung dividierte) Maße aus dem stationären, 1-sekundigen Mittelteil gehaltener Vokale, wie sie in Tabelle 5.4 aufgelistet sind. Da einige akustischen Maße, wie Jitter oder

akust. Maß	Beschreibung	Transformation	Einheit
MWMC	wmc	$\log(1-x)$	
Jitter	$j3$: PPQ mit K=3	$\log x$	%
Shimmer	$s3$: EPQ mit K=3	$\log x$	%
	$s15$: EPQ mit K=15	$\log x$	%
GNE	$gne3$: 3000Hz Bandbreite	$\log(1-x)$	

Tabelle 5.4: *Darstellung der dem GHD zugrunde liegenden akustischen Maße, deren Berechnung und Transformation (um eine annähernde Normalverteilung zu erhalten), die in die Irregularitäts- und Rauschkomponente einfließen.*

Shimmer eine deutliche Häufung bei geringen Werten und einen langen „Schwanz" zu hohen Werten hin aufweisen, ist die Transformation der Werte auf eine logarithmische Skala, auf der sie annähernd normalverteilt sind, ratsam. Eine zusätzliche Translation um 5 Einheiten der Irregularitäts- und $1{,}5$ Einheiten der Rauschkomponente soll sicherstellen, bei natürlichen Signalen lediglich positive Werte der beiden Komponenten im GHD zu erhalten, die sich wie folgt berechnen [MFS98]:

Irregularitätskomponente I_{Vokal} bei Analyse gehaltener Phonation:

$$I_{\text{Vokal}} = 5 + \frac{1}{\sqrt{3}} \left(\frac{\log(1-wmc)+1{,}614}{0{,}574} + \frac{\log(j3)+0{,}374}{0{,}645} + \frac{\log(s15)-0{,}757}{0{,}368} \right)$$

Rauschkomponente R_{Vokal} bei Analyse gehaltener Phonation:

$$R_{\text{Vokal}} = 1{,}5 + \frac{(0{,}695 - gne3)}{0{,}242}$$

Da diese Skalierung der akustischen Maße auf Ergebnissen der Analyse gehaltener Phonation basiert, stellt sie nicht zwangsläufig auch eine gute Repräsentation für fortlaufende Sprache dar, da sich in deren Analyse im Vergleich zu gehaltener Phonation bspw. erhöhte Werte von Jitter und Shimmer zeigen [S89], [SG95]. Dies resultiert in einer eingeschränkten Darstellungsmöglichkeit der Ergebnisse im Göttinger Heiserkeits-Diagramm auf Basis gehaltener Phonation.

Erweiterung des GHD für fortlaufender Sprache

Die in Abschnitt 5.1 vorgestellte Klassifikationsmethode ermöglicht eine automatisierte Bestimmung der dem GHD zugrunde liegenden akustischen Maße aus fortlaufender Sprache und liefert somit die Grundlage für eine Erweiterung des GHD für fortlaufende Sprache, im Folgenden auch als GHDT (Göttinger Heiserkeits-Diagramm Textanalyse) bezeichnet. Die Auswahl der akustischen Maße wird im Hinblick auf eine Vergleichbarkeit der Analyseergebnisse mit denen gehaltener Phonation weitestgehend übernommen. Eine Änderung ist allerdings bei der Linearkombination der Irregularitätskomponente notwendig, da für gehaltene Phonation der Shimmer dort in einer lokalen Umgebung von 15 Grundperioden (s15) berechnet eingeht und dieses Intervall bei tiefen Grundfrequenzen und kurzen stimmhaften Teilsegmenten in fortlaufender Sprache zu Problemen führen kann. Aus diesem Grund wird der Perturbation Quotient des Shimmers – ebenso wie beim Jitter – über 3 aufeinander folgende Perioden bestimmt.

Analog der Entwicklung des GHD für gehaltene Phonation stellen Aufnahmen von Sprechern des gesamten Spektrums an Stimmgüte – in diesem Fall die Aufnahmen fortlaufender Sprache der vorliegenden Gruppen aus 430 Sprechern mit Stimmstörung und 50 mit normaler Stimmfunktion – die Datenbasis dar. Mit der entwickelten Klassifikationsmethode sind aus diesen Aufnahmen stimmhafte Segmente extrahiert und analysiert worden. Eine vorgegebene Mindestlänge der Analysesegmente von 7 aufeinander folgenden Signalfenstern (= 100 ms) soll sicherstellen, den Einfluss eventueller Ausreißer in Kurzsegmenten – auf Grund der angesprochenen möglichen Fehlerquellen an Phonationsübergängen – zu minimieren.

Signal	**Sprechergruppe**	**Analysesegmente** Anzahl	**Segmentlänge** Mittelwert
Vokal	normal	88	1000 ms
Vokal	pathologisch	447	1000 ms
Text	normal	4598	316 ms
Text	PHONDAT	5409	267 ms
Text	pathologisch	41161	333 ms

Tabelle 5.5: *Zusammensetzung des Analysematerials zur Skalierung des GHD auf Basis gehaltener Phonation (Vokal) und GHDT für fortlaufende Sprache (Text).*

Ein Vergleich der Zusammensetzung des Analysematerials für die Bestimmung der Normierungsgrößen der Irregularitätskomponente aus gehaltener Phonation und fortlaufender Sprache ist in Tabelle 5.5 dargestellt. Die Bestimmung der akustischen Maße Jitter, Shimmer und MWMC in ihrer vorgegebenen Transformation aus den 41161 Segmenten der pathologischen Gruppe (durchschnittliche Analysesegmentlänge in fortlaufender Sprache von 333 ms) liefert die Normierungs-Mittelwerte und -Standardabweichungen für die Neuskalierung der Irregularitätskomponente des GHDT für fortlaufende Sprache.

Für die Betrachtung der Rauschkomponente des GHDT ist keine Neuskalierung vorgenommen worden, da die Berechnung des GNE nicht auf der Variation einzelner Schwingungsperioden basiert (die in fortlaufender Sprache nachweislich höhere Werte erlangen), sondern ein integrierendes Maß zur Beschreibung eines additiven Rauschanteils in der Stimme darstellt. Allerdings ist die dem GHD für gehaltene Phonation zugrunde liegende Analysefensterlänge von 500 ms in fortlaufender Sprache nicht praktikabel und wird auf 100 ms reduziert. Dieses Fensterlänge deckt sich mit der Mindestanzahl von 7 stimmhaften Segmenten (vergleiche vorletzten Absatz) für die akustische Analyse.

Maß	**Mittel (SD)** Vokal normal	**Mittel (SD)** Vokal patho	**Mittel (SD)** Text normal	**Mittel (SD)** Text patho	**Maß**
$j3$	-0,792 (0,246)	-0,374 (0,645)	0,267 (0,375)	0,343 (0,434)	$j3$
$s3$			1,029 (0,272)	1,065 (0,297)	$s3$
$s15$	0,531 (0,204)	0,757 (0,368)			$s15$
wmc	-2,021 (0,335)	-1,614 (0,574)	-1,139 (0,298)	-1,044 (0,345)	wmc

Tabelle 5.6: *Normierungsgrößen (Mittelwert und Standardabweichung) für die Irregularitätskomponente des GHD auf Basis gehaltener Phonation (Vokal)[MFS98] und GHDT für fortlaufende Sprache (Text).*

Beim Vergleich der Ergebnisse aus der Analyse fortlaufender Sprache von Sprechern mit normaler (*Text normal*) und denen mit gestörter Stimmfunktion (*Text patho*) in Tabelle 5.6 zeigt sich die Konsistenz der erhaltenen Mittelwerte. Die pathologische Gruppe weist bei allen berechneten Maßen im Mittel schlechtere Werte als die Normalsprecher auf, was zu erwarten war. Im Vergleich der Gruppen Vokal und Text zeigen sich – entsprechend der Literatur – ebenfalls erhöhte Werte aus

fortlaufender Sprache. Die zu den Mittelwerten angegebenen Standardabweichungen spiegeln eine ähnliche Verteilung der jeweiligen Datenmengen wider.

Für die Erweiterung des Göttinger Heiserkeits-Diagramms auf fortlaufende Sprache ergeben sich zur Bestimmung der Irregularitäts- und Rauschkomponente folglich die Gleichungen 5.5 und 5.6:

Irregularitätskomponente I_{Text} bei Analyse fortlaufender Sprache:

$$I_{\text{Text}} = 5 + \frac{1}{\sqrt{3}} \left(\frac{\log(1 - wmc) + 1{,}044}{0{,}345} + \frac{\log(j3) - 0{,}343}{0{,}434} + \frac{\log(s3) - 1{,}065}{0{,}298} \right) \tag{5.5}$$

Rauschkomponente R_{Text} bei Analyse fortlaufender Sprache:

$$R_{\text{Text}} = 1{,}5 + \frac{(0{,}695 - gne3)}{0{,}242} \tag{5.6}$$

Um sicherzustellen, dass das trainierte Neuronale Netz auch eine Generalisierungsfähigkeit erlangt hat und nicht lediglich eine Abbildung des Trainingsdatensatzes aus PHONDAT-Sprechern erlernt hat, ist u. a. eine vergleichende Gruppenanalyse der 50 Normalstimmen aus der Sprachdatenbank mit den Analyseergebnissen der 32 PHONDAT-Trainingssprecher und den Ergebnissen aus deren Sollklassifikation im GHDT durchgeführt worden. Die berechneten Gruppenmittelwerte und Standardabweichungen der Irregularitäts- und Rauschkomponente sind Tabelle 5.7 zu entnehmen und zeigen eine hohe Übereinstimmung.

Gruppe	**Irregularitätskomponente** Mittelwert (SD)	**Rauschkomponente** Mittelwert (SD)
Datenbank	4,67 (0,53)	1,46 (0,30)
PhonDat-I	4,69 (0,41)	1,48 (0,16)
PhonDat-II	4,18 (0,53)	1,45 (0,15)

Tabelle 5.7: *Vergleichende Gruppenanalyse von Normalsprechern: Datenbank (50 Sprecher aus Sprachdatenbank), PhonDat-I (Klassifikation der 32* PHONDAT*-Trainingssprecher) und PhonDat-II (Soll-Klassifikation der 32* PHONDAT*-Trainingssprecher) im GHDT für fortlaufende Sprache.*

Da die Berechnung der akustischen Maße des GHDT – im Gegensatz zur Analyse gehaltener Phonation im GHD – auf unterschiedlichen, unzusammenhängenden Abschnitten des gesamten fortlaufenden Sprachsignals basiert, werden die Werte der Irregularitäts- und Rauschkomponente der Einzelsegmente zur Bestimmung des Ellipsenmittelpunktes und der Halbachsen in der grafischen Darstellung gemittelt. Um der Aussagekraft von langen zusammenhängenden Segmenten im Gegensatz zu Kurzsegmenten für die Ergebnismittelung mehr Gewicht zu verleihen, fließen die Analyseergebnisse mit der zugrunde liegenden Segmentlänge bewertet in die Mittelung ein. Diese Vorgehensweise soll den Einfluss von möglichen „Ausreißern" der akustischen Maße in kurzen Segmenten minimieren. Sollte ein langes Analysesegment – dessen Ergebnis stärkeres Gewicht bei der Gesamtmittelung erhält – ein fälschlicherweise klassifiziertes stimmloses Teilsegment mit überdecken, so wird dessen u. U. „verzerrender" Einfluss bereits während der Mittelung innerhalb des Analysesegmentes ausgeglichen und die Gesamtaussagekraft des langen Teilsegmentes bleibt erhalten.

Zur weiteren Prüfung der Konsistenz der Analyseergebnisse dieser Neuskalierung des Göttinger Heiserkeits-Diagramms für fortlaufende Sprache sind verschiedene Vergleichsuntersuchungen durchgeführt worden. Eine Auswahl der Ergebnisse ist in den Abbildungen 5.13 - 5.17 dargestellt.

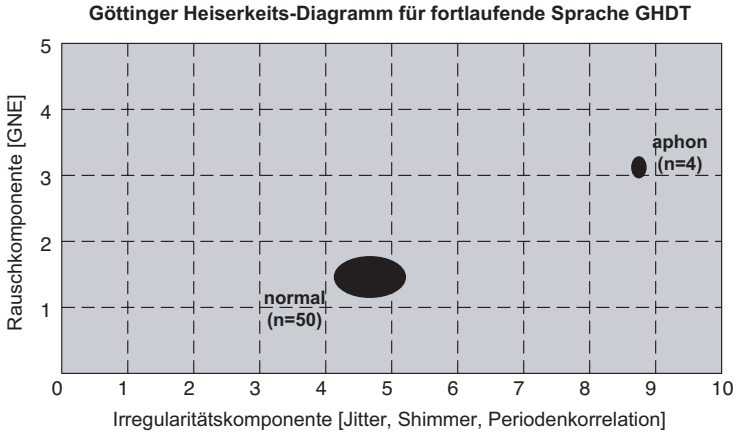


Abb. 5.13: *Gruppenmittelung mehrerer Einzelsprecher mit normaler Phonation (normal) und aphoner Phonation (aphon) als die beiden Extreme der Stimmgüte.*

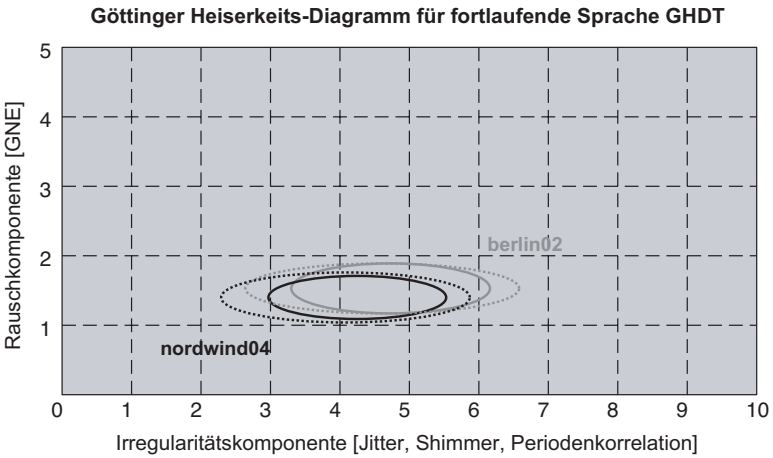


Abb. 5.14: *Ergebnisse nach Klassifikation desselben PhonDat-Sprechers bei unterschiedlichen Texten (durchgezogene Linie) und jeweils zugehöriges Ergebnis aus Sollklassifikation (gestrichelte Linie).*

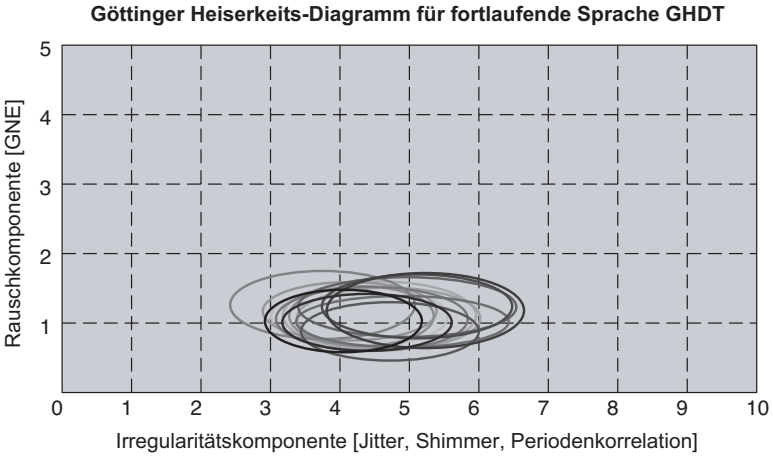


Abb. 5.15: *Verlauf der Stimmgüte eines Sprechers über den Zeitraum von 4 Jahren von 1997 (hellste Linie) bis 2001 (dunkelste Linie).*

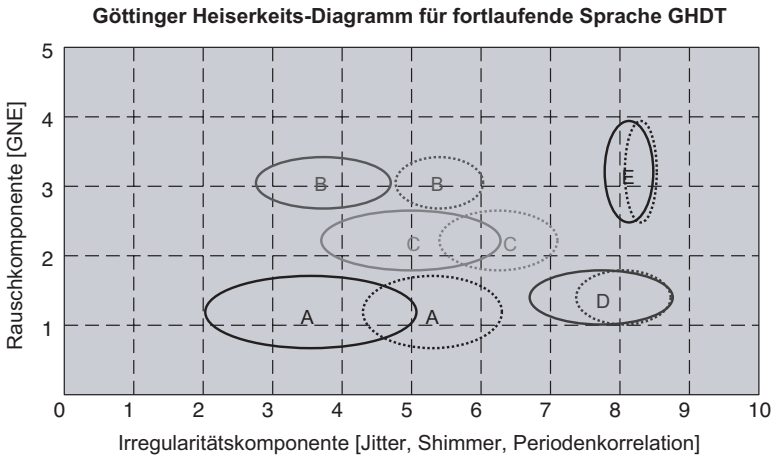


Abb. 5.16: *Vergleich der Analyseergebnisse fortlaufender Sprache in der Skalierung des GHDT (durchgezogene Linie) und des GHD (gestrichelte Linie) für verschiedene Sprecher (A, B, C, D, E) unterschiedlicher Stimmqualität.*

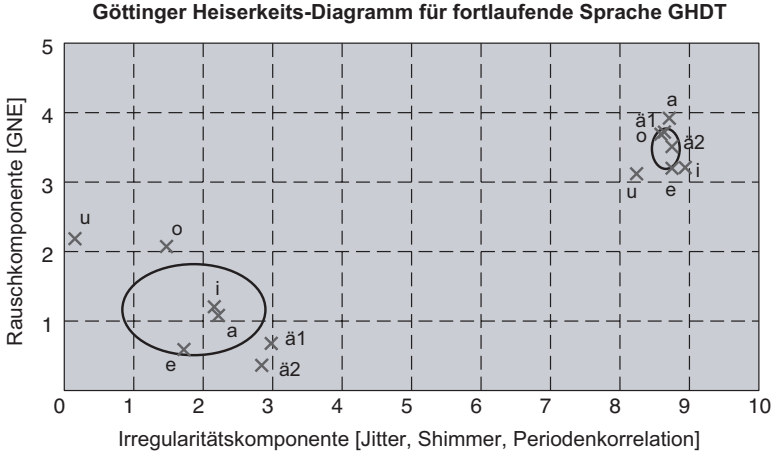


Abb. 5.17: *Serie von gehaltenen Vokalen eines Normalsprechers (links unten) und aphonen Sprechers (rechts oben) als Spezialfall der fortlaufenden Sprache.*

In Abbildung 5.13 sind die Ergebnisse einer Gruppenanalyse von 50 Normalstimmen und 4 aphonen Sprechern im GHDT dargestellt. Die Ellipse – als Darstellung des Ergebnisses – liegt bei Sprechern mit keiner oder geringer Stimmstörung im unteren Wertebereich von Irregularitäts- und Rauschkomponente. Sprecher mit hochgradiger Stimmstörung – das andere Extrem – weisen je nach Pathologie sehr hohe Werte in den akustischen Maßen auf und dadurch auch hohe Irregularitäts- und Rauschkomponenten. Die Verteilung deckt sich mit den Ergebnissen aus der Analyse gehaltener Phonation im GHD.

Die Analyse fortlaufender Sprache nach der vorgestellten Methode ist unabhängig vom Kontext und Inhalt des gesprochenen Textes. Der Sprecher könnte auch bedeutungslose Kunstworte im Kontext fortlaufender Sprache äußern, ohne das Analyseergebnis dadurch zu verfälschen. In Abbildung 5.14 ist das Ergebnis der akustischen Analyse im GHDT für unterschiedliche Äußerungen („Nordwind und Sonne"-Text und „Buttergeschichte") eines Sprechers visualisiert. Zusätzlich sind die jeweiligen Ergebnisse unter Verwendung der in diesem Fall vorliegenden Sollklassifikation (gestrichelte Linie) im Vergleich zur tatsächlich geleisteten (durchgezogene Linie) dargestellt und zeigen eine gute Übereinstimmung.

Die Konsistenz der Analyseergebnisse über eine Vielzahl an Untersuchungen eines Sprechers aus einem mehrjährigen Zeitraum ist in Abbildung 5.15 zu erkennen. Trotz natürlicher Variationen der Stimmgüte auf Grund von temporärer Heiserkeit, Belastung oder Erkältung bspw. liegen die Analyseergebnisse durchweg in einem Bereich.

Der angesprochene Nachteil in der Skalierung auf Basis von gehaltener Phonation im GHD zeigt sich deutlich in Abbildung 5.16. Gestrichelt dargestellt ist jeweils das Analyseergebnis unter Verwendung der GHD-Skalierung aus Gl. 2.32 und 2.33 und durchgezogen das der Neuskalierung des GHDT nach Gl. 5.5 und 5.6. Die Basis für die vergleichende Berechnung war jeweils dasselbe Klassifikationssignal.

Die Analyse gehaltener Phonation als „Spezialfall" der fortlaufenden Sprache ist in Abb. 5.17 für die Vokalserie (ä1, a, e, i, o, u, ä2) zweier Sprecher mit normaler (unten links) und aphoner (oben rechts) Stimmfunktion dargestellt. Zusätzlich zum Gesamtergebnis (Ellipse mit Halbachsen) ist auch die Lage der Irregularitäts- und Rauschkomponente der Einzel-Vokale mit angegeben. Wie beim Einzelergebnis des Vokals /u/ des Normalsprechers zu erkennen, kann es u. U. zu negativen Werten in der Irregularitätskomponente kommen. Dies ist leider – genau wie beim GHD für gehaltene Phonation – nicht zu vermeiden, da ein breiter Wertebereich der Einzelmaße abgebildet werden soll.

In Abbildung 5.18 sind abschließend die Analyseergebnisse des gesamten Datenmaterials an unterschiedlicher Stimmgüte dargestellt, um die breite Abdeckung des neu skalierten Göttinger Heiserkeits-Diagramms für fortlaufende Sprache darzustellen.

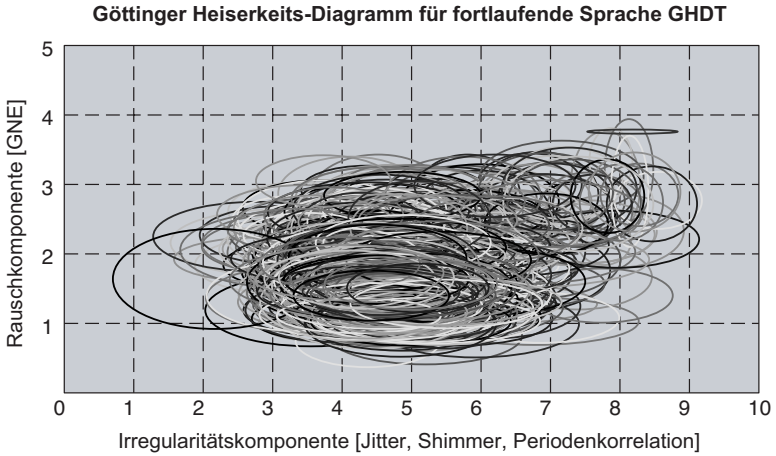


Abb. 5.18: *Darstellung der Analyseergebnisse des gesamten fortlaufenden Sprachdatenmaterials jeglicher Stimmgüte im neu skalierten GHDT.*

Das im Rahmen dieser Arbeit entwickelte GHDT – auf Basis des GHD für gehaltene Phonation – bietet eine aussagekräftige Möglichkeit zur Beurteilung von Stimmstörungen und stellt somit eine Erweiterung der akustischen Analysemöglichkeiten fortlaufender Sprache auf Basis der entwickelten Klassifikationsmethode dar.

Das Hauptaugenmerk bei der Beschreibung akustischer Maße für fortlaufende Sprache liegt in der vorliegenden Arbeit eindeutig auf der Erweiterung des Göttinger Heiserkeits-Diagramms, da diese Darstellung durch Verknüpfung mehrerer Einzelmaße eine sehr gute Beschreibung ermöglicht und durch die Analyse gehaltener Phonation gestützt werden kann. Darüber hinaus existieren natürlich noch weitere Maße, die in der aktuellen Literatur Verwendung finden und in Kapitel 2 beschrieben sind. Sofern für deren Bestimmung eine Selektion stimmhafter Segmente aus dem Sprachsignal notwendig ist, liefert die entwickelte Klassifikationsmethode unter Verwendung Neuronaler Netze die erforderliche Basis.

Exemplarisch seien deshalb die Ergebnisse der Bestimmung der akustischen Maße *Spectral Flatness Ratio* (SFR, siehe Abschnitt 2.3.2) und *Pitch Amplitude* (PA, siehe Abschnitt 2.3.3) im folgenden Abschnitt dargestellt, da diese nach Untersuchungen von Parsa et al. [PJ01] im Vergleich der Analyse gehaltener Phonation und fortlaufender Sprache die besten Resultate zur Differenzierung unterschiedlicher Stimmqualitäten geliefert haben.

5.2.3 Pitch Amplitude und Spectral Flatness Ratio

Die Bestimmung der *Pitch Amplitude* (PA) und des *Spectral Flatness Ratio* (SFR) aus dem Residualsignal der Linearen Prädiktion (siehe Abschnitt 2.3) haben sich in mehreren Arbeiten unterschiedlicher Autorengruppen (bspw. [PMWH87], [ECH90], [PJ01]) als aussagekräftige akustische Maße zur Beurteilung stimmhafter Phonation herausgestellt. Die PA ist definiert als Amplitudenwert des ersten Nebenmaximums (das nicht bei 0 liegt) der normierten Autokorrelationsfunktion des Residualsignals der Linearen Prädiktion und stellt somit ein Maß für die Regelmäßigkeit der Grundperioden in stimmhafter Phonation dar. Ein hoher Wert im Intervall $[0; 1]$ kennzeichnet eine Regelmäßigkeit in der Anregung und sollte bei normaler Stimmfunktion in stimmhaften Segmenten festzustellen sein, ein entsprechend niedrigerer Wert in stimmloser Phonation oder bei gestörter Stimmfunktion.

Der SFR kann als ein Maß für die Maskierung der Harmonischen der Grundfrequenz durch Rauschen betrachtet werden und wird üblicherweise als Logarithmus des Verhältnisses aus geometrischem und arithmetischem Mittel der spektralen Energieverteilung eines Signalfensters in dB berechnet (vgl. Gl. 2.31) und weist bei gesunder stimmhafter Phonation große negative Werte $[-\infty; 0]$ auf.

	Pitch Amplitude		**Spectral Flatness Ratio**	
Datenmaterial	normal	patho	normal	patho
Vokal	0,566	0,122	-17,36	-6,47
Text	0,291	0,130	-15,76	-7,45

Tabelle 5.8: *Vergleich der PA und des SFR für die Analyse gehaltener Phonation (Vokal) und fortlaufender Sprache (Text) für Sprecher mit normaler Stimmfunktion (normal) und gestörter Stimmfunktion (patho).*

Da die Bestimmung beider Maße lediglich in stimmhaften Teilsegmenten erfolgt, ist eine Klassifikation des fortlaufenden Sprachsignals notwendig, für die die entwickelte Methode eine geeignete Wahl darstellt. Die Ergebnisse einer exemplarischen Vergleichsuntersuchung gehaltener Phonation und fortlaufender Sprache für Sprecher mit normaler und gestörter Stimmfunktion sind in Tab. 5.8 dargestellt.

Innerhalb der einzelnen Vergleichsgruppen decken sich die Ergebnisse der Bestimmung der Pitch Amplitude und des Spectral Flatness Ratio aus dem Residualsignal der LP. Sowohl PA als auch SFR erlauben eine eindeutige Differenzierung zwischen Normalstimme und gestörter Stimmfunktion sowohl auf Basis gehaltener Phonation (Vokal) als auch fortlaufender Sprache (Text). Beide Maße weisen für Normalsprecher in gehaltener Phonation höhere Werte als in fortlaufender Sprache auf. Dies ist auch sicherlich zu erwarten gewesen, da es sich bei der PA um die Amplitude eines normierten Korrelationswertes handelt, der empfindlich auf Variationen reagiert und im fortlaufenden Text – trotz einer Analysefensterlänge von lediglich 60 ms (in Anlehnung an [PJ01]) – mit geringen Koartikulationseffekten zu rechnen ist.

Insgesamt zeigt diese Analyse eine weitere Anwendungsmöglichkeit und das breite Spektrum der in dieser Arbeit entwickelten automatisierten Klassifikationsmethode stimmhafter/stimmloser Teilsegmente eines fortlaufenden Sprachsignals.

5.3 Vokalerkennung

Aus dem dargestellten Ansatz zur automatischen Klassifizierung der Stimmhaftigkeit der Phoneme aus fortlaufender Sprache ist eine Methode abgeleitet worden, die eine automatische Erkennung der, im Rahmen der Stimmaufnahmen ebenfalls akquirierten, Vokale aus den Vokalsequenzen ermöglicht. Jede Vokalsequenz umfasst – laut Aufnahmeprotokoll für die Stimmanalyse in der Abteilung Phoniatrie und Pädaudiologie – die Vokalabfolge /ɛ/, /a/, /e/, /i/, /o/, /u/ und /ɛ/, jeweils in den Tonlagen „normal“, „tief“ und „hoch“ sowie „belastet“ nach Vorlesen des „Nordwind und Sonne“-Textes. Eine automatische Erkennung und Markierung der einzelnen Vokale erlaubt einen effizienteren Ablauf der Stimmaufnahme und -analyse und erleichtert die Arbeit und den Zeitaufwand des Aufnahmeleiters deutlich. Ein flüssiger Ablauf der Stimmaufnahme steigert auch die Akzeptanz bei den Patienten. Des Weiteren stellt diese automatische Klassifikation für die Stimmanalyse eine Methode zur Beurteilung der Sprachäußerung an die Hand, die eine Bewertung des „tatsächlich“ phonierten Vokals in Bezug auf den durch das Aufnahmeprotokoll „vorgegebenen“, zu phonierenden Vokal erlaubt. Ein /ɛ/ sollte auch als solches gesprochen werden und nicht ein /e/ oder /a/ sein.

Diese Anwendung der „Spracherkennung“ unterscheidet sich im Ansatz von den gängigen Methoden der Spracherkennung insofern, als dass das zu analysierende Sprachmaterial lediglich 6 unterschiedliche Phoneme umfasst und diese jeweils mindestens 2 Sekunden lang phoniert werden sollen. Des Weiteren ist die Stimmaufnahme durch den speziell akustisch gedämmten, reflexionsarmen Aufnahmeraum von sämtlichen Hintergrundstörgeräuschen befreit, wobei andererseits das gesamte Spektrum an Stimmstörungen analysiert werden soll und nicht lediglich Normalstimmen. Modelle wie *Markov-Ketten* oder *Time Delayed Neural Networks (TDNN)* bspw. kommen deshalb hier nicht zum Einsatz, da jeweils lediglich ein Phonem klassifiziert werden muss. Statische Multi-Layer-Perceptrons bieten dabei eine gute Wahl.

Vorverarbeitung

Entsprechend der Vorgehensweise aus der akustischen Analyse fortlaufender Sprache (vgl. Kapitel 5.1) werden mittels Linearer Prädiktion Kurzzeitspektren von Signalsegmenten bestimmt, die nach Barkskalierung, Dynamikkompression und Normierung einem Neuronalen Netz zur Klassifikation zugeführt werden. Dieses

NN besitzt – im Gegensatz zu dem der stimmhaft/stimmlos-Klassifikation – 6 Ausgangszellen. Die Aktivität jedes dieser Ausgangsneurone spiegelt dabei die Übereinstimmung mit einem der Vokale wider. Es erfolgt eine Eins-aus-N-Kodierung. Die Ausgangszelle mit der maximalen Aktivität bestimmt einerseits den analysierten Vokal, andererseits ist über die Höhe der Ausgangsaktivität und die Werte der anderen Ausgangsneurone auch eine Aussage über die Zuverlässigkeit der Klassifikation des einzelnen Vokals möglich. Liegt der Netzausgabewert aller Ausgangszellen unterhalb eines vorgegebenen Schwellwertes, so wird die analysierte Sprachäußerung gar nicht als Vokal gedeutet und für die akustische Analyse verworfen. Es kann sich dabei bspw. um Husten oder Räuspern des Sprechers handeln, oder auch um eine mit aufgezeichnete Konversation mit dem Aufnahmeleiter.

Ein mit einem höheren Prozentsatz als bei der Analyse fortlaufender Sprache gewähltes Energiekriterium dient der Exklusion von Sprechpausen und Hintergrundinstruktionen des Aufnahmeleiters zwischen den einzelnen Vokalen und beschränkt den zu analysierenden Phonationsbereich. Die einzelnen Vokale werden in der Regel deutlich getrennt voneinander phoniert, sodass keine Phonemübergänge berücksichtigt werden müssen. Metainformationen über die – durch das Aufnahmeprotokoll vorgegebene – Reihenfolge und Anzahl der Vokale gehen nicht in die Erkennung ein, um den Ablauf möglichst unabhängig von individuellen Aufnahmeprotokollen zu gestalten.

Training des Neuronalen Netzes

Während des Trainings des Neuronalen Netzes werden Kurzzeitsignalfenster einzelner Vokale, entsprechend der zu Beginn des Abschnitts skizzierten Vorgehensweise, transformiert und dem NN als Trainingsmenge zugeführt. Die Sollausgabe an den 6 Ausgangszellen setzt sich dabei aus der Sollausgabe $s_o = 0{,}9$ an der Zelle des entsprechenden Vokals und $s_o = 0{,}1$ an den anderen 5 Ausgangszellen zusammen. Über die Aktivierung der jeweiligen Netzausgabezellen bei Klassifikation eines unbekannten Vokals lässt sich somit eine Erkennung der Vokale vornehmen.

Das Training ist mit den berechneten Merkmalsvektoren von 8192 mindestens 2 Sekunden andauernden Vokalen aus der Sprachdatenbank der Abteilung Phoniatrie und Pädaudiologie durchgeführt worden. Diese Trainingsauswahl umfasst die jeweils gleiche Anzahl an Vokalen /ɛ/, /a/, /e/, /i/, /o/, /u/ aus dem gesamten Bereich der Stimmgüte, wobei für jede einzelne Sprachäußerung der jeweilige Vokal, die Tonlage sowie Beginn und Ende des stationären Phonationsbereichs aus

der Datenbank bekannt sind. Anhand dieser detaillierten Informationen ist eine hohe Klassifikationsgüte des trainierten Neuronalen Netzes erreicht worden.

Die Dimensionalität der Zwischenschicht spielt bei dieser Anwendung keine gravierende Rolle, wie entsprechende Untersuchungen gezeigt haben. Die Klassifikationsgüte ist bei allen untersuchten Neuronenzahlen annähernd gleich. Fehlklassifikationen treten – unabhängig von der Netzdimensionalität – insbesondere bei nicht exakter Artikulation der Vokale durch den Sprecher auf. Eine klare Differenzierung zwischen /ɛ/ und /e/, bzw. /ɛ/ und /a/ ist selbst bei perzeptiver Beurteilung mitunter bei einzelnen Aufnahmen nur schwer möglich. Dieses Ergebnis spiegelt sich auch in der Netzausgabe an den entsprechenden Neuronen wider.

Die Netzausgabewerte des trainierten Neuronalen Netzes zur Vokalklassifikation sind exemplarisch in der Abbildung 5.19 für den Vokal /ɛ/ eines Normalsprechers aufgezeigt. Die Vokalbeschriftungen an den Kurven dienen der Identifizierung des Neurons, das auf den entsprechenden Vokal trainiert worden ist.

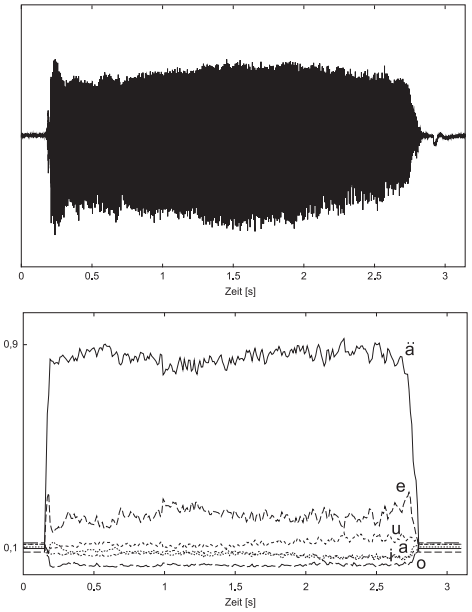


Abb. 5.19: *Ausgabe der sechs Ausgangsneurone des NN bei Klassifikation des Vokals /ɛ/.*

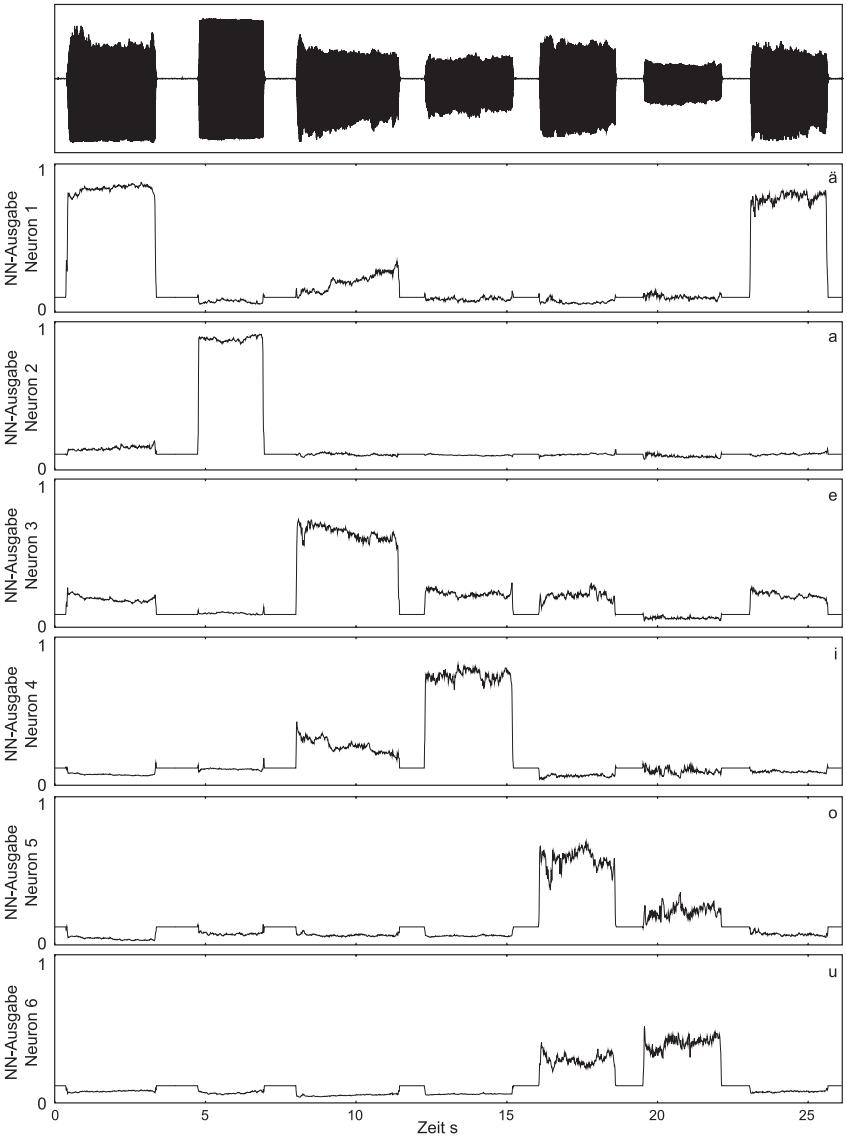


Abb. 5.20: *Ausgabe der sechs Ausgangsneurone des NN bei Klassifikation einer Vokalserie aus /ɛ/, /a/, /e/, /i/, /o/, /u/, /ɛ/.*

Diese Vokalerkennung ermöglicht in Zusammenhang mit der automatischen Phonemklassifikation der fortlaufenden Sprache eine computergestützte, nahezu vollautomatische Bestimmung der Stimmgüte des Sprechers. Die Interaktion des Aufnahmeleiters wird dabei auf ein Minimum reduziert, was den Ablauf deutlich beschleunigt und die Akzeptanz beim Patienten erhöht. Durch die datenbankgestützte Archivierung der Stimmaufnahmen ist es möglich, mehrere Untersuchungsergebnisse aus verschiedenen Stadien einer Therapie oder Rehabilitation miteinander zu vergleichen und so fundierte Aussagen über den Therapieverlauf eines Sprechers vornehmen zu können.

6 Zusammenfassung

Die Sprache stellt das wichtigste Kommunikationsmittel des Menschen dar und hat wesentlich zu seiner Entwicklung in der Evolution beigetragen. Der Träger der Sprache ist die Stimme, die aus der Sicht der Akustik als ein an den schwingenden Stimmlippen modulierter Luftstrom zu verstehen ist, der den Vokaltrakt (Resonanzfilter) durchläuft und an den Lippen als Schalldruckwelle abgestrahlt wird.

Die akustische Stimmanalyse dient der Quantifizierung von Störungen im Sprechapparat und von Irregularitäten der Stimmanregung auf glottaler Ebene. Es wird dabei zwischen der Analyse gehaltener Phonation (lang anhaltend gesprochene Vokale) und fortlaufend gesprochener Sprache differenziert. Gehaltene Vokale sind annähernd stationär in ihrer Struktur und dadurch attraktiv als Grundlage für eine akustische Analyse. Fortlaufende Sprache ist dagegen gekennzeichnet durch eine wechselnde Abfolge von stimmhaften bzw. stimmlosen Phonemen und Sprechpausen – und damit häufigen und schnellen Wechseln der Artikulatorstellungen –, was die akustische Analyse erschwert. Die gehaltene Phonation stellt im Alltag allerdings nur einen Randbereich dar und spiegelt eher den Einsatz als Singstimme wider. Den natürlichen und alltäglichen Gebrauch der Stimme stellt die fortlaufende Sprache dar und ist folglich ein wichtiger Bestandteil einer umfassenden Beurteilung von Stimmgüte.

Ein wesentlicher Teil der publizierten akustischen Maße zur Beschreibung der Stimmgüte quantifiziert Irregularitäten im Schwingungsverhalten der Stimmlippen oder additives Rauschen während der Phonation stimmhafter Phoneme. Um diese Maße der Analyse fortlaufender Sprache zugänglich zu machen, ist eine Klassifikation des Sprachsignals in Bereiche stimmhafter und stimmloser Phonation nötig. Aus Mangel an zuverlässigen automatisierten Klassifikationsmethoden findet eine Beurteilung der Stimmgüte zumeist auf Basis gehaltener Phonation statt. Umso wichtiger ist die Entwicklung einer Klassifikationsmethode für jegliche Stimmgüte, die diese kurzen quasistationären Teilsegmente selektiert und einer weiteren Analyse zugänglich macht.

Die im Rahmen dieser Arbeit entwickelte Klassifikationsmethode ermöglicht eine solche Differenzierung des Sprachsignals jeglicher Stimmgüte in Bereiche stimmhafter und stimmloser Phonation sowie Sprechpausen. Die Selektion stimmhafter Teilbereiche des akustischen Gesamtsignals erfolgt bei dieser Methode auf Basis der Beurteilung transformierter spektraler Vokaltraktparameter kurzer Signalfenster quasistationärer Phonation. Eine Differenzierung wird dabei anhand der eigentlichen Stimmhaftigkeit des artikulierten Phonems und nicht der Stimmhaftigkeit auf Grund seiner Stimmanregung auf glottaler Ebene – der zumeist in der Literatur anzutreffenden Vorgehensweise – vorgenommen. Der hier gewählte Ansatz ermöglicht eine stimmhaft/stimmlos-Klassifikation weitestgehend unabhängig vom glottalen Anregungssignal und dadurch auch unabhängig von Störungen der Stimmlippenschwingung in Form von Schwingungsirregularitäten oder additivem Rauschen durch inkompletten Glottisschluss. Aus diesem Grund ist nunmehr eine Beurteilung von Stimmgüte aus fortlaufender Sprache für eine große Bandbreite an Stimmstörungsbildern – normale bis aphone Stimmfunktion – möglich.

Für diese Art der Klassifikation bietet sich die Verwendung Neuronaler Netze (NN) mit nichtlinearer Aktivierungsfunktion der Neuronen an, die es ermöglicht, auch nicht linear separierbare Klassifikationsaufgaben – wie die vorliegende – adäquat zu lösen. Für die stimmhaft/stimmlos-Klassifikation der Kurzzeitsegmente kommt ein vollständig verbundenes, vorwärts gekoppeltes Multi-Layer-Perzeptron (MLP) mit sigmoider Kennlinie zum Einsatz, dessen Gewichts- und Schwellwertmatrizen in einem vorausgehenden Trainingsprozess auf die zu leistende Klassifikationsaufgabe adaptiert werden müssen. Da eine analytische Bestimmung der optimalen Topologie und Trainingsparameter eines NN bisher nicht möglich ist, sind umfangreiche Untersuchungsreihen zur Bestimmung der bestmöglichen Struktur des NN durchgeführt worden. Als Trainingsmaterial kam dabei die phonemgelabelte Sprachdatenbank PHONDAT und die umfangreiche Datenbank der Abteilung Phoniatrie & Pädaudiologie mit Aufnahmen gehaltener Phonation und fortlaufender Sprache jeglicher Stimmgüte zum Einsatz. Die entwickelte Methode umfasst die Schritte der Vorverarbeitung, spektralen Transformation und eigentlichen Klassifikation mittels eines trainierten NN und ermöglicht eine Selektion stimmhafter Teilsegmente aus dem fortlaufenden Sprachsignal jeglicher Stimmgüte.

Diese Klassifikationsmethode stellt die Basis für eine akustische Analyse fortlaufender Sprache dar. In den zusammenhängenden, als stimmhaft klassifizierten Teilsegmenten können unterschiedliche akustische Maße zur Beschreibung der Stimmgüte berechnet werden, die bei Beurteilung des unklassifizierten Gesamtsignals we-

nig Aussagekraft besäßen. Das Göttinger Heiserkeits-Diagramm für gehaltene Phonation (GHD) stellt ein sehr aussagekräftiges Instrument dar, da es auf mehreren akustischen Einzelmaßen basiert, die sowohl Schwingungsirregularitäten als auch additives Rauschen quantifizieren und unter informationstheoretischen Gesichtspunkten aus einer Vielzahl bekannter Maße als Kombination mit dem größten unkorrelierten Informationsgehalt selektiert worden sind. Die Skalierung des GHD basiert auf der Analyse gehaltener Phonation und stellt demzufolge nicht zwangsläufig auch eine gute Repräsentation für die Beurteilung fortlaufender Sprache dar, da Schwingungsirregularitätsmaße wie Jitter und Shimmer – die beide in die Irregularitätskomponente des GHD einfließen – in fortlaufender Sprache höhere Werte als in gehaltener Phonation aufweisen. Die Erweiterung des GHD auf fortlaufende Sprache zum Göttinger Heiserkeits-Diagramm Textanalyse (GHDT) stellt einen zweiten wesentlichen Bereich in dieser Arbeit dar. Grundlage für die Bestimmung der dem GHDT zugrunde liegenden Einzelmaße ist die entwickelte Klassifikationsmethode und deren Anwendung bei der Analyse von 480 akustischen Aufnahmen fortlaufender Sprache des „Nordwind und Sonne"-Textes. Vergleichsuntersuchungen zwischen GHD und GHDT zeigen die Validität und Aussagekraft der Ergebnisse im neu skalierten GHDT und untermauern die Zuverlässigkeit der entwickelten Klassifikationsmethode. Darüber hinaus erlaubt die Klassifikation die Bestimmung auch anderer akustischer Maße aus fortlaufender Sprache, von denen einige exemplarisch dargestellt sind.

Die in dieser Arbeit entwickelte Klassifikationsmethode stellt somit eine wesentliche Erweiterung des Spektrums der akustischen Analyse fortlaufender Sprache zur Beschreibung der Stimmgüte dar und findet direkte Anwendung in der ebenfalls entwickelten Neuskalierung des Göttinger Heiserkeits-Diagramms Textanalyse (GHDT) für fortlaufende Sprache. Der breite Einsatzbereich für Sprachäußerungen jeglicher Stimmgüte hebt diese Arbeit von den bisherigen Verfahren ab und bietet mit dem GHDT ein dem erfolgreichen GHD verwandtes Maß zur Beschreibung der Stimmgüte aus fortlaufender Sprache.

6.1 Diskussion und Ausblick

Der Erfolg einer Klassifikation mittels Neuronaler Netze hängt stark von der Zusammensetzung und Verteilung des Trainingsmaterials ab sowie von der gewählten Parametrisierung der Eingangsdaten. Der breite Variationsbereich an Stimmstörungen – von normaler gesunder Stimmfunktion bis hin zu aphoner Flüster-

stimme ohne jegliche periodische Schwingung auf glottaler Ebene – erschwert die Entwicklung einer zuverlässigen Methode. Die hier geleistete Klassifikation bietet eine bestmögliche Differenzierung zwischen stimmhaften und stimmlosen Teilsegmenten, die bei Normalsprechern Klassifikationsleistungen von über 95% erreicht. Je gesunder die Stimmfunktion, desto zutreffender letztendlich auch die gemachten Modellannahmen und desto höher die Klassifikationsrate. Mit zunehmendem Grad der Stimmstörung lässt die Klassifikationsleistung etwas nach und die Ausgabe des Neuronalen Netzes erreicht nicht immer den notwendigen Klassifikationsschwellwert in stimmhafter Phonation. Dadurch werden aber insbesondere einzelne stimmhafte Segmente nicht mit selektiert und äußerst selten nur stimmlose Teilsegmente fälschlicherweise als stimmhaft klassifiziert, die einen verzerrenden Einfluss auf die berechneten akustischen Maße haben könnten. Die Anzahl der selektierten Analysesegmente fällt bei starken Stimmstörungen ab, erlaubt aber dennoch eine zuverlässige Beurteilung der Stimmgüte, wie Vergleiche der berechneten akustischen Maße zeigen.

Eine wie in diesem Fall vorliegende Sprachverarbeitung mittels MLPs bietet den Vorteil der Unabhängigkeit vom Kontext und Inhalt des gesprochenen Textes. Der einzeln zu klassifizierende Parametersatz eines Kurzzeitsegmentes steht während der Klassifikation in keinerlei Bezug zu seinen vorherigen und nachfolgenden Teilsegmenten. Eine Steigerung der Klassifikationsleistung könnte u. U. erreicht werden, wenn Kontextinformationen oder Metainformationen über den Inhalt des gesprochenen Textes mit in die Klassifikation einfließen würden. Diese Zusatzinformationen würden allerdings die Variabilität des Einsatzbereichs einschränken und die Komplexität der Verarbeitung deutlich erhöhen. Der zusätzliche Aufwand müsste in Bezug zur Steigerung der Erkennungsleistung bewertet werden, wobei eine Steigerung der Klassifikationsrate immer anzustreben ist.

Akustische Maße, die einen bestimmten Stimmstörungstyp bestmöglich beschreiben und differenzieren, sind nicht zwangsläufig auch die besten für einen anderen Stimmstörungstyp. Das Göttinger Heiserkeits-Diagramm hat durch die Verknüpfung mehrerer Einzelmaße eine höhere Robustheit und ermöglicht einen qualifizierten Einsatz für jegliche Form der Stimmgüte. Da die akustische Analyse fortlaufender Sprache bisher aus Mangel an zuverlässigen Klassifikationsmethoden nur eine Randstellung bei der Beurteilung von Stimmgüte einnimmt, sind entsprechend wenig spezielle akustische Maße für die Beschreibung fortlaufender Sprache entwickelt worden. Durch ein wachsendes Angebot an Klassifikationsmethoden eröffnet sich auch ein breiterer Anwendungsbereich.

Die Analyse fortlaufender Sprache stellt eine wesentliche Erweiterung einer umfassenden Beurteilung der Stimmgüte anhand von akustischen Maßen dar und bietet zum einen eine Validierungsmöglichkeit der erhaltenen Ergebnisse aus gehaltener Phonation, kann aber darüber hinaus auch wichtige Zusatzinformationen liefern, die bei der Phonation gehaltener Vokale nicht in Erscheinung treten. Eine akustische Analyse ist gegenüber einer perzeptiven Beurteilung durch ein Gutachterkollektiv meist objektiver, apparativ kostengünstiger, einfacher reproduzierbar und effizienter in der Durchführung. Die erhaltenen Ergebnisse anderer im Dritten Physikalischen Institut entwickelter Stimmanalysemethoden finden direkte Anwendung in der täglichen Arbeit der Abteilung Phoniatrie & Pädaudiologie, und auch die Analyse fortlaufender Sprache könnte dort Gewinn bringend eingesetzt werden.

7 Anhang

7.1 „Nordwind und Sonne“

Einst stritten sich Nordwind und Sonne,
wer von ihnen beiden wohl der Stärkere wäre,
als ein Wanderer,
der in einen warmen Mantel gehüllt war,
des Weges kam.

Sie wurden einig,
dass derjenige für den Stärkeren gelten sollte,
der den Wanderer zwingen würde,
seinen Mantel abzunehmen.

Der Nordwind blies mit aller Macht,
aber je mehr er blies,
desto fester hüllte sich der Wanderer
in seinen Mantel ein.
Endlich gab der Nordwind den Kampf auf.

Nun erwärmte die Sonne die Luft
mit ihren freundlichen Strahlen,
und schon nach wenigen Augenblicken
zog der Wanderer seinen Mantel aus.

Da musste der Nordwind zugeben,
dass die Sonne von ihnen beiden die Stärkere war.

7.2 „Buttergeschichte“

Es war in Berlin zu einer Zeit,
als Lebensmittel nicht genügend vorhanden waren.
Vor einem Laden stand bereits um sieben Uhr
eine beachtliche Menschenmenge,
denn man hatte dort am Abend vorher auf einem Schild schon lesen können,
dass frische Butter eingetroffen sei.

Jeder wusste, daß die Butter schnell ausverkauft sein würde
und dass man ganz früh kommen müsse, um noch etwas zu erhalten.
Da das Geschäft erst um acht geöffnet wurde,
stellten sich die Leute vor der Ladentür in einer Reihe an.
Wer später kam, musste sich hinten anschließen.

Je näher der Zeiger auf acht kam, desto unruhiger wurden die Leute.
Da kam endlich ein kleiner Mann mit grauem Haar
und drängte sich ziemlich rücksichtslos nach vorn.
Die wartenden Menschen waren empört über solches Verhalten
und forderten ihn auf, sich ebenfalls hinten anzustellen.

Aber auch als schon mit der Polizei gedroht wurde,
ließ sich der Mann nicht beirren, sondern drängte sich weiter durch.
Er bat, man solle ihn doch durchlassen, oder glaubte man,
dass diese Drängelei für ihn vielleicht ein Vergnügen sei?
Das war für die Leute nun doch zu viel! Alle kochten bereits vor Wut,
und der Mann konnte jetzt von allen Seiten Schimpfwörter hören.

Er aber zuckte resigniert mit den Schultern und bemerkte:
„Nun gut, wie Sie wollen. Wenn Sie mich nicht vorlassen,
dann kann ich die Tür nicht aufschließen,
und Sie können meinetwegen hier stehen bleiben,
bis die Butter ranzig geworden ist.“

7.3 „Regenbogen-Passage“

When the sunlight strikes raindrops in the air,
they act as a prism and form a rainbow.
The rainbow is a division of white light into many beautiful colors.
These take the shape of a long round arch, with its path high above,
and its two ends apparently beyond the horizon.

There is, according to legend, a boiling pot of gold at one end.
People look, but no one ever finds it.
When a man looks for something beyond his reach,
his friends say he is looking for the pot of gold at the end of the rainbow.

Throughout the centuries people have explained the rainbow in various ways.
Some have accepted it as a miracle without physical explanation.
To the Hebrews it was a token that there would be no more universal floods.
The Greeks used to imagine that it was a sign from the gods
to foretell war or heavy rain.
The Norsemen considered the rainbow as a bridge
over which the gods passed from earth to their home in the sky.

Others have tried to explain the phenomenon physically.
Aristotle thought that the rainbow was caused
by reflection of the sun’s rays by the rain.
Since then physicists have found that it is not reflection,
but refraction by the raindrops which causes the rainbows.
Many complicated ideas about the rainbow have been formed.

The difference in the rainbow depends considerably upon the size of the drops,
and the width of the colored band increases as the size of the drops increases.
The actual primary rainbow observed is said to be the effect
of super-imposition of a number of bows.
If the red of the second bow falls upon the green of the first,
the result is to give a bow with an abnormally wide yellow band,
since red and green light when mixed form yellow.
This is a very common type of bow, one showing mainly red and yellow,
with little or no green or blue.

7.4 Phonemliste

Die für das Training der Neuronolen Netze notwendige Segmentierung der Sprachaufnahmen erfolgt in der PHONDAT-Datenbank entsprechend den SAMPA-Konventionen. SAMPA (*Speech assessment methods phonetic alphabet*) ist als phonetisches Alphabet des ESPRIT-Projekts entstanden und u. a. in den Projekten PHONDAT und VERBMOBIL für das Deutsche adaptiert und ergänzt worden. Anhand einer leicht modifizierten Liste von Larry M. Hyman [H75] (Abweichungen mit * gekennzeichnet) sind die einzelnen Phoneme jeweils einer der beiden Klassen – *stimmhaft* oder *stimmlos* – zugeordnet worden.

stimmhaft:

Vokale:	a, a:, e, e:, E(ε), E:(ε:), i, i:, I, o, o:, O, u, u:, U, y, y:, Y, 2, 2:, 9, @, 6
Diphthonge:	AI, aI, aU, a6, E6, e6, I6, i6, OI, OY, O6, U6, y6, Y6,
Konsonanten:	j, l, r, v, z, Z
Nasale:	m, n, N

stimmlos:

Konsonanten:	b, d, g, f, h, k, p, q*, s, t, x, C, Q*, S

7.5 PHONDAT-Sprecherzuordnung

Zuordnung der PHONDAT-Sprachaufnahmen zur internen Nomenklatur mit den Bezeichnungen *Nordwind* und *Berlin*:

Phondat-Sprecher	Sex	interne Bezeichnung
ERL D 458 0	w	nordwind01
ESN D 458 0	w	nordwind02
HDB D 458 0	m	nordwind03
HEI D 458 0	m	nordwind04
HOR D 458 0	m	nordwind05
HSB D 458 0	m	nordwind06
JAN D 458 0	m	nordwind07
JEH D 458 0	m	nordwind08
LIN D 458 0	m	nordwind09
MXB D 458 0	w	nordwind10
OBL D 458 0	m	nordwind11
PTZ D 458 0	w	nordwind12
SPI D 458 0	m	nordwind13
WAG D 458 0	w	nordwind14
WEL D 458 0	w	nordwind15
WIN D 458 0	w	nordwind16
EGG D 459 0	w	berlin01
HEI D 459 0	m	berlin02
HUD D 459 0	m	berlin03
JAE D 459 0	w	berlin04
JNE D 459 0	w	berlin05
KPR D 459 0	w	berlin06
MAI D 459 0	w	berlin07
MXB D 459 0	w	berlin08
ROT D 459 0	m	berlin09
SCD D 459 0	m	berlin10
SEG D 459 0	w	berlin11
SFR D 459 0	w	berlin12
SMD D 459 0	w	berlin13
SMH D 459 0	m	berlin14
SWT D 459 0	m	berlin15
WEI D 459 0	w	berlin16

7.6 Phoniatrische Diagnose-Codes

Code	Bezeichnung	Code	Bezeichnung
0	normale Kehlkopffunktion	50	Z.n. partieller Chordektomie
1	hypofunktionelle Dysphonie	51	Z.n. subtotaler Chordektomie
2	hyperfunktionelle Dysphonie	52	Z.n. kompletter Chordektomie
3	psychosomatische Dysphonie	53	Z.n. endolaryng. Teilresektion
4	funktionelle Dysphonie	54	Z.n. partieller Taschenf.-Resekt.
5	Vorstadium Kontaktgranulom	55	Stimmlippen-Motilitätsstörung
6	Kontaktgranulom	56	Stimmlippen-Fixation
7	spasmodische Dysphonie	57	Z.n. Larynx-Teilresektion
8	spastische Dysphonie	58	Z.n. Kontroll-Mikrolaryngoskopie
9	zentrale Dysphonie	59	Z.n. Korrektur-Mikrolaryngosk.
10	Dysarthrophonie	60	akute Laryngitis
14	hormonelle Dysphonie	61	chronische Laryngitis
15	postoperative Dysphonie	62	Monochorditis
16	dysplastische Dysphonie	63	spezifische Laryngitis
20	Stimmlippen-Stillstand	64	subakute Laryngitis
21	Stl.-Lähmung paramedian	65	Reflux-Laryngitis
22	Stl.-Lähmung intermediär	66	Kehlkopf-Sarkoidose
23	Stl.-Lähmung (Vagus)	69	Mutation
24	Lähmung des M. cricothyr.	70	prolongierte Mutation
25	Lähmung N. laryng. superior	71	larvierte Mutation
26	Taschenfaltenaktivierung	72	inkomplette Mutation
27	Z.n. Medianverlagerung	73	Mutationsfistelstimme
28	Z.n. Glottiserweiterung	74	supraglott. Ersatzphon. (Sonderform)
29	bulbäre Stimmlippenlähmung	75	ventrikuläre Ersatzphonation
30	Stimmlippen-Knötchen	76	glottische Ersatzphonation
31	Phonationsverdickung	77	ary-epiglottische Ersatzphonation
32	Stimmlippen-Varix	78	glotto-ventrikuläre Ersatzphonation
33	Stimmlippen-Polyp	79	pseudo-glottische Ersatzphonation
34	Larynx-Papillomatose	80	traumatische Dysphonie
35	Laryngocele	81	Aryknorpel-Luxation
36	Stimmlippen-Cyste	82	Myopathie M. cricothyreoideus
37	Reinke-Oedem	83	Myasthenia gravis
38	Wundgranulation	84	Kehlkopf-Amyloidose
39	Intubationsgranulom	85	Turner-Syndrom
41	Sulcus glottidis	86	Transsexualität
42	Taschenfalten-Hyperplasie	87	M. Parkinson
43	Stimmlippen-Epidermoidcyste	90	PROVOX
44	intravocale Stimmlippencyste	93	Multiple Sklerose (MS)
45	glottischer Tumor	94	Amyotrophe Lateralsklerose (ALS)
46	Leukoplakie	95	Stimmlippen-Haematom
47	Hyperkeratose	96	Kehlkopf-Haemangiom
48	supraglottischer Tumor	97	intralaryngeale Synechie
49	Hypopharynx-Tumor	98	Aphonie
		99	unbekannte Diagnose

7.7 Phoniatrische Diagnose-Code-Zusätze

Code-Zusatz	Bezeichnung
A	Z. n. Mikrolaryngoskopie (MLE)
B	Wundheilungsphase
C	Abschluß Wundheilung
D	Kontrolle nach Wundheilung
E	bilateral
F	rechts
G	links
H	Beginn Therapie/Reha
I	während Therapie/Reha
K	Abschluß Therapie/Reha
L	Kontrolle nach Abschluß Therapie/Reha
M	postoperativ nach Wundheilung (EP)
N	Kontrolle postoperativ
O	inkomplett
P	aktive Restbeweglichkeit
Q	Wiederbeweglichkeit
R	Verdacht auf
S	zur Beobachtung
T	Differentialdiagnose (DD)
U	nicht differenzierbar
V	Z. n. Rezidiv-Operation
W	Z. n. Radiatio
Z	Z. n. Therapie auswärts

Literaturverzeichnis

[AH71] B. S. Atal, S. L. Hanauer: *Speech Analysis and Synthesis by Linear Prediction of the Speech Wave*; J. Acoust. Soc. Am. **50**, 637 – 655 (1971)

[AH86] A. Askenfelt, B. Hammarberg: *Speech waveform perturbation analysis: A perceptual-acoustical comparison of seven measures*; J. Speech Hear. Res. **29**, 50 – 64 (1986)

[AS70] B. S. Atal, M. R. Schroeder: *Adaptive predictive coding of speech signals*; Bell Sys. Tech. J. (1970)

[B58] J. W. van den Berg: *Myoelastic-aerodynamic theory of voice production*; J. Speech Hear. Res. **1**, 227 – 244 (1958)

[B87] R. J. Baken: *clinical Measurement of Speech and Voice*; Boston, College Hill Press (1987)

[BGS05] F. Bettens, F. Grenez, J. Schoentgen: *Estimation of vocal dysperiodicities in disordered connected speech by means of distant-sample bidirectional linear predictive analysis*; J. Acoust. Soc. Am. **117**, 328 – 337 (2005)

[BKGD96] S. Bielamowicz, J. Kreiman, B. R. Gerratt, M. S. Dauer, G. S. Berke: *Comparison of voice analysis systems for perturbation measurement*; J. Speech and Hear. Res. **39**, 126 – 134 (1996)

[BO00] R. J. Baken, R. F. Orlikoff: *Clinical Measurement of Speech and Voice*; San Diego, CA (2000)

[C71] R. F. Coleman: *Effect of waveform changes upon roughness perception*; Folia Phoniatrica **23**, 314 – 322 (1971)

[CKRT99] D. E. Callan, R. D. Kent, N. Roy, S. M. Tasko : *Self-Organizing Map for the Classification of Normal and Disordered female voices*; J. Speech and Hear. Res. **42**, 355-366 (1999)

[CL91] D. Childers, C. Lee: *Vocal quality factors: Analysis, synthesis, and perception*; J. Acoust. Soc. Am. **90**, 2394 – 2410 (1991)

[CSAL71] F. S. Cooper, M. Sawashima, A. S. Abramson, L. Lisker: *Looking at the Larynx during Running Speech*; Ann. Otol. **80**, 678 – 682 (1971)

[D76] S. B. Davis: *Computer evaluation of laryngeal pathology based on inverse filtering of speech*; SCRL Monograph Number 13, Speech Communications Research Laboratory, Santa Barbara, CA (1971)

[D81] S. B. Davis: *Acoustic characteristics of normal and pathological voices*; Proceedings of the Conference on the Assessment on Vocal Pathology, ASHA Report **11**, 97 – 112 (1981)

[DMKM89] F. F. Deem, W. H. Manning, J. V. Knack, J. S. Matesich: *The automatic extraction of pitch perturbation using microcomputers: Some methodological considerations*; J. Speech and Hear. Res. **32**, 689 – 697 (1989)

[DPH93] J. R. Deller, J. G. Proakis, J. H. L. Hansen: *Discrete-time processing of speech signals*; MacMillian Publishing, New York (1993)

[ECH90] L. Eskenazi, D. G. Childers, D. M. Hicks: *Acoustic correlates of vocal quality*; J. Speech and Hear. Res. **33**, 298 – 306 (1990)

[ED05] T. L. Eadie, P. C. Doyle: *Classification of dysphonic voice: acoustic and auditory-perceptual measures*; J. Voice **19** 11, 1 – 14 (2005)

[F57] P. Fabre: *Un procède électrique parcutane d'inscription de l'accolement glottique au cours de la phonation: glottographie de haute fréquence; premiers résultats*; Bull Acad Nat Med **3–4**, 66 – 69 (1957)

[F60] G. Fairbanks: *Voice and Articulation Drillbook*; Harper & Row, 2nd Edition, New York (1960)

[F81] G. Fant: *The source filter concept in voice production*; STL-QPSR **1** 21 – 37 (1981)

[FFHSKK06] M. Fuchs, M. Fröhlich, B. Hentschel, I. W. Stuermer, E. Kruse, D. Knauft: *Predicting Mutational Change in the Speaking Voice of Boys*; J. Voice **11** (2006)

[FMK98] M. Fröhlich, D. Michaelis, E. Kruse: *Objektive Beschreibung der Stimmgüte unter Verwendung des Heiserkeits-Diagramms*; HNO **46**, 684 – 689 (1998)

[FMLSK00] M. Fröhlich, D. Michaelis, J. Lessing, H. W. Strube, E. Kruse: *Breathiness measures in acoustic voice analysis*; Advances in Quantitative Laryngoscopy, Voice and Speech Research, Proc. 4th International Workshop, Jena (2000)

[FMLSK01] M. Fröhlich, D. Michaelis, J. Lessing, H. W. Strube, E. Kruse: *Inverse filtering and simultaneous model matching*; Advances in Quantitative Laryngoscopy, Voice and Speech Research, Proc. 5th International Workshop, Groningen (2001)

[FMSK97] M. Fröhlich, D. Michaelis, H. W. Strube, E. Kruse: *Case studies for different regions of the hoarseness diagram*; Advances in Quantitative Laryngoscopy, Erlangen, 143 – 150 (1997)

[FMSK98] M. Fröhlich, D. Michaelis, H. W. Strube, E. Kruse: *Stimmgütebeschreibung mit Hilfe des Heiserkeits-Diagramms: Untersuchungen verschiedener pathologischer Gruppen*; Aktuelle phoniatrisch-pädaudiologische Aspekte, Median Verlag Heidelberg (1998)

[FMSK00] M. Fröhlich, D. Michaelis, H. W. Strube, E. Kruse: *Acoustic voice analysis by means of the hoarseness diagram*; J. Speech, Lang. and Hear. Res. **43**, 706 – 720 (2000)

[FSK97] M. Fröhlich, H. W. Strube, E. Kruse: *Akustische Parameter zur Stimmgütebeschreibung aus fortlaufender Sprache*; Aktuelle phoniatrisch-pädaudiologische Aspekte, Median Verlag Heidelberg, 22 – 24 (1997)

[GG04] J. I. Godino-Llorente, P. Gomez-Volda: *Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors*; IEEE Trans. Biomed. Eng **51/2**, 380 – 384 (2004)

[GMZ80] R. Gubrynowicz, W. Mikiel, P. Zarnecki: *An acoustic method for the evaluation of the state of the larynx source in cases involving pathological changes in the vocal folds*; Archives of Acoustics **5**, 3 – 30 (1980)

[H63] J. N. Holmes: *The effect of simulating natural larynx behaviour of the quality of synthetic speech*; in G. Fant: Speech Communication Seminar,

Speech Transmission Laboratory, Royal Institute of Technology Stockholm (1963)

[H65] H. Hollien: *Stroboscopic laminagraphy of the vocal folds*; Proc. 5th Intl. Cong. Phon. Sci. Munster, Basel, New York, Karger, 362 – 364 (1965)

[H75] L. M. Hyman: *Phonology, Theory and Analysis*; Holt, Rinehart and Winston (1975)

[H81] M. Hirano: *Psycho-Acoustic Evaluation of Voice: GRBAS Scale for Evaluating the Hoarse Voice*; Clinical Examination of Voice, Springer Verlag Wien, 81 – 84 (1981)

[H83] W. J. Hess: *Pitch Determination on Speech Signals: Algorithms and Devices*; Springer Verlag Berlin, Heidelberg, New York (1983)

[H91] R. Hecht-Nielsen: *Neurocomputing*; Addison-Wesley (1991)

[H95] W. J. Hess: *Determination of Glottal Excitation Cycles in Running Speech*; Phonetica **52**, 196 – 204 (1995)

[HFGS80] B. Hammarberg, B. Fritzell, J. Gauffin, J. Sundberg: *Perceptual and acoustic correlates of abnormal voice qualities*; Acta Otolaryngol. **90**, 441 – 451 (1980)

[HK71] M. H. L. Hecker, E. J. Kreul: *Descriptions of the speech of patients with cancer of the vocal folds. Part I: Measures of fundamental frequency*; J. Acoust. Soc. Am. **49**, 1275 – 1282 (1971)

[HMD73] H. Hollien, J. Michael, E. T. Doherty: *A method for analyzing vocal jitter in sustained phonation*; Journal of Phonetics **1**, 85 – 91 (1973)

[IL70] S. Iwata, H. von Leden: *Pitch perturbation in normal and pathological voices*; Folia Phoniatrica **22**, 413 – 424 (1970)

[IPA49] International Phonetic Association: *The Principles of the International Phonetic Association*; International Phonetic Association (1949)

[IS69] F. Itakura, S. Saito: *Speech Analysis-Synthesis System Based on the Partial Autocorrelation Coefficient*; Acoust. Soc. of Japan Meeting (1969)

[J64] E. I. Jury: *Theory and Application of the Z-Transform Method*; John Wiley and Sons, New York (1964)

[K71] Y. Koike: *Application of Some Acoustic Measures for the Evaluation of Laryngeal Dysfunction*; Studia Phonologica (Kyoto University) **7**, 45 – 50 (1971)

[K87] F. Klingholz: *The measurement of the signal-to-noise-ratio (SNR) in continuous speech*; J. Speech Comm. **6**, 15 – 26 (1987)

[K90] F. Klingholz: *Acoustic recognition of voice disorders: A comparative study of running speech versus sustained vowels*; J. Acoust. Soc. Am. **87**, 2218 – 2224 (1990)

[K92] Kay Elemetrics Corp: *Computerized Speech Lab Operations Manual*; Pine Brook, NJ (1992)

[K93] G. de Krom: *A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals*; J. Speech Hear. Res. **36**, 224 – 266(1993)

[K95] G. de Krom: *Some spectral correlates of pathological breathy and rough voice quality for different types of vowel fragments*; J. Speech Hear. Res. **38**, 794 – 811 (1995)

[KES93] H. Kasuya, Y. Endo, S. Saliu: *Novel acoustic measurements of jitter and shimmer characteristics from pathological voice*; Eurospeech **3**, 1973 – 1976 (1993)

[KOME86] H. Kasuya, S. Ogawa, K. Mashima, S. Ebihara: *Normalized noise energy as an acoustic measure to evaluate pathologic voice*; J. Acoust. Soc. Am. **80**, 1329 – 1334 (1986)

[KFM98] E. Kruse, M. Fröhlich, D. Michaelis: *Phonatory conditions and acoustic analysis of pathologic voices. Is there a correspondence?*; 24th IALP, Amsterdam 127 (1998)

[KL03] E. Kruse, J. Lessing: *The Göttinger hoarseness diagram*; Workshop WK1/18 Pan European Voice Conference PeVOC V, Graz Österreich (2003)

[KM75] Y. Koike, J. Markel: *Application of inverse filtering for detecting laryngeal pathology*; Annals of Otology, Rhinology and Laryngology, **84**, 117 – 124 (1975)

[KW79] H. Kasuya, H. Wakita: *An approach to segmenting speech into vowel- and nonvowel-like intervals*; IEEE Transactions on Acoustics Speech and Signal Processing **27**, 319 – 327 (1979)

[L61] P. Liebermann: *Perturbation in vocal pitch*; J. Acoust. Soc. Am. **33**, 597 – 603 (1961)

[L63] P. Liebermann: *Some acoustic measures fo the fundamental periodicity of normal and pathologic larynges*; J. Acoust. Soc. Am. **35**, 344 – 353 (1963)

[L87] R. P. Lippmann: *An Introduction to Computing with Neural Nets*; IEEE ASSP Magazine April 1987, 4 – 22 (1987)

[L98] J. Lessing: *Methoden zur Detektion stimmhafter Phoneme in fortlaufender Sprache unterschiedlicher Stimmgüte*; Diplomarbeit, Göttingen (1998)

[LFMS99] J. Lessing, M. Fröhlich, D. Michaelis, H. W. Strube, E. Kruse: *Akustische Stimmanalyse aus fortlaufender Sprache – Untersuchung von Tumorgruppen*; Aktuelle phoniatrisch-pädaudiologische Aspekte 1998/99, M. Gross, Median Verlag Heidelberg, 126 – 130 (1999)

[LFMS99b] J. Lessing, M. Fröhlich, D. Michaelis, H. W. Strube, E. Kruse: *Verwendung Neuronaler Netze zur Stimmgütebeschreibung pathologischer Stimmen*; Aktuelle phoniatrisch-pädaudiologische Aspekte 1998/99, M. Gross, Median Verlag Heidelberg, 39 – 43 (1999)

[LFMSK00] J. Lessing, M. Fröhlich, D. Michaelis, H. W. Strube, E. Kruse: *A Neural Network Based Method to assess Voice Quality from Continuous Speech for Different Voice Disorders*; Proc. of the 4th Int. Workshop: Advances in Quantitative Laryngoskopy, Voice and Speech Research, 124 – 131 (2000)

[LH94] W. E. Lucas III, W. Hudson: *An adaptive algorithm for the automatic segmentation of continuous stuttered speech*; ISA 94-098 (1994)

[LMFK01] J. Lessing, D. Michaelis, M. Fröhlich, E. Kruse: *A Neural Network Based Method to Assess Voice Quality from Continuous Speech for Different Voice Disorders*; Pan-European Voice Conference PeVOC IV, Stockholm, Schweden (2001)

[LSK98] J. Lessing, H. W. Strube, E. Kruse: *Akustische Analyse pathologischer Stimmen aus fortlaufender Sprache*; Aktuelle phoniatrisch-pädaudiologische Aspekte 1997/98, M. Gross, Median Verlag Heidelberg, 53 – 59 (1998)

[M71] J. D. Markel: *Formant Trajctory Estimation from a Linear Least-Squares Inverse Filter Formulation*; SCRL Monograph **7**, Speech Communication Research Laboratory, Santa Barbara, California (1971)

[M75] J. Makhoul: *Linear Prediction: A Tutorial Review*; Proceedings of the IEEE **63**, 561 – 580 (1975)

[M87] P. Milenkovic: *Least mean square measures of voice perturbation*; J. Speech and Hear. Res. **30**, 529 – 538 (1987)

[MBWMF88] H. Muta, T. Baer, K. Wagatsuma, T. Muraoka, H. Fukuda: *A pitch-synchronous analysis of hoarseness in running speech*; J. Acoust. Soc. Am. **84**, 1292 – 1301 (1988)

[MD80] T. Murry, E. T. Doherty: *Selected acoustic characteristics of pathologic and normal speakers*; J. Speech and Hear. Res. **23**, 361 – 369 (1980)

[MFLK01] D. Michaelis, M. Fröhlich, J. Lessing, E. Kruse: *Leaf: Ein (Near) Real Time System zur akustischen Stimmanalyse*; Aktuelle phoniatrisch-pädaudiologische Aspekte 2000/01, M. Gross, Median Verlag Heidelberg, **8**, 61 – 64 (2001)

[MFS98] D. Michaelis, M. Fröhlich, H. W. Strube: *Selection and combination of acoustic parameters for the description of pathologic voices*; J. Acoust. Soc. Am. **103**, 1628 – 1639 (1998)

[MG76] J. D. Markel, A. H. Gray jr. : *Linear Prediction of Speech*; Springer-Verlag Berlin, Heidelberg, New York (1976)

[MGS97] D. Michaelis, T. Gramss, H. W. Strube: *Glottal to noise excitation ratio – a new measure for describing pathological voices*; Acustica / acta acustica **83**, 700 – 706 (1997)

[MKH68] F. D. Minifie, C. A. Kelsey, T. J. Hixon: *Measurement of vocal fold motion using an ultrasonic Doppler velocity monitor*; J. Acoust. Soc. Am. **43**, 1165 – 1169 (1968)

[MP43] W. S. McCulloch, W. Pitts: *A Logical Calculus of the Ideas Immanent in Neural Nets*; Bull. Math. Biophys **5**, 115 – 133 (1943)

[MP69] M. Minsky, S. Papert: *Perceptrons*; The MIT Press, Cambridge (Mass.) (1969)

[MR90] B. Müller, J. Reinhardt: *Neural Networks: An Introduction*; Springer-Verlag Berlin, Heidelberg, New York (1990)

[MS95] D. Michaelis, H. W. Strube: *Empirical study to test the independence of different acoustic voice parameters on a large voice database*; Eurospeech 95 **3**, 1891 – 1894 (1995)

[MSZK94] D. Michaelis, H. W. Strube, P. Zwirner, E. Kruse: *Frequenzabhängige Korrelationen der Stimmschallanregung als akustisch-diagnostischer Stimmgüteparameter*; Aktuelle phoniatrisch-pädaudiologische Aspekte 1994, M. Gross, Median Verlag Heidelberg, **2**, 128 (1994)

[MW72] J. Makhoul, J. Wolf: *Linear Prediction and the Spectral Analysis of Speech*; NTIS No. AD-749066, BBN Report No. 2304, Bolt Beranek and Newmann Inc., Cambridge, Massachusetts (1972)

[MYC91] Y. Medan, E. Yair, D. Chazan: *Super Resolution Pitch Determination of Speech Signals*; IEEE Trans. Sig. Process. **39**, 40 – 48 (1991)

[N64] A. M. Noll: *Short-time spectrum and cepstrum techniques for vocal-pitch detection*; J. Acoust. Soc. Am. **36**(2), 296 – 302 (1964)

[N67] A. M. Noll: *Cepstrum Pitch Determination*; J. Acoust. Soc. Am. **41**, 293 – 309 (1967)

[NAW94] T. Nawka, L. C. Anders, J. Wendler: *Die auditive Beurteilung heiserer Stimmen nach dem RBH-System*; Sprache Stimme Gehör 18 (1994)

[NS64] A. M. Noll, M. R. Schroeder: *Short-time cepstrum pitch detection*; J. Acoust. Soc. Am. **36**(2), 1030 (1964)

[OB89] R. F. Orlikoff, R. J. Baken: *The Effect of the Heartbeat on Vocal Fundamental Frequency Perturbation*; J. Speech and Hear. Res. **32**:3, 576 – 582 (1989)

[ON92] A. van Ooyen, B. Niehaus: *Improving the Convergence of the Back-propagation Algorithm*; Neural Networks, **5**, 465 – 471 (1992)

[P93] PHONDAT Sprachdatenkorpora auf CD-ROM; Institut für Phonetik und Sprachliche Kommunikation, Schellingstr. 3/II, D-80799 München (1993)

[PFTV86] W. H. Press, B. P. Flannery, S. A. Teukolsky, W. T. Vetterling: *Numerical Recipes: The Art of Scientific Computing*; Cambridge University Press (1986)

[PHW97] R. L. Plant, A. D. Hillel, P. F. Waugh: *Analysis of voice changes after thyroplasty using linear predictive coding*; Laryngoscope **107**, 703 – 709 (1997)

[PJ99] V. Parsa, D. G. Jamieson: *A Comparison of High Precision F0 Extraction Algorithms for Sustained Vowels*; J. Speech and Hear. Res. **42**, 112 – 126 (1999)

[PJ00] V. Parsa, D. G. Jamieson: *Identification of Pathological Voices Using Glottal Noise Measures*; J. Speech, Lang., and Hear. Res. **43**, 469 – 485 (2000)

[PJ01] V. Parsa, D. G. Jamieson: *Acoustic Discrimination of Pathological Voice: Sustained Vowels Versus Continuous Speech*; J. Speech, Lang., and Hear. Res. **44**, 327 – 339 (2001)

[PMWH87] R. A. Prosek, A. A. Montgomery, B. E. Walden, D. N. Hawkins: *An Evaluation of Residue Features as Correlates of Voice Disorders*; J. Commun. Disord. **20**, 105 – 117 (1987)

[PSIJN06] M. Ptok, C. Schwemmle, C. Iven, M. Jessen, T. Nawka: *On the auditory evaluation of voice quality*; HNO **54** 10, 793 – 802 (2006)

[PT90] Pinto, I. R. Titze: *Unification of perturbation measures in speech signals*; J. Acoust. Soc. Am. **87**, 1278 – 1289 (1990)

[QH93] Y. Qi, B. R. Hunt: *Voiced-unvoiced-silence classification of speech using hybrid features and a network classifier*; IEEE Transactions on Speech and Audio Processing **1**, 250 – 255 (1993)

[QH97] Y. Qi, R. E. Hillman: *Temporal and spectral estimations of harmonics-to-noise ratio in human voice signals*; J. Acoust. Soc. Am. **71**, 537 – 543 (1997)

[QHM99] Y. Qi, R. E. Hillman, C. Milstein: *The estimation of signal-to-noise ratio in continuous speech for disordered voices*; J. Acoust. Soc. Am. **105**, 2532 – 2535 (1999)

[R59] F. Rosenblatt: *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*; Spartan Books, New York (1959)

[RHW86] D. E. Rumelhart, G. E. Hinton, R. J. Williams: *Learning Representations by Back-Propagating Errors*; Nature **323**, 533 – 536 (1986)

[RK89] R. P. Ramachandran, P. Kabal: *Pitch prediction filters in speech coding*; IEEE Trans. Acoust., Speech, Signal Process. **37**, 467 – 478 (1989)

[RLM97] J. Rouat, Y. C. Liu, D. Morissette: *A pitch determination and voiced/unvoiced decision algorithm for noisy speech*; Speech Commun. **21**, 191 – 207 (1997)

[RS77] R. W. Rabiner, M. Sambur: *Application of an LPC distance measure to the Voiced-Unvoiced-Silence detection problem*; IEEE Trans. Acoust., Speech, Signal Process. **25**, 338 – 343 (1977)

[RS78] R. W. Rabiner, L. R. Schafer: *Digital Processing of Speech Signals*; Prentice-Hall, Inc., Englewood Cliffs, New Jersey (1978)

[S34] M. Swadesh: *The phonemic principle*; Language **10**, 117 – 129 (1934)

[S60] V. Sonesson: *On the anatomy and vibratory patterns of the human vocal folds*; Acta Oto-laryng. Suppl. **156** (1960)

[S68] M. R. Schroeder: *Period histogram and product spectrum: New methods for fundamental-frequency measurement*; J. Acoust. Soc. Am. **43**, 829 – 834 (1968)

[S68a] M. M. Sondhi: *New Methods of Pitch Extraction*; IEEE Trans. Audio and Electroacoustics **16**, 262 – 266 (1968)

[S81] M. R. Schroeder: *Direct (nonrecursive) relations between cepstrum and predictor coefficients*; IEEE Trans. Acoust., Speech, Sig. Process. **ASSP-29** 297 – 301 (1986)

[S83] K. Sickert: *Automatische Spracheingabe und Sprachausgabe*; Verlag Markt & Technik (1983)

[S89] J. Schoentgen: *Jitter in sustaines vowels and isolated sentences produced by dysphonic speakers*; Speech Commun. **8**, 61 – 79 (1989)

[S00] O. Schreiner: *Ein Vergleich verschiedener Frequenzzerlegungen zur Modulationsfilterung von Sprache*; Diplomarbeit, Göttingen (2000)

[S04] M. R. Schroeder: *Computer Speech – Recognition, Compression, Synthesis*; Springer 2nd Edition, Berlin Heidelberg New York (2004)

[SA85] M. R. Schroeder, B. S. Atal: *Code-Excited Linear Prediction (CELP): High Quality Speech at Very Low Bit Rates*; Proc. ICASSP-85, Tampa, 937 – 940 (1985)

[SB82] L. Siegel, A. Bessey: *Voiced/Unvoiced/Mixed excitation classifcation of speech*; IEEE Trans. Acoust., Speech, Signal Processing **30** 451 – 460 (1982)

[SG91] J. Schoentgen, R. de Guchteneere: *An algorithm for the measurement of jitter*; Speech Commun. **10**, 533 – 538 (1991)

[SG95] J. Schoentgen, R. de Guchteneere: *Time series analysis of jitter*; Journal of Phonetics **23**, 189 – 201 (1995)

[SH94] S. D. Stearns, D. R. Hush: *Digitale Verarbeitung analoger Signale*; R. Oldenbourg Verlag München Wien (1994)

[SMLA03] H. W. Strube, D. Michaelis, J. Lessing, S. Anderson: *Akustische Analyse pathologischer Stimmen in fortlaufender Sprache* in Fortschritte der Akustik, DAGA 03, 760 – 761 (2003)

[SQ02] O. Schreiner, H. Quast: *Grundfrequenzbestimmung aus dem Modulationsspektrum*; *Fortschritte der Akustik (DAGA)*, DEGA, Berlin (2002)

[SS96] J. G. Svec, H. K. Schutte: *Videokymography: high-speed line scanning of vocal fold vibration*; Journal of Voice **10**, 201 – 205 (1996)

[T94] I. R. Titze: *Principles of Voice Production*; Prentice Hall, New Jersey (1994)

[T94a] I. R. Titze: *Toward standards in acoustic analysis of voice*; J. of Voice **8**, 1 – 7 (1994)

[TK75] H. Takahashi, Y. Koike: *Some perceptual dimensions and acoustic correlates of pathological voices*; Acta Otolaryngologica, **338**, 2 – 24 (1975)

[TL87] H. Traunmüller, F. Lacerda: *Perceptual relativity in identification of two-formant vowels*; Speech Commun. **6**, 143 – 157 (1987)

[TL92] I. R. Titze, H. Liang: *Comparison of F0 extraction methods for high precision voice perturbation measurement*; NCVS Status and Progress Report **3**, 97 – 115 (1992)

[TL93] I. R. Titze, H. Liang: *Comparison of F0 extraction methods for high-precision voice perturbation measurements*; J. Speech and Hear. Res. **36**, 1120 – 1133 (1993)

[UKPJ05] K. Umapathy, S. Krishnan, V. Parsa, D. G. Jamieson: *Discrimination of Pathological Voices Using a Time-Frequency Approach*; IEEE Transactions on Biomedical Engineering **52** No. 3, 421 – 430 (2005)

[VFMD93] J. Verstraete, G. Forrez, P. Mertens, F. Debruyne: *The effect of sustained phonation at high and low pitch on vocal jitter and shimmer*; Fol. Phoniatrica **45**, 223 – 228 (1993)

[W66] N. Wiener: *Extrapolation Interpolation and Smoothing of Stationary Time Series*; M. I. T. Press, Cambridge, Massachusetts (1966)

[W73] H. Wakita: *Direct Estimation of the Vocal Tract Shape by Inverse Filtering of Acoustic Speech Waveforms*; IEEE Trans. on Audio and Electroacoustics **21**, 417 – 427 (1973)

[W74] P. J. Werbos: *Beyond Regression: New Tools for Prediction and Analysis in the Behavioural Science*; PhD Arbeit, Harvard Universität (1974)

[WRK86] J. Wendler, A. Rauhut, H. Krüger: *Classification of voice qualities*; Journal of Phonetics **14**, 483 – 488 (1986)

[WSDHEL06] T. Wurzbacher, R. Schwarz, M. Dollinger, E. Hoppe, U. Eysholdt, J. Lohscheller: *Model-based classification of nonstationary vocal fold vibrations*; J. Acoust. Soc. Am. **120**, 1012 – 1027 (2006)

[Y67] N. Yanagihara: *Significance of harmonic changes and noise components in hoarseness*; J. Speech Hear. Res., **10**, 531 – 541 (1967)

[YGB82] E. Yumoto, W. J. Gould, T. Baer: *The harmonics-to-noise ratio as an index of the degree of hoarseness*; J. Acoust. Soc. Am. **71**, 1544 – 1550 (1982)

[Z82] E. Zwicker *Psychoakustik*; Springer-Verlag Berlin Heidelberg New York (1982)

[Z86] D. E. Zimmer: *So kommt der Mensch zur Sprache*; Haffmans Verlag (1986)

Danksagung

Danken möchte ich zuallererst meinen Eltern, Tatjana und meinem Bruder für die große Unterstützung in allen Belangen.

Für die wissenschaftliche Betreuung dieser Arbeit gilt mein Dank insbesondere Prof. Manfred R. Schroeder für die Möglichkeit der Promotion, der mit seiner persönlichen Art die Arbeitsgruppe seit Jahren prägt und bereichert. Des Weiteren gilt Prof. Eberhard Kruse – ohne den dieses Projekt nicht zustande gekommen wäre – großer Dank, der mich über den gesamten Zeitraum unserer gemeinsamen Arbeit immer unterstützt hat und keine Mühen und Mittel gescheut hat, den gemeinsamen Kooperationsprojekten des Dritten Physikalischen Instituts und der Abteilung Phoniatrie & Pädaudiologie zum Erfolg zu verhelfen. Nicht minderer Dank gilt Dr. Hans Werner Strube, der die kontinuierlichste Größe in der Abteilung *Sprache und Neuronale Netze* des Dritten Physikalischen Instituts darstellt und mit seiner fachlichen Kompetenz und seinem Rat unersetzbar ist.

Aus der Arbeitsgruppe möchte ich insbesondere Dirk und Matthias danken, die das Projekt in weiten Teilen begleitet haben und mir immer mit Rat und Hilfe zur Seite standen. In der Forschungsgruppe *Stimme und Sinnesentwicklung* habe ich mit Ingo, Iris, Sven, Cordula, Thorsten und Corinna Freunde gefunden, die mich einen Großteil meiner wissenschaftlichen Zeit begleitet haben und die ich nicht missen möchte. Mein Dank gilt nicht zuletzt auch Knut, Olaf, Heiko und Joachim und dem gesamten Dritten Physikalischen Institut in Göttingen.

Gerade für die letzte, sehr intensive Zeit an dieser Arbeit möchte ich Holk danken, der mich immer wieder ermuntern konnte und mit seinem fachlichen und persönlichen Rat mir jederzeit zur Seite stand.

Jan Lessing

Index

Lebenslauf

Jan Lessing

Geboren am 10. Juni 1972 in Göttingen

Eltern: Prof. Dr. Volker Lessing und Roswitha Lessing (geb. Bode)

Staatsangehörigkeit: deutsch

Schulbildung

1978 – 1991	Grundschule Wilhelm Busch, Hannover Orientierungsstufe Martensplatz, Hannover Gymnasium Humboldtschule, Hannover
Juni 1991	Abitur Humboldtschule, Hannover Leistungskurse: Mathematik und Physik

Studium, Promotion und wissenschaftliche Tätigkeit

Sep. 1991 - Jun. 1998	Physikstudium Diplom, Universität Göttingen
Juni 1998	Diplom im Dritten Physikalischen Institut (DPI), Universität Göttingen
Sep. 1998 – Sep. 1999	Zivildienst im Universitätsklinikum Göttingen Abteilung Röntgendiagnosik I
Okt. 1999 – Aug. 2000	Wissenschaftlicher Mitarbeiter im Projekt: *„Entwicklung einer objektiven Bewertung der pathologischen Stimmqualität in natürlichsprachlicher Situation (fortlaufende Sprache) für die klinische Praxis“* der Deutschen Forschungsgemeinschaft
Aug. 2000 – Mai 2001	Wissenschaftlicher Mitarbeiter, Universitätsklinikum Göttingen Abteilung HNO sowie Phoniatrie & Pädaudiologie
Sep. 2001 – Dez. 2005	Wissenschaftlicher Mitarbeiter, Uniklinikum Göttingen Abteilung Klinische Neurophysiologie
Okt. 2002 – Mrz. 2006 Mrz. 2007 –	Promotionsstudiengang Physik, Universität Göttingen Promotions-Betreuer: Prof. M. R. Schroeder