Ein sprachgestütztes Trainingssystem zur Evaluierung der Nasalität

Inauguraldissertation
zur Erlangung des akademischen Grades
eines Doktors der Naturwissenschaften
der Universität Mannheim

vorgelegt von

Dipl.-Wirtsch.-Inf. Anto Zečević

aus Drenovci / Kroatien

Mannheim, 2002

Dekan: Professor Dr. Herbert Popp, Universität Mannheim

Referent: Professor Dr. Reinhard Männer, Universität Mannheim

Korreferent: Professor Dr. Wolfgang Effelsberg, Universität Mannheim

Tag der mündlichen Prüfung: 18. Juli 2003

Für Andrea

Vorwort

Die vorliegende Arbeit entstand während meiner Tätigkeit als wissenschaftlicher Mitarbeiter am Lehrstuhl für Informatik V der Universität Mannheim. Sie hat ihren Ursprung in dem Forschungsvorhaben „Sprachliche Rehabilitation von Lippen-Kiefer-Gaumenspalt-Patienten“ der Medizinischen Fakultät Heidelberg, das gemeinsam von Prof. Dr. Gerda Komposch und Prof. Dr. Reinhard Männer beantragt wurde. Beiden danke ich sehr herzlich.

Danke sage ich insbesondere Prof. Dr. Reinhard Männer für die ständige tatkräftige Unterstützung sowohl bei allen wissenschaftlichen als auch organisatorischen Fragen, ohne welche die vorliegende Arbeit nicht möglich gewesen wäre.

Prof. Dr. Wolfgang Effelsberg danke ich herzlich für die Übernahme des Korreferats.

Ganz spezieller Dank gilt Dr. Steffen Noehte, der mich während der gesamten Zeit betreute und sich stets weit über das selbstverständliche Maß hinweg für mich eingesetzt hat. Auf seine zahlreichen Ratschläge, seine hervorragende Unterstützung in fachlicher und organisatorischer Hinsicht sowie sein Vertrauen durfte ich jederzeit zählen.

Bedanken möchte ich mich bei der Mund-Zahn-Kieferklinik der Universität Heidelberg für die tatkräftige Unterstützung. Dabei bin ich vor allem Dr. Angelika Stellzig-Eisenhauer sehr zu Dank verpflichtet. Sie hat dazu beigetragen, dass die Kooperation mit der Universität Mannheim möglich war, und mich jederzeit bei medizinischen Fragestellungen unterstützt. Weiterhin danke ich allen Logopäd(inn)en für ihre Bereitschaft zur Aufnahme der Sprachdaten. Insbesondere gilt mein Dank Frau Strate und Frau Becker für die Bewertung der Sprachdatenbank NASAL hinsichtlich der Nasalität.

Bei Prof. Dr. Steidl und Dr. Tanja Karp bedanke ich mich für die Unterstützung bei Fragen zur Signalverarbeitung. Prof. Dr. Stenger danke ich für die Ratschläge zur Linearen Diskriminanzanalyse.

Danken möchte ich meinem Studien- und Diplomarbeiter Holger Kögel, dessen Ergebnisse zum Fortschritt der Trainingsumgebung beitrugen.

Dr. Tanja Karp, Dr. Rainer Himmeröder, Matthias Gerspach und meine Frau Andrea haben als kritische Korrekturleser sehr zum Verständnis der Arbeit beigetragen. Vielen Dank dafür.

Bedanken möchte ich mich auch bei allen Kolleginnen und Kollegen am Lehrstuhl für Informatik V. Die vielen fruchtbaren Diskussionen und die intensiven Gespräche haben meinen Horizont stetig erweitert und mir bewusst gemacht, welches Privileg es ist, in einem solchen Umfeld arbeiten zu dürfen.

Danke sage ich auch der Sekretärin unseres Lehrstuhles, Andrea Seeger, für die tatkräftige Unterstützung bei vielen organisatorischen Fragen.

Mein privates Umfeld musste oft darunter leiden, weil meine Prioritäten zu Gunsten meiner Promotion verlagert waren. Für das Verständnis, das mir meine Freunde und meine Familie dabei entgegenbrachten, danke ich allen.

Meinen innigsten Dank widme ich abschließend meiner Frau Andrea. Sie musste auf so manches verzichten und mich ertragen, wenn ein Problem mal wieder schwieriger erschien, als es vielleicht tatsächlich war. Sie hat es dabei geschafft, mich immer wieder aufzubauen, und half mir dadurch auch, schwierige Phasen zu überstehen. Insbesondere hat sie mir in den letzten Jahren, in denen unsere Kinder Nina und Jan auf die Welt kamen, immer wieder den notwendigen Freiraum verschafft, meine Arbeit zum erfolgreichen Ende zu führen.

Anto Zečević Sinsheim, im August 2002

Zusammenfassung

In der vorliegenden Arbeit wurde erstmalig ein sprachgestütztes Trainingssystem zur Evaluierung der Nasalität aus Mikrophonaufnahmen konzipiert und prototypisch implementiert. Die wichtigste Voraussetzung für dieses Vorhaben war die automatische Bestimmung der Nasalität mit einer hohen Güte.

Hierzu wurde die Sprachdatenbank NASAL mit verschiedenen Sprechergruppen, Passagen und Nasalitätsausprägungen aufgebaut und logopädisch bewertet. Aufbauend auf dieser erfolgten umfangreiche Untersuchungen zur Quantifizierung der Nasalität. Als ein wesentliches Ergebnis dieser Untersuchungen wird ein automatisches Verfahren zur Quantifizierung der Nasalität mit einer hohen Güte vorgestellt. Dieses beruht auf den Frequenzbandintensitäten und Sprachmodellparametern stimmhafter Laute. Um Aussagen über die beeinflussenden Faktoren der Nasalität gewinnen zu können, wurde als Klassifizierungsverfahren die Lineare Diskriminanzanalyse gewählt.

Die im Trainingssystem zu Zuge kommende Spracherkennungstechnologie basiert auf dem kommerziellen Produkt „ViaVoice“ der Firma IBM. Über das Mikrophon aufgenommene Sprachdaten werden sowohl bzgl. der Nasalität als auch der Spracherkennungsleistung bewertet. Gerade in der sehr niedrigen Spracherkennungsrate zeigte sich die Komplexität bei der Verarbeitung sprechgestörter Aussprache. Die Integration der kommerziellen Spracherkennung ist daher zum jetzigen Zeitpunkt nicht empfehlenswert.

A Inhaltsverzeichnis

B Abbildungsverzeichnis

C Tabellenverzeichnis

D Definitions- und Abkürzungsverzeichnis

Definitionen

Klassifikationsaufgaben

N01 Klassifikation Nasal: nonnasal (0) - nasal (1)

N123 Klassifikation Nasal: leicht (1), mittel (2), stark (3)

N0123 Klassifikation Nasal: nonnasal (0), leicht (1), mittel (2), stark (3)

Sprechergruppen

FRAU Frauen ab 16 Jahren

MANN Männer ab 16 Jahren

KIND Jungen und Mädchen von 6 bis 16 Jahren

Parametergruppen: Frequenzband

U100_N Uniform 100 Hz Bandbreite, keine Intensitätsnormierung

U100_I Uniform 100 Hz Bandbreite, Intensitätsnormierung: Gesamtintensität

U100_S Uniform 100 Hz Bandbreite, Intensitätsnormierung: Sone

U400_N Uniform 400 Hz Bandbreite, keine Intensitätsnormierung

U400_I Uniform 400 Hz Bandbreite, Intensitätsnormierung: Gesamtintensität

U400_S Uniform 400 Hz Bandbreite, Intensitätsnormierung: Sone

BARK_N Bark, keine Intensitätsnormierung

BARK_I Bark, Intensitätsnormierung: Gesamtintensität

BARK_S Bark, Intensitätsnormierung: Sone

Parametergruppen: Formanten und Antiformanten

GF Lage, Intensität und Bandbreite der Grundfrequenz

FREQ Lage der Formanten F_1 bis F_4 bzw. Antiformanten AF_1 bis AF_4

INT Energie der Formanten F_1 bis F_4 bzw. Antiformanten AF_1 bis AF_4

BAND Bandbreite der Formanten F_1 bis F_4 bzw. Antiformanten AF_1 bis AF_4

ALLE Lage, Intensität und Bandbreite der Grundfrequenz, Formanten und Antiformanten, sowie die Gesamtintensität der Aufnahme

Abkürzungen

ADV	Amplitudendichteverteilung18
AF_i	der i-te Antiformant
AR	Auto Regressive
ARMA	Auto Regressive Moving Average
dB	Dezibel
DFT	Diskrete Fourier Transformation
DLL	Dynamic Link Library
F_i	der i-te Formant i, i = 0 entspricht der Grundfrequenz
FFT	Fast Fourier Transformation
HMM	Hidden Markov Model
HNO	Hals-Nasen-Ohren
HNR	Harmonic-Noise-Ratio
Hz	Hertz
kHz	Kilohertz
LDA	Lineare Diskriminanzanalyse
LKG	Lippen-Kiefer-Gaumen
LPC	Linear Predictive Coding
MA	Moving Average
MEM	Maximum Entropy Method
NN	Neuronales Netz
PC	Personal Computer
PLP	Perceptual Linear Prediction
RASTA	Relative Spectral
SAQ	Summe der Abweichungsquadrate
VQ	Vektorquantisierung

1 Einleitung

Die vorliegende Arbeit hat ihren Ursprung in einem interdisziplinären Projekt der Medizinischen Fakultät Heidelberg und der Universität Mannheim. Ziel des Projektes war es, neue bzw. neuartig angewandte diagnostische und therapeutische Verfahren zur Optimierung der sprachlichen Rehabilitation von Patienten mit Lippen-Kiefer-Gaumenspalten einzusetzen.

Die Aufgabenstellung dieser Arbeit ist zunächst die analytische Untersuchung möglicher Verfahren zur Erkennung von Sprechstörungen. Insbesondere geht es um die Entwicklung und Etablierung eines automatischen Verfahrens zur quantitativen Bestimmung der Nasalität und dessen Einbettung in ein sprachgestütztes Trainingssystem.

Neben der rein anwendungsbezogenen Motivation steht hier die wissenschaftliche Fragestellung der Sprecherkennung und der Möglichkeit einer automatischen Analyse durch eine geeignete Computeranalyse. Die Methoden der immer erfolgreicher werdenden Sprachanalyse können nicht direkt auf die Sprechanalyse übertragen werden. In der Sprachanalyse versucht man mit minimalem Lernaufwand eine hohe Erkennungsrate des gesprochenen Wortes zu erkennen. In der Sprechanalyse geht es nicht darum Worte zu erkennen, sondern Eigenarten aus dem Gesprochenen zu extrahieren. So werden zum Beispiel Spracherkennungssystemen Sprachraum-Erkennungsmodule vorangestellt. Das bedeutet, dass zunächst aus globalen Informationen längerer Passagen die Landessprache oder ein Dialekt, wie z. B. bayrisch, erkannt werden soll, auch wenn dieser schwach ausgeprägt ist.

In der hier gestellten Aufgabe der Erkennung eines Sprechfehlers und dessen qualitative Beurteilung ist das Problem noch wesentlich komplexer. Es gibt nur sehr begrenzte Trainingsdatensätze, die Variationsbreite ist sehr groß, häufig ist die Nasalität von weiteren funktionalen Sprechstörungen begleitet und ein Training an der einzelnen Person ist nicht möglich. Eine Ausnahme hierzu bildet lediglich die Kontrolle eines Lernerfolges in den Übungen über längere Zeit. So war zu Beginn der Arbeit auch nicht klar ob eine solche Sprechanalyse für die gewählte Aufgabenstellung überhaupt erfolgreich sein wird.

1.1 Gang der Arbeit

Um die Problemstellung besser erfassen zu können, wird in den folgenden Unterkapitel dieses Abschnittes auf die medizinische Motivation für diese Arbeit und den Begriff der Nasalität eingegangen.

Kapitel 2 „Stand der Technik“ skizziert die Probleme bei der Klassifizierung der Nasalität und zeigt den Stand der Forschung auf den Gebieten der Sprachproduktion, Sprachwahrnehmung und der automatischen Spracherkennung.

In Kapitel 3 „Verfahren der Merkmalsextraktion“ werden die in dieser Arbeit angewandten Vorverarbeitungen am Sprachsignal sowie Merkmalextraktionsverfahren beschrieben. Das Ziel dieser Verfahren ist eine parametrische Repräsentation des kontinuierlichen Zeitsignals. Unter Vorverarbeitung werden im Kontext dieser Arbeit alle Verfahren zusammengefasst, die auf dem Zeitsignal operieren. Von diesen werden die Verfahren zur Gewinnung der spektralen und cepstralen Parameter unterschieden.

In Kapitel 4 „Klassifikation“ wird das in dieser Arbeit angewandte Verfahren zur Klassifikation der Parameter, die Lineare Diskriminanzanalyse, vorgestellt. Bei der Linearen Diskriminanzanalyse handelt es sich um ein statistisches Verfahren, das für Fragestellungen mit bekannter Gruppenzugehörigkeit geeignet ist. Ziel der Diskriminanzanalyse ist es, die Linearkombinationen der Variablen zu finden, die eine möglichst gute Differenzierung dieser Gruppen ermöglichen.

Kapitel 5 „Sprachdatenbank NASAL“ beinhaltet eine Beschreibung der im Rahmen dieser Arbeit erstellten Sprachdatenbank. Zurzeit enthält diese Sprachdaten auf der Basis eines standardisierten Sprachbogens von 116 Sprechern mit unterschiedlichen Nasalitätstypen und –ausprägungen. Alle Sprachdaten sind logopädisch hinsichtlich der Nasalität beurteilt. Die Sprachdatenbank bildet die Basis der umfangreichen Untersuchungen und wird zur Steigerung der Signifikanz der Ergebnisse stetig erweitert.

In Kapitel 6 „Computergestützte Trainingsumgebung“ wird der Stand der implementierten sprachgestützten Trainingsumgebung beschrieben. Diese besteht aus den Komponenten „Eingangsuntersuchung“, „Spracherkennung“, „Trainingsmodul“ und „Sprachdatenbank“. In diesem Kapitel wird insbesondere auf das Zusammenspiel aller Module eingegangen. Zur Spracherkennung erfolgen weiterhin Aussagen über ihre Güte.

Auf die Methodik und Schwierigkeiten bei der Bestimmung der Sprachmodell-Parameter wird in Kapitel 7 „Bestimmung der Parameter“ eingegangen. Der Schwerpunkt dieser Beschreibung besteht dabei in einer genauen Darstellung der Randbedingungen, unter denen die Parameter bestimmt werden.

In Kapitel 8 „Klassifikationsergebnisse“ werden die Ergebnisse sämtlicher Versuchsreihen zur Ermittlung der Klassifizierungsraten und der relevanten Parameter der untersuchten Verfahren präsentiert. Bei den Parametern wurde dabei nach Frequenzbandparametern und Sprachmodellparametern unterschieden. Eine Zusammenfassung dieser Ergebnisse schließt das Kapitel ab.

Im Kapitel 9 „Zusammenfassung und Ausblick“ werden die Resultate dieser Arbeit zusammengefasst. Weiterhin zeigt der Abschnitt einige sinnvolle und notwendige Ansätze zur Weiterentwicklung auf.

1.2 Medizinische Motivation

Lippen-Kiefer-Gaumenspalten sind die zweithäufigste angeborene Fehlbildung mit einer Häufigkeit von 1 auf 500 Geburten[1]. Sie müssen als ein Krankheitsbild angesehen werden, bei dem zahlreiche Funktionen wie Mimik, Atmung, Ernährung, Sprache und Gehör gestört sind. Es kann z. B. zu Trinkstörungen kommen, da Saugen und Schlucken manchmal beeinträchtigt sind. Beide Vorgänge stellen gleichzeitig wichtige Grundlagen zum Sprechen dar, so dass hier eine mögliche Fehlentwicklung zu sehen ist. Aber auch die eigentlichen Sprechbewegungen gelingen nicht so gut, wenn Narbengebilde feinste Koordinierungen nur bedingt zulassen. Wenn der Patient aufgrund der Spalte Hörstörungen hat, kann er Geräusche und Laute nur in undeutlicher oder abgeschwächter Form aufnehmen; der normale Reifungsprozess der Hörbahnen ist dann nicht gewährleistet. Aus diesen Gründen ist es möglich, dass er später zu sprechen beginnt, Sätze fehlerhaft oder unvollständig bildet, oder dass es Schwierigkeiten hat, genaue Lautunterscheidungen zu treffen.

Schließlich ist bei einer Spalte die Trennung von Mund- und Nasenraum meist nur unvollständig möglich. Durch den unvollkommenen Nasen-Rachenverschluss dringt der Luftstrom beim Sprechen auch in die Nasenhöhle. Dadurch erhalten insbesondere Vokale einen nasalen Klang. Aber auch die Artikulation der Konsonanten ist beeinträchtigt. So klingen sie unscharf, weil bei ihnen in der Mundhöhle nur ein geringer Luftstau entstehen kann. Oft kommt es zu funktionellen Sprachstörungen, da die Lauterzeugung an benachbarte Stellen im Rachen- und Kehlkopfbereich verlagert wird. Zu hören sind mitunter zischende, blasende oder knackende Geräusche, ein Stimmknall oder ein harter Stimmeinsatz, der sich nachteilig auf die Stimme auswirken kann. Durch die Überanstrengung setzt das Kind auch manchmal mimische Hilfsbewegungen ein. Sie zeigen sich durch die Verengung der Nasenflügel, Runzeln der Nase und Heben der Oberlippe [WRG89].

---

[1] Es ist immer noch ungeklärt, warum es zu Spaltbildungen kommt. Neben genetischen Faktoren sollen auch Umweltfaktoren eine Rolle bei der Fehlbildung spielen. Hierzu ein Artikel aus der Bild der Wissenschaft [BdW99]:

„Eines von 500 Neugeborenen kommt mit einer Hasenscharte auf die Welt. Woher die rätselhafte Lippen-Kiefer-Spalte kommt, enthüllen Rike Derynck und Zena Web von der Universität San Francisco in der aktuellen Ausgabe von Nature. Sie konnten zeigen, dass bei der normalen Embryonalentwicklung der Wachstumsfaktor T*GF*-alpha an den E*GF*-Rezeptor andockt und so die Bildung von Matrix-Metallproteinasen (MMPs) auslöst. „Knock-out“-Mäuse, die keine E*GF*-Rezeptor besitzen, konnten auch keine MMPs herstellen und so kam es bei den Versuchstieren häufig zu Hasenscharten.“

Damit wird klar, dass LKG-Spaltpatienten neben der nasalen Aussprache noch weitere überlagerte Sprechstörungen aufzeigen, so dass sie manchmal nahe der Verständlichkeitsgrenze sind. Dies erschwert die automatische Quantifizierung der Nasalität sehr, für einige Patienten ist sie nicht durchführbar.

Bislang beruht die Beurteilung der Sprecherverständlichkeit im deutschsprachigen Raum in der Mehrzahl der Fälle auf dem subjektiven Höreindruck der Logopäden oder Verfahren mit informellem Charakter. Von den vielfältigen in der Literatur beschriebenen Verfahren zur Objektivierung der klinischen Diagnose Näseln, konnte sich bislang keine Methode in der Routinediagnostik etablieren. Speziell für den deutschen Sprachraum existieren nur sehr wenige Arbeiten zur Sprecheranalyse bei nasaler Aussprache. Gesucht ist ein einfaches, schnell durchführbares, wenig belastendes, zuverlässiges und insbesondere praktikables Verfahren zur Nasalitätsevaluierung. Die Bewertung soll mindestens auf einer ordinalen[2] Skala erfolgen, um individuelle Veränderungen registrieren zu können.

Alle diese Anforderungen an ein Verfahren, realisiert die menschliche Wahrnehmung und Verarbeitung durch einen Logopäden am besten. Da diesem der bloße Höreindruck zur Bewertung der Nasalität ausreicht, haben wir uns in diesem Projekt für die Objektivierung der Nasalität aus einer Mikrophonaufnahme entschieden. Ziel war es durch eine geeignete Auswahl von Parametern und ihrer Optimierung, hohe und stabile Erkennungsraten bei der Quantifizierung des offenen und geschlossenen Näselns zu erreichen.

Um den Patienten nicht stark zu belasten bzw. ihn sogar zu motivieren, soll dieses Verfahren in eine computergestützte, sprachgesteuerte Trainingsumgebung integriert werden. Dieses soll neben der Diagnostik der Nasalitätsausprägung insbesondere auch zur Verlaufsbeobachtung, z. B. vor und nach operativen Gaumensegeleingriffen, eingesetzt werden. Die Motivation der Patienten lässt sich neben der Lösung von Teilaufgaben vor allem durch ein Biofeedback steigern, indem ihm seine individuellen Lernerfolge visuell angezeigt werden. Dank der Leistungsfähigkeit heutiger Computer, lassen sich aufwendige Berechnungen nahezu in Echtzeit durchführen. Durch die unmittelbare Rückmeldung ist ein Patient bspw. mit Gaumensegelinsuffizienz in der Lage, während der Therapie, seine aktuelle Nasalität zu registrieren und unter visueller Kontrolle zu verändern. Über die individuelle Einstellung eines Schwellenwertes seitens des Therapeuten kann der Patient auch schon geringfügige Trainingserfolge erkennen und somit in jeder Therapiestunde neu motiviert werden.

Da dieses Verfahren nicht invasiv ist, kann es bereits bei Kindern ab 5 Jahren problemlos eingesetzt werden. Sinnvoll ist der Einsatz in der Verlaufskontrolle vor und nach HNO-ärztlichen Eingriffen, um Veränderungen des Nasalitätsgrades aufzuzeigen. Dies ermöglicht

[2] In einer Ordinal- oder Rangskala besteht eine größer-kleiner-Beziehung der Merkmale, ohne dass der Abstand zwischen den Rängen ein bestimmtes Ausmaß hätte (z. B. die Platzierung von Sportwettkämpfern) [MT99].

eine objektive Feststellung von Sprechverständlichkeitsveränderungen. Nach der klinischen Testphase ist auch ein Einsatz in logopädischen Praxen und zu Hause denkbar.

1.3 Begriffsklärung

Der Begriff Nasalität stammt aus der Linguistik und dient zur Charakterisierung der nasalen natürlichen Klangfarbe eines Lautes. Er wird bestimmt durch den Lautbestand einer Sprache und ihren Anteil an Nasalkonsonanten und nasalierten Vokalen. Die einzigen nasalen Laute im Deutschen sind die auch als Rhinophone bezeichneten Nasalkonsonanten /m/, /n/, und /ŋ/. Eine Nasalierung von Vokalen kommt in der deutschen Hochlautung nicht vor. Verschiedene deutsche Mundarten (z. B. Saarländisch, Schwäbisch) und Fremdsprachen wie Französisch sind dagegen durch das Vorhandensein nasalierter Vokale geprägt, die ihnen ihren unverwechselbaren nasalen Charakter verleihen [HWS91]. Werden abweichend vom normalen Nasalitätsmaß der jeweiligen Sprache die suprapalatinalen Räume (Nasenrachen, Nasenhöhlen) zu stark (Hyper-) oder zu schwach (Hyporhinophonie) an der Phonation und Artikulation beteiligt, spricht man vom Näseln. Die verschiedenen Formen des Näselns werden in Abhängigkeit von Lokalisation und Genese (= Entstehung), heute nach dem allgemein gebräuchlichen Schema von Arnold in offenes, geschlossenes und gemischtes sowie organisches und funktionales Näseln untergliedert. Infolge des fehlenden Abschlusses zwischen Mund und Nase kommt es beim offenen Näseln zu einer fehlerhaften Aussprache aller im Mund erzeugten Laute, vor allem der Vokale. Das geschlossene Näseln dagegen äußert sich in einer fehlerhaften, durch einen resonanzmindernden Verschluss der Verbindung Gaumen/Nase bedingten Aussprache der drei Nasallaute /m/, /n/, und /ŋ/. Solches Näseln kann schon bei Schwellungen oder Verschleimung des Nasenraumes infolge eines Schnupfens auftreten [Hab86].

Näseln stellt eine Störung des Stimmklanges dar, die bei verschiedenen funktionellen und organischen Krankheitsbildern wie z. B. Gaumenspalten, aber auch bei gesunden Menschen situationsbedingt oder gewohnheitsmäßig auftreten kann. Wenngleich die verschiedenen Rhinophonie-Formen durch klinische Untersuchungsbefunde genau charakterisiert sind, ist das akustische Phänomen Nasalität bis heute nicht ausreichend verstanden.

2 Stand der Technik

Wie eingangs erwähnt, stellt die pathologische Nasalität eine Störung der Sprecherverständlichkeit, oder allgemeiner des Kommunikationsprozesses, dar. Der lautsprachliche Kommunikationsprozess gliedert sich in drei Teile (vgl. Abbildung 1):

Ein Sprecher verleiht einem Gedanken Ausdruck, indem er durch eine Folge artikulatorischer Bewegungen einen zeitlich variierenden Sprachschall erzeugt.

Diese Schallwelle wird anschließend in Gestalt von Druckschwankungen durch die Luft (oder aber auch in Form einer zeitabhängigen elektrischen Spannung über ein Telefonkabel) zum Empfänger der Botschaft übertragen.

Beim Hörer wird die Nachricht unter anderem mit den Mechanismen menschlicher Schall-, Laut- oder Wortwahrnehmung entschlüsselt.

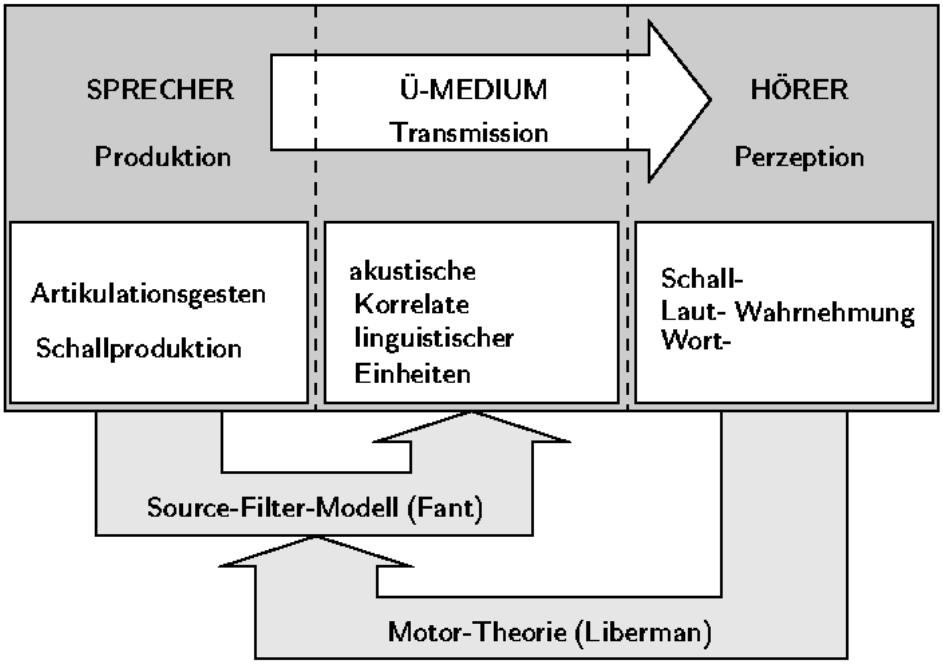

Abb. 1: Der Prozess menschlicher Sprachkommunikation [St95]

Hinsichtlich der Relation zwischen produzierten bzw. perzeptierten Sprachlauten und der akustischen Signalform gibt es bereits gut ausgebaute Theorien, darunter das Produktionsmodell von Fant [Fan60] und die Befunde zur menschlichen Lautheits- und Tonhöhenwahrnehmung (Standardwerke zu diesem Thema [ZF90, Moo89]).

Dieses Kapitel beginnt mit dem aktuellen Stand der Forschung bei der Klassifizierung der Nasalität und geht auch ferner auf die Klassifizierungsproblematik ein.

Da Sprachsignale akustische Signale sind und ihre Eigenschaften wesentlich durch die akustischen Eigenschaften der Sprechorgane bestimmt sind, folgt danach ein Unterkapitel über die menschliche Sprachproduktion. Dies erfolgt auch vor dem Hintergrund, dass viele Näselnde an anatomisch verursachten Sprechstörungen, z. B. infolge angeborener Spaltenbildung im Gaumenbereich, leiden und die Sprechorgane die Sprachsignale somit nicht einwandfrei produzieren bzw. diese verzerren können.

Danach werden einige wichtige Erkenntnisse der Sprachperzeption vorgestellt. Dies ist insbesondere unter dem Aspekt zu sehen, dass das menschliche Gehör und Gehirn den Beweis für eine, wenn auch subjektive, Klassifizierung der Nasalität liefern. Sowohl für die Sprachproduktion als auch für die Perzeption werden die führenden Funktionsmodelle vorgestellt.

Um die Arbeit in den Kontext der Sprachverarbeitung einordnen und von Spracherkennungssystemen abgrenzen zu können, wird zum Schluss dieses Kapitels die prinzipielle Funktionsweise sowie eine mögliche Kategorisierung von Spracherkennungssystemen vorgestellt.

Weiterhin spielen insbesondere die Merkmalgewinnungsverfahren der Sprachsignalverarbeitung eine wichtige Rolle. Diese werden in einem separaten Kapitel beschrieben.

2.1 Klassifizierungsproblem der Nasalität

Bis zum heutigen Tage beruht die Beurteilung der Nasalität in den meisten Fällen auf dem Höreindruck von Logopäden und Phoniatern[3]. Die Zuverlässigkeit dieser subjektiven Einschätzung ist weitgehend vom Erfahrungsgrad der Untersucher abhängig. So besteht die Möglichkeit, dass eine geringe bis milde Hyperrhinophonie[4] von weniger erfahrenen Einschätzern übersehen wird. Nach Counihan und Cullinan nimmt die Reliabilität zwischen verschiedenen Bewertern von Vokalen über Silben zu Textpassagen zu [CC70]. Ebenso konnten sie feststellen, dass sich die Nasalitätseinschätzungen der Vokale mit zunehmender Vokalintensität verbesserten. Insgesamt wird die subjektive logopädische Einschätzung als variabel und nicht ausreichend genau angesehen [CC72]. Im Gegensatz dazu konnten Paynter et al. sowie Hardin et al. eine hohe Übereinstimmung zwischen verschiedenen logopädischen Beurteilungen beobachten [PEJ91, HVM92].

---

[3] Logopädie = Sprachheilkunde.
Phoniatrie = Lehre von den Erkrankungen des Stimmapparates.

[4] Zur Definition der medizinschen Begriffe siehe Kapitel 1.3 „Begriffsklärung“.

Darüber hinaus wurden eine Vielzahl unterschiedlicher Techniken und Instrumente entwickelt, mit dem Ziel die velopharyngeale[5] Funktion bzw. den Stimmklang während des Sprechens zu analysieren. Neben einfachen Funktionstests wie des Czermak´schen Spiegeltests[6] oder der A-I-Probe von Gutzmann[7] wurden manometrische[8] Untersuchungsmethoden [HM60], sonagraphische Analysen [Bau63, Kyt69, RHH90] und aerodynamische Gaumensegelfunktionsprüfungen [WD64] angewandt. Als weitere Verfahren zur Analyse der strukturellen Beziehungen bzw. der physiologischen Bewegungen kommen der Ultraschall, die Kephalometrie[9], die Tomographie[10], die Video-Fluoroskopie, akustische Verfahren wie z. B. die Nasometrie und die Video-Nasopharyngoskopie[11] zur Anwendung. Viele dieser Verfahren erwiesen sich aufgrund des meist großen Zeitaufwands und der erheblichen Belastung für die Patienten während der Mess-Situation für die klinische Routinediagnostik als ungeeignet.

Insgesamt ist festzustellen, dass von den vielfältigen in der Literatur beschriebenen Verfahren zur Objektivierung der Nasalität sich bislang keine Methode in der Routinediagnostik etablieren konnte. Lediglich das von Fletcher entwickelte Nasometer fand als einfache, den Patienten nicht allzu stark bedrückende Methode im angloamerikanischen Sprachraum im Rahmen der Therapie des Näseln bei Gaumenspalten breitere Anwendung [Fle76]. Das Nasometer stellt ein mikrocomputergestütztes System dar, welches das Verhältnis von oraler zu nasaler akustischer Energie mit Hilfe zweier an jeder Seite einer Trennplatte befestigter Richtmikrophone beim Sprechen bestimmter Sprechpassagen bestimmt. Dieses Verhältnis wird in der Literatur als Nasalanz bezeichnet. Während entsprechende Studien für den deutschen Sprachraum den klinischen Nutzen des Nasometers bestätigen [SHK94, Ste98], wird die Eignung in anderen Studien als eingeschränkt betrachtet. So konnten Watterson et. al. nur eine bescheidene Korrelation zwischen der Nasalanz und der Nasalität bei Passagen ohne Nasalkonsonanten feststellen [WMW93]. Unter Berücksichtigung der Nasalkonsonanten zeigte sich gar keine Signifikanz zur Beurteilung der Hypernasalität. Mögliche Gründe könnten in den Aufnahmebedingungen zu finden sein, da durch das Anlegen der Trennplatte eine mechanische Irritation und somit eine

---

[5] velopharyngeal leitet sich ab von den Begriffen Velum (weicher Gaumen) und Pharynx (Rachen).

[6] Dem Sprechenden wird ein Spiegel zwischen Mund und Nase gehalten. Bei einer bestehenden Nasalität läuft der Spiegel beim Aussprechen der Vokale aufgrund des nasalen Luftflusses leicht an.

[7] Hält man dem Sprechenden mit zwei Fingern die Nase zu, während er im Wechsel a und i sagt, so bemerkt man bei Bestehen einer Verschlussstörung zum oberen Rachen zu, einen deutlichen Klangunterschied der Laute gegenüber den mit offener Nase gesprochenen gleichen Lauten.

[8] Manometer = Druckmesser; Gerät zum Messen des Drucks von Gasen oder Flüssigkeiten.

[9] Kephalo-: auch Keph-, Ceph-, Cephalo-, Zephalo-; Wortteil mit der Bedeutung Kopf, Haupt.

[10] Tomographie: Schichtaufnahmeverfahren der Röntgendiagnostik.

[11] Nasopharynx = Nasenrachenraum.

Beeinflussung denkbar wäre. Eine andere Störquelle haben Zajac et. al. in der Mikrophonsensitivität in ihrer Studie ausgemacht [ZLM96].

Ein viel versprechender Ansatz zur Bestimmung eines Nasalitätsmaßes ist derzeit die Analyse der spektralen Eigenschaften der Sprache, welcher auch in dieser Arbeit verfolgt wird. Die praktischen Vorteile einer solchen Messungsmethode liegen in der einfachen Handhabung und Wiederholbarkeit, insbesondere bei kleinen Kindern, da solch ein Verfahren nicht invasiv ist. Während uns die prinzipielle Eignung des spektralen Ansatzes das menschliche Gehör und seine anschießende neuronale Verarbeitung beweist, führen die bereits ausgeführten Untersuchungen zu keinen übereinstimmenden Ergebnissen. So misst Bauer zusätzlichen spektralen Frequenzen im Spektrum der Vokale eine entscheidende Bedeutung für das Wesen der Nasalität bei [Bau63]. Diese zusätzliche Frequenzen, welche in unterschiedlichen Bereichen auftreten, können seiner Meinung nach neue Resonanzbezirke darstellen oder eine relative Verbreiterung bereits vorhandener Formantgebiete. Als weiteres Merkmal nennt er die Reduktion der Intensität des 1. Formanten. Kyttä beschreibt als charakteristische Merkmale für die Nasalierung von Vokalen eine Verstärkung der Formanten bei 250 Hz, Abschwächung bei 500 Hz sowie das Auftreten schwach ausgeprägter Teilfrequenzbänder um 1000-2500 Hz zwischen den Vokalformanten [Kyt69].

Kataoka et al. stellen am Beispiel einer Studie des isoliert gesprochenen Vokals /i/ auf der Basis von 33 japanischen Sprechern, eine hohe Korrelation zwischen wahrgenommener Nasalität und der Zunahme der Intensität zwischen 1. und 2. Formanten, sowie der Abnahme der Intensitäten des 2. und 3. Formanten fest [KMO96]. Garnier et al. untersuchten an Kindern einige spektrale Methoden (Formantanalyse, Cepstrum, FFT[12]) an den Vokalen /a/, /i/ und /u/ bzgl. ihrer Eignung zur Charakterisierung der Hypernasalität [GGC96]. Bei ihren Untersuchungen zeigten sich die FFT und das Cepstrum der Formantenanalyse überlegen. Unerwarteterweise erzielten sie durch die Trennung der Kinder in Jungen und Mädchen bessere Ergebnisse bei der Klassifizierung. Eine Benutzung der Barkskala[13] ergab keine Verbesserung.

Fragt man sich nach den die Nasalität beeinflussenden Faktoren und insbesondere nach ihrem Ausmaß auf die Nasalität, findet man keinen Konsens in der Literatur. Übereinstimmung besteht darin, dass sowohl strukturelle Defizite (kurzes Gaumensegel, etc.), funktionelle Störungen (Artikulationsfehlbildungen) als auch dynamische Einschränkungen (schlecht bewegliche Gaumensegel) zur velopharyngealen Insuffizienz und Hyperrhinophonie führen können und dass der nasale Charakter nicht durch einfache physikalische

[12] Zur Definition der Merkmalsextraktionsverfahren siehe Kapitel 3 „Verfahren der Merkmalsextraktion“.
[13] Zur Definition der Barkskala siehe Kapitel 2.3.2 „Frequenzgruppen- und Tonhöhenwahrnehmung“.

Resonanz im Sinne von Mitschwingen der Nasenräume in der Grundfrequenz[14] der Stimmlippen erzeugt wird. Inwieweit die Nasalität vom Geschlecht der Patienten abhängt, wird bislang kontrovers diskutiert. Während Seaver et al. [SDL91] bei Frauen signifikant höhere Werte beim Sprechen nasaler Textpassagen nachweisen konnten, wurde kein geschlechtsspezifischer Unterschied in der Arbeit von Litzaw und Dalston festgestellt [LD92]. Auch hinsichtlich des Einflusses des Alters auf die Nasalität bestehen erhebliche Unstimmigkeiten in der Literatur. Nach Haapanen scheint das Alter der Patienten einen bestimmten Effekt auf die Nasalanzwerte bei den Plosiven[15] zu besitzen [Haa91]. So wurde mit zunehmendem Alter eine Verringerung dieser Werte in geringem Umfang beobachtet, die auf die ausgereiftere Gaumensegelbewegung bei diesen Patienten zurückgeführt werden. Keine Korrelation zwischen Alter und Nasalanzwerten wurde hingegen bei Sätzen ohne Plosive gefunden. Fletcher konnte keinen Alterseffekt bei Kindern zwischen dem 6. und 13. Lebensjahr nachweisen [Fle89]. Hörstörungen, die durch wiederkehrende Mittelohrentzündungen entstehen können, sind ein weiterer Aspekt bei der Frage nach den Einflussfaktoren auf die Sprache. Nach Wirth sind Hörstörungen eine der wichtigsten Ursachen für eine Sprachentwicklungsstörung, da das beeinträchtigte Gehör das Erfassen des Lautbildes und die sprachliche Selbstkontrolle behindert [Wir94].

Zusammenfassend lässt sich sagen, dass die objektive Bestimmung der Nasalität sehr stark erschwert wird von den individuellen Eigenschaften eines Sprechers wie Alter, Geschlecht oder Stimmqualität, sowie der subjektiven perzeptuellen Einschätzung der Schätzer. Die Bestimmung eines spektralen Maßes ist aufgrund ihrer Praktikabilität und Eignung derzeitig die attraktivste Forschungsrichtung. Die Diskrepanzen in der Literatur sind neben der Unsicherheit, welche Faktoren die Nasalität bestimmen, insbesondere auch methodischer Art, insofern, dass die Aussagen der Studien häufig auf zu wenig Sprechern, nur einzelnen bzw. einigen Vokalen und unterschiedlichen Ländern beruhen. Da die Nasalität aber stark vom Dialekt und Soziolekt[16] abhängt, sind die Aussagen aus anderen Ländern nicht ohne weiteres auf den deutschsprachigen Raum übertragbar.

Daher ist das Ziel dieser Arbeit, die aktuellen Forschungsergebnisse zur spektralen Eignung auf den deutschen Raum zu untersuchen und eigene Parameter zu entwickeln. Die Nasalität soll dabei nach dem Typ (offenes, geschlossenes) und der Ausprägung auf einer ordinalen Skala gemessen werden. Dies gelingt in dieser Arbeit durch die Bestimmung von spezifischen Parametersätzen für die stimmhaften Laute. Es zeigt sich, dass die Nasalität durch

---

[14] Die Sprachparameter werden in Kapitel 2.2 „Sprachproduktion“ und Kapitel 2.3 „Sprachwahrnehmung“, die Grundfrequenz insbesondere im Kapitel 2.2.1 „Stimmhafte Laute“, definiert.

[15] Plosive = Verschlusslaute: Der Luftstrom wird durch Zunge oder Gaumen gestoppt, so dass er weder durch den Mund noch durch die Nase entweichen kann (oraler und nasaler Verschluss). Nach einem Druckaufbau wird der Verschluss plötzlich freigegeben.

[16] Soziolekt = einer sozialen Schicht eigentümliche Sprachform.

diesen Ansatz gut messbar ist. Da die Güte dieses Ansatzes sehr stark von der Größe der zur statistischen Untersuchung verfügbaren Sprachdaten abhängt, wird die Sprachdatenbank NASAL (vgl. Kapitel 5 „Sprachdatenbank NASAL“) stetig ausgebaut.

2.2 Sprachproduktion

Das Sprechen ist eine Kombination aus der Stimmgebung, einer durch Ausatmen bewirkten Schallanregung an der Stimmritze (Glottis), und der darauf folgenden Resonanzbildung, einer durch kontinuierliche Muskelbewegung verursachten Ausformung des Anregungsschalls im Vokaltrakt[17], zu welchen wir den Rachenraum (Pharynx), den Mundraum und den Nasenraum zählen. Die wesentlich zur Klangfarbe des Sprachsignals beitragenden Komponenten im Vokaltrakt sind die Stellung der Zunge, der Grad der Mundöffnung sowie die Stellung der Lippen. Die Gesamtheit der Stellungen und Bewegungen der Organe des Sprechtrakts und ihre zeitliche Aufeinanderfolge wird als Artikulation bezeichnet.

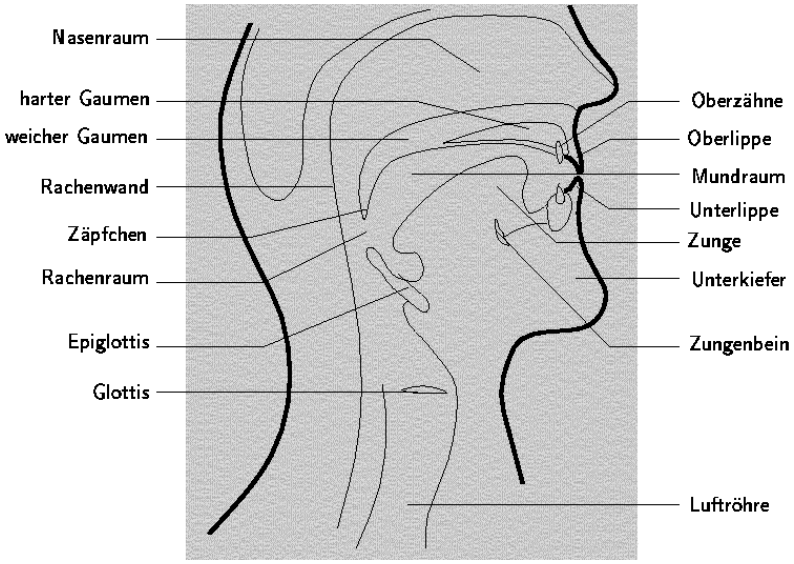

Abb. 2: Das menschliche Artikulationssystem [St95]

Beim Stimmbildungsprozess, der auch als Phonation bezeichnet wird, passiert ein Luftstrom aus den Lungen die Luftröhre (Trachea) und gelangt in den Kehlkopf (Larynx). Je nach Öffnungsgrad der Stimmritze (Glottis) unterscheidet man in der deutschen Sprache hauptsächlich zwei Phonationsarten. Steht die Glottis weit offen, bewirkt die vorbeiströmende Luft Turbulenzen an den Stimmbändern, und es entstehen stimmlose Laute wie /t/ oder /s/. Im zweiten Fall geraten die Stimmbänder durch Verengung der Stimmritze in eine quasiperio-

[17] Häufig wird für den Vokaltrakt synonym der Begriff Ansatzrohr verwendet.

dische Öffnungs- und Schließbewegung, aus der stimmhafte Laute wie die Vokale hervorgehen[18].

Der so gewonnene Anregungsschall wird beim Durchlaufen des Vokaltraktes zu einer Fülle verschiedener Sprachlaute ausgeformt. Für das dabei wirksame charakteristische Resonanzverhalten sowie das Auftreten von Reibe-, Vibrations- und Sprengungsgeräuschen sind die jeweils eingenommene Form, die Verengungen und die Verschlüsse an diversen Positionen des Vokaltraktes verantwortlich. Passive Artikulatoren sind die Oberlippe, die Oberzähne, der Gaumen, der weiter unterteilt wird in Zahndamm (Aveolen), harten (Palatum) und weichen (Velum) Gaumen und Zäpfchen (Uvulum). Aktive Artikulatoren sind die Unterlippe, die Zungenspitze und der Zungenrücken.

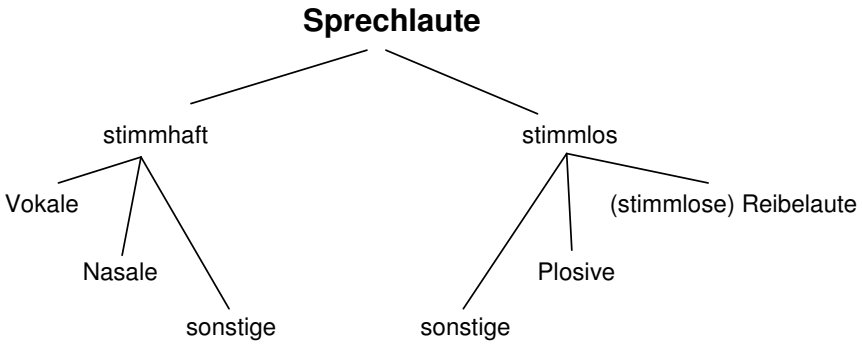

Abb. 3: Einteilung der Laute in stimmhafte und stimmlose Laute

Während die Konsonanten mit oben genannten Klassifizierungsmethoden beschrieben werden können, greift man zur Unterscheidung der Vokale häufig auf die Zungenposition zurück[19]. Hierbei ist wichtig, wo und wie hoch sich der Zungenrücken befindet (Horizontal- bzw. Vertikalposition). So unterscheidet man in der Horizontallage zwischen Vorderzungenvokalen, wie z. B. /i/, zentralen Vokalen und Hinterzungenvokalen wie z. B. /u/. Die Vertikalposition betreffend unterscheidet man zwischen einer hohen Lage und damit einem hohen Vokal, z. B. /i/, und einer tiefen Lage, z. B. /a/. Der Wert dieses Merkmals (hoch, halbhoch, halbtief, tief) korreliert mit dem Öffnungsgrad des Mundes, der die Vokale in geschlossene, halbgeschlossene, halb offene und offene unterteilt. Diese Klassifizierung gibt das Vokalviereck wider.

[18] Eine weitere Phonationsart ist der totale Verschluss, nebst anschließendem explosivem Öffnen der Glottis. Das dabei hörbare Schlaggeräusch bezeichnen wir als Glotisschlag.

[19] Üblich ist auch eine Unterscheidung zwischen gerundeten und ungerundeten Vokalen, wobei sich das gerundet auf die Lippen bezieht.

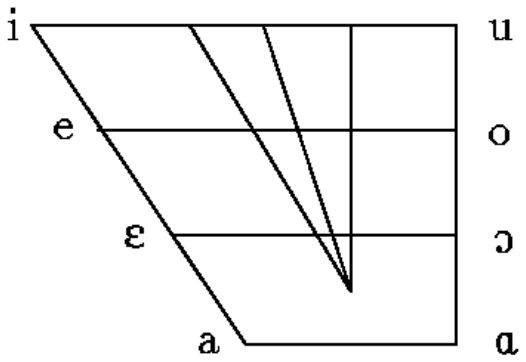

Abb. 4: Vokalviereck mit 8 Kardinalvokalen, nach [Koh77]

2.2.1 Stimmhafte Laute

Dieses Kapitel beschreibt die Charakteristik von stimmhaften Lauten. Da aus isolierten stimmlosen Lauten eine Nasalitätsbewertung seitens der Logopäden nicht möglich war (s. Kapitel 5.2.3 „Logopädische Beurteilung“), wird auf die Abhandlung stimmloser Laute nicht eingegangen und daher auf die Standardliteratur verwiesen (vgl. [Zwi82, Wir94]).

Das Anregungssignal, welches durch den Luftstrom der Lunge und die Stimmbänder erzeugt wird, ist näherungsweise eine periodische Dreiecksimpulsfolge. Daraus ergibt sich ein Linienspektrum, bei dem die Spektrallinien einen Frequenzabstand von

$$f_g = \frac{1}{t_g} \qquad (1)$$

haben; f_g bezeichnet man als Grundfrequenz, t_g als Grundperiode.

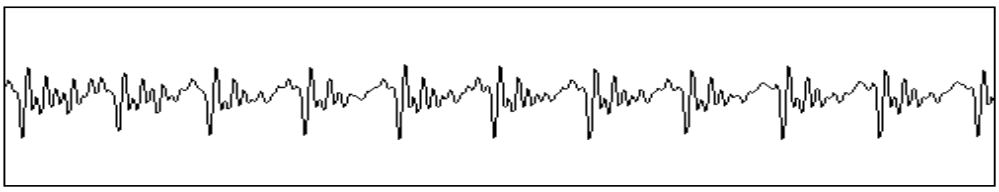

Abb. 5: Der Vokal /a/ im Zeitbereich

Die Grundfrequenz liegt etwa zwischen 80 Hz (tiefe Männerstimme) und 400 Hz (Kinderstimme). Infolge der Dreieckform nimmt die Einhüllende des Spektrums in Richtung höherer Frequenzen um 12 dB pro Oktave ab.

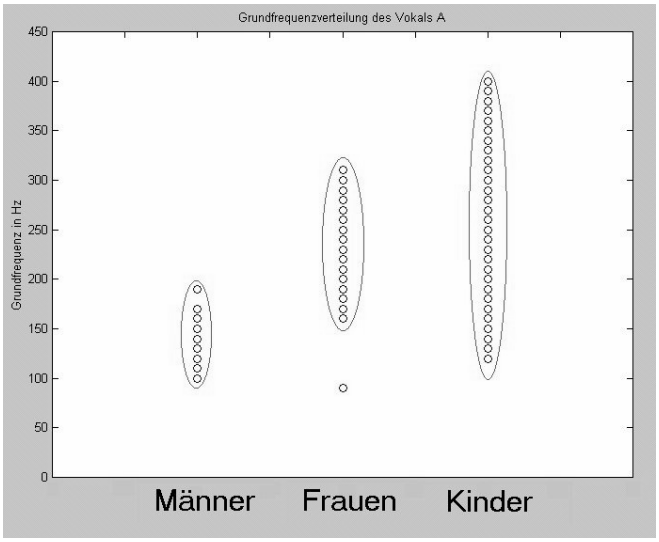

Abb. 6: Grundfrequenzverteilung der Sprachdatenbank NASAL für /a/

Betrachtet man die Spektralverläufe von Vokalen, kann man charakteristische Linien für jeden Vokal feststellen. Diese Linien sind die Resonanzfrequenzen des Vokaltraktes und werden als Formanten bezeichnet. Die Feinstruktur in den Spektren rührt von den Harmonischen der Grundfrequenz.

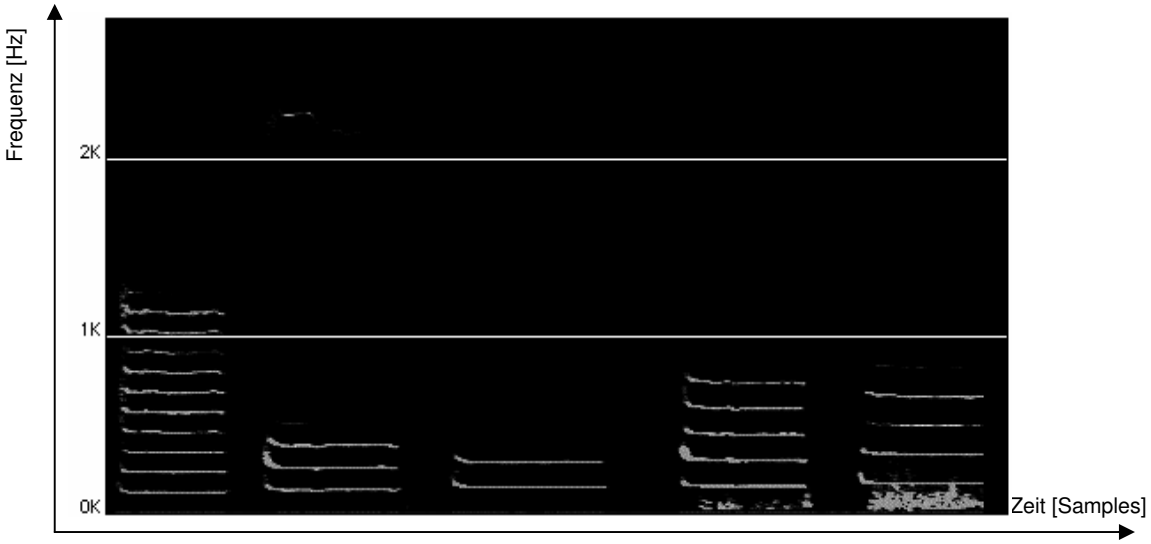

Abb. 7: Sonagramm der Vokale /a/, /e/, /i/, /o/, /u/[20]

Vokale werden hauptsächlich durch die ersten beiden Formanten F_1 und F_2 sowie mit gewissen Einschränkungen durch den 3. Formanten F_3 charakterisiert. Von F_4 aufwärts müssen die Formanten jedoch zu den sprecherspezifischen, d. h. nicht lautabhängigen Merkmalen gezählt werden. Dies trifft mit Einschränkungen bereits auf F_3 zu.

[20] Die Intensität der auftretenden Frequenzen ist in der Färbung kodiert. Die Aufnahme zeigt die Vokale in der Reihenfolge a, e, i, o, u mit Pausen zwischen den Vokalen gesprochen

	Men			Women			Children		
Vowel	F1	F2	F3	F1	F2	F3	F1	F2	F3
[i]	270	2,300	3,000	300	2,800	3,300	370	3,200	3,700
[I]	400	2,000	2,550	430	2,500	3,100	530	2,750	3,600
[ɛ]	530	1,850	2,500	600	2,350	2,000	700	2,600	3,550
[ae]	660	1,700	2,400	860	2,050	2,850	1,000	2,300	3,300
[a]	730	1,100	2,450	850	1,200	2,800	1,030	1,350	3,200
[ɔ]	570	850	2,400	590	900	2,700	680	1,050	3,200
[U]	440	1,000	2,250	470	1,150	2,700	560	1,400	3,300
[u]	300	850	2,250	370	950	2,650	430	1,150	3,250
[ʌ]	640	1,200	2,400	760	1,400	2,800	850	1,600	3,350
[ɜ·]	490	1,350	1,700	500	1,650	1,950	560	1,650	2,150
Mean	500	1,420	2,400	575	1,700	2,800	670	1,900	3,250
F2/F1		2.84			2.96			2.84	
F3/F2			1.69			1.65			1.71

Tab. 1: Formantfrequenzen der ersten 3 Formanten bei Vokalen in Hz (nach [PB52])

Vergleicht man die aus der Literatur bekannt gewordenen Untersuchungen über Formantanalysen, zeigt sich, dass die Formantfrequenzen für den gleichen Laut erheblich differenzieren können. Eine Formantkarte, in welcher die beiden ersten Formantfrequenzen (F_1 und F_2) der Vokale gegeneinander aufgetragen werden, zeigt wie groß der Streubereich der Formanten für die einzelnen Laute ist. So sind bei Kindern aufgrund ihrer höheren Grundfrequenz auch höhere Formantfrequenzen zu beobachten. Bemerkenswert sind auch die Überschneidungsbereiche. Hier lassen sich die einzelnen Laute auf Basis von Formanten alleine nicht unterscheiden.

Was die Formanten genau darstellen, wird in der Literatur kontrovers diskutiert. Einige Autoren vertreten die Auffassung, dass die Formanten Harmonische der Grundfrequenz seien. Das impliziert aber, dass die Formanten im Vokalspektrum immer genau harmonisch zu liegen kommen, auch bei unterschiedlicher Grundfrequenz. Andere Autoren vertreten die These, dass die Formantfrequenzen durch eine Resonanz des Vokaltrakts, gleichsam einer angeschlagenen Glocke, hervorgerufen werden. Das Bild von der Glocke ist nicht von ungefähr gewählt: Auch bei einer Glocke liegen die Resonanzfrequenzen nicht harmonisch; genauso, wie man es auch bei den Formantfrequenzen beobachten kann [TS99].

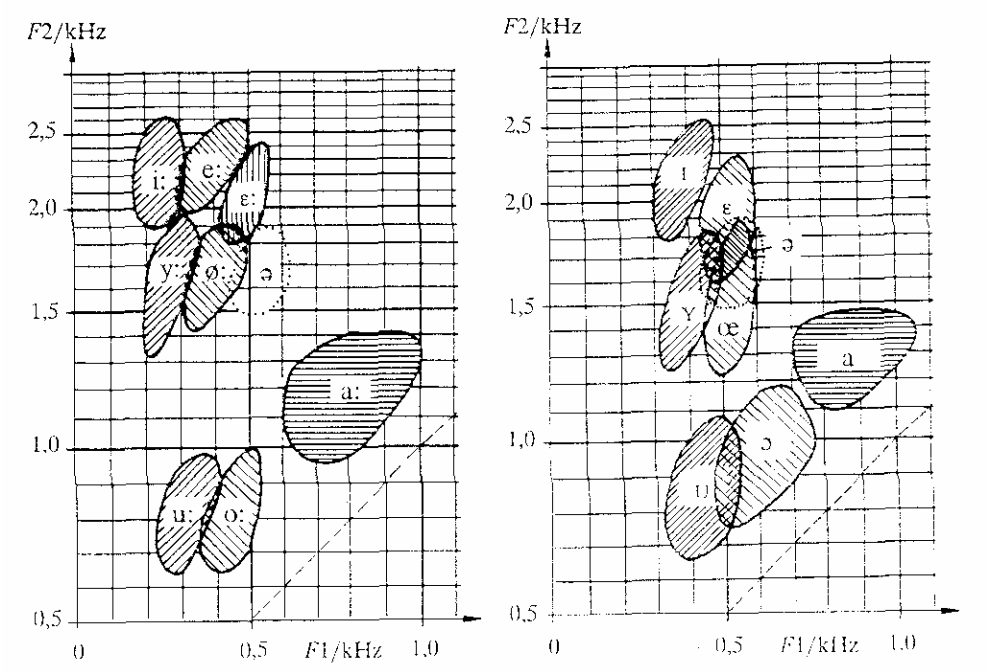

Abb. 8: Formantkarte der deutschen Vokale aus 16 Sprecherinnen und Sprecher; Links: Langvokale, rechts: Kurzvokale (nach [Hes76])

Stimmhafte Konsonanten können genau wie Vokale eine ausgeprägte Formantstruktur aufweisen. Dies gilt vor allem für die Nasalkonsonanten /m/, /n/, und /ŋ/. Anders als bei Vokalen entweicht hier jedoch die Luft nicht durch den Mund sondern nur durch die Nase. Als Resonanzraum wirkt hier zusätzlich zur Mundhöhle auch die Nasenhöhle. Für die Übertragungsfunktion bedeutet das Zusammenwirken beider Resonanzräume, dass neben den Formanten (spektrale Maxima) auch Antiformanten (spektrale Minima) auftreten[21]. Nach Untersuchungen von Fujimura sind die Nasale /m/, /n/, und /ŋ/ gekennzeichnet durch eine niedrige (750 Hz bis 1250 Hz), mittlere (1450 Hz bis 2200 Hz) bzw. hohe (über 3000 Hz) Lage des Antiformanten. Weiterhin wird festgestellt, dass der Antiformant einen erheblichen Einfluss auf den in der Nähe liegenden Formanten hat; die übrigen Formanten bleiben jedoch relativ konstant[22].

Bis jetzt haben wir die Akustik des Vokaltrakts ohne den Nasenraum betrachtet, d. h. wir setzten immer voraus, dass das Gaumensegel (Velum) angehoben sei und so den Nasenraum vom übrigen Vokaltrakt abtrennt.

---

[21] Zur Definition und Bestimmung der Formanten und Antiformanten siehe Kapitel 7.2.3.3 „Antiformanten“ und Kapitel 7.2.3.2 „Formanten“.

[22] Aus der Tatsache, dass Nasale im wesentlichen durch den Nasenhohlraum gebildet werden und dieser Hohlraum – anders als der Mundraum - praktisch nicht veränderbar ist, ergibt sich für diese Laute ein sehr interessanter Anwendungsfall im Zusammenhang mit der Sprechererkennung. Verwendet man nämlich einen Testsatz mit möglichst vielen Nasallauten, so werden Täuschungsversuche in hohem Maße erschwert.

Beim Senken des Velums wird der Nasenraum bis hin zu den Nasenlöchern akustisch an den Vokaltrakt angekoppelt. Je nach Stellung des übrigen Vokaltrakts ergeben sich – bei stimmhafter Anregung - die folgenden Möglichkeiten:

Freie Passage im Mundraum

Durch die Ankoppelung des Nasenraums entstehen analog zum Modell des Ansatzrohres stehende Wellen im abzweigenden Nasenraum. Durch Resonanz dieser stehenden Wellen wird dem Schallfluss im restlichen Ansatzrohr Energie bei bestimmten Frequenzen entzogen. Im Spektrum des schließlich von den Lippen abgestrahlten Sprachsignals zeigt sich dies in Form von Einbrüchen bei diesen Frequenzen, den so genannten Antiformanten. Zwar wird dabei auch von den Nasenlöchern - parallel zum Schallfeld der Mundöffnung - ein Schallfeld abgestrahlt, welches genau diese Antiformanten betont, aber dieses ist sehr viel schwächer und manifestiert sich daher kaum im gemessenen Schalldrucksignal. Das auditive Ergebnis dieses Vorgangs sind nasalierte Vokale, die im deutschen Phonemsystem eigentlich nicht vorkommen, aber in der Realität – und nicht nur im Zusammenhang mit Fremdwörtern – natürlich nachgewiesen werden können.

Lippen geschlossen – Zunge gesenkt

Bei geschlossenen Lippen findet die akustische Abstrahlung nur noch von den Nasenlöchern statt. Die Situation kehrt sich nun quasi um: Die vorn abgeschlossene Mundhöhle wirkt als Resonanzhohlraum, der dem Ansatzrohr von Glottis bis Nasenlöchern bei bestimmten Antiformanten Energie entzieht. Sind nur die Lippen geschlossen, ist die ganze Mundhöhle Resonanzraum. Das auditive Ergebnis dieser akustischen Konfiguration ist der Nasal /m/.

Lippen geöffnet – Verschluss Zunge/Zahndamm

Der Verschluss liegt nun tiefer in der Mundhöhle und verringert somit den angekoppelten Resonanzraum der Mundhöhle. Die Lage der Antiformanten wird dadurch verändert, es entsteht der auditive Eindruck des Nasales /n/.

Lippen geöffnet – Verschluss Zunge/weicher Gaumen

Liegt der Verschluss an der Stelle dicht vor dem gesenkten Velum, auf dem weichen Gaumen, entsteht der auditive Eindruck des Eng-Lauts, Nasal /ŋ/, wie in ‚lang'.

2.2.2 Lineares Modell der Spracherzeugung

Die Modellierung der Spracherzeugung hat die wesentlichen Komponenten des Artikulationssystems angemessen zu berücksichtigen. Dazu zählen die Schallanregung an der Glottis, die zeitliche Veränderung der Vokaltraktform und die Abstrahlung von den Lippen.

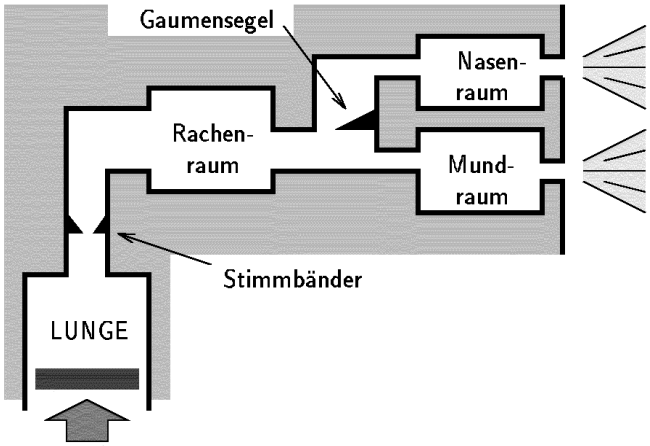

Abb. 9: Schematische Darstellung des menschlichen Artikulations-systems [St95]

Das Resonanzverhalten bei der Schallformung wird ferner durch Verluste an den Vokaltraktwänden beeinflusst, die durch Wärmeleitung, Elastizität und Reibung entstehen. Schließlich ist der Einfluss des Nasenraumes zu beachten, dessen Zuschaltung die Unterdrückung gewisser Frequenzkomponenten des Anregungsschalls bewirkt (sog. „Antiresonanzen“)[23].

G. Fant’s source-filter-Modell [Fan60, RS78] geht von einem Signalerzeugungsprozess aus, welcher zwei bzw. drei hintereinander geschaltete lineare zeitinvariante Systeme in sich vereinigt. Das diskrete Signal lässt sich als Faltung $f_n = u_n * v_n * r_n$ schreiben, so dass sich für die z-Transformierten[24] die Darstellung

$$F(z) = U(z) \cdot V(z) \cdot R(z) \qquad (2)$$

ergibt. Unter Faltung versteht man eine Verknüpfungsoperation zweier Folgen h(n) und x(n) nach der Vorschrift:

$$y(n) = h(n) * x(n) = \sum_{m=-\infty}^{\infty} h(n - m)x(m) \qquad (3)$$

[23] Die bekannten akustischen Sprachmodelle wurden ursprünglich mit dem Ziel der maschinellen Sprachsynthese entwickelt.

[24] Zur Definition der z-Transformierten siehe Kapitel 3.2 „Transformationen“.

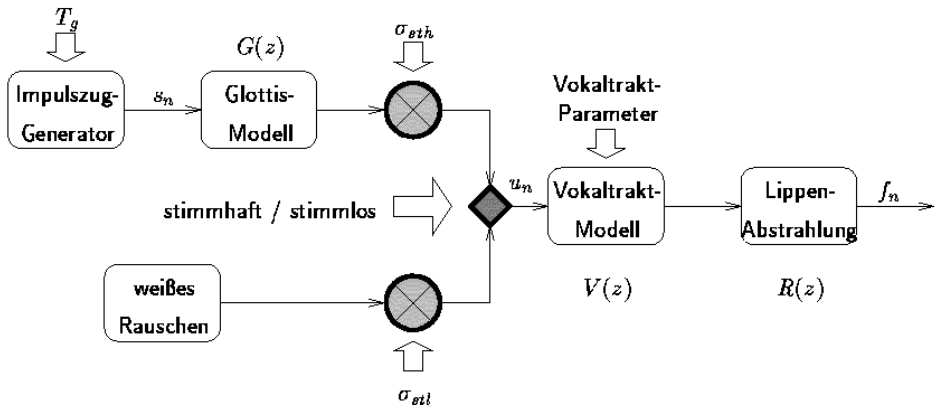

Abb. 10: Das source-filter-Modell der Spracherzeugung[St95]

2.2.2.1 Glottismodell

Die Schwingung u_n am Glottisausgang kann bei stimmlosen Lauten durch ein Rauschsignal mit flachem Spektrum angenähert werden, welches mit dem Verstärkungsfaktor σ_{stl} skaliert wird (vgl. vorherige Abbildung). Bei stimmhaften Lauten besteht die Glottisschwingung aus einem periodischen Signal von 50 Hz bis 400 Hz. Dieser Umstand wird durch die Faltung eines Impulszuges s_n geeigneter Periode T_g mit der Impulsantwort g_n des Glottismodells ausgedrückt. Die Grundschwingung kann im Zeitbereich durch zusammengesetzte Kosinusfunktionen, in der z-Ebene durch ein System

$$G(z) = \frac{1}{(1 - e^{-cT} z^{-1})^2} \qquad (4)$$

mit zwei Polen angenähert werden [MG76]. Dabei ist T die Abtastperiode und c << 1/T eine geeignete Konstante.

2.2.2.2 Lippenabstrahlung

Die Abstrahlung von den Lippen wird physikalisch als Druckwellenaustritt durch eine kleine Öffnung in einer sehr großen Schallwand modelliert. Idealisiert erhalten wir ein Hochpass-filter

$$R(z) = R_0(1-z^{-1}) \quad , R_0 \text{ konstant} \qquad (5)$$

das im Zeitbereich auf die Bildung der ersten Ableitung hinausläuft.

2.2.2.3 Vokaltraktmodell

Bei der physikalischen Erfassung des Resonanzsystems bleiben der gesamte Nasenraum sowie Verluste an den Vokaltraktwänden unberücksichtigt. Das aus Mund- und Rachenraum bestehende Ansatzrohr betrachten wir vereinfachend als akustisches Rohr der Länge L, zusammengesetzt aus M gleichlangen Zylinderabschnitten mit den Querschnittsflächen A_i (i = 1, ... ,M). Diese Verhältnisse sind in Abbildung 11 veranschaulicht. L beträgt typischerweise 170 mm.

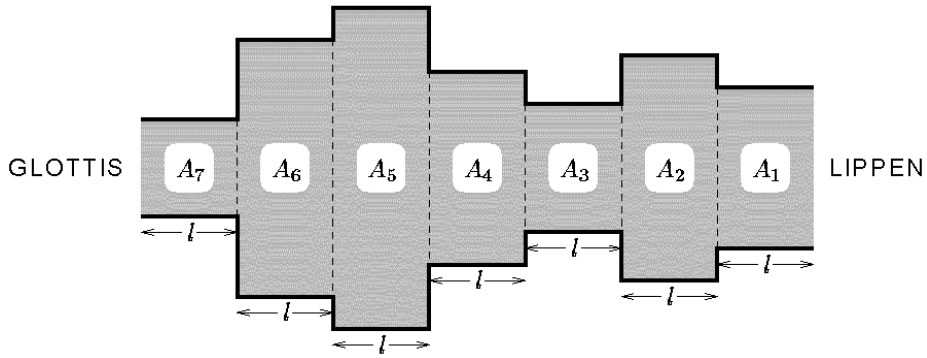

Abb. 11: Die verlustfreie akustische Röhre gleichlanger Zylinderabschnitte [St95]

Da die Länge l = L/M der Zylinderscheiben im allgemeinen weit unterhalb der Wellenlänge von Sprachsignalen liegt, können wir im wesentlichen eine ebene Wellenausbreitung in Achsenrichtung der Röhre annehmen. Unter Beachtung der einschränkenden Stetigkeitsbedingungen für die Zylinderübergänge können wir den Schallfluss der vor- und rücklaufenden Wellen iterativ aus den Reflexionskoeffizienten

$$k_i = \frac{A_i - A_{i+1}}{A_i + A_{i+1}} \qquad i = 0, \ldots, M \qquad (6)$$

berechnen. Die Fläche des „Außenweltzylinders“ vor den Lippen setzen wir zweckmäßigerweise mit $A_0 \rightarrow \infty$ an, so dass $k_0 = 1$ wird. Der Abschlusswiderstand A_{M+1} an der Glottis kann willkürlich gewählt werden, da er keinen Einfluss auf das Resonanzverhalten hat [Reg88].

Die Wellenausbreitung innerhalb der Röhre wird nur zu äquidistanten Zeitpunkten gestört, nämlich infolge der Durchmesseränderung beim Wechsel des Zylinderabschnitts. Daher ergibt sich bei bandbegrenzten, abgetasteten Glottissignalen ein besonders einfacher Ausdruck für das Übertragungsverhalten:

$$V(z) = \frac{\prod_{i=0}^{M}(1+k_i)}{1-\sum_{i=1}^{M} a_i z^{-1}} \qquad (7)$$

Die Koeffizienten des Nennerpolynoms lassen sich mit $a_i = a_i^{(M)}$ iterativ berechnen:

$$a_i^{(m)} = \begin{cases} 1 & i = 0 \\ a_i^{(m-1)} + k_m a_{m-i}^{(m-1)} & 0 < i < m \\ k_m & i = m \end{cases} \qquad (8)$$

Die Funktion V(z) hat in der komplexen Ebene keine Nullstellen, jedoch M/2 Paare konjugiert komplexer Polstellen, nämlich die Wurzeln des Nennerpolynoms.

$$1-\sum_{i=1}^{M} a_i z^{-i} = \prod_{i=1}^{M/2}(1-2e^{-c_i T}\cos(b_i T)z^{-1}+e^{-2c_i T}z^{-2}) \qquad (9)$$

Diese Pole markieren mit $F_i = b_i/2\pi$ und $B_i = c_i/2\pi$ die Mittenfrequenz und die Bandbreite der Vokaltraktresonanzen, die als Formanten bezeichnet werden. Wie im Kapitel 2.2 „Sprachproduktion“ bereits erwähnt, charakterisieren die durchschnittlichen Frequenzen F_1, F2 der beiden unteren Formanten ziemlich genau die deutschen Vokalphoneme.

2.2.2.4 Das autoregressive Modell

Die Gesamtübertragungsfunktion der Sprachproduktion – genau genommen trifft dies nur auf stimmhafte Laute zu - ist $H(z) = \sigma \cdot G(z) \cdot V(z) \cdot R(z)$ und wird für praktische Belange durch $H(z) \approx \sigma/A(z)$ angenähert, wobei A(z) ein Polynom $\sum_{i=0}^{M} a_i z^{-1}$ in z^{-1} mit $a_0 = 1$ ist; σ bezeichnet wieder den Verstärkungsfaktor. Lineare Systeme dieser Form heißen wegen ihrer Zeitbereichseigenschaften autoregressive Systeme oder im englischen all-pole-Systeme.

Eine Schwäche des autoregressiven Modells besteht darin, dass die Einwirkung des Nasaltraktes nicht berücksichtigt wird. Bei nasaler Artikulation werden gewisse Frequenzbereiche während der Lautformung gedämpft. Diese Antiresonanzen oder Antiformanten würden durch kompliziertere, beliebig rationale Übertragungsfunktionen der Form $H(z) = B(z) / A(z)$ angemessener wiedergegeben. Filter dieser Form werden als ARMA-Systeme (autoregressive moving average) bezeichnet.

2.3 Sprachwahrnehmung

Wesentliche Vorteile ergeben sich, wenn bei der Sprachsignalverarbeitung die Funktionsweise des menschlichen Gehörs berücksichtigt wird. Die messtechnische Auswertung des Gehörs, das für die akustische Analyse sprachlicher wie nichtsprachlicher Schallsignale verantwortlich ist, dient der Erforschung des Zusammenhangs zwischen akustischen Reizen (Frequenz, Schallintensität) und den dadurch ausgelösten subjektiven Empfindungen (Tonhöhe, Lautheit). Die Befunde zur menschlichen Lautheits- und Tonhöhenwahrnehmung sowie zum zeitlichen wie spektralen Auflösungsvermögen haben zur Entwicklung mathematischer Funktionsmodelle und zur Definition psychoakustischer Skalen geführt, die aus den Merkmalsextraktionsstufen heutiger Spracherkenner kaum wegzudenken sind. Die spektrale Zusammensetzung eines komplexen Klanges ist durch die Amplituden der auftretenden Sinuskomponenten und deren Phaseneigenschaften charakterisiert. Die Phasenbeziehungen machen sich durch eine unterschiedliche Empfindung von Rauhigkeit [Ter74] bemerkbar, jedoch nur, wenn sie spektrale Komponenten innerhalb einer Frequenzgruppe betreffen. Der Einfluss der Phasenverschiebung auf die Lautwahrnehmung scheint eher gering zu sein. Vertiefende Literatur zur Sprachperzeption ist unter anderem [ZF90, Moo89]. Es soll eine kurze Zusammenfassung der für die Nasalitätsmessung relevanten Grundlagen gegeben werden.

2.3.1 Lautheitswahrnehmung

Die physikalische Stärke eines Schallsignals wird als Schalldruck p_s in Pascal oder als Intensität I_s in N/m^2 angegeben; es ist I_s proportional zu p_s^2. Mit der willkürlich festgelegten Bezugsgröße $p_0 = 2*10^{-5}$ Pa wird der *Schalldruckpegel* definiert als

$$\text{Schalldruckpegel [dB]} = 20*\log(p_s/p_0) = 10*\log(I_s/I_0) \qquad (10)$$

Eine Intensitätsverdopplung entspricht somit einem Pegelzuwachs von 3 dB.

Der Pegel des leisesten gerade noch wahrnehmbaren reinen Sinustons von 1000 Hz beträgt im Mittel 6 dB. Diese Ruhehörschwelle ist allerdings frequenzabhängig; generell werden Schallwellen gleichen Pegels bei unterschiedlichen Frequenzen subjektiv nicht als gleich lautstark empfunden. Das Diagramm in Abbildung 12 zeigt die Linien gleicher Lautstärkewahrnehmung (Isophone) in der Pegel-Frequenz-Ebene für reine Töne. Als psychoakustisches Maß für die menschliche Lautstärkewahrnehmung wurde von Barkhausen das *phon* eingeführt [ZF67]. Die Phonzahl eines Teststimulus ergibt sich demzufolge als Pegel (in dB) des als gleichlaut beurteilten 1 kHz Tones. Für einen Standardschall dieser Frequenz stimmen Pegel und Lautstärke quantitativ überein.

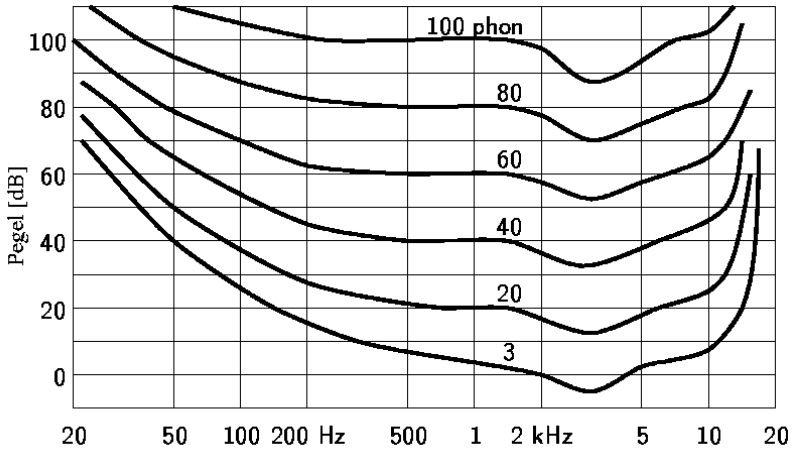

Abb. 12: Linien gleicher Lautstärke in der Schallebene [Zwi82]

Der Lautstärkepegel, der aus den Kurven gleicher Lautstärke ermittelt wird, ist noch keine Empfindungsgröße. Eine weitere psychoakustische Größe, die *Lautheit* (engl. loudness), berücksichtigt das wahrgenommene Lautstärkeverhältnis zweier Töne. Zwei Schalle verschiedener Intensität lassen sich so einstellen, dass der eine als halb bzw. doppelt so laut empfunden wird wie der andere. Die Lautheit eines 1 kHz Tones von 40 phon beträgt 1 *sone*; ein L-mal so laut wahrgenommener Vergleichston wird auf L sone festgelegt. Nach [Ste57] verhält sich die Lautheit proportional zu $I^{0.3}$. Ähnliche Potenzgesetze finden sich auch bei anderen Autoren [Vog75]. Die Lautheit ist eine komplizierte Empfindungsgröße, die insbesondere auch von der spektralen Zusammensetzung und der Bandbreite eines Schallreizes abhängt. Breitbandige Schalle werden als lauter empfunden als schmalbandige Schalle gleichen Pegels[25].

2.3.2 Frequenzgruppen- und Tonhöhenwahrnehmung

Wahrnehmungsexperimente legen nahe, dass das Gehör in eng begrenzten Frequenzbändern Intensitäten von verschiedenen Schallreizen zusammenfasst. Diese Frequenz-

[25] Die Intensität von Tönen kürzerer Dauer (t<200ms) wird vom Gehör zeitlich integriert. Für die Hörschwelle I eines Tonimpulses der Dauer t gilt mit guter Näherung $(I-I_L)*t$ = const, wenn I_L das Intensitätsminimum des entsprechenden Dauertons ist [Gar47]. Ist das Gehör einem Schallreiz für längere Zeit ausgesetzt, spielt der Integrationseffekt keine Rolle mehr, dafür nimmt die scheinbare Reizstärke spürbar ab, und die Hörschwelle wird kurzzeitig heraufgesetzt. Diese Phänomene werden als Adaption bzw. Ermüdung bezeichnet. Die kleinste wahrnehmbare Intensitätsänderung ΔI folgt dem Weberschen Gesetz $\Delta I/I$ = const, falls wir es mit einem breitbandigen Rauschen zu tun haben. Die entsprechende Pegeldifferenz $10\log((I+\Delta I)/I)$ ist damit auch näherungsweise eine Konstante und beträgt knapp 1 dB. Bei reinen Tönen wächst unsere Diskriminierungsleistung mit der Lautstärke. Bereits bei 40 Phon genügt eine Pegeldifferenz von 0.7 dB zur Unterscheidung und bei 80 Phon sind es 0.3 dB [Rie28].

bänder werden als *Frequenzgruppen* (engl. critical bands) bezeichnet[26]. Die Bandbreite B_g einer Frequenzgruppe ist eine Funktion ihrer Mittenfrequenz F_g; sie beträgt etwa 100 Hz falls $F_g <= 1000$ Hz; oberhalb dieser Grenze ungefähr 15% der Mittenfrequenz. Reiht man über den gesamten Hörbereich alle Frequenzgruppen auf, ergibt sich eine gehörorientierte nichtlineare Frequenzskala, die als *Tonheit* (engl. critical band rate) bezeichnet wird und die Einheit *Bark* besitzt. Der Bereich der wahrnehmbaren Frequenzen von 16 Hz bis 20000 Hz lässt sich in 24 nichtüberlappende Frequenzgruppen kritischer Bandbreiten aufteilen.

Der Zusammenhang zwischen der Tonheit Ω in Bark und der Frequenz ω in rad/s ist durch Gleichung (11) gegeben [Sch77]

$$\Omega(\omega) = 6 \ln\left(\frac{\omega}{1200\pi} + \sqrt{\left(\frac{\omega}{1200\pi}\right)^2 + 1}\right) \qquad (11)$$

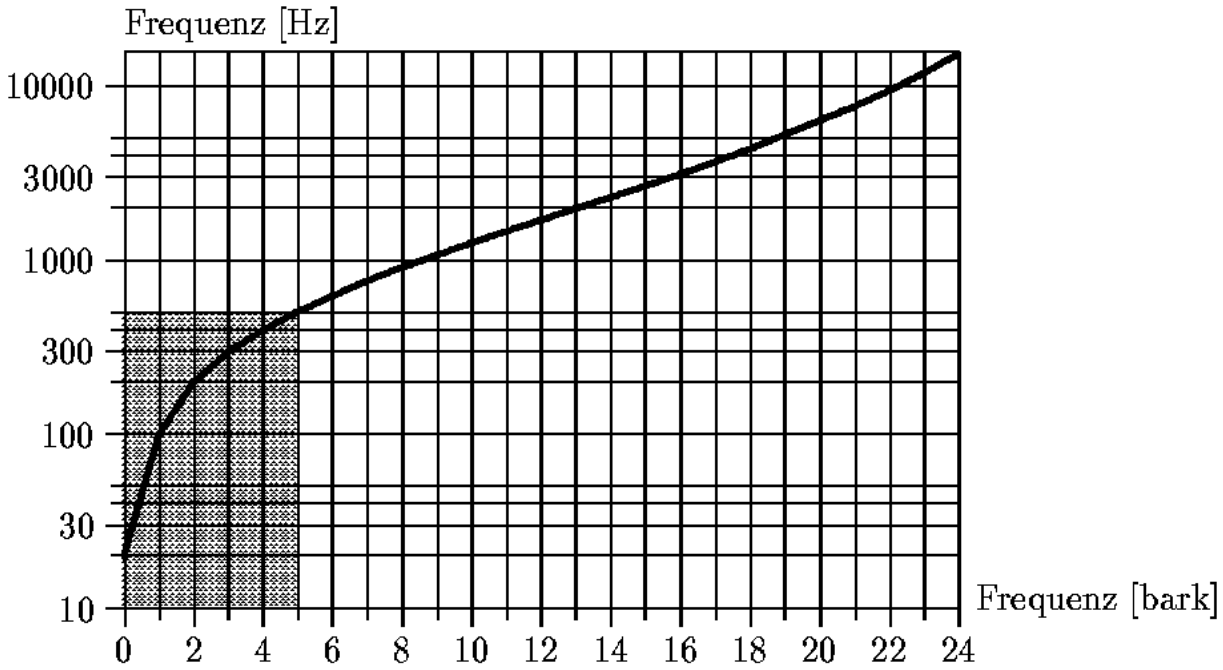

Abb. 13: Frequenzskala in Bark [Zwi82]

[26] Die Schallfrequenz wird vom Gehör durch deren bevorzugte Auslenkungsregion auf der Basilarmembran kodiert. Für nahe beieinanderliegende Frequenzen überschneiden sich diese Regionen, und die Schallkomponenten verschmelzen perzeptiv zu einem Gesamteindruck. Die Wahrnehmung komplexer Schalle fällt unterschiedlich aus, je nachdem ob die spektralen Reizkomponenten innerhalb einer kritischen Bandbreite (diese werden auch Frequenzgruppen genannt) liegen oder aber über mehrere solcher Bänder verstreut sind. Diese Sichtweise wird von einer Fülle psychoakustischer Befunde untermauert; die Verdeckung reiner Töne durch schmalbandiges Rauschen [Fle40], die Sensitivität gegenüber Phasenverschiebungen [Zwi52] und die Integration der Schallintensität von Klängen oder Schmallbandrauschen finden wir ausschließlich innerhalb von Frequenzgruppen.

Analog zur Herleitung der Lautheit, die das Lautstärkeverhältnis zweier Töne beschreibt, können wir auch Kennlinien zum Verhältnis von Tonhöhen gewinnen. So können wir einen Testton derart einstellen, dass er als halb hoch oder doppelt so hoch wie ein Vergleichsschall mit eindeutig zuordnungsfähiger Tonhöhe empfunden wird. Führt man dies für Töne zahlreicher Frequenzen durch, erhält man eine Kennlinie, welche die Frequenz des Reizes der *Verhältnistonhöhe* zuordnet. Einem Ton von 131 Hz, entsprechend der Note c_0, wird willkürlich die Einheit 131 mel zugewiesen. Töne anderer Frequenzen erhalten die doppelte mel-Zahl wie ein halb so hoch perzipierter Ton. Die Verhältnistonhöhe ist eng mit der Tonheit liiert; über den gesamten wahrnehmbaren Bereich hinweg gilt zwischen der Frequenzgruppenskala und der mel-Skala die mathematische Beziehung [ZF67][27]:

```
1 bark = 100 mel                                                    (12)
```

Die Mittenfrequenzen und Bandbreiten der Teilfilter der Filterbank legen die Frequenzgruppen fest, deren Kurzzeitenergien zum Merkmalvektor z(m) zusammengefasst werden. In der vorliegenden Arbeit werden zwei verschiedene Aufteilungen des Frequenzbereichs von 50 Hz bis 8 kHz verwendet:

- äquidistante Bandbreiten der Kanäle mit Bandbreiten von 100 Hz und 400 Hz
- Aufteilung der Bandbreiten nach den Frequenzgruppen der Psychoakustik

Alle Teilfilter haben dieselbe Bandbreite B, der Abstand zwischen den Mittenfrequenzen benachbarter Teilfilter ist ebenfalls B. Das Filter an der unteren Grenze des Übertragungsbereichs ist ein Bandpass mit der kleineren Bandbreite von 50 Hz bzw. 350 Hz. Dies bewirkt die Unterdrückung von niederfrequenten Störungen zwischen 0 und 50 Hz. Die Durchlassbereiche der Filter überlappen nicht. Für eine Bandbreite von 400 Hz gilt:

- Der erste Bandpass von 50 Hz bis 400 Hz erfasst die Grundwelle der Sprachgrundfrequenz bei stimmhaften Lauten.
- Der erste Formant von Sprachsignalen liegt im Durchlassbereich der ersten beiden Teilfilter. Der zweite Formant kann im Durchlassbereich der ersten sechs Teilfilter liegen. Meist liegen die beiden Formanten in den Durchlassbereichen verschiedener Teilfilter.

[27] Die Frequenzgruppenbildung kann vom Gehör an jeder beliebigen Stelle der Frequenzskala vorgenommen werden und hat daher keinen direkten Einfluss auf die Fähigkeit zur Tonhöhenunterscheidung. Im mittleren Frequenzbereich beträgt die Unterschiedsschwelle $\Delta F/F$ für Dauertöne etwa 0.1% - 0.3%, z. B. ist bei 1000 Hz eine Tonhöhendifferenz von 3 Hz gerade noch wahrnehmbar. Diese überraschend hohe Leistung ist mit der Ortsgenauigkeit der überschwellig erregten Nervenfasern allein nicht zu erklären; wenigstens im Bereich niedriger Frequenzen scheint auch das Periodizitätshören [Lic52] mit Hilfe der Schallphasenkodierung eine Rolle zu spielen. Tatsächlich bricht unser Differenzierungsvermögen immer dann dramatisch ein, wenn der phase-locking-Mechanismus unscharf wird, nämlich bei sehr hochfrequenten oder bei sehr kurzdauernden Tönen. Die Unterschiedsschwellen für reine Töne dürfen keinesfalls auf komplexere Schallereignisse verallgemeinert werden. Formantfrequenzen sind beispielsweise rund 40 mal schlechter diskriminierbar als diejenigen reiner Töne [Hol91].

Nur für Laute mit hohem ersten und tiefen zweiten Formanten liegen beide Formanten im Durchlassbereich desselben (zweiten) Filters, z. B. beim /o/ und /u/. Für den Merkmalvektor bedeutet dies, dass die Energie der stimmhaften Anregung meist zur ersten Koordinate beiträgt. Die Signalenergie der beiden ersten Formanten verteilt sich meist auf zwei der ersten sechs Koordinaten.

2.3.3 Differentielle Wahrnehmbarkeitsschwellen

Unter differentiellen Wahrnehmbarkeitsschwellen versteht man Schwellen für die Änderung einer Reizgröße. Zwei dieser Schwellen sind für die Sprachwahrnehmung von Bedeutung: die differentielle Wahrnehmbarkeitsschwelle für Pegeländerungen und die entsprechende Schwelle für Frequenzänderungen.

Als Faustformel für die Amplitude gilt: Eine Amplitudenänderung ist dann wahrnehmbar, wenn sie (innerhalb einer Frequenzgruppe) 1 dB[28] überschreitet.

Die Wahrnehmbarkeitsschwelle für Frequenzänderungen von Sinustönen ist die empfindlichste Schwelle dieser Art, die das Gehör überhaupt besitzt: sie liegt bei etwa 0.7% für Frequenzen oberhalb 500 Hz und bei etwa 3.5 Hz für Frequenzen unterhalb 500 Hz. Damit ist sie an die Frequenzgruppe gekoppelt und beträgt etwa 1/27 Bark.

2.4 Sprachverarbeitung

Sprachverarbeitung ist ein Aufgabenbereich mit vielgestaltigen interdisziplinären Bezügen, unter anderem zur Physiologie, Psychologie, Phonetik, Linguistik, Akustik, Psychoakustik und insbesondere zur digitalen Signalverarbeitung. Der Grad der Interdisziplinarität hängt dabei natürlich von der jeweiligen Aufgabenstellung ab. Gegenstand des Interesses kann dabei sowohl der Sprecher als auch der Inhalt der vorliegenden Nachricht sein. Beispiele für das erste Anliegen sind die automatische Ermittlung der Sprecheridentität bzw. die Bestätigung einer behaupteten Sprecheridentität aufgrund einer dargebotenen Beispielsäußerung (Sprecheridentifikation und –verifikation). Das bekannteste Anwendungsgebiet des zweiten Anliegens stellt die Spracherkennung dar. Gegenstand der automatischen Spracherkennung ist die Transformation einer als Zeitsignal vorliegenden sprachlichen Äußerung unbekannten Inhalts in die Rechtschriftform. Bei Sprachsynthesesystemen steht neben der Verständlichkeit des Inhaltes auch die Erhaltung sprechertypischer Elemente bei drastisch reduzierter Datenrate des Audiosignals im Vordergrund.

---

[28] 1 dB entspricht einer Änderung der Amplitude um ungefähr 12.5% und der Intensität um ca. 26%.

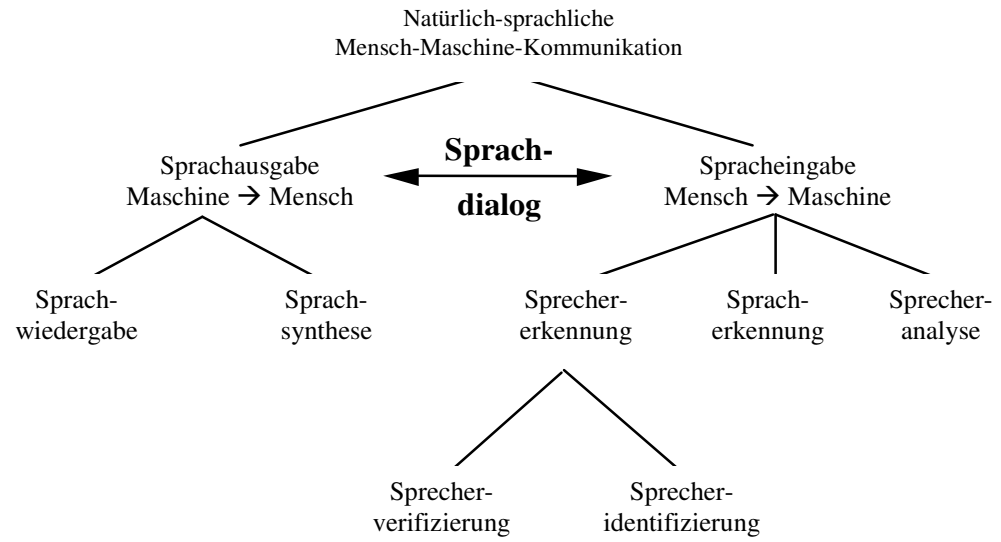

Abb. 14: Natürlich-sprachliche Mensch-Maschine-Kommunikation

Die vorliegende Arbeit hat die Objektivierung der Nasalität aus digitalisierten Daten als Ziel. Die Ergebnisse dieser Arbeit sollen in eine computergestützte, sprachgesteuerte Trainingsumgebung einfließen. Eine kurze Einführung über das Gebiet der automatischen Spracherkennung wird das Umfeld skizzieren. Es wird jedoch betont, dass das Anliegen dieser Arbeit weniger die Spracherkennung ist, d. h. *was* wurde gesprochen, sondern vielmehr die Sprecheranalyse, d. h. die Art und Weise der Aussprache. Damit führt der gemeinsame Weg bis zu den Verfahren der Merkmalsextraktion (vgl. Kapitel 3 „Verfahren der Merkmalsextraktion“) und scheidet sich dann an den Klassifikationsverfahren. Gute Einführungen in das Gebiet der Sprachverarbeitung im allgemeinen und Spracherkennung im speziellen sind unter anderem in [RJ93, Rus94, St95] zu finden.

2.4.1 Automatische Spracherkennung

2.4.1.1 Kategorisierung von Spracherkennungssystemen

Ein sehr wichtiger Punkt bei der Beschreibung und Bewertung von Spracherkennungssystemen besteht in der genauen Eingrenzung der Funktionalität und der Randbedingungen, unter denen die beschriebenen Systeme trainiert, getestet und eingesetzt werden. In der Literatur werden vor allem folgende Klassifikationskriterien für die verschiedenen Systeme verwendet (s. Abbildung 15) [Fel84, WDM88, WMG90, HBA93].

- Komplexität der zu erkennenden sprachlichen Äußerungen
- Abhängigkeit von Sprechereigenschaften
- Größe des Vokabulars

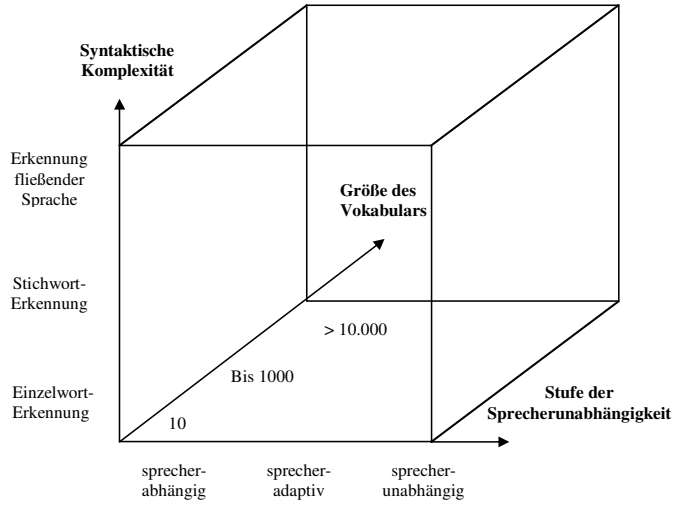

Abb. 15: Aufgabenraum der Spracherkennung [Hau93]

Bezüglich der Komplexität der zu erkennenden sprachlichen Äußerungen lassen sich z. B. die folgenden Formen unterscheiden, die nach zunehmendem Schwierigkeitsgrad geordnet sind [Fel84]

- Isolierte Wörter (diskrete Spracherkennung): Einzelwörter, die in ausreichendem Umfang von Pausenintervallen umgeben sind. Diese Pausenintervalle enthalten keine weitere Sprache, sondern lediglich Hintergrundgeräusche. Aufgrund dieser Definition fallen auch Wortketten, die deutlich erkennbare Pausen zwischen den einzelnen Wörtern enthalten, auch unter diese Kategorie.
- Verbundwörter: Folge von Einzelwörtern, die ohne eindeutig erkennbare Pausen zwischen den einzelnen Wörtern gesprochen werden.
- Schlüsselwörter: Wörter, die im Kontext weiterer fließend gesprochener Äußerungen auftreten.
- Kontinuierliche Sprache: Sie umfasst den weitaus größten Teil menschlicher Sprache, der nicht den beiden obigen Kategorien zugeordnet werden kann und in natürlicher Art und Weise von einem menschlichen Sprecher geäußert wird.

Seitens der Abhängigkeit der Spracherkennungssysteme von den Sprechereigenschaften lassen sich prinzipiell die folgenden Systeme unterscheiden:

- Sprecherabhängige Systeme: Sie erfordern ein für jeden Sprecher spezifisches Training des zu erkennenden Vokabulars, um eine Anpassung an die sprachlichen Charakteristika des jeweiligen Sprechers zu ermöglichen.

- Sprecheradaptive Systeme: Sie führen während der Spracherkennung eines Sprechers eine Anpassung an die spezifischen Merkmale dieses Sprechers durch, wodurch bessere Erkennungsraten des Spracherkennungssystems ermöglicht werden [HBA93].
- Sprecherunabhängige Systeme: Sie erkennen ohne weiteres Training die sprachliche Äußerungen einer Vielzahl von verschiedenen Sprechern, die durchaus extrem unterschiedliche sprachliche Charakteristika aufweisen können. Um eine sichere Erkennung zu gewährleisten, sind in der Regel umfangreiche Trainingsverfahren mit einer sehr großen Anzahl verschiedener Sprecher notwendig.

Weitere sehr wichtige Klassifikationskriterien von Spracherkennungssystemen stellen einerseits die Aufnahmebedingungen, unter denen die Trainings- und Testdaten erstellt werden, und andererseits die von dem eventuell verwendeten Übertragungskanal den Sprachdaten hinzugefügte Störungen dar. Zu den wichtigsten Klassifikationskriterien aus beiden Bereichen zählen:

- Bandbreite des verwendeten Sprachsignals
- Existenz und Art von Hintergrundgeräuschen
- Qualität und Variabilität der verwendeten Aufnahmegeräte und der jeweils benutzten Übertragungskanäle
- Stärke, Art und Dauer von Störungen des verwendeten Übertragungskanals

Insbesondere die Aufnahmebedingungen bei der Erstellung der jeweils verwendeten Sprachdatenbank spielen beim Vergleich verschiedener Spracherkennungssysteme eine dominierende Rolle. Nur bei genauer Angabe dieser Bedingungen ist ein objektiver Vergleich verschiedener Spracherkennungssysteme möglich. Speziell auf dem Gebiet der Simulation von Spracherkennungssystemen auf Rechnern ist daher die Verwendung einer für viele Forschungseinrichtungen verfügbaren Sprachdatenbank die wohl beste Möglichkeit, die geforderte Objektivität zu gewährleisten.

2.4.1.2 Allgemeines Prinzip von Spracherkennungssystemen

Das Ziel beim Einsatz eines Spracherkennungssystems besteht darin, menschliche Sprache in einer vorgegebenen Art und Weise zu klassifizieren. Um dies zu erreichen wird meist ein mehrstufiger Ansatz verwendet, indem die in Form von Signalen vorliegende menschliche Sprache in Merkmale und diese wiederum in Klassen umgewandelt werden.

Die einzelnen Verarbeitungsstufen eines Spracherkennungssystems lassen sich entsprechend Abbildung 16 wie folgt charakterisieren:

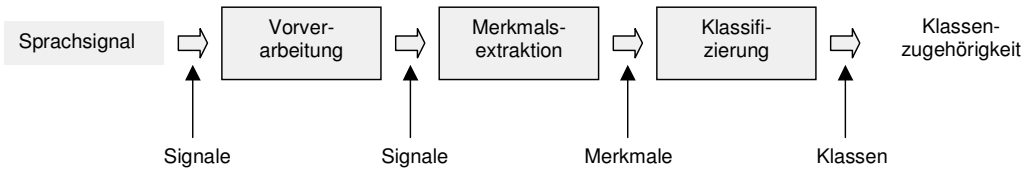

Abb. 16: Typischer Aufbau eines Spracherkennungssystems [Sch95]

Das digitalisierte Sprachsignal wird verschiedenen Vorverarbeitungsstufen unterzogen. Ein sehr wichtiges Verarbeitungsverfahren ist dabei die Pausenerkennung, deren Aufgabe die Unterscheidung zwischen Hintergrundgeräuschen bzw. Pausenintervallen und Sprachintervallen des zu analysierenden Sprachsignals ist. Das Ziel der Pausenerkennung variiert dabei zwischen der Detektion von Wort-, Satz- oder Absatzgrenzen, wobei der hauptsächliche Unterschied zwischen diesen Formen in der zeitlichen Lage der zu detektierenden Pausenabschnitte liegt. Weitere Verfahren der Vorverarbeitung dienen der Aufbereitung des Sprachsignals für die nachfolgend angewendeten Merkmalextraktionsverfahren.

Die Merkmalsextraktion dient dazu, die für die jeweilige Anwendung wichtigen Merkmale aus dem vorverarbeiteten Sprachsignal zu ermitteln. Das Ziel besteht dabei in einer möglichst großen Datenreduktion bei gleichzeitiger Erfassung der für die jeweilige Anwendung prägnanten Merkmale des Sprachsignals, welche auch noch möglichst robust gegen Störungen und Schwankungen in der Aussprache sind.

Die Klassifikation wertet die von der Merkmalsextraktion bereitgestellten Merkmale aus und ordnet diesen Klassen zu. Diese Zuordnung kann als weitere Datenreduktion interpretiert werden. Das Ergebnis der Klassifikation ist in der Regel ein Wort. Die dabei verwendete Unterteilung des Sprachsignals in die verschiedenen Klassen wird vorab festgelegt. Bei der Klassifikation hat sich die Technik der Hidden-Markov-Modelle (HMM) als das Standardverfahren etabliert[29]. Mit dieser statistischen Modellierung ist es möglich, die schwer durchschaubaren Zusammenhänge zwischen Spracheinheiten und ihren akustischen Gegenstücken in ein Wahrscheinlichkeitsmodell zu verpacken, dessen freie Parameter aus vorgelegten Sprachproben geschätzt werden konnten. Die außerordentliche Flexibilität und Generalisierungsfähigkeit der Markovmodelle löste in den letzten Jahren eine Lawine von Fortschritten hinsichtlich Sprecheranpassung, Erkennung großer Wortschätze, Grammatikmodellierung[30] und Echtzeitverhalten aus. Im Kampf gegen die verschleifungsbedingte

[29] Diese Markov-Ketten werden als verborgen bezeichnet, weil die Zustandsfolge selbst nicht beobachtet werden kann, sondern nur die durch zustandsabhängige Emissionsverteilungen generierten Merkmalsvektoren.

[30] Ein grammatisches Sprachmodell, bezieht allgemeinsprachliche oder anwendungsbedingte syntaktische, semantische oder pragmatische Restriktionen in die Wortfolgenerkennung ein. Bei umfangreichem Wortschatz wird dadurch der kombinatorischen Explosion des Lösungsraumes (hervorgerufen durch die uneingeschränkten Satzbildungsmöglichkeiten) wirkungsvoll begegnet.

Variabilität gesprochener Sprache wurde mit der Technik kontextabhängiger Wortuntereinheiten ein durchbruchartiger Erfolg erzielt.

Eine weitere Spielart parametrischen Lernens, die mittels gewöhnlichen Gradientenabstiegs trainierbaren Neuronalen Netzwerke, gewannen in jüngster Zeit der Signalverarbeitung zunehmend an Bedeutung. Neuronale Netze bestehen aus einer parallelen, vernetzten Struktur von primitiven Schaltelementen, die echten Nervenzellen nachempfunden sind. Gewisse Parameter dieser Elemente können sich in einer Lernphase automatisch so einstellen, dass bestimmte am Eingang anliegende Merkmalsvektoren ein bestimmtes Ergebnis am Ausgang liefern. Neuronale Netze sind sehr gut für die Spracherkennung geeignet; sie erweisen sich vor allem dann als besonders erfolgreich, wenn die Testmuster durch Störungen, etwa Umgebungsgeräusche, verfälscht sind. Neuronale Netze benötigen jedoch eine intensive Trainingsphase und sind, wenn die Eingangsvektorlänge nicht erheblich reduziert werden kann, sehr rechenintensiv [Kra90].

Der dritte Ansatz ist der traditionelle Mustererkennungsansatz (Pattern Recognition), indem ein gesprochenes Testmuster mit allen Referenzmustern verglichen wird. Das Referenzmuster, das dem vorliegenden am „ähnlichsten" ist, wird ausgewählt. Für den Vergleich, müssen die Muster auf eine gemeinsame Zeitachse gebracht werden, weil kein Mensch selbst das gleiche Wort zweimal mit der gleichen Geschwindigkeit ausspricht. Hier hat sich die dynamische Zeitanpassung (dynamic time warping) als sehr wirkungsvolles Verfahren erwiesen.

2.4.1.3 Stand der Spracherkennung

Eines der größten Probleme der Spracherkennung heutzutage stellt die Variabilität der Sprache und die damit einhergehende Kombinatorik dar[31]. Unter Variabilität versteht man im Kontext der Spracherkennung, dass ein und dieselbe Spracheinheit akustisch auf vielfältigste Weise realisiert sein kann. Ursachen für die Variabilität sind in spezifischen Eigenschaften des Aufnahmekanals (Typ und Position des Mikrophons, räumliche Reflektionseigenschaften, Diskretisierungsrauschen) oder akustische Störquellen (Stimmen weiterer Personen, Verkehrsgeräusch, Büro- oder Fertigungsumgebung) zu suchen. Ferner stellen sich Variationen aufgrund der Sprechweise (Tempo, Artikulationsdruck, Anspannung,

[31] Die rechenaufwendigste Komponente bei diesem Vorgehen stellt im allgemeinen die Suche mit einem Anteil von 60-90% dar. Da die Suche nach der wahrscheinlichsten Wortfolge ein rechenintensiver Prozess ist, empfiehlt es sich, sogenannte Pruning-Techniken in den Suchprozess aufzunehmen. Des weiteren hängt die Effizienz eines Suchverfahrens von der Strukturierung des Suchraums ab. Entscheidenden Einfluss auf die Suchraumstruktur nehmen die Organisation des Aussprachelexikons sowie das verwendete Sprachmodell. Insbesondere bei großem Wortschatz (20000 Wörter oder mehr) ist ein baumorganisiertes Lexikon effizienter als ein lineares Lexikon. Grund hierfür ist, dass viele Wörter mit derselben Phonemsequenz beginnen. Zudem konzentriert sich der Suchaufwand in den ersten drei Phonemen eines Wortes. Allerdings benötigt die Kombination eines Baumlexikons mit komplexen, linguistischen Modellen wie etwa Bigramm- oder Trigramm-Sprachmodelle Kopien des lexikalischen Baumes. Diese sogenannten Baumkopien sind entweder von der Worthistorie (wortabhängige Baumsuche) oder von der Startzeit des Baumes (zeitabhängige Baumsuche) abhängig.

Emotionen, Kooperativität) individueller Sprechermerkmale (Alter, Geschlecht, Vokaltraktanatomie, Gesundheitszustand) und habitueller Sprechermerkmale (Dialekt, Soziolekt) ein. Der vielleicht massivste Einfluss ist in der kontextuellen Aussprachevariation zu sehen. Abhängig von ihrer unmittelbaren lautlichen Umgebung und der Satzprosodie erhalten wir akustische Ausprägungen mit unterschiedlichsten Klang-, Intensitäts- und Rhythmuseigenschaften.

Seit Anfang der 80er Jahre dominiert der wahrscheinlichkeits- bzw. informationstheoretische Zugang zum Spracherkennungsproblem, der durch den weitgehenden Verzicht auf die Verwendung „handgefertigter" Aussprache- und Grammatikmodelle zugunsten effizienter Strategien des maschinellen Lernens aus akustischen und textuellen Sprachdaten geprägt ist [St95]. Die automatische Erkennung kontinuierlich diktierter Texte basiert im wesentlichen auf dem statistischen Ansatz der Bayes'schen Entscheidungsregel. Dabei werden zur Bestimmung der optimalen Wortfolge Wissensquellen wie Sprachmodell, Akustik und Phonetik in den Suchprozess integriert. Das Optimierungsproblem beruht auf der Maximierung des Produkts aus linguistischen und akustischen Wahrscheinlichkeiten. Das akustische Modell schätzt die Wahrscheinlichkeit, dass der gesprochene Satz eine bestimmte akustische Ausprägung hat, während das linguistische Modell (=Sprachmodell) die Wahrscheinlichkeit dieser Wortfolge erfasst. Mit beiden Wissensquellen zusammen lässt sich entscheiden, welche Satzhypothese die wahrscheinlichste ist. Die Ende der 80er Jahre wieder in Mode gekommenen neuronalen Netze eigneten sich zunächst nicht als Spracherkenner. Sie waren außerstande, dynamische Muster unterschiedlicher Länge effektiv zu verarbeiten. Erst 1991 zeigte der Belgier Herve Boulard[32] wie neuronale Netze mit HMM zu so genannten hybriden Modellen zu verschmelzen sind [BWM97]. Das Cambridger ABBOT-System hat diese Nische für sich entdeckt. In der Liga der Spitzenprodukte hält es einen vorderen Platz [Rob98]. Hybride Systeme könnten den lang ersehnten Technologieschub einläuten, durch den die sprecherunabhängige Fließtexterkennung in greifbare Nähe rückt.

Stand der Technik bei der automatischen Spracherkennung sind Diktiersysteme zur sprecherunabhängigen Erkennung kontinuierlicher Sprache. Für ein aktives Vokabular von ungefähr 100000 Wörtern bei anwendungsbezogener grammatischer Einschränkung des Lösungsraumes und einer Anpassung an den Sprecher gewährleisten diese eine sehr hohe Worterkennungsrate mit >95%. Verschiedene kommerzielle Systeme[33] ringen in diesem Markt um die Gunst der Anwender, und hier insbesondere um die Anwender mit einge-

---

[32] Die Beschreibung dieses Ansatz von Boulard und des Mitforscher Nelson Morgan gewann 1996 den IEEE Signal Processing Magazine Award.

[33] Alle derzeit erhältlichen Diktierprogramme für die deutsche Sprache basieren auf den Erkennungsmaschinen von VoiceXpress/Lernout&Hauspie, Freespeech/Philips, Naturally Speaking/Dragon Systems, und ViaVoice/IBM [Kuh00].

schränkten Fachvokabularen wie Radiologen, Juristen, etc. [Kuh00]. Konfrontiert man solche Systeme aber mit spontanen Äußerungen, ungeübten Benutzern und realistischen Übertragungskanälen fällt die Leistung jedoch in aller Regel drastisch ab [St95]. Wichtige Forschungszweige sind gegenwärtig die Verbesserung der Erkennung von Spontansprache[34] sowie die Erkennung in verrauschten Umgebungen. Um hier erfolgreich zu sein, konzentrieren sich die aktuellen Forschungsprojekte auf domänenspezifisches Wissen und die Integration noch weiterer Informationsquellen wie der Prosodie. Beispielhaft sei hier das deutsche Forschungsprojekt VERBMOBIL vorgestellt[35]. Im VERBMOBIL[36]-Vorhaben sollen spontansprachliche Äußerungen in den Domänen Terminabsprache, Reiseplanung und PC-Fernwartung in den Sprachen Deutsch, Englisch und Japanisch analysiert, in eine Zielsprache übersetzt und ausgesprochen werden. Das übergeordnete Ziel dieses Vorhabens ist die Entwicklung eines portablen Übersetzungsgerätes, welches auf Konferenzen mit Teilnehmern unterschiedlicher Muttersprachen die Dolmetscherfunktion übernimmt. Die multilinguale Verständigung erfolgt dabei in einer quasi-neutralen und weltweit verbreiteten Zwischensprache, konkret Englisch. Zurzeit kann Deutsch, Englisch und Japanisch verarbeitet werden. Im Vordergrund steht dabei die robuste und bidirektionale Übersetzung spontansprachlicher Dialoge aus den Domänen Reiseplanung und Hotelreservierung für die Sprachpaare Deutsch-Englisch (ca. 10.000 Wörter) und Deutsch-Japanisch (ca. 2.500 Wörter). Die Erkennungsrate beträgt aktuell 73.3 % bei klarer Artikulation und einem Umfang von ca. 2300 Wörtern [VM99][37].

---

[34] Spontansprache ist frei formulierte Alltagssprache, bei der ein Sprecher nicht etwa vorbereitete Texte vorliest. Gedankengänge werden fortlaufend in Sprache umgesetzt, wobei sehr häufig auch ungrammatische Sätze entstehen. Ein Spracherkennungssystem muss deshalb mit abgebrochenen Sätzen, Einschüben, und Selbstkorrekturen umgehen können. Nicht bedeutungstragende Äußerungselemente wie Räuspern, Schmatzen, ‚äh' und ‚ehm' müssen erkannt und für die weitere Analyse entfernt werden.

[35] Weitere deutsche Forschungsprojekte sind EVAR und JANUS. Bei EVAR handelt es sich um ein System des Lehrstuhls für Mustererkennung der Universität Nürnberg-Erlangen zur automatischen Bahnauskunft über Intercityverbindungen. JANUS hat ähnlich VERBMOBIL die Übersetzung spontangesprochener Äußerungen von einer Quellsprache in eine Zielsprache als Ziel. Gegenwärtig werden Deutsch, Englisch, Spanisch, Japanisch und Koreanisch unterstützt [Wai99].

[36] VERBMOBIL ist ein langfristig angelegtes, interdisziplinäres Leitprojekt des Bundesministeriums für Bildung, Wissenschaft, Forschung und Technologie (BMBF) im Bereich der Sprachtechnologie. Projektträger ist die Deutsche Forschungsanstalt für Luft- und Raumfahrt (DLR). In diesem Verbundvorhaben kooperieren Unternehmen der Informationstechnologie (IBM, Philips, Siemens, DaimlerChrysler, etc.), 19 Universitäten und Forschungszentren miteinander. Mit dem Projekt wird angestrebt, durch Fokussierung und Zusammenschluss möglichst vieler Know-How-Träger, Deutschland im nächsten Jahrtausend eine internationale Spitzenposition in der Sprachtechnologie und ihrer wirtschaftlichen Umsetzung zu sichern.

[37] Die Ingenieure von AT&T haben sich - laut einem Bericht von RTL- eine überraschende Computerlösung für Telefonie ausgedacht. Telefonieren soll ab jetzt mit Fremdsprachen-Übersetzung möglich sein. Wie es dazu heißt, soll eine Simultanübersetzungssoftware bei Reisen das Telefonieren in der jeweiligen Landessprache ermöglichen, ohne dass man dazu fundierte Kenntnisse der jeweiligen Sprache haben muss. Dank eingebauter Simultanübersetzung im Telefon will AT&T so das Einchecken in Hotels und das Bestellen von Tickets wesentlich erleichtern. Das Telefon erreicht eine Übersetzungsgenauigkeit von 91%. Die Übersetzungen in Englisch, Mandarin-Chinesisch und Spanisch entsprechen dieser Genauigkeit. Die Techniker wollen nun ein System entwickeln, dass Übersetzungen in 11 Sprachen ermöglicht, darunter Deutsch, Französisch, Japanisch, Hindi, Tamilisch, Koreanisch und Vietnamesisch. Das Telefon wird in diesem Jahr fertiggestellt werden, und AT&T denkt daran, die Technik in den nächsten Jahren zu kommerzialisieren [BdW-Agent 10.3.98].

3 Verfahren der Merkmalsextraktion

Eine wichtige Rolle in jedem Sprachverarbeitungssystem kommt der Merkmalsextraktion zu. Die Merkmalsgewinnung dient der Transformation der akustischen Spracheingabe in eine für die weiteren Verarbeitungsschritte geeignete parametrische Darstellung. Sie stellt der nachfolgenden Klassifikation diejenigen Größen zur Verfügung, die für den Entscheidungsprozess besonders gut geeignet sind. In unserem Fall werden Variabilitäten hervorgehoben, die den Sprecher und seine Sprechweise charakterisieren. Idealerweise sollten Merkmale die vom Äußerungsinhalt, Umgebungsgeräuschen sowie dem Übertragungsmedium herrühren, ausgeblendet werden. Ziel ist es nur solche Merkmale anzubieten, die für die Klassentrennung relevant sind[38]. Dies erleichtert die Konstruktion des Klassifikators erheblich, da nur wenige Merkmale verarbeitet werden müssen und das Problem der Trennung von wichtigen und unwichtigen Merkmalen bereits durchgeführt wurde.

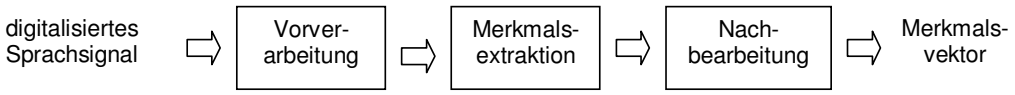

Abb. 17: Prinzipieller Datenfluss vom Sprachsignal zum Datenvektor

Direkt aus dem Zeitsignal berechenbare Merkmale spielen heute in der Sprachverarbeitung eine untergeordnete Rolle. Eine Ausnahme ist die Kurzzeitenergie des Signals, die ebenso wie die Nulldurchgangrate [RS78] brauchbare Hinweise zur Unterscheidung stimmhafter und stimmloser Laute liefert (s. Kapitel 3.1.3 „Pausenerkennung“). Bei der Sprachverarbeitung wird im allgemeinen nicht das Zeitsignal des Mikrophons (Schalldruckverlauf) selbst ausgewertet, sondern die spektrale Zusammensetzung des Sprachlautes. Zur Gewinnung des Spektrums setzen die meisten Verfahren dabei auf der diskreten (Fast-) Fourier Transformation und der z-Transformation auf. Der Unterschied bei den verschiedenen Merkmalextraktionsverfahren besteht letztlich in der Art und Weise, wie aus den durch die jeweiligen Transformationen erzeugten Koeffizienten möglichst kompakte, das Sprachsignal repräsentierende Merkmalsvektoren erzeugt werden. Hier setzen die gegenwärtigen Techniken der Merkmalgewinnung vorwiegend auf Funktionsmodelle für die Sprachproduktion und -perzeption auf.

[38] In der Regel erfolgt mit der Merkmalsextraktion auch eine Datenreduktion, so dass diese Verfahren auch bei Kompressionsalgorithmen Anwendung finden.

Dieses Kapitel beginnt mit der Beschreibung der für diese Arbeit relevanten Vorverarbeitungen. Unter Vorverarbeitung werden im Kontext dieser Arbeit alle Verfahren zusammengefasst, die vor der Berechnung des Spektrums erfolgen, also Arbeiten auf dem Zeitsignal. Anschließend werden die gängigen Transformationen, FFT und z-Transformation, und ihre Anwendungen vorgestellt. Dies sind Spektral- und Cepstralanalyse sowie Lineare Prädiktion. Danach erfolgt die Beschreibung der auf das Spektrum angewandten Verfahren zur Berücksichtigung der gehörorientierten Eigenschaften (=Nachbearbeitung). Neben diesen spektralen Methoden, weisen einige Vorabuntersuchungen darauf hin, dass die Nasalität auch über die Grundfrequenz bestimmbar ist. Daher wird am Ende noch auf die Grundfrequenzanalyse eingegangen.

Um die verschiedenen Methoden der Merkmalsextraktion und ihrer gewonnenen Parameter hinsichtlich ihrer Güte zur Nasalitätsmessung bewerten zu können, werden in Kapitel 8 „Klassifikationsergebnisse“ die Verfahren in ihrer Wirksamkeit bei der Bestimmung der Sprachparameter von stimmhaften Lauten gegenübergestellt.

3.1 Vorverarbeitung

3.1.1 Digitalisierung

Nach der Aufnahme von Sprache mit einem Mikrophon liegt die Schallwelle in Form eines elektronischen Signals vor, das durch die reellwertige, kontinuierliche Zeitfunktion f(t) beschrieben wird. Dieses Analogsignal ist somit zeit- und wertkontinuierlich. Zur Weiterverarbeitung auf einem Digitalrechner müssen Definitions- und Wertebereich des Signals diskretisiert werden. Hierzu wird das Signal zuerst an einer endlichen Zahl äquidistanter Stützstellen $n \cdot T$ ($n = 1,2,...$). abgetastet; wir bezeichnen T[s] als Abtastperiode bzw. $f_A = 1/T$ [Hz] als Abtastfrequenz. Nach dem Shannon’schen Abtasttheorem lässt sich das Originalsignal vollständig rekonstruieren, wenn das Analogsignal bandbegrenzt ist durch eine höchste vorkommende Frequenz F_B. Die Abtastfrequenz muss dabei mindestens doppelt so hoch wie F_B sein. Reale Sprachsignale müssen daher vor der Abtastung tiefpassgefiltert werden, da die Anwesenheit von Frequenzkomponenten oberhalb $f_A/2$ sonst zu einer Überschneidung führt („spectral aliasing“) [Cou86]. Für Sprache ist eine Bandbreite von 10 kHz hinreichend [Fur89]; es kann daher mit 20 kHz abgetastet werden[39]. Die resultierende Folge von Abtastwerten $f_n := f(n \cdot T)$ wird anschließend quantisiert. Der kontinuierliche Amplituden-

[39] Die analoge Übertragung im öffentlichen Telefonnetz ist dagegen auf das Frequenzband 300 Hz – 3.4 kHz beschränkt, so dass eine Abtastung mit 8 kHz genügt.

bereich der Abtastwerte wird durch eine begrenzte Anzahl von Amplitudenstufen ersetzt und jeder Amplitudenwert durch seinen Repräsentanten ausgetauscht. Soll jeder Abtastwert mit b Bits kodiert werden, stehen 2^b Stufen zur Verfügung. Ein Sprachsignal von einer Sekunde Dauer wird somit in $f_A b$ Bits gespeichert[40]. Durch die Quantisierung entsteht ein Fehler, der im Maximalfall gleich der halben Intervallbreite ist. Bei grober Quantisierung enthält das Fehlersignal deterministische vom Eingangssignal abhängige Anteile. Ist die Quantisierung hinreichend fein, d. h. die Intervallbreite klein genug, hat das Fehlersignal einen rauschförmigen Charakter. Für die bei heutigen Systemen verwendete Quantisierung von mindestens 8 Bit je Abtastwert ist dies weitgehend der Fall. Die Intervallbreite ist hierbei so klein, dass eine gleichmäßige Verteilung des Quantisierungsfehlers angenommen werden kann [Fel84]. Die Sprachpassagen der Datenbank NASAL wurden mit 22 kHz Samplerate und 16 Bit-Kodierung aufgenommen, so dass die Abtast- und Quantisierungsfehler vernachlässigbar sind.

3.1.2 Eliminierung des Gleichspannungsanteils

Untersuchungen der Amplitudendichteverteilung (ADV) von menschlicher Sprache ergaben, dass diese durch eine Gamma-ADV mit dem Mittelwert Null und der Varianz σ_x^2 angenähert werden kann [Fel84, TS95]:

$$p(x) = \sqrt{\frac{\sqrt{3}}{8\pi\sigma_x|x|}}\, e^{-\frac{\sqrt{3}|x|}{2\sigma_x}} \qquad (13)$$

Der Verlauf von p(x) ist in Abbildung 18 grafisch dargestellt, wobei entlang der Abszisse des Verhältnis von Sprachsignalamplitude x und der Standardabweichung σ_x aufgetragen ist.

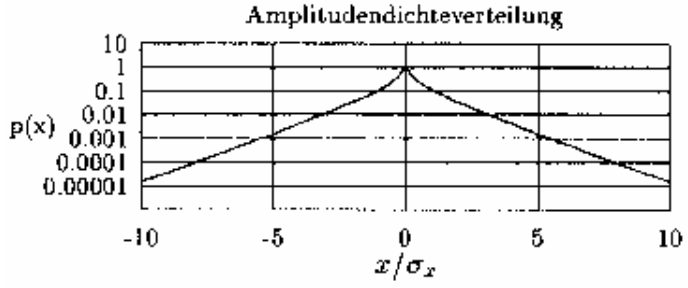

Abb. 18: Approximation der Amplitudendichteverteilung von Sprache durch eine Gamma-ADV [Fel84]

[40] Sind die Stufen gleichmäßig verteilt, spricht man von einer uniformen Quantisierung. Weit verbreitet ist auch die logarithmische Codierung der Amplitudenwerte wie z. B. beim µ-Law-Format.

Da diese ADV symmetrisch zum Mittelwert 0 ist, muss die Summation aller Abtastwerte x_i über der Zeit den Wert 0 ergeben:

$$\sum_i x_i = 0 \qquad (14)$$

3.1.3 Pausenerkennung

Die Anwendung von Sprachverarbeitungssystemen, die auf isoliert gesprochenen Wörtern basieren, erfordert sehr häufig eine relativ genaue Bestimmung der Anfangs- und Endpunkte des zu erkennenden Wortes. Diese Bestimmung wird in der Literatur entweder Pausenerkennung oder Endpunktdetektion genannt, wobei beide Begriffe das gleiche Prinzip beschreiben. In der Literatur existiert eine Reihe von Verfahren, mit deren Hilfe die Pausenerkennung innerhalb eines Sprachsignals realisiert werden kann. Zu den sehr häufig verwendeten Verfahren zählen das Energieschwellverfahren und das Nulldurchgangsratenverfahren, welche auch im Rahmen dieser Arbeit untersucht wurden. Insbesondere das Energieschwellverfahren erwies sich als sehr geeignet[41].

3.1.3.1 Energieschwellverfahren

Das Energieschwellverfahren stützt sich darauf, dass Sprachintervalle in der Regel eine höhere Energie besitzen als Pausenintervalle [LRR81, WRM84]. In der Praxis treten jedoch oft Probleme z. B. durch hohe Hintergrundgeräuschpegel auf. Nach Rabiner [Rab75] ergeben sich Schwierigkeiten, wenn die phonetische Zusammensetzung des zu detektierenden Sprachsignals in eine der folgenden Gruppen eingeordnet werden kann:

- schwach gesprochene Frikative[42] zu Beginn oder Ende einer Äußerung (im Deutschen z. B. /f/, /s/, /x/, /z/)
- schwach gesprochene Plosive (im Deutschen /p/, /b/, /t/, /d/)

---

[41] In Spracherkennungssystemen mit großen Vokabular wird häufig die Mahalobis-Klassifikation eingesetzt. Die Grundidee des Verfahrens besteht in der Verwendung verschiedener Merkmale des Sprachsignals (z. B. Energie, Autokorrelationskoeffizienten, LPC-Koeffizienten) zur Klassifikation in Sprach- oder Pausenintervalle. Dabei findet die Mahalanobis-Distanz zwischen dem jeweiligen Merkmalsvektor und den während des Trainings ermittelten Referenzvektoren jeder Klasse als Klassifikationskriterium Anwendung. Um hinreichend genaue Klassifikationsergebnisse zu erzielen können, ist das Training der Referenzvektoren mit umfangreichen Sprachmaterial notwendig, das zuvor in Pausen- und Sprachintervalle unterteilt wurde. Da der Rechenaufwand bei der Mahalanobis-Klassifikation gegenüber den verwendetem Energieschwellenverfahrens jedoch um einiges höher ist, und eine bessere Pausenerkennung bei isolierten Vokalen in der Literatur nicht beschrieben wurde, wurde diese hier nicht verwendet. Eine genauere Beschreibung des Verfahrens ist in [Sch95] zu finden. Eine erfolgreiche Verwendung des Mahalanobis-Verfahren im Rahmen eines Sprecherverifizierungssystems wird bspw. in [Fli91] beschrieben.

[42] Frikative = Reibelaute: Der Luftstrom wird im Mund- oder Rachenraum eingeengt, so dass ein rauschartiger Laut entsteht.

- finale Nasale (im Deutschen z. B. /m/, /n/).

Hingegen eignet sich dieses Verfahren sehr gut für die Detektion von Vokalen, insbesondere wenn sie, wie in unserem Fall, isoliert vorliegen.

Die Kurzzeitenergie ergibt sich nach derselben Berechnungsvorschrift wie die Langzeitenergie

$$E = \sum_{n=-\infty}^{\infty} \left| f_n \right|^2 \qquad (15)$$

nur dass die jeweilige Kurzzeitvariante des Signals eingesetzt wird, d. h. für ein endliches Fenster der Länge N ergibt sich die Form:

$$E^{(m)} = \sum_{n=0}^{N-1} a_n \left| f_{m+n} \right|^2 \qquad (16)$$

In Abbildung 19 sind die Abtastwerte und die Signalenergie des Vokals /a/ dargestellt, wobei die Signalenergie jeweils über eine Fenstergröße von 400 Abtastwerten berechnet wurde und die Abtastfrequenz gleich 22050 Hz war. Dabei ist deutlich zu erkennen, dass der Hauptteil des Wortes anhand der Energiekontur zwischen ca. 11000 Samples und 19000 Samples erkannt werden kann. Zwischen 1000 Samples und 2000 Samples trat ein Mikrophonknacken ein, um die 10000 Samples ein Schmatzgeräusch.

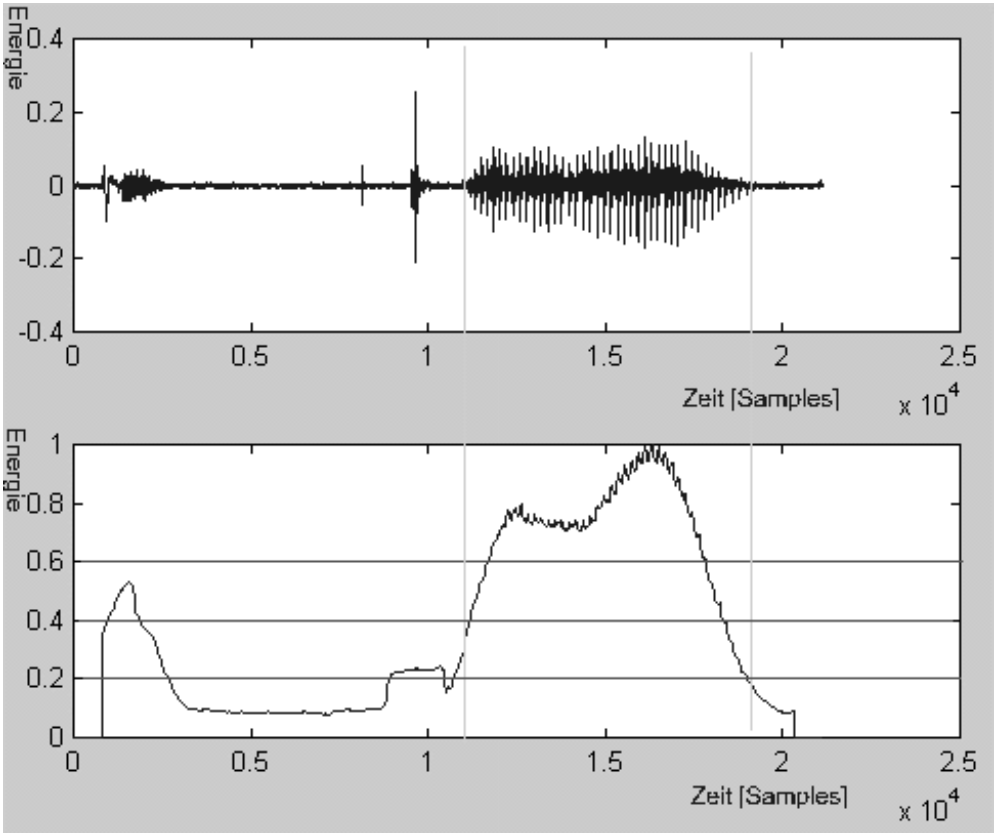

Abb. 19: Grafische Beschreibung des Energieschwellverfahrens

Das bei den Versuchen verwendete Pausenerkennungsverfahren orientiert sich in der Implementierung eng an [TS95]. Es wurden einige Erleichterungen aufgrund der isolierten Vokale vorgenommen. Das Verfahren beruht im wesentlichen auf einem Top-Down-Algorithmus der zweistufig aufgebaut ist:

Die Aufgabe der ersten Stufe besteht darin, auf der Basis von Signalabschnitten sehr hoher Energie eine grobe Abschätzung der Anfangs- und Endpunkte des zu detektierenden Wortes zu finden.

Die zweite Stufe des Verfahrens berechnet anhand der von der ersten Stufe erzeugten initialen Werte den möglichst exakten Anfangs- bzw. Endpunkt des Wortes anhand der gefundenen Energieverteilung. Dabei werden auch verschiedene Randbedingungen bzgl. der zeitlichen Dauer von Sprach- und Pausenintervallen beachtet.

Anhand der in Abbildung 19 exemplarisch dargestellten Energieverteilung soll das implementierte Verfahren nachfolgend kurz beschrieben werden, wobei E_{speech} die mittlere Energie der Sprachintervalle und E_{noise} die der Pausenintervalle bezeichnet.

Erster Verarbeitungsschritt

Für jedes zu bearbeitende Sprachintervall wird die Energie aller Abtastwerte x_i dieses Intervalls berechnet:

$$E = \frac{1}{n}\sum_{i=1}^{n}(x_i^2) \qquad (17)$$

Danach werden alle Intervalle auf die maximale Energie der Sprachaufnahme normiert. Ein Schwellwert zwischen den Klassen Sprache und Pause wird auf einen experimentell ermittelten Schätzwert gesetzt (in der Abbildung 19 gilt E_{change} = 0.6). Damit wird ein unbekanntes Intervall wird wie folgt klassifiziert:

$E > E_{change}$ $\Rightarrow$ Sprachintervall

$E \leq E_{change}$ $\Rightarrow$ Pausenintervall

Zweiter Verarbeitungsschritt

Das Sprachintervall aus dem ersten Verarbeitungsschritt wird durch zwei initiale Abtastwerte ($begin_{initial}$) und ($end_{initial}$) beschrieben. Um die endgültigen Begrenzungswerte zu erhalten, wird der Anfangspunkt nach links verschoben, bis seine Energie = 0.4 beträgt. Der Endpunkt wird nach rechts verschoben, bis seine Energie = 0.2 beträgt. Die Werte für die Schwellen wurden experimentell ermittelt.

Um zu verhindern, dass entweder zu kurze Sprachintervalle bzw. länger dauernde Hintergrundgeräusche als Sprache klassifiziert werden, muss der Abstand zwischen (begin) und (end) eine Mindestlänge besitzen. In meinen Untersuchungen betrug dieser 0.3 Sekunden. Eine visuelle und auditive Überprüfung von über 2000 Vokalen ergab eine Erfolgsrate mit Einbezug der Mindestlänge von 99.4%. Die fehlerhaften Sprachdateien wurden manuell korrigiert und konnten daher für die Untersuchungen verwendet werden. Insgesamt kann man das Energieschwellverfahren für isolierte Vokale bei guter Sprachaufnahmequalität als sehr präzise bezeichnen.

3.1.3.2 Nulldurchgangsratenverfahren

Eine hohe Nulldurchgangsrate (und damit eine hohe Frequenz) ist ein Kriterium für stimmlose Laute, während eine niedrige Rate auf einen stimmhaften (quasistationären) Laut hinweist [Fel84]. Das Nulldurchgangsratenverfahren kann eigenständig nicht direkt zur Pausenerkennung eingesetzt werden. Es existiert eine Reihe von stimmlosen Lauten, deren Charakteristika denen des Rauschens relativ ähnlich sind. In der Literatur wird die Nulldurchgangsrate daher lediglich zur Korrektur der durch die Energieschwellverfahren detektierten Grenzen verwendet. Die in [Rab75] beschriebenen Untersuchungen zeigen, dass unter zusätzlicher Verwendung der Nulldurchgangsrate die Ergebnisse des Energieschwellverfahrens verbessert werden können. Analoge Untersuchungen zum Einsatz der Nulldurchgangsrate als Nachklassifikator für von den Energieschwellverfahren inkorrekt erkannte Pausenintervalle beschreibt Savoji in [Sav89]. Laut [TS95] ist eine Verbesserung

aber nur bei Sprachaufnahmen hoher Qualität zu beobachten. Bei nicht idealen Bedingungen, wie z. B. bei Telefonkanälen, deren Bandbreite auf 3 kHz beschränkt ist, waren durch die Verwendung des Nulldurchgangsratenverfahrens keine Verbesserungen zu verzeichnen. Eigene Untersuchungen zur Güte des Nulldurchgangrateverfahrens wurden aufgrund der guten Ergebnisse des Energieschwellverfahrens nicht vorgenommen.

In Abbildung 20 wurde in einem Zeitfenster von jeweils 100 Samples die Nulldurchgangsrate an dem Wort „Fahrrad“ bestimmt. Deutlich hebt sich der Vokal /a/ durch die geringe Rate von den Hintergrundgeräuschen ab[43].

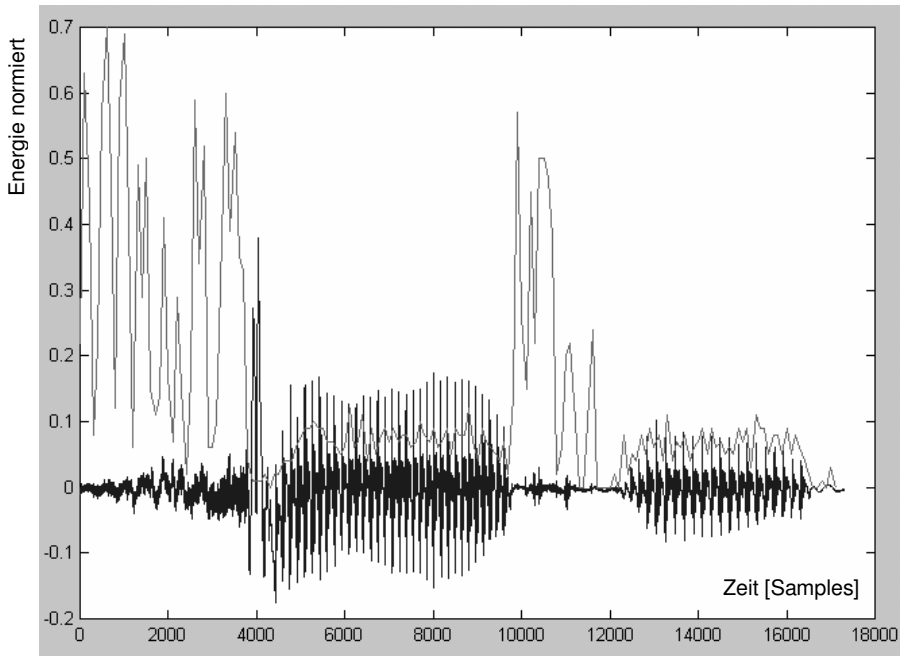

Abb. 20: Nulldurchgangsrate am Beispiel des Wortes „Fahrrad“

In der Abbildung 21 wird zur Abgrenzung des Vokals von den Hintergrundgeräuschen die Anzahl der positiven Amplituden pro Fenster berechnet. Bei periodischen Signalen wie bei dem dargestellten Vokal /a/ sollte diese um 50 % liegen. Auch dieses Maß würde sich zum Herausfiltern der Hintergrundgeräusche eignen.

[43] Streng genommen, stellen bei diesem Beispiel die Konsonanten /f/ und /r/ die Störgeräusche dar.

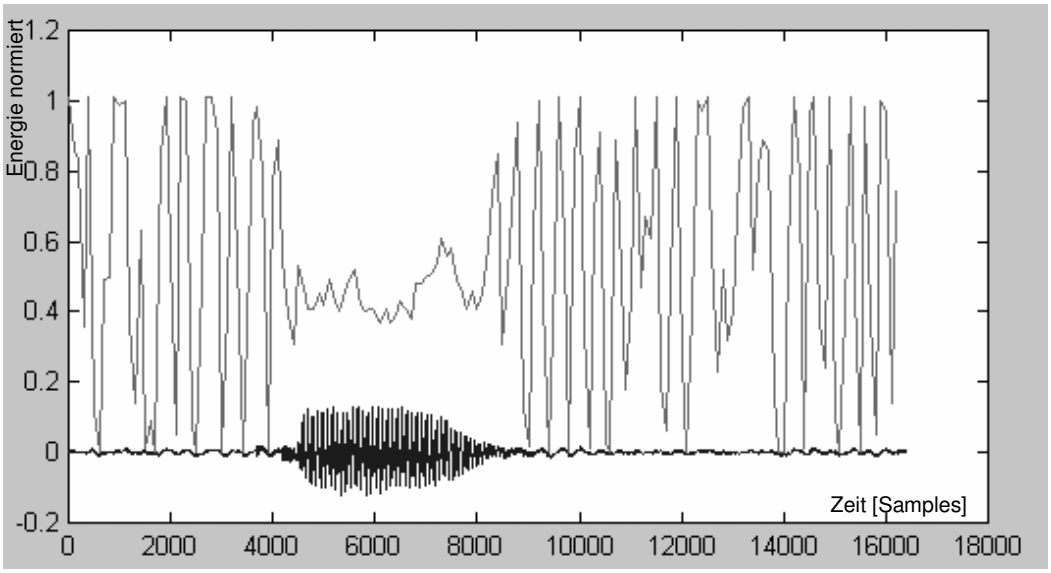

Abb. 21: Anzahl der positiven Amplitudenwerte (Fensterbreite = 150 Samples)

3.1.4 Preemphase

Wie im Kapitel 2.2.1 „Stimmhafte Laute“ erwähnt, nimmt die Einhüllende des Spektrums in Richtung höherer Frequenzen um 12 dB pro Oktave ab. Um diesen Abfall zu kompensieren, erfolgt häufig eine Anhebung der höheren Frequenzen (Preemphase) mit einem digitalen Hochpass der Übertragungsfunktion:

$$H(z) = 1 - \mu z^{-1} \qquad 0.9 \leq \mu \leq 1.0 \qquad (18)$$

Der Preemphasefaktor μ wird dabei aus dem Bereich $0.9 < \mu < 0.98$ gewählt [Fel84, RJ93].

3.2 Transformationen

In der digitalen Signalverarbeitung nehmen die Frequenztransformationen eine zentrale Stellung ein. Durch sie ist es möglich, die in einem Zeitsignal enthaltenen Frequenzen und ihre Energie zu ermitteln. Der prominenteste Vertreter dieser Gattung ist die Fouriertransformation, welche eine Zeitfunktion als Summe von Sinus- und Cosinus-Anteilen darstellt. Die Fouriertransformierte S(y) eines zeitkontinuierlichen Signals s(t) ist definiert als[44]

[44] Die trigonometrischen Funktionen und die Exponentialfunktion sind durch die Euler’sche Formel $\cos\varphi + i*\sin\varphi = e^{i\varphi}$ miteinander verknüpft.

$$S(y)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}s(t)\cdot e^{-iyt}dt \qquad (19)$$

Durch diese Transformation entsteht kein Informationsverlust. Es handelt sich um eine andere, oft nützlichere Darstellung der zu analysierenden Funktion. Aus dem transformierten Signal kann, aufgrund der Bijektivität, das Ursprungssignal mittels der inversen Fouriertransformation rekonstruiert werden [BS89].

$$s(t)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}S(y)e^{iyt}dy \qquad (20)$$

Die Fouriertransformierte ist eine komplexe Funktion, bestehend aus einem Realteil und einen Imaginärteil. Den Betrag der Fouriertransformierten nennt man Amplitudenspektrum. Das Quadrat des Amplitudenspektrums wird als Leistungsdichtespektrum bezeichnet.

Die kontinuierliche Fouriertransformation verknüpft kontinuierliche Funktionen im Zeitbereich und Frequenzbereich miteinander. Bei Ausführung dieser Transformation mit einem Digitalrechner treten zwei Probleme auf:

- Es können nur endlich viele Werte s(n) verarbeitet werden, weil der Speicherplatz in einem Rechner endlich ist.
- Neben der Zeitvariablen muss auch die Frequenzvariable diskretisiert werden, weil der Rechner nur diskrete Zahlenwerte verarbeiten kann.

Man bedient sich bei Digitaldaten der sog. Diskreten Fouriertransformation (DFT). Zur Herleitung der DFT aus der kontinuierlichen Fouriertransformation wird auf die Literatur verwiesen (unter anderem [Bri87, Rus94, Sch92]).

Für N gegebene Abtastwerte s(0), s(1),..., s(N-1) ist die DFT definiert als:

$$S(k)=\sum_{n=0}^{N-1}s(n)e^{-ik\frac{2\pi}{N}n} \quad \text{für } k=0,1,\ldots,N-1 \qquad (21)$$

die inverse DFT als:

$$s(n)=\frac{1}{N}\sum_{k=0}^{N-1}S(k)e^{ik\frac{2\pi}{N}n} \quad \text{für } k=0,1,\ldots,N-1 \qquad (22)$$

Die praktische Berechnung der DFT erfordert $O(N^2)$Multiplikationen. Durch den Einsatz der schnellen Fouriertransformation (FFT, fast fourier transform) [CT65] reduziert sich der Aufwand auf $O(N\log N)$[45].

Eine nachteilige Eigenschaft der Fouriertransformation ist, dass keine zeitliche Lokalisierung der auftretenden Frequenzen vorhanden ist. Sprachsignale stellen nichtstationäre Signale dar; ihre spektralen Eigenschaften ändern sich wenigstens von Laut zu Laut. Nur über sehr kurze, etwa 5-30 ms andauernde Zeitabschnitte hinweg kann das Signal näherungsweise stationär angesehen werden. Will man Aussagen über das zeitliche Auftreten von Frequenzen machen, ist eine Fensterung notwendig. Dabei muss sichergestellt werden, dass alle ursprünglichen Sprachsignalwerte in ausreichender Art und Weise in das Fensterungsresultat eingehen. Aus diesem Grund wird die jeweils verwendete Fensterfunktion überlappend auf das Sprachsignal angewendet (s. Abbildung 22). Typische Werte der Fenstergröße bewegen sich in dem Bereich zwischen 10 ms bis 50 ms; entsprechend werden oft Verschiebungsgrößen zwischen 10 ms und 20 ms verwendet.

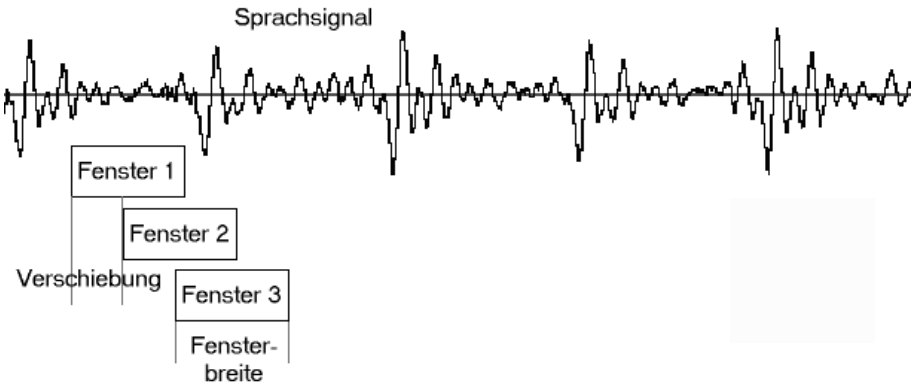

Abb. 22 Fensterung

Zu jedem Zeitpunkt m ergibt sich ein modifiziertes Signal $f^{(m)}$ mit $f_n^{(m)} = f_n \cdot w_{m-n}$, welches in einer kleinen Umgebung von m eine gewichtete Version von f_n darstellt und außerhalb gewöhnlich verschwindet. Dies entspricht einer Faltung des auf den Rahmen eingeschränkten Signals mit der Fensterfunktion. Beispiele für endliche Fensterfunktionen sind (für $n < 0$ und $n >= N$ sei jeweils $w_n = 0$):

[45] Eine weitere Einsparung erbringt das Ausnutzen der Reellwertigkeit von k mit Hilfe der schnellen Hartley-Transformation [Bra90].

Fenster	**Zeitfunktion**
Rechteckfenster	$w_n^R = 1$
Hamming-Fenster	$w_n^M = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right)$
Hanning-Fenster	$w_n^N = 0.50 - 0.50\cos\left(\frac{2\pi n}{N-1}\right)$
Gauß-Fenster, $\sigma = 3$	$w_n^G = e^{-0.5\left(\frac{n-N/2}{\sigma N/2}\right)^2}$
Parabel-Fenster	$w_n^P = 4\frac{n}{N}\left(1 - \frac{n}{N}\right)$

Tab. 2: Fensterfunktionen

Die Wahl der Fensterfunktion w_n hat einen entscheidenden Einfluss auf die nachfolgenden Analysen. Meist wird von der Fensterfunktion gefordert, dass sie im Frequenzbereich möglichst schmal, d. h. lokalisiert ist und im Zeitbereich schnell abklingt. Einen guten Kompromiss bietet das so genannte Hamming-Fenster [RS78, Fel84, DPH93]. Der Frequenzauflösung einer Fensterfunktion w_n sind aufgrund eines Unschärfeprinzips enge Grenzen gesetzt. Dieses Prinzip besagt, dass die Ausdehnungen von w_n und $W(e^{iw})$ (im Sinne der Varianzen σ_w^2 und σ_W^2) umgekehrt proportional sind. Insbesondere gilt die untere Schranke $1/(4\pi) <= \sigma_w \sigma_W$, die z. B. vom Gaußfenster erreicht wird. Wegen des bereits erwähnten Unschärfeprinzips lässt sich die erstere nur auf Kosten der letzten erhöhen. Dies wird in Abbildung 23 an der Frequenzverschmierung des Breitband- und an der zeitlichen Verschmierung des Schmalbandspektogramms dargestellt. Die Analyse von Sprachsignalen erfordert entweder eine hohe Zeitauflösung, etwa zur Detektion der kurz dauernden Plosiven, oder eine hohe Frequenzauflösung, um nahe beieinander liegende Formanten zu unterscheiden – jedoch nie beides zugleich. Eine gute Lösung bieten Spektralrepräsentationen deren Frequenzauflösung proportional zu ω verläuft. Zeit-Frequenz-Darstellungen mit solch einer variablen Auflösungscharakteristik sind die Wavelet-Transformationen [RV91, Dou92]. Wavelet-Transformationen bauen auf nicht sinusförmigen Basiskomponenten auf. Der große Vorteil von Wavelets ist ihre inhärent von der analysierenden Frequenz abhängige lokale zeitliche Auflösung, d. h. sie benötigen keinen zusätzlichen Aufwand, um die gewünschte Frequenzauflösung mit einer guten zeitlichen Auflösung zu verbinden.

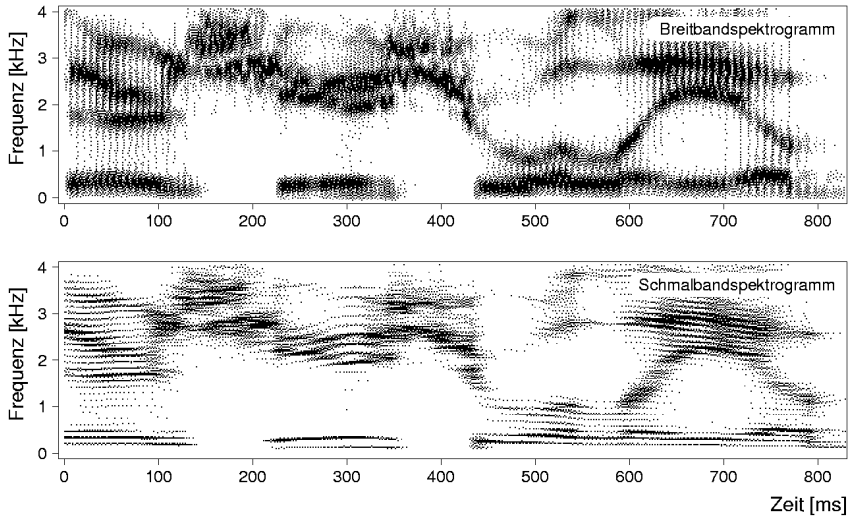

Abb. 23: Breitband- und Schmalbandspektogramm (Frequenzauflösung 125 Hz bzw. 19 Hz) [St95]

Eine weitere häufig angewandte Transformation ist die z-Transformation [BS89]. Eine Folge y(n) hat eine z-Transformierte Y(z) = Z{y(n)} mit:

$$Y(z) = \sum_{n=-\infty}^{\infty} y(n) \cdot z^{-n} \qquad (23)$$

Setzt man für z das Argument $e^{i\omega T}$ ein, so ist die z-Transformierte identisch mit der zeitdiskreten Fouriertransformation, d. h. die z-Transformation enthält die DFT als Spezialfall [Sch92].

Die z-Transformation hat eine große praktische Bedeutung für die Beschreibung von digitalen Filtern. Digitale Filter stellen lineare Systeme dar, durch welche ein Signal transformiert und umgewandelt wird. Digitale Filter werden häufig zum Ausblenden unerwünschter Störgrößen entworfen.

Im allgemeinen, steht die z-Transformierte Y eines Ausgangs eines digitalen Filters zu der z-Transformierten X des Eingangssignals in folgender Beziehung:

$$Y(z) = H(z)X(z) = \frac{b_0 + b_1 z^{-1} + \ldots + b_N z^{-N}}{a_0 + a_1 z^{-1} + \ldots + a_M z^{-M}} X(z) \qquad (24)$$

H(z) stellt die Transferfunktion des Filters, die b_i und a_i die Koeffizienten des Filters dar. Die Ordnung des Filters ist gleich dem Maximum aus {M, N}. Anhand der Ordnung des Zählers und Nenners kann man digitale Filter grob unterscheiden. Gilt N = 0 und M > 0, d. h. es gibt nur ein skalares b_0, nennt man das Filter ein autoregressives (AR-) Filter. Bei N > 0 und

M = 0, d. h. es gibt nur ein Skalar a_0, spricht man von moving average (MA-) Filtern. Sind sowohl M und N größer 0, nennt man das Filter autoregressive moving average (ARMA)-Filter[46].

Bei MA-Filtern hängt das Ausgangssignal nur von den Werten des Eingangssignals ab. Dadurch ist das MA-Filter und mit ihm sein Lineares System eindeutig bestimmt:

$$y(n) = b_0x(n) + b_1x(n-1) + \ldots + b_Mx(n-M) \qquad (25)$$

Ein MA-Filter kann wegen der fehlenden Rückkopplung nicht schwingen und ist somit immer stabil. Bei AR-Filtern gehen in das Ausgangssignal auch zurückliegende Werte des Ausgangssignals mit ein. Anders formuliert bedeutet dies: y(n) ist direkt abhängig von genau einem x(n) und mindestens einem vorhergehenden y(n-j). Es gilt:

$$y(n) + a_1y(n-1) + \ldots + a_My(n-M) = x(n) \qquad (26)$$

mit der z-Transformierten:

$$Y(z) = \frac{1}{1+\sum_{j=1}^{M} a_j z^{-j}} X(z) \qquad (27)$$

ARMA-Filter sind analog zu den AR-Filtern. Hier kommt aber noch eine Kombination mit den MA-Filtern hinzu, so dass gilt:

$$y(n)+a_1y(n-1)+\ldots+a_My(n-M) = b_0x(n)+b_1x(n-1)+\ldots+b_Nx(n-N) \qquad (28)$$

mit der zugehörigen z-Transformierten:

$$Y(z) = \frac{\sum_{j=0}^{N} b_j z^{-j}}{\sum_{j=0}^{M} a_j z^{-j}} X(z) \qquad (29)$$

Betrachten wir dies als Modell für die Spracherzeugung in den Sprachorganen des Menschen, so erhalten wir ein sehr gutes Modell. Es ist recht aufwendig und rechenintensiv. Der untere Teil des Bruchs liefert dabei den des normalen Wegs der Spracherzeugung,

[46] Andere Benennungen in der Literatur für ARMA-Filter sind rekursive Filter, Pol-Zero-Filter oder IIR- (infinite impulse response). AR-Filter werden auch als all-pole-Filter, MA-Filter als nicht-rekursive Filter, all-zero-Filter oder FIR (finite impulse response) bezeichnet.

außer im Nasenraum, also von den Stimmbändern bis zu den Lippen (vgl. Kapitel 3.5 „Lineare Prädiktion“). Der Zähler liefert bei den Frequenzen, welche im Nasenraum verschluckt werden, die entsprechenden Nullstellen und modelliert somit die nasalen Laute.

Zusammenfassend lässt sich sagen, dass die „beste“ Transformation von der Anwendung abhängt. Im Rahmen dieser Arbeit werden stimmhafte Laute untersucht, welche über den ganzen betrachteten Zeitraum quasistationär sind. Die Zeitauflösung ist daher vernachlässigbar. Wichtig ist eine optimale Frequenzauflösung, welche man durch eine diskrete Fouriertransformation oder eine z-Transformation des gesamten Signals enthält. Diese beiden Transformationen waren daher auch die Grundlage der Untersuchungen.

3.3 Spektralanalyse

Die spektrale Energieverteilung einer Lautrealisierung resultiert, wie in Kapitel 2.2 „Sprachproduktion“ diskutiert, aus der glottalen Anregung des Sprachschalls und seiner weiteren Ausformung im Vokaltrakt. Der (für stimmhafte Laute) periodische Anregungsschall geht idealerweise als Linienspektrum ein, dessen nichtverschwindende Komponenten die Stimmbandfrequenz F_0 und deren Obertöne („Harmonische“) $j \cdot F_0$ identifizieren. Diesen Anteil bezeichnen wir als Feinstruktur des Spektrums; er macht sich im Leistungsdichtespektrum in Gestalt schnell aufeinander folgender, äquidistanter Spitzen bemerkbar (s. Abbildung 24).

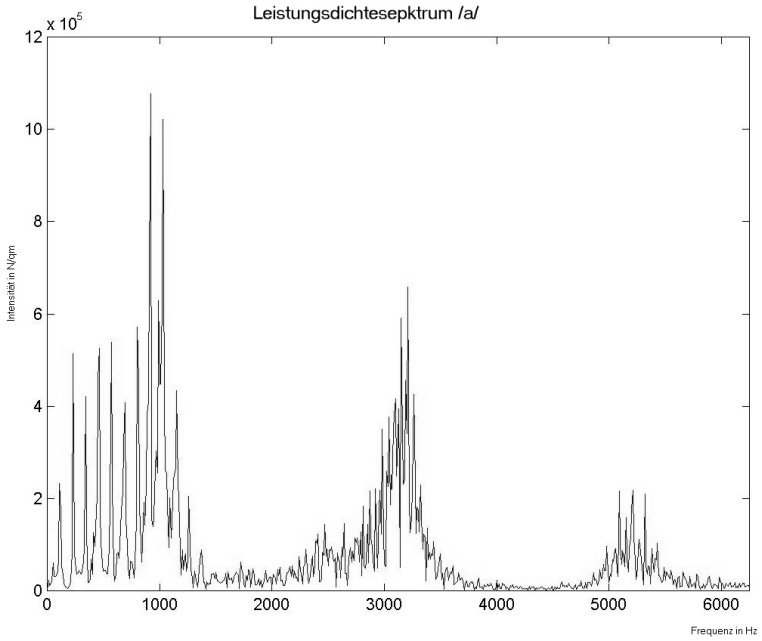

Abb. 24: Leistungsdichtespektrum des Lautes /a/

Die Vokaltraktkonfiguration ist durch ihre charakteristischen Resonanzen und Antiresonanzen gekennzeichnet, die sich im Spektrum als relative Maxima oder Minima manifestieren. Ihre ungefähre Lage ergibt sich aus der Umhüllenden bzw. dem geglätteten Verlauf des Kurzzeitspektrums. Die wesentliche Komponente dieses groben Spektralverlaufs ist die Formantstruktur. Ziel der Spektralanalyse ist es diese spektralen Charakteristika sowie ihre zeitliche Aufeinanderfolge zu beschreiben.

Ausgangspunkt für die weiteren Überlegungen ist das aus der DFT berechnete Leistungsdichtespektrum. Es enthält zu viele Informationen; die Frequenzauflösung ist zu genau. Wie beim menschlichen Gehör werden deshalb nahe beieinander liegende Frequenzen zu Frequenzbändern zusammengefasst. Im Kontext der Signalverarbeitung spricht man dabei von Filterbänken. Die Filterbank besteht aus einer Reihe von Bandpässen, welche die Energie des Sprachsignals in einzelnen Frequenzbändern messen. Die Frequenzbänder sind durch ihre Mittenfrequenzen und Bandbreiten charakterisiert. Der gebräuchlichste Typ einer Filterbank ist die uniforme Filterbank, für welche die Mittenfrequenz f_i des i-ten Bandpassfilters wie folgt definiert ist:

$$f_i = \frac{F_s}{N} i \qquad 1 \le i \le Q \tag{30}$$

Dabei stellt F_s die Abtastrate des Sprachsignals und N die Anzahl äquidistanten Filter, welche den Frequenzbereich überspannen, dar. Für die aktuelle Anzahl der benötigten Filter gilt die Beziehung:

$$Q <= N/2 \tag{31}$$

Somit bedeutet Gleichheit der obigen Gleichung, dass der gesamte Frequenzbereich für die Analyse benutzt wird. Die Bandbreite b_i des i-ten Filters besitzt im allgemeinen die Eigenschaft:

$$b_i \ge \frac{F_s}{N} \tag{32}$$

Hierbei bedeutet Gleichheit, dass keine Überlappung zwischen benachbarten Filtern existiert und Ungleichheit ein Überlappen benachbarter Filter.[47]

[47] Bei $b_i<F_s/N$ gilt, dass einige Frequenzbereiche des Spektrums bei der Analyse fehlen würden, was bei als nicht wichtig erachteten Frequenzbereichen auch sinnvoll sein kann.

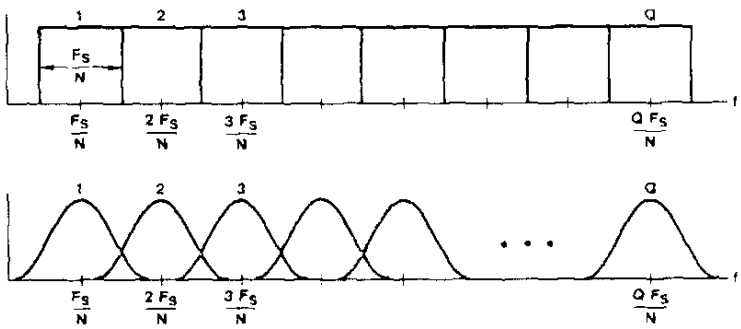

Abb. 25: Eine ideale (nichtüberlappende) und eine realistische Menge von Filtern einer Filterbank [RJ93]

Für die Zusammenfassung der verschiedenen Frequenzbänder gibt es verschiedene Filter, wie z. B. Dreiecks-, Rechtecks- oder Trapezfilter. Beim Dreiecksfilter besitzt jedes Band die gleiche Länge und endet bei den Mittenfrequenzen seiner Nachbarbänder. Dieses Glätten des Spektralverlaufes zerstört dabei die harmonische Struktur der stimmhaften Laute und lässt Vokaltraktresonanzen deutlicher hervortreten.

Um die spektralen Eigenschaften des Gehörs nachzubilden, ist eine Transformation der Frequenz in die Tonheit sowie des Schalldrucks in die psychoakustische Größe Lautheit notwendig. Eine brauchbare Modellierung der Tonheit liefert eine Filterbank mit 24 Bandfiltern[48], die in Abständen von jeweils 1 Bark angeordnet sind. Hier erhält man statt einer linearen Skala, eine Skala, die bei Frequenzen unterhalb 500 Hz nahezu identisch mit der linearen ist, oberhalb aber tendenziell logarithmisch. Einige Varianten dieser kritischen Bandbreiten-Skalierung wurden in der Literatur bereits beschrieben, wie z. B. die Melskala. Die Unterschiede zwischen ihnen sind meistens gering. Daher werden sie auch häufig durch eine logarithmische Skalierung implementiert. Für eine Anzahl Q an Bandpassfiltern mit Mittenfrequenzen f_i und Bandbreiten b_i gilt dabei:

$$\begin{aligned} b_1 &= C \\ b_i &= \alpha\, b_{i-1} \qquad 2 \le i \le Q \\ f_i &= f_1 + \sum_{j=1}^{i-1} b_j + \frac{(b_i - b_1)}{2} \end{aligned} \tag{33}$$

wobei C und f_1 die beliebige Bandbreite und Mittenfrequenz des 1. Filters darstellen, und α der logarithmische Wachstumsfaktor ist. Die in der Literatur am häufigsten benutzten Werte von α sind $\alpha = 2$ (octave band spacing) und $\alpha = 4/3$ (1/3 octave filter spacing).

[48] Für die Spracherkennung kann dabei die Zahl der Kanäle auf 22 beschränkt werden, was einem erfassten Frequenzbereich bis 8.5 kHz entspricht.

Das Funktionsmodell der Lautheit gibt an, wie laut ein Schall vom Menschen wahrgenommen wird. Es bildet in jedem Kanal ν durch Logarithmierung den Erregungspegel $E_\nu(t)$ des Gehörs nach und leitet daraus die Lautheitskomponente $N_\nu(t)$ ab. Diese dient als Näherung für die spezifische Lautheit N'(z, t). $N_\nu(t)$ ist der Anteil der Lautheitsempfindung für die entsprechende Frequenzgruppe ν. Wenn der Einfluss der Ruhehörschwelle vernachlässigt wird, gilt das Potenz-Gesetz:

$$N_\nu(t) \sim E_\nu(t)^{0.23} \tag{34}$$

Die Summe aller Kanäle $N_\nu(t)$ zu einem Zeitpunkt t ergibt die Gesamtlautheit N(t):

$$N(t)=\sum_{\nu=1}^{24} N_\nu(t) \tag{35}$$

Die Transformation des Schalldrucks in die Lautheit hat den zusätzlichen Vorteil, dass eine sinnvolle Skalierung erfolgt.

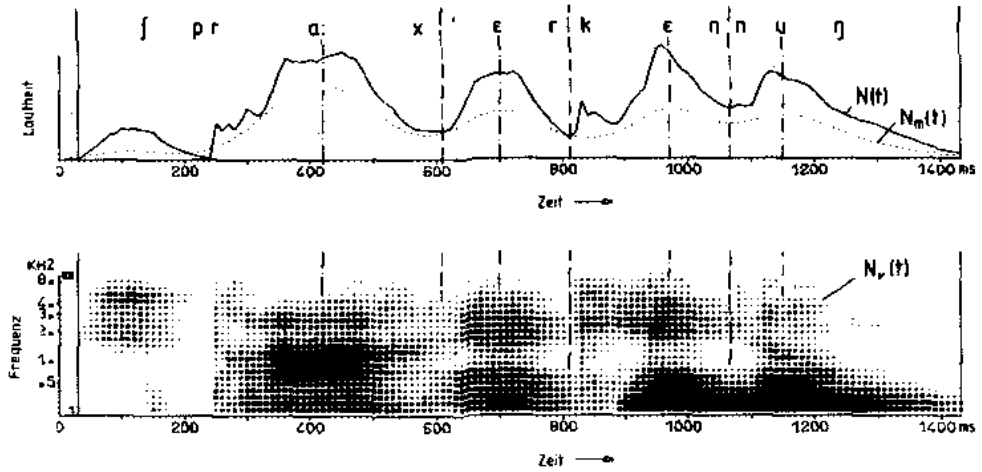

Abb. 26: Spektrale Vorverarbeitung des Wortes „Spracherkennung" mit Hilfe des Lautheitsmodells [Rus94]

Abbildung 27 zeigt für die Verteilung der Lautheitskomponenten N_ν als Funktion der Frequenzgruppe ν. Aufgrund der großen Bandbreite der Frequenzgruppen bei höheren Frequenzen und aufgrund der Maskierungseffekte können die Formanten F_2 und F_3 oft nicht getrennt werden; sie bilden ein gemeinsames Maximum, das in der Literatur auch als F_2' bezeichnet wird. Abhängig vom Vokal kann F_2' von F_2 selbst, von F_3 oder von beiden zusammen gebildet werden. Die psychoakustische Repräsentation in Form der

Lautheitsspektren zeigt damit, dass als markante Merkmale der Vokale sowohl F_1 als auch F_2‘ anstelle von F_2 verwendet werden sollten.

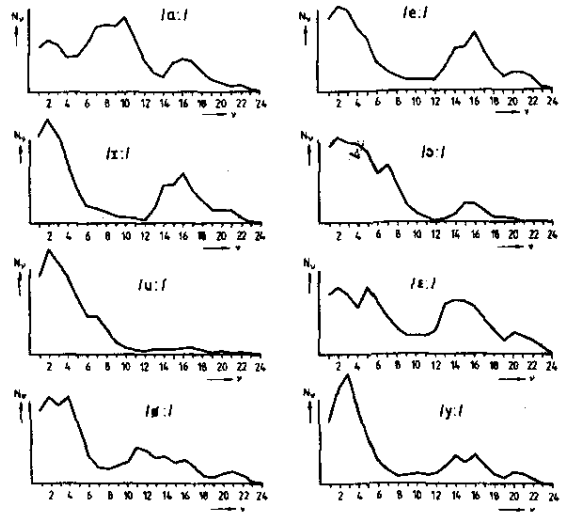

Abb. 27: Typische Lautheitsspektren deutscher Vokale [Rus94]

3.4 Cepstral-Analyse

Die Cepstral-Analyse liefert eine Methode zur Separierung der Vokaltraktinformation von der Anregungsinformation. Sie basiert auf dem Fant’schen Source-Filter-Modell der Spracherzeugung (vgl. Kapitel 2.2.2 „Lineares Modell der Spracherzeugung“). Diese geht von einer Faltungsdarstellung $f_n = e_n * h_n$ des Sprachsignals aus, wobei e_n die Anregung und h_n die Impulsantwort eines linearen Systems repräsentiert, das gleichzeitig der Modellierung der Vokaltraktresonanzen sowie der Glottissignalform und der Lippenabstrahlung diente. Um die durch Faltung vermengten Anregungs- und Übertragungskomponenten des Sprachschalls zu weitgehend zu entfalten, macht die Cepstralanalyse Gebrauch von den Zusammenhängen:

$$\begin{aligned} FT\{f_n\} &= FT\{e_n\} \cdot FT\{h_n\} \\ \log FT\{fn\} &= \log FT\{en\} + \log FT\{hn\} \\ FT^{-1}\{\log FT\{f_n\}\} &= FT^{-1}\{\log FT\{e_n\}\} + FT^{-1}\{\log FT\{h_n\}\} \end{aligned} \qquad (36)$$

mit FT als zeitdiskreter Fouriertransformierte. Die in dem durch Faltung entstandenen Signal enthaltenen Anteile sind im Spektrum multiplikativ verknüpft. Eine elegante Methode zur Dekonvolution (=Entfaltung) bildet der Logarithmus. Dadurch geht die multiplikative Verknüpfung in eine additive über. Wendet man hierauf eine inverse Fouriertransformation an, bleibt wegen deren Linearität die additive Überlagerung erhalten, wodurch eine Trennung

der Einflüsse von Anregungstrakt und Vokaltrakt möglich ist. Diese Transformation führt an sich wieder in den Zeitbereich zurück. Um das Resultat aber vom eigentlich zugrunde liegenden Zeitsignal abzuheben, führt man eigene Bezeichnungen ein, und zwar auch für die unabhängige Variable der Inverstransformation und ebenso für Operationen in diesem Bereich. Man nennt die Transformierte

$$c_x^{(c)}(n)=F^{-1}\left\{\ln X(e^{j\Omega})\right\} \qquad (37)$$

das komplexe Cepstrum[49], seine Variable bezeichnet man als „quefrency“. Die Buchstabenvertauschungen in den Namen sollen darauf hindeuten, dass man in vielen Anwendungsfällen in Spektrum und Cepstrum Zusammenhänge wie sonst in Zeit- und Frequenzbereich quasi „gespiegelt“ wiederfindet. Die Dekonvolutionseigenschaft der Transformation bleibt offensichtlich erhalten, wenn FT{f_n} durch das Betragsspektrum |FT{f_n}| und der komplexe durch den reellen Logarithmus ersetzt wird. Die resultierende Abbildung FT^{-1}{log|FT{f_n}|} wird als (reelles) Cepstrum bezeichnet [BHT63].

Die praktische Berechnung der (Näherungswerte für) Cepstrumskoeffizienten geschieht durch Anwendung der inversen DFT auf das logarithmierte Betragsspektrum:

$$c_q^{(m)} = \frac{1}{N}\sum_{\nu=0}^{N-1} \log\left|F_\nu^{(m)}\right| e^{i2\pi\nu q/N} \qquad q=0,\ldots,N-1 \qquad (38)$$

Das Betragsspektrum bzw. seine periodische Fortsetzung ist reell und symmetrisch, das Gleiche gilt daher auch für die $c_q^{(m)}$.

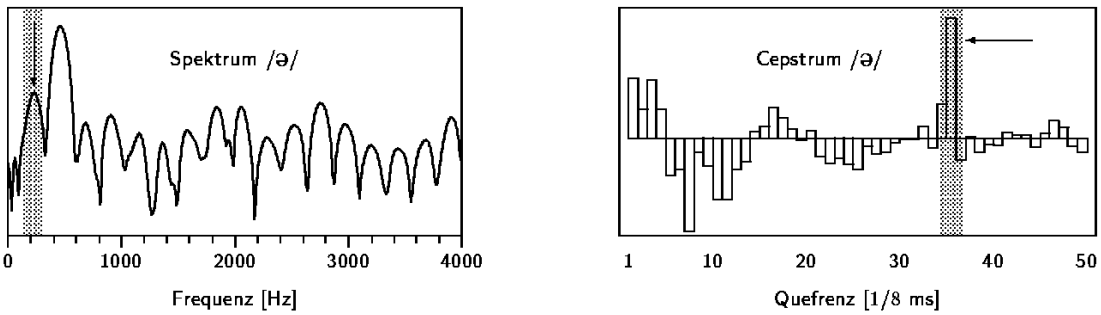

Abb. 28: Logarithmiertes Leistungsdichtespektrum $\log|F_\nu^{(m)}|$ und Cepstralkoeffizienten $c_q^{(m)}$ einer vokalischen Sprachprobe [St95]

Abbildung 28 demonstriert das Dekonvolutionsvermögen des reellen Cepstrums am Beispiel einer Realisierung des Lautes /e/. Das logarithmierte Vokalspektrum zur Linken weist gut

[49] Das komplexe Cepstrum ist ein Spezialfall der Klasse homomorpher Analyseverfahren [Opp68a,Opp68b].

erkennbare Resonanzen bei 500 Hz, 2000 Hz und 2700 Hz auf; die Formantstruktur ist jedoch von den periodischen Spitzen der zur Grundfrequenz von etwa 230 Hz gehörenden Harmonischen überlagert. Die Gleichung (3.30) formuliert aber gerade eine „Spektralanalyse des Spektrums“: die Koeffizientenfolge $\log|F_v^{(m)}|$ (von der aus Symmetriegründen nur eine halbe Periode abgebildet ist) wird mit Hilfe der inversen DFT auf sinusförmige Anteile hin analysiert, und ein cepstraler Gipfel bei $c_q^{(m)}$ ist ein Indiz für eine spektrale Schwingungsbewegung mit der Hertz-Periode f_A/q.

So finden wir die langsamen Anteile des Spektralverlaufes in den Cepstrumgliedern niedriger Ordnung repräsentiert, während die anregungsbedingte harmonische Struktur unter günstigen Bedingungen in einem prägnanten cepstralen Gipfel kulminiert [Nol67] – im abgebildeten Beispiel bei einer Quefrenz (so wird traditionell der Definitionsbereich des Cepstrums bezeichnet) von 35 Einheiten zu 1/8 ms, entsprechend einer Grundfrequenz von rund 230 Hz[50]. Für die automatische Spracherkennung sind daher nur die Koeffizienten niedriger Quefrenzen von Belang; die restlichen Parameter werden nicht als Merkmale verwendet. Bezeichnet $\hat{c}_q^{(m)}$ die nach Ausblendung, d. h. Nullsetzen der Anteile mit den höheren Quefrenzen, verbleibende Koeffizientenfolge, können wir mittels diskreter Fouriertransformation in den Spektralbereich zurückkehren, um ein geglättetes Spektrum zu erhalten. Wir führen also eine Hochpassfilterung durch, die im Kontext der Cepstralanalyse als Lifterung bekannt ist[51].

Die beiden Lifterungsergebnisse in Abbildung 29 stellen eine additive Zerlegung von $\log|F_v^{(m)}|$ in die spektrale Grob- und Feinstruktur dar, sie wurden durch Rücktransformation der unterhalb (Tiefpass-Lifter) bzw. oberhalb (Hochpass-Lifter) einer Grenzquefrenz von 20 ms – das entspricht einer Schwingungsfrequenz von 400 Hz –gelegenen Cepstrumskoeffizienten gewonnen. Aufgrund der Tiefpasslifterung weist das links stehende Spektrum sehr deutliche Maxima bei den Vokaltraktresonanzen, während die Einflüsse der Anregungskomponente weitgehend unterdrückt wurden.

---

[50] Sei q der q-te Koeffizient des Cepstrums. Ihm entspricht die Frequenz $f= f_A/q$, mit f_A als Abtastfrequenz.

[51] Dieser Vorgang wird auch als homomorphe Filterung bezeichnet.

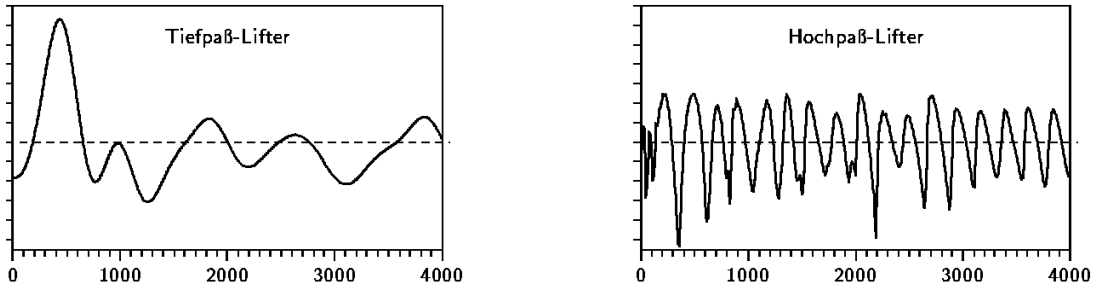

Abb. 29: Gelifterte Leistungsdichtespektren der Sprachprobe aus Abb. 28; Grenzquefrenz war q = 20 ms [St95]

Die heute meistverwendete Form cepstraler Merkmale sind seit ihrer Einführung um 1980 die mel-Cepstrum-Parameter [DM80]. Wir erhalten sie mittels Kosinustransformation aus dem logarithmierten mel-Spektrum.

$$c_q^{(m)} = \sum_{k=1}^{K} \log e_k^{(m)} \cos\frac{\pi q(2k+1)}{2K} \qquad q = 1,2,\ldots \qquad (39)$$

Für weitere Ausführungen zum Mel-Cepstrum wird auf die Literatur verwiesen [ZF67, Zwi80, DM80, Rab93, DPH93].

3.5 Lineare Prädiktion

Weit verbreitet sind Verfahren zur parametrischen Beschreibung des Zeitsignals s(n) durch Linearkombination vorangegangener Werte. Rabiner führt die folgenden Ursachen für die dominante Stellung der LPC (linear predictive coding) -Analyse auf [Rab93]:

- Sie liefert ein sehr gutes Modell für das menschliche Sprachsignal. Insbesondere in relativ statischen Sprachbereichen, wie z. B. innerhalb stimmhafter Laute, können die spektralen Eigenschaften des Artikulationstraktes sehr gut durch das AR-Modell der linearen Prädiktion dargestellt werden.
- Durch ihre Anwendung ergibt sich eine sehr sinnvolle Separation zwischen der Anregung des Sprachsignals und den Eigenschaften des Artikulationstraktes.
- Die LPC-Analyse kann einfach implementiert werden.
- Es existiert eine Reihe von Anwendungen, in denen LPC-basierte Merkmalsextraktionsverfahren sehr erfolgreich zum Einsatz kommen.
- Sie ermöglicht eine äußerst kompakte und für die Sprachverarbeitung günstige Repräsentation bei vergleichsweise bescheidenem Rechenaufwand [DM80, Rab93].

In diesem Kapitel wird das Verfahren kurz erläutert. Eine sehr ausführliche Darstellung der mathematischen Grundlagen dieser Technik bietet die Monographie von J. D. Markel [MG76].

Unter linearer Prädiktion versteht man ein lineares System, das einen Ausgangswert s als Summe endlich vieler vorausgegangener Ausgangswerte s(n-i) mit i = 1...p wiedergibt:

$$\hat{s}(n) = \sum_{i=1}^{p} a_i s(n-i) \qquad (40)$$

Dabei wird angenommen, dass die Abtastwerte hinreichend stationärer Signalabschnitte durch eine lineare Vorhersageformel der Ordnung p abgeschätzt werden können. Selbst bei geeignet gewählten Vorhersagekoeffizienten a_i wird sich im allgemeinen eine Abweichung bzw. eine Fehlergröße e(n) zwischen dem wahren Wert s(n) und dem vorhergesagten $\hat{s}(n)$ ergeben:

$$e(n) = s(n) - \hat{s}(n) \qquad (41)$$

Eingesetzt:

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + e(n) \qquad (42)$$

Nach z-Transformation der Gleichung ergibt sich die Übertragungsfunktion H(z) zu

$$H(z) = \frac{S(z)}{E(z)} = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-1}} \qquad (43)$$

Dieser Ansatz stellt ein einfaches Modell der Spracherzeugung dar, indem der Vokaltrakt als lineares Filter angesehen wird, das von den Impulsen der Stimmbänder in bestimmten Zeitabständen angeregt wird. Die Koeffizienten a_i werden für jedes Zeitfenster so bestimmt, dass die Abweichung zwischen dem Ausgang dieses linearen Modells und dem tatsächlichen Sprachsignal minimal wird [MG76].

Wenn H(z) als Modell des Vokaltraktes allein gelten soll, müssen vorher die Einflüsse der Glottiswellenform und der Lippenabstrahlung eliminiert werden. Das Dämpfungsverhalten der Glottis (ca. 12 dB/Oktave) und die Hochpassfilterung an den Lippen (ca. -6 dB/Oktave)

können durch ein Preemphase des Sprachsignals mit dem System $1\text{-}z^{-1}$ kompensiert werden [Wak73], was im Zeitbereich der Differenzbildung $f_n' = f_n - f_{n-1}$ entspricht[52].

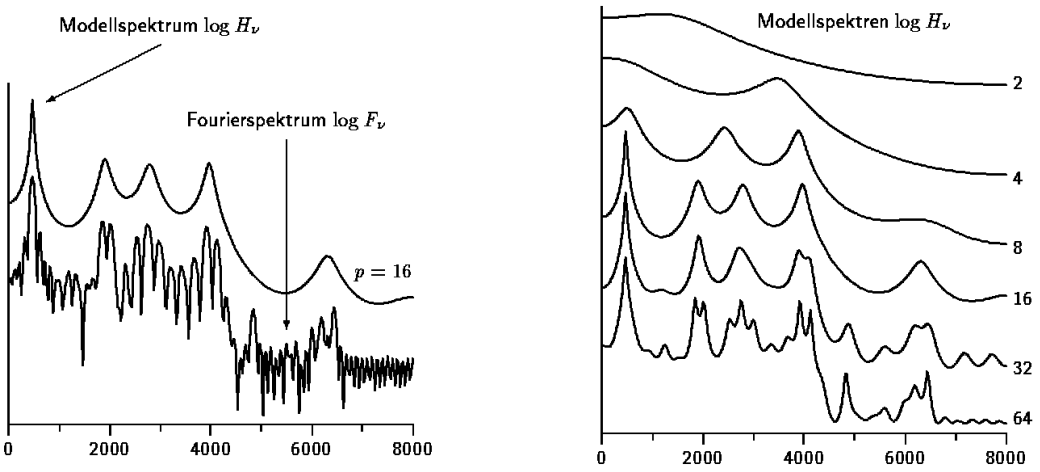

Abb. 30: DFT-Spektrum und Modellspektren verschiedener Ordnungen zu einem gesprochenem Vokalphonem [St95]

Wie die Beispiele in Abbildung 30 illustrieren, bildet das Modellspektrum eine geglättete Näherung an das DFT-Spektrum; mit zunehmender Ordnung p schmiegt sich H_ν näher an F_ν an. Offensichtlich hängt die LP-Analyse stark von der Wahl des Parameters p ab – wird die Ordnung zu niedrig angesetzt (p = 8), kommt es zu einer Verschmelzung benachbarter Formanten; ist p zu groß (p = 64), wird bereits die harmonische Struktur mitmodelliert. Zur Wahl von p wurde unter anderem die Faustformel $p = f_A+4$ vorgeschlagen; die Abtastfrequenz f_A ist in kHz anzugeben [Mar72]. Die Vorteile gegenüber der Berechnung der Kurzzeitspektren mit Hilfe der Fouriertransformation liegen vor allem in der Tatsache, dass sehr „glatte“ Spektren entstehen, die keine Welligkeit infolge der Sprachgrundfrequenz aufweisen, da Anregungsfunktion und Resonanzen des Vokaltrakts in diesem Modell voneinander getrennt sind[53].

Da dieses Filter nur Pole haben kann, lassen sich im Sprachsignal zwar die Formanten von Vokalen gut nachbilden aber keine Nullstellen im Spektrum erzeugen, die z. B. für die Nasallaute /m/ und /n/ charakteristisch sind.

---

[52] Interessant sind auch Umrechnungen der LPC-Koeffizienten in eine Darstellung, die als Röhren-Modell des Vokaltrakts aufgefasst werden kann und damit die Gestalt des Mund- und Rachenraums wiederspiegelt (PARCOR-Koeffizienten). Diese Modellverteilung bietet vor allem für den Einsatz in der Sprachsynthese eine günstige Ausgangsbasis [MG76].

[53] Für Rauschmuster muss das Filter mit Rauschen anstelle der Grundfrequenz angeregt werden. Hierfür ist eine exakte stimmhaft/stimmlos-Entscheidung notwendig. Die flach verlaufenden Rauschmuster der Frikativlaute /f/,/s/ werden meist nicht sehr gut wiedergegeben, da auch hier das Spektrum durch p Pole angenähert wird; ähnliche Probleme treten bei den Plosivlauten auf.

Aus den Polstellen der Vokaltraktübertragungsfunktion, also näherungsweise aus den Nullstellen des Polynoms A(z), lassen sich die Frequenzen und Bandbreiten der Formanten bestimmen [McC74, Mar72, DJ88]. Zieht man je Formant vereinfachend nur einen Partner $z_j = p_j e^{i\omega j}$ eines komplex-konjugierten Polpaares zur Berechnung heran, ergeben sich die Mittenfrequenzen f_j und die 3 dB-Bandbreiten β_j wie folgt:

$$f_j = f_A \frac{\omega_j}{2\pi} \tag{44}$$

$$\beta_j = \left| \frac{f_A}{2\pi} \ln \rho_j \right| \tag{45}$$

Die praktische Berechnung der LPC erfolgte im Rahmen der Arbeit über Autokorrelation, und zwar über die MEM-Methode (Maximum Entropy Methode).

3.6 Sonstige Ansätze

Neben den erwähnten Verfahren sind eine Fülle von Studien zur Gewinnung von Kurzzeitmerkmalen bekannt, in denen teils die signaltheoretischen Grundlagen, teils die Funktionsmodelle der physiologischen und neuronalen Sprachverarbeitung weiterentwickelt werden. So bezieht die perzeptuelle lineare Prädiktion (PLP) die Tonheitmodellierung bereits in die Vorhersageanalyse ein [Her90].

Die im letzten Abschnitt eingeführten Verfahren beschreiben momentane spektrale Eigenschaften des Signals. Für die menschliche Sprachwahrnehmung scheinen aber auch die zeitlichen Veränderungen relevante Indikatoren zu sein [Str82, Rus82]. Einfache Verfahren zur Berücksichtigung des zeitlichen Kontextes sind die Erweiterung des Merkmalsvektors um einen oder mehrere Nachbarvektoren [Ney80] oder die Näherung der 1. Ableitung durch die Bildung der 1. Differenz. Diskrete Approximationen für die Ableitungen höherer Ordnungen werden mittels iterativer Anwendung durch Einsetzen der Ableitungen niedrigerer Ordnung gewonnen [Fur86]. Ein komplexeres Verfahren stellt die von Hermansky vorgestellte *R*el*a*tive *S*pec*tra*l (RASTA) – Methode dar. Die grundsätzliche Idee des RASTA-Vefahrens besteht darin, bisher verwendete Vorverarbeitungsverfahren um Filterungsalgorithmen zu erweitern, deren Aufgabe es ist, sowohl langsame (quasi-statische) als auch sehr schnelle Variationen des auditorischen Spektrums eines gestörten Sprachsignals auszufiltern. Da sehr viele in der Praxis beobachtete Störungen (insbesondere der Übertragungskanäle) relativ statisch in Bezug auf die Geschwindigkeit menschlicher Sprache sind, können derartige Störungen

durch die oben erwähnten Filterungsalgorithmen unterdrückt werden [HMB91, DGH93, SV93].

Eine ganze Reihe neuerer Untersuchungen teilen das Anliegen, über die Energieverteilung hinsichtlich der Frequenzgruppen hinaus, auch den Phasengang in die Merkmalberechnung mit einzubeziehen [IU87, MY91, TKM91].

3.7 Zusammenfassung

Zur digitalen Weiterverarbeitung muss das kontinuierliche Sprachsignal tiefpassgefiltert und abgetastet werden; die Amplitudenwerte sind in endlich viele Stufen zu quantisieren.

Anschließend werden in schneller Fortschaltung (etwa alle 10 ms) überlappende Datenfenster aus der Abtastwertfolge ausgeblendet und auf ihre spektralen oder periodischen Eigenschaften hin analysiert. Das Ziel der meisten Merkmalsextraktionsverfahren ist eine parametrische Repräsentation der momentanen Vokaltraktübertragungsfunktion unter Ausschaltung der störenden Einflüsse des Anregungsspektrums. Ein Weg besteht darin, mit Hilfe einer diskreten Fouriertransformation das Betragsspektrum zu berechnen und durch die Integration der spektralen Energieanteile innerhalb kritischer Frequenzbänder zu glätten. Durch eine weitere Transformation in den Cepstralbereich hinein können die durch Faltung vermengten Anregungs- und Übertragungskomponenten des Sprachsignals weitgehend entflochten werden.

Ein alternatives Verfahren ist die lineare Vorhersage; sie setzt ein autoregressives Modell für den Vorgang der Spracherzeugung an und bestimmt dessen freie Parameter. Auch vom Vorhersagepolynom führt ein direkter Weg zu spektralen und cepstralen Repräsentationen [LRP90, AS91]. Die im Verlaufe der Kurzzeitanalyse gewonnenen Merkmalvektoren verkörpern erst die statische spektrale Information; sie wird durch dynamische Parameter ergänzt, welche den zeitlichen Verlauf der Merkmale charakterisieren.

In einer ausführlichen Untersuchung von [WN76] wurde herausgefunden, dass die Verwendung einer Filterbank mit 20 Kanälen bei der Erkennung isoliert gesprochener Wörter dieselben Ergebnisse lieferte wie die Verwendung von LPC-Koeffizienten (mit $p = 14$). Beide Verfahren können daher letztlich als gleichwertig für die Spracherkennung angesehen werden.

Sehr gute Erkennungsleistungen wurden mit der Verwendung des Cepstrums erzielt. Besonders günstig ist es, wenn das Spektrum zur Berechnung des Cepstrums über der Bark-Skala oder Mel-Skala gebildet wird; es ergeben sich dann die sogenannten mel-Cepstrum-Koeffizienten [DM80]. Wie bei der spektralen und cepstralen Analyse hat sich

auch bei den Vorhersageverfahren die Modellierung der Tonhöhenempfindung und Lautheit als vorteilhaft erwiesen. Eine Untersuchung von Jungua ergab, dass bei ungestörten Sprachaufnahmen die cepstrale LPC-Analyse und bei mit weißem Rauschen gestörten Sprachaufnahmen die PLP-Analyse zu den besten Erkennungsergebnissen bei der sprecherabhängigen Erkennung isoliert gesprochener Wörter führten [JW89]. Sind die zu klassifizierenden Sprachaufnahmen stark gestört, zeigt sich die RASTA-PLP den anderen Verfahren überlegen.

4 Klassifikation

Aufgabe der Klassifikation ist die Einteilung von Objekten/Individuen in Klassen. Diese Einteilung wird mit Hilfe von m extrahierten Merkmalen durchgeführt, von denen nicht alle zwingend zur Einteilung benötigt werden. Damit beinhaltet die Klassifikationsaufgabe auch eine Merkmalsauswahl, in der die Zahl der m Merkmale auf charakteristische n reduziert wird. Die Merkmale werden zu einem n-dimensionalen Merkmalsvektor zusammengefasst, der die Position der Objekte im n-dimensionalen Merkmalsraum beschreibt. Jede Objektklasse weist eine für sich kennzeichnende Verteilung der Merkmalsvektoren aus, die im Merkmalsraum durch so genannte Muster repräsentiert werden. Werden nur signifikante Merkmale verwendet, ergeben sich deutlich getrennte Musterklassen, wie dies in Abbildung 31a an einem zweidimensionalen Merkmalsraum schematisch dargestellt ist. Die Konstellationen werden als Cluster bezeichnet. Werden korrelierte Merkmale verwendet, überlappen sich die Häufigkeitsverteilungen. Eine fehlerfreie Klassifizierung ist damit nicht mehr möglich (s. Abbildung 31b) [Nor96].

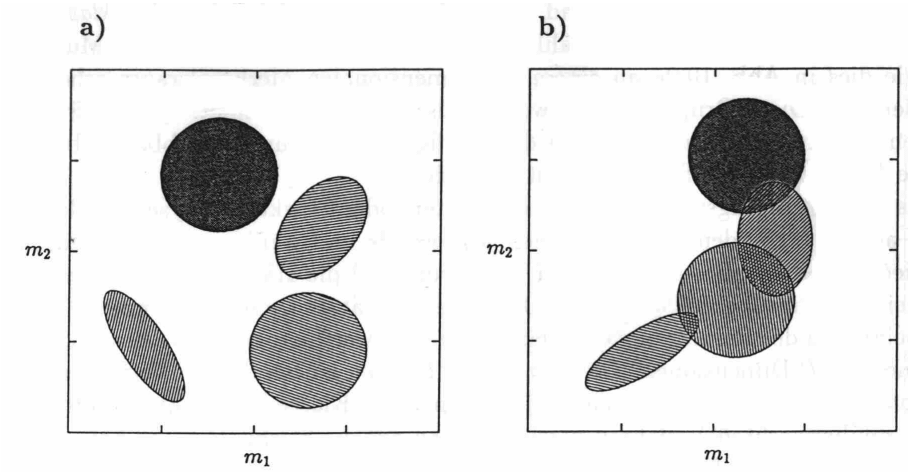

Abb. 31: Schematische Darstellung eines zweidimensionalen Merkmalsraum mit vier Klassen von Objekten: (A) die Merkmalsvektoren der Objekte fallen in deutlich getrennte Bereiche, (B) die Häufigkeitsverteilungen überlappen sich; [Jäh95]

Für die Zuordnung von Objekten zu Gruppen werden in der Praxis hauptsächlich klassische statistische Verfahren wie die Clusteranalyse oder Diskriminanzanalyse eingesetzt. Die Clusteranalyse kommt zum Einsatz, wenn die Gruppenzugehörigkeit nicht bekannt ist. Sie dient zur Identifikation von Gruppierungen bei Daten. Liegen die Gruppenzugehörigkeiten wie in unserem Falle vor, ist die Diskriminanzanalyse geeigneter. Ziel der Diskriminanzanalyse ist es, die Linearkombinationen der Variablen zu finden, die eine möglichst gute Differenzierung dieser Gruppen ermöglichen [BEP96]. Eine Alternative zur Behandlung von

Fragestellungen der Diskriminanzanalyse stellen die in den letzten Jahren wieder intensiv diskutierten Neuronale Netze dar (vgl. Kapitel 2.4.1 „Automatische Spracherkennung“). Zwischen den statistischen Verfahren und Neuronalen Netzen bestehen einige Beziehungen. Formal betrachtet bilden beide Ansätze im allgemeinen Eingabevektoren eines höherdimensionierten Eingaberaumes in Ausgabevektoren niedrigerer Dimension ab. Lineare, assoziative Netze stellen jeden Ausgabewert eines Neurons in der Ausgabeschicht als lineare Funktion seines Inputs dar. Dies entspricht im Falle diskreter Ausgabevariablen der linearen Diskriminanzanalyse [Hag97]. Aber auch nichtlineare Zusammenhänge lassen sich mit ihnen gut approximieren [Kra90]. Ein in der Praxis wichtiger Vorteil der statistischen Verfahren ist, dass sich der Einfluss der einzelnen Variablen auf die abhängige Variable ermitteln lässt. Demgegenüber ist in einem Neuronalen Netz die funktionale Abbildung der Eingabeneuronen auf die Netzausgabe eine „black box“, was zur Folge hat, dass der Einfluss einer unabhängigen Variablen nicht interpretiert werden kann. Dementsprechend ist das Verfahren in der Anwendung für den Benutzer nur schwer oder nicht nachvollziehbar. Da das Studium der Literatur sowie eigene Vorabstudien den Schluss nahe legten, dass sich die Nasalität an Parametern messen lässt, wurde in dieser Arbeit der Diskriminanzanalyse der Vorzug gegeben. Die Diskriminanzanalyse wird daher im nächsten Kapitel vorgestellt. Für vertiefende Literatur zur Diskriminanzanalyse sei auf [BEP96], zur Clusteranalyse sei auf [Ber80] und zur Anwendung von Neuronalen Netzen zur statistischen Datenanalyse auf [Hag97] verwiesen.

Um für alle stimmhaften Laute die typischen spektralen Eigenschaften sowie ihre Variationen wiederzugeben, muss die Stichprobe groß genug, d. h. repräsentativ, sein. Darüber hinaus ist es für einen sinnvollen Vergleich der verschiedenen Verfahren erstrebenswert, dass die Klassen gleich häufig vorkommen, d. h. für die Auftrittswahrscheinlichkeit p(k) der Klasse k sollte gelten:

$$p(k) = \text{const} = 1/K \qquad \text{für } k = 1 \ldots K \tag{46}$$

Ist dies nicht der Fall, können die unterschiedlichen Auftrittswahrscheinlichkeiten einen großen Einfluss auf das Klassifikationsergebnis haben.

Bezüglich der Eignung der Sprachdatenbank NASAL an die beiden Forderungen der Gleichverteilung der Klassen und der Stichprobenmindestanzahl vgl. Kapitel 8.1 „Eignung der Sprachdatenbank NASAL“.

4.1 Lineare Diskriminanzanalyse

Das Ziel der Diskriminanzanalyse ist es, Merkmale linear so miteinander zu kombinieren, dass sie gegebene unabhängige Gruppen möglichst gut voneinander trennen. Dabei ist zu bestimmen, ob sich die Gruppen bezüglich der Eigenschaften prägnant unterscheiden oder welche Eigenschaften zur Unterscheidung geeignet sind. Hat man die Diskriminanzfunktionen bestimmt, welche die Unterschiede unter Gruppen am besten charakterisieren, lassen sich neue unbekannte Fälle klassifizieren.

Die Diskriminanzanalyse berechnet nur so viele Diskriminanzachsen, wie man zur optimalen Trennung der Gruppen benötigt. Allgemein gilt, dass man bei G Gruppen mit G-1 Diskriminanzachsen auskommt – unabhängig von der Anzahl der beteiligten abhängigen Variablen. Sind etwa nur 2 Gruppen gegeben, reicht eine Achse immer aus, um diese beiden Gruppen zu trennen. Die Diskriminanzanalyse geht zum einem von mehreren intervallskalierten Merkmalsvariablen $X_{i,\ i\ =\ 1...m}$ aus, zum anderen von einer die Gruppenzugehörigkeit des Individuums definierenden Gruppierungsvariable. Diese muss eine begrenzte Anzahl unterschiedlicher Kategorien besitzen, die als ganzzahlige Werte kodiert werden[54]. Aus den Merkmalsvariablen werden durch Linearkombination eine oder mehrere Diskriminanzfunktionen gebildet:

$$Y = v_0 + v_1X_1+v_2X_2+ \ldots + v_mX_m \qquad (47)$$

Die Diskriminanzfunktion ordnet jedem Individuum einen Wert auf der Diskriminanzvariable Y zu. Die Parameter v_i müssen so bestimmt werden, dass sich die Gruppen auf der Diskriminanzvariable Y maximal unterscheiden. Sei SAQ die Summe der Abweichungsquadrate von einem Mittelwert. Dann unterscheidet die Diskriminanzanalyse zwischen den SAQ innerhalb der einzelnen Gruppen (SAQ_{inn}) und den SAQ der Mittelwerte vom Gesamtmittelwert (SAQ_{ZW}). Den Gesamtmittelwert erhält man, indem man die Gruppenzugehörigkeit ignoriert. Nimmt man an, dass alle Erwartungswerte in den Gruppen gleich sind und die Mittelwerte in den Gruppen zufällig um diese Erwartungswerte fluktuieren (Nullhypothese), stellen sowohl SAQ_{inn} als auch SAQ_{zw} Schätzungen für die Variation der Daten um den Gesamtmittelwert dar. Insbesondere sollte das Verhältnis

$$\lambda = \frac{SAQ_{zw}}{SAQ_{inn}} \qquad (48)$$

[54] Basiert die Gruppenzugehörigkeit auf den Werten einer kontinuierlichen Variablen, sollte lieber die lineare Regression verwendet werden, um von den reichhaltigeren Informationen zu profitieren, die in der kontinuierlichen Variablen selbst enthalten sind.

zufällig um den Wert 1 variieren [GK98]. Große positive Abweichungen vom Wert 1 sprechen für einen signifikanten Effekt. Das durch diese Beziehung erreichte Maximum wird als der Eigenwert λ der Diskriminanzfunktion, die benutzten Parameter v_i als Eigenvektor bezeichnet. Zu der Berechnung der Diskriminanzkoeffizienten sei auf [BEP96] verwiesen. Die Zuordnung der Objekte zu den Gruppen erfolgt über Schwellenwerte, die sich aus den Diskriminanzmittelwerten der Gruppen ableiten.

Um Aussagen über die Trennfähigkeit der Diskriminanzfunktion zu machen, ist eine Skalierung des Eigenwertes notwendig.

$$\Lambda = \frac{SAQ_{inn}}{SAQ_{inn} + SAQ_{zw}} = \frac{1}{1+\lambda} \qquad (49)$$

Λ wird als Wilks‘ Lambda bezeichnet. Ein kleiner Wert, d. h. SAQ_{inn} ist klein im Verhältnis zur Gesamtstreuung, bedeutet eine gute Trennleistung. Sei N die Größe der Gesamtstichprobe, M die Zahl der Merkmalsvariablen und G die Zahl der Gruppen. Dann lässt sich durch die folgende Transformation Wilks‘ Lambda als eine probabilistische Variable deuten.

$$\chi^2 = -\left[N - \frac{M+G}{2} - 1\right] \ln \Lambda \qquad (50)$$

Diese Transformation liefert eine annähernd χ^2-verteilte Variable, so dass sich aus einer χ^2-Tabelle das Signifikanzniveau (Irrtumswahrscheinlichkeit) bestimmen lässt. Die Diskriminanzleistung aller k Diskriminanzfunktionen kann mit dem multivariaten Wilks‘ Lambda überprüft werden (vgl. [BEP96]).

Da nicht alle Merkmale für die Diskriminierung bedeutend sind, versucht man die relevanten Merkmale ausfindig zu machen. Neben der theoretischen Motivation, die Unterschiede zwischen den Klassen erklären zu können, spielen praktische Gründe ebenfalls eine Rolle. Es wird aus Gründen des Rechenaufwands angestrebt, die Dimensionalität niedrig zu halten. Für die Bestimmung der optimalen Untermenge der Merkmale ist aufgrund der Vielzahl von Möglichkeiten ein Auswahlverfahren notwendig. Die sog. schrittweise Diskriminanzanalyse bestimmt in jedem Schritt, anhand eines Gütemaßes, welche Variable neu aufgenommen werden soll. Im Rahmen dieser Arbeit wurde als Gütemaß das Wilks' Lambda verwendet. In jedem Schritt wird diejenige Variable mit dem kleinsten Wilks‘ Lambda aufgenommen. Es gibt eine ganze Reihe von schrittweisen Methoden, welche die Suche steuern. Nach [GK98] sind die Unterschiede zwischen diesen Suchprozeduren relativ unwichtig. Viel schwerer wiegt, dass die schrittweise Diskriminanzanalyse bei jedem Schritt statistische Fehler erster

und zweiter Art begeht, die nicht korrigiert werden. Daher sind die Ergebnisse immer ein wenig mit Unsicherheit belastet.

Die Güte der Diskriminierungsanalyse wird im Rahmen dieser Arbeit mit Klassifikationsmatrizen gemessen. Diese enthalten in den Hauptdiagonalen die Anzahl (und Prozentsätze) der korrekt klassifizierten Fälle jeder Gruppe, in den übrigen Feldern die falsch klassifizierten Elemente. Außerdem wird der Gesamtprozentsatz der richtigen Klassifizierungen angegeben. Da diese Klassifikationen mit den trainierten Daten durchgeführt werden, sind sie in der Regel zu optimistisch. Um realistischere Klassifikationsergebnisse zu erhalten, wird die Klassifikation daher nach der Jackknife („Taschenmesser“)-Methode durchgeführt [Efr82]. Die Idee hierbei ist, statistische Verfahren durch einen einfachen Mechanismus zu simulieren. In der Jackknife-Methode wird jeweils eine Beobachtung ausgeschlossen und das jeweilige statistische Verfahren unter Ausschluss dieser Beobachtung durchgeführt. Am Ende wird die ausgeschlossene Beobachtung aus dem ermittelten Modell „vorhergesagt“. Wenn man dies für alle Fälle durchführt, kann man beurteilen, wie gut und wie stabil die Klassifikation ist [GK98]. Die Wirkung der Jackknife-Methode lässt sich an einem Extrembeispiel sehr gut verdeutlichen, indem man die Anzahl der Merkmale dem Stichprobenumfang anpasst. Der Klassifikationsfehler ohne Jackknife-Methode geht gegen 0 Prozent, aber bei der Klassifikation der ausgelassenen Beobachtung gegen 100 Prozent.

Zur Durchführung der Diskriminanzanalyse wurde das Statistikprogramm SPSS 7.5 verwendet. SPSS bietet die schrittweise Diskriminanzanalyse unter Anwendung der Jackknife-Methode an. Dabei lassen sich auch die a-priori-Wahrscheinlichkeiten für die verschiedenen Gruppen einstellen. Neben der gleichen a-priori-Wahrscheinlichkeit für alle Gruppen, kann man eine aus der beobachteten Gruppengröße in der Stichprobe berechnete Wahrscheinlichkeit verwenden. Im Rahmen der nachfolgenden Untersuchungen wurde für alle Gruppen die gleiche Auftrittswahrscheinlichkeit gewählt.

5 Sprachdatenbank NASAL

Im Rahmen der vorliegenden Arbeit wurde die Sprachdatenbank NASAL erstellt. Der Aufbau der Datenbank war notwendig, da für den deutschen Sprachraum keine nasalitätsspezifische Sprachdaten vorhanden sind. Die Aufnahmen erfolgen an der HNO-Klinik in Heidelberg an einer eigens dafür errichteten Aufnahmestation. Zurzeit enthält die Datenbank Sprachdaten auf der Basis eines standardisierten Sprachbogens von 116 Sprechern mit unterschiedlichen Nasalitätstypen und –ausprägungen. Alle Sprachdaten wurden logopädisch hinsichtlich der Nasalität beurteilt. Die Sprachdatenbank bildete die Basis der umfangreichen Untersuchungen zur geeigneten Parameterbestimmung für die Nasalitätsevaluierung. Zusätzlich ist sie auch die Grundlage zur Bewertung der Güte der hier herausgearbeiteten Klassifikationsparameter[55].

5.1 Aufnahmeumgebung

Aufnahmestation

Die Aufnahmestation wurde in der HNO-Klinik Heidelberg errichtet. Der Standort in der Klinik ist insofern sehr wichtig, da er die Gewinnung von Patientensprachdaten stark erleichtert.

Alle Sprachaufnahmen der Sprachdatenbank NASAL wurden an dieser Aufnahmestation aufgenommen. Dadurch dass die exakt gleiche Hard- und Software eingesetzt wurde, ist die Vergleichbarkeit der Daten gewährleistet. Unterschiedliche Frequenzbereiche von Mikrophonen oder Audiokarten kommen somit nicht vor.

Um eine hohe Rauschunterdrückung bei den Aufnahmen im klinischen Alltag zu erreichen, wurde eine Schallbox gebaut. Diese besteht aus drei gleichen, rechteckigen und in einem „U“ angebrachten Schall absorbierenden Platten. Die Schallbox steht auf einem Tisch, und ist zu dem Sprecher hin offen. Der davor sitzende Patient ist somit von links, rechts und vorne gegenüber akustischen Störquellen (Stimmen weiterer Personen, Raumumgebung, Verkehrslärm) abgeschottet. Eine weitere Störquelle, der unterschiedliche Abstand zum Mikrofon beim Sprechen mit seinen Auswirkungen auf die Lautstärke, schlimmstenfalls

[55] Für die Anpassung an andere Sprachräume wäre eine neue Sprachdatenbank Voraussetzung, um die Referenzwerte neu zu bestimmen.

Verzerrungen, wurde gemildert, indem das Mikrofon[56] in der linken hinteren Ecke befestigt wurde[57].

Abb. 32: Aufnahmestation

Das Mikrofon ist über einen Klinkenstecker mit einem Mischpult[58] verbunden. Am Mischpult erfolgt die Regelung der Lautstärke, so dass zu laute Aufnahmen und dadurch entstehende Verzerrungen vermieden werden. Es wird dabei versucht die Aussteuerung der Aufnahmen auf ca. 40 dB zu halten. Über den „Line out"-Ausgang ist das Mischpult mit der Soundkarte[59] eines Power Macintosh 7600 verbunden. Die Steuerung der Aufnahme erfolgte in einem Audioeditor namens „SoundEdit v.2" mit einer Abtastrate von 22.05 kHz und 16 Bit Auflösung.

---

[56] Technische Daten: Sennheiser MD 735, dynamisches Mikrofon mit Supernierenrichtcharakteristik. Der Übertragungsbereich beträgt 50-18000 Hz, die elektrische Impedanz bei 1 kHz 350 Ohm.

[57] Der konstante Abstand vom Sprecher zum Mikrophon lässt sich auch über ein Headset erreichen. Durch die Nähe und die Ausrichtung zum Mund, hätte dieses den Vorteil, dass es auch unter Normalbedingungen eingesetzt werden kann. Damit keine Atemgeräusche aufgenommen werden, ist das Mikrophon etwas seitlich vom Mund zu positionieren.

[58] Technische Daten: Stereo-Mischpult SA-1000V der Firma Conrad Electronic. Der Frequenzbereich beträgt 10 Hz – 110 kHz, der Klirrfaktor 0.02%. Es können Mikrophone mit max. 600 Ohm angeschlossen werden, da es sonst zu Fehlanpassungen kommt.

[59] Technische Daten des Audiosystem des Power Macintosh 7600: 16-Bit-Stereoeingang und –ausgang, Frequenzbereich von 10 Hz bis 18 kHz bei einer Abtastrate von 44.1 kHz, Audioeingang 3 kOhm Impedanz.

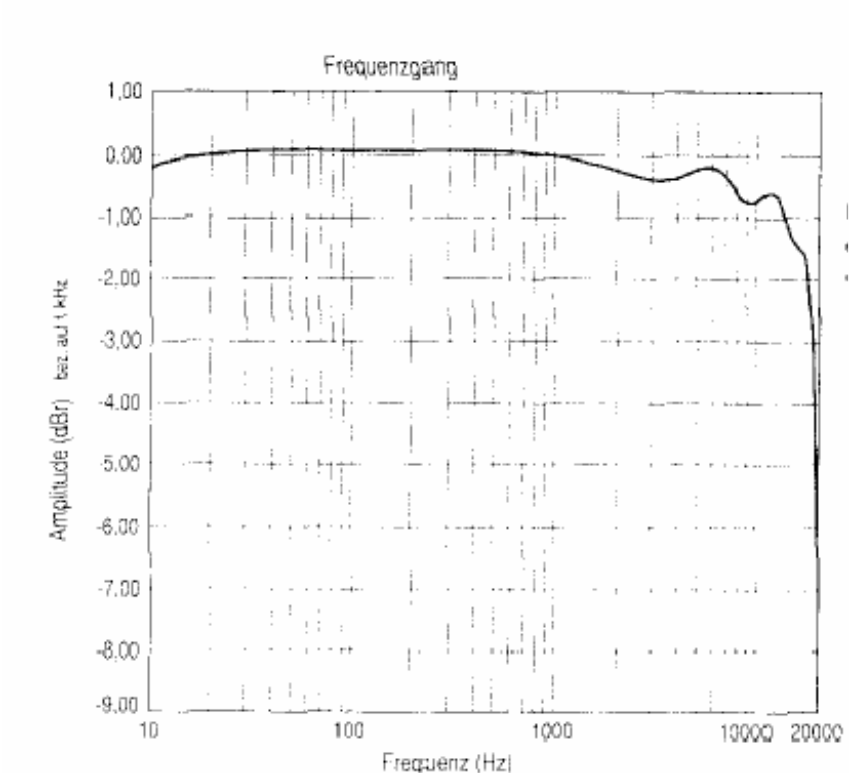

Abb. 33: Frequenzgang der Soundkarte [PM96]

Aufnahmeablauf

Der Aufnahmeablauf während der Erstellung der Sprachdatenbank NASAL war derart, dass jeder Sprecher aufgefordert wurde, einzeln die verschiedenen Phrasen nachzusprechen. Dies barg den Nachteil in sich, dass die Probanden zum Teil die Sprechweise des Vorsprechers nachahmten und daher oft besonders deutlich und überbetont sprachen. Um die Sprecher nicht allzu stark zu belasten, wurden alle Aufnahmen hintereinander, an einem Stück aufgenommen und in einer Datei im Macintosh AIFF-Format gespeichert. Der Name der Datei wurde so gewählt, dass aus ihm die Sprecherbezeichnung, sein Alter und das Geschlecht hervorgehen.

Kennzeichnung und Sicherung

Bevor die Daten in einer Datenbank gesichert werden konnten, waren das „Herausschneiden“ der einzelnen Passagen aus der Datei notwendig. Als erstes erfolgte eine manuelle Kennzeichnung der Sprachpassagen. Hierbei wurden für jede Passage der Aufnahme die Anfangs- und Endpunkte grob bestimmt, und die Passage markiert. Dieser Markierung konnte man im Audioeditor einen Namen zuordnen (=Label) und die Passage später über diesen Namen auswählen. War sie ausgewählt, wurde sie in die Zwischenablage kopiert, um sie in einer Datenbank einzufügen. Bei der verwendeten Datenbank handelt es

sich um Claris Filemaker 4.0, welche eine der wenigen Datenbanken ist, die auf dem Macintosh das Speichern von multimedialen Inhalten erlaubt. FileMaker verwaltet die Datensätze in Form von Seiten, welche unter anderem Felder besitzen. Jeder Sprecher ist auf einer Seite gespeichert, seine einzelnen Sprachpassagen in Feldern auf dieser Seite. Existieren von einem Sprecher mehrere Aufnahmen, wie z. B. bei den Aufnahmen von Logopäden, sind diese Aufnahmen in separaten Spalten auf dieser Seite gespeichert. Im aktuellen Seitenlayout sind bis zu 7 Aufnahmen pro Sprecher möglich. Neben den Sprachpassagen werden auch die Personeneigenschaften Name, Alter und Geschlecht gespeichert. Für jede Sprachpassage existiert weiterhin ein dazugehöriges Bewertungsfeld, das nach der logopädischen Bewertung ausgefüllt wird. Es wird betont, dass die Sprachdatenbank alle Aufnahmen in „roher“ Form enthält, d. h. es wurden keine Vorverarbeitungsverfahren wie Energie- und Längennormierung durchgeführt.

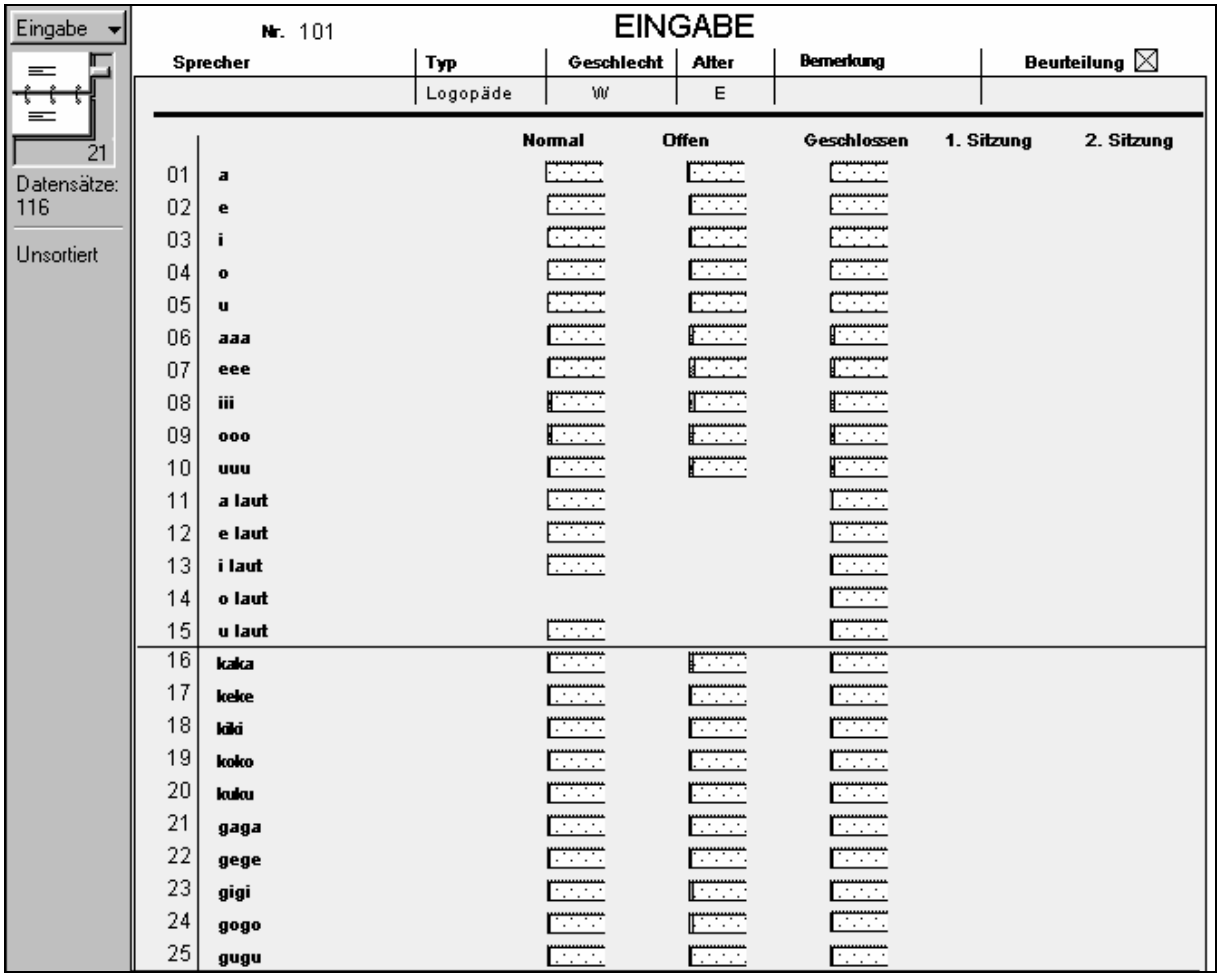

Abb. 34: Filemaker: Verwaltung der Sprachdaten eines Sprechers

5.2 Datenbestand

5.2.1 Sprachbogen

Grundlage für die Sprechmuster war der von Prof. Dr. W. Heppt, in Heidelberg entwickelte Rhinophoniebogen [HWS91]. Da in dem Projekt isoliert gesprochene Laute untersucht werden sollten, wurde der Sprachbogen leicht modifiziert.

Vokale kurz	a,e,i,o,u
Vokale lang	aaaaa, eeeee, iiiii, ooooo, uuuuu
Vokale laut	a, e, i, o, u
Plosiv+Vokal	kaka, keke, kiki, koko, kuku gaga, gege, gigi, gogo, gugu
Plosive	p, b, t, d, k, g
Zischlaute	x, s, z
Nasale	m,n
Sätze	Fritz geht zur Schule. Der Vater liest ein Buch. Die Schokolade ist sehr lecker. Peter spielt auf der Straße. Das Pferd steht auf der Weide. Nenne meine Mama Mimi.

Tab. 3: modifizierte Heidelberger Rhinophonie-Bogen

Die zum Messen der Plosive, Zischlaute und Nasale vorgesehenen Wörter wurden durch die entsprechenden Laute ersetzt. Der Aufbau dieses modifizierten Sprachbogens geht aus Tabelle 3 hervor. Die Vokale wurden dreimal pro Sprecher aufgenommen, einmal kurz (ca. ½ s), einmal lang (ca. 2 s) und einmal laut phoniert. Damit sollte untersucht werden, ob die Phonationsart Auswirkungen auf die Nasalität hat.

5.2.2 Sprechergruppen

Die Datenbank enthält zurzeit Sprachdaten von Patienten, Kindern und Logopäden. Während die Patienten unterschiedliche Nasalitätsausprägungen aufweisen, wurden als Referenz nicht nasal sprechende Kinder und Logopäden aufgezeichnet. Von den Logopäden wurden neben den nonnasalen Passagen zusätzlich Aufnahmen gemacht, in denen sie das Näseln simulieren. Diese simulierten Passagen stellen für das offene Näseln die Vokale, und

für das geschlossene Näseln die Nasallaute /m/ und /n/, sowie der Nasalsatz „Nenne meine Mama Mimi“ dar. Aufgrund ihrer guten Eignung zur Klassifizierung der offenen bzw. geschlossenen Nasalität wurden diese Passagen ausgewählt.

Gruppe	**Anzahl**	**Geschlecht**		**Alter**	
		M	**W**	**Kind**	**Erw.**
Patienten	28	16	12	22	6
Kinder	53	23	30	53	-
Logopäden	35	4	31	-	35

Tab. 4: Zusammensetzung der Sprechergruppen.

Die Verteilung der Gruppen auf Alter und Geschlecht kann Tabelle 4 entnommen werden. Dabei erfolgte beim Alter eine Einteilung als „Kind“ wenn das Alter im Bereich von 6-16 Jahren lag. Während ein fast ausgewogenes Verhältnis bei den Patienten und Kindern zwischen weiblichen und männlichen Sprechern vorherrscht, dominieren bei den Logopäden berufsstandspezifisch die Frauen. Weiterhin sind die Patienten vorwiegend Kinder. Dies hängt damit zusammen, dass Erwachsene hinsichtlich ihrer Gaumenspalte in der Klinik seltener bzw. gar nicht mehr behandelt werden, ein operativer Gaumenverschluss fand normalerweise schon statt.

5.2.3 Logopädische Beurteilung

Nachdem die Sprachdaten in der Datenbank aufgenommen wurden, erfolgte in mehreren Sitzungen von 2 Logopädinnen ihre Bewertung hinsichtlich der Nasalität. Die Sprachaufnahmen wurden bezüglich ihres Typs in kein, offenes und geschlossenes Näseln eingestuft. Lag ein Näseln vor, wurde dies auf einer 3-stufigen Skala mit den Ausprägungen gering, mittelgradig und ausgeprägt eingeordnet.

In der ersten Bewertungssitzung wurden den Logopädinnen die Sprachaufnahmen in zufälliger Reihenfolge präsentiert. Dabei kristallisierten sich folgende Punkte heraus, welche die nachfolgenden Bewertungssitzungen prägten:

- Das offene Näseln konnte nur an den Vokalen und den langen Sätzen diagnostiziert werden.
- Das geschlossene Näseln konnte nur an den Nasallauten /m/, /n/ sowie dem Nasalsatz „Nenne meine Mama Mimi“ bewertet werden.

- Die Nasalität konnte bei den isolierten Plosiv- und Zischkonsonanten nicht diagnostiziert werden. Hier ist lediglich eine Beurteilung der Konsonantensätze möglich.
- Die Logopädinnen hatten Mühe die Nasalität an einzelnen Vokalen zu bewerten. Vielmehr benötigten sie die Folge der 5 Vokale hintereinander, um ein Gesamtmaß für alle 5 Vokale abzugeben.
- bei den Vokalen konnten keine Nasalitätsunterschiede aufgrund ihrer Phonation, d. h. kurz, lang oder laut, festgestellt werden.

Aufgrund dieser Ergebnisse wurden in den nachfolgenden Sitzungen lediglich die Vokale und Nasallaute /m/ und /n/ bewertet.

Die Verteilung der Datensätze pro Sprachlaut, Sprechergruppe und Nasalitätsausprägung gibt Tabelle 5 wieder. Dabei wurde die Phonationsart bei den Vokalen nicht berücksichtigt. Die leichten Unterschiede in der Anzahl der Passagen kommen dadurch zustande, dass nicht alle Aufnahmen brauchbar waren, z. B. aufgrund von Verzerrungen oder Störungen bzw. nicht aufgenommen wurden.

Gruppe	Grad	A	E	I	O	U	M	N	A	E	I	O	U	M	N
		Absolut							**Relativ in %**						
Frauen	**Gesamt**	201	198	200	197	198	59	59	100	100	100	100	100	100	100
	Nonnasal	119	119	120	119	120	30	30	59.2	60.1	60.0	60.4	60.6	50.8	50.8
	Nasal	82	79	80	78	78	29	29	40.8	39.9	40.0	39.6	39.4	49.2	49.2
	Nasal - 1	15	15	15	15	15	6	6	18.3	19.0	18.8	19.2	19.2	20.7	20.7
	Nasal - 2	46	44	45	42	44	9	9	56.1	55.7	56.3	53.8	56.4	31.0	31.0
	Nasal - 3	21	20	20	21	19	14	14	25.6	25.3	25.0	26.9	24.4	48.3	48.3
Männer	**Gesamt**	34	33	33	33	33	11	11	100	100	100	100	100	100	100
	Nonnasal	26	25	24	25	25	3	3	76.5	75.8	72.7	75.8	75.8	27.3	27.3
	Nasal	8	8	9	8	8	8	8	23.5	24.2	27.3	24.2	24.2	72.7	72.7
	Nasal - 1	-	-	-	-	-	3	3	-	-	-	-	-	37.5	37.5
	Nasal - 2	8	8	9	8	8	3	4	100	100	100	100	100	37.5	50.0
	Nasal - 3	-	-	-	-	-	2	1	-	-	-	-	-	25.0	12.5
Kinder	**Gesamt**	206	202	205	204	203	74	74	100	100	100	100	100	100	100
	Nonnasal	175	172	176	175	174	59	59	85.0	85.1	85.9	85.8	85.7	79.7	79.7
	Nasal	31	30	29	29	29	15	15	15.0	14.9	14.1	14.2	14.3	20.3	20.3
	Nasal - 1	22	21	20	21	20	11	11	71.0	70.0	69.0	72.4	69.0	73.3	73.3
	Nasal - 2	9	9	9	8	9	4	4	29.0	30.0	31.0	27.6	31.0	26.7	26.7
	Nasal - 3	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Tab. 5: Verteilung der Sprachdaten je Sprachlaut, Sprechergruppe und Nasalitätsausprägung[60]

[60] Die relativen Angaben den nasalen und der non-nasalen Gruppe beziehen sich jeweils auf die Gesamtdatenanzahl, so dass beide aufsummiert 100 % ergeben. Die Relativangaben der Klassen 1, 2 und 3 beziehen sich auf die Gesamtanzahl der Nasaldaten.

6 Computergestützte Trainingsumgebung

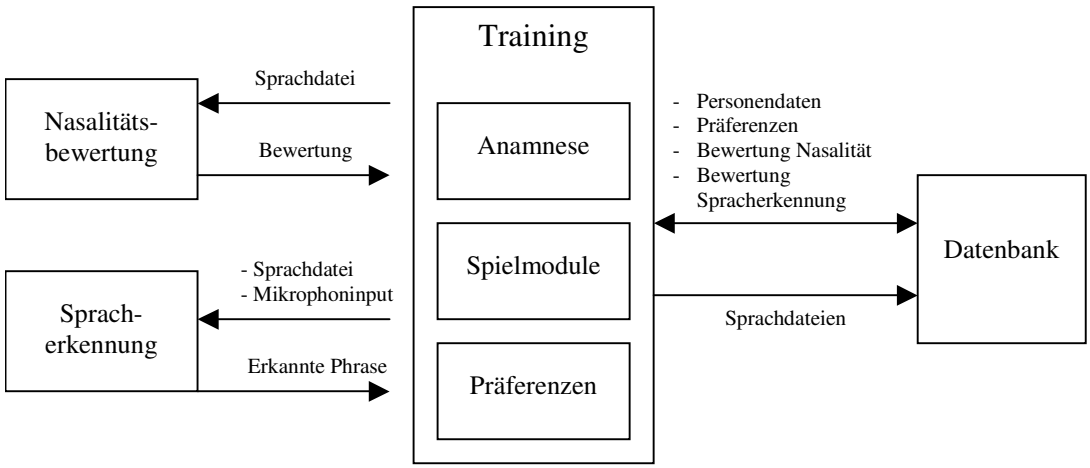

Abb. 35: Computergestützte Trainingsumgebung zur Evaluierung der Nasalität

Die Trainingsumgebung besteht aus den 4 Modulen:

- Nasalitätsbewertung
- Spracherkennung
- Training
- Datenbank

In diesem Kapitel sollen die zwei Module „Steuerung“ und „Spracherkennung“ sowie das Zusammenspiel aller Module wiedergegeben werden. Das Modul „Datenbank“ wird in Kapitel 5 „Sprachdatenbank NASAL“ beschrieben. Mit der „Nasalitätsbewertung“ beschäftigen sich die Kapitel 7 „Bestimmung der Parameter“ und 8 „Klassifikationsergebnisse“. Diese Einteilung erfolgt vor dem Hintergrund, dass die Bestimmung der Nasalität mit einer hohen Güte das wichtigste Kriterium für die Trainingsumgebung ist. Hier soll lediglich festgestellt werden, dass die Messung der Nasalität zum jetzigen Zeitpunkt für eine große Klasse von Sprechern mit einer hohen Güte realisiert ist.

6.1 Ablauf einer Sitzung

Das Zusammenspiel der verschiedenen Module soll am Ablauf einer Beispielsitzung verdeutlicht werden.

In einer ausführlichen Eingangsuntersuchung (Anamnese) erfasst der behandelnde Sprachtherapeut alle relevanten Daten, als da wären:

- Patientendaten (Name, Alter, Geschlecht, Sprechstörung)
- Auditive Bewertung der Nasalität des Patienten durch den Sprachtherapeuten

Die auditive Bewertung basiert auf dem Nachsprechen der Passagen des Heidelberger Rhinophoniebogens durch den Patienten. Diese Aufnahmen werden zusätzlich durch das Nasalitätsmodul hinsichtlich des Nasalitätsgrades und durch das Spracherkennungsmodul hinsichtlich der Spracherkennungsgüte bewertet. Der Sprachtherapeut hat somit die Möglichkeit die Auswertungen des Nasalitätsmoduls mit seiner eigenen Bewertung zu vergleichen. Da bei den Aufnahmen die nachzusprechenden Wörter bekannt sind, lassen sich hier weiterhin Aussagen über die Güte der Spracherkennung für den einzelnen Probanden gewinnen. Am Ende der Sitzung wird festgehalten, wie viele Wörter das System richtig, falsch oder gar nicht erkannte. Diese Informationen dienen dem Sprachtherapeuten als Entscheidungsstütze für den weiteren Gebrauch des Trainingssystems. So kann man anhand der Anzahl der nicht erkannten Phrasen, die Eignung der Sprachsteuerung überprüfen. Ist die Anzahl zu hoch sollte auf jeden Fall eine Anpassung der Spracherkennung an den Probanden mit Generierung eines persönlichen Sprachprofils durchgeführt werden. Führt auch diese nicht zum Erfolg (z. B. aufgrund vieler übergelagerter Sprechstörungen), sollte von einem Einsatz der Spielmodule abgesehen werden und zwecks Verlaufskontrolle lediglich eine Anamnese durchgeführt werden.

Die Patienten-, Bewertungs- und Sprachdaten dieser Eingangssitzung werden in der Datenbank gespeichert. Mit der Auswertung der Daten (Bestimmung der Nasalität) und der Einschätzung durch den Sprachtherapeuten wird der anschließende Trainingsverlauf bestimmt. Ganz wichtig ist hierbei die Einschätzung des Sprachtherapeuten. Sie entscheidet über den Einsatz der Spielmodule. Als Entscheidungsunterstützung kann der Sprachtherapeut hierbei die Auswertungen des Nasalitätsmoduls sowie die Statistiken des Spracherkennungsmoduls bezüglich der Spracherkennungsgüte verwenden. Entscheidet der Sprachtherapeut sich für den Einsatz, wählt er die in Frage kommenden Spielmodule aus. Basierend sowohl auf seiner, als auch auf der automatischen Bestimmung der Nasalität des Probanden, stellt er einen Schwellwert ein. Dieser Schwellwert gibt die maximal zulässige Nasalitätsstärke an, die bei der Spracheingabe vom Trainingsprogramm akzeptiert wird. Die Analyse des Nasalitätsgrades kann dabei sowohl absolut im Vergleich mit nonnasalen Sprachdaten als auch relativ im Vergleich zu den eigenen bisher aufgenommenen Sprachdaten erfolgen. Somit kann die relative Nasalitätsbestimmung auch dann noch Aussagen machen wo sprecherspezifische störende Überlagerungen im Vergleich zu den Normdaten keine Aussage über einen tendenziellen Lernerfolg zulassen würden.

Vor jeder folgenden Trainingssitzung wird dann eine kürzere Untersuchung (kleine Anamnese) vorangestellt. In dieser kann anhand eines kleinen Satzes an aufzunehmenden Sprachdaten die Tagesform ermittelt werden. Damit soll die Trainingsumgebung an die Schwankungen der Aussprache bzgl. Sprechgeschwindigkeit, Lautstärke, Tonhöhe und Nasalität angepasst werden. Der Proband kann z. B. über die Darstellung einer Schalldruckkurve oder durch Wiedergabe seiner aufgenommenen Sprache seine Lautstärke und sein Tonhaltevermögen überprüfen. Wie bei der ausführlichen Eingangsuntersuchung werden die Aufnahmen bzgl. der Nasalität und Spracherkennungsgüte bewertet. Damit kann vor jeder Sitzung erneut über den Einsatz der Spielmodule entschieden werden.

Die Trainingsphase wird in Form eines Spieles durchgeführt. Für die Steuerung eines Spieles über Sprache, sollen zwei Spielarten beispielhaft dargestellt werden. Bei der einen Variante handelt es sich um die Steuerung einer Spielfigur über Kontroll- und Navigationsbefehle durch ein Labyrinth. Da hier lediglich Kontroll- und Navigationsbefehle wie „Ende“ oder „links“ nötig sind, reicht in diesem Fall ein kleiner Wortschatz von ca. 20 Wörtern für die Spracherkennung aus. Solch eine Spielidee, allerdings ohne Sprachsteuerung, wurde im Rahmen einer Studienarbeit am Lehrstuhl bereits implementiert [Kög98]. Eine andere Variante wäre das Konzept eines Adventures. Hier können neben den Kontroll- und Navigationsbefehlen spezifische Befehle pro Szene gesprochen werden. Die Größe des Vokabulars in diesem Szenario hängt natürlich von der Komplexität des Adventures ab, wird aber normalerweise mehr als 50-100 Wörter betragen. Da sich der Benutzer aber von Bild zu Bild bewegt, sind nicht alle Wörter in dem jeweiligen Kontext relevant. Somit bietet es sich an, ein Vokabular pro Szene sowie ein Vokabular für die Navigation und Kontrolle des Spieles zu erstellen. Nur das Vokabular der aktuellen Szene sowie das Vokabular zur Navigation und Kontrolle werden aktiv geschaltet. Demzufolge sollte auch bei diesem Spieltyp ein aktives Vokabular von 20-30 Wörtern ausreichen.

Beim Spielverlauf werden die Befehle über ein Mikrophon gesprochen. Die gesprochene Phrase wird als Audiodatei gespeichert. Das Spracherkennungsmodul liest die Datei, wertet sie aus und liefert das erkannte Wort an das Trainingssystem[61]. Das erkannte Wort wird sowohl in Textform als auch als Audiodatei an das Nasalitätsmodul weitergegeben. Aufgrund des übergebenen Wortes, generiert das Nasalitätsmodul aus der Audiodatei die relevanten Parameter zur Nasalitätsmessung. Der aus den Parametern ermittelte Nasalitätsgrad wird an das Trainingssystem zurückgegeben und mit dem eingestellten Schwellwert verglichen. Nur wenn der Nasalitätsgrad unterhalb des Schwellwertes ist, wird der Befehl ausgeführt.

---

[61] Die Erkennung kann natürlich auch sofort aus der Mikrophoneingabe erfolgen. Die Erkennung aus einer Datei erfolgt vor dem Hintergrund, dass die Datei für die Nasalitätsbewertung und der Speicherung in der Datenbank benötigt wird (vgl Kapitel 6.2 Modul „Spracherkennung).

Pro Spiel werden am Ende folgende statistische Informationen in eine Datei geschrieben:

- Gesamtanzahl aller gesprochenen Phrasen
- Anzahl der nicht erkannten Phrasen (zur Ermittlung der Spracherkennungsgüte)
- Anzahl nicht akzeptierter Befehle, aufgrund des Schwellwertes.

Die Anzahl der nicht erkannten Phrasen im Verhältnis zur Gesamtanzahl, gibt Rückschlüsse über die Spracherkennungsgüte. Interessant ist hier noch der dritte Wert. Die Anzahl nicht akzeptierter Befehle aufgrund des Schwellwertes, lässt sich zu seiner Feinjustierung verwenden. Ist die Anzahl zu hoch, sollte der Schwellwert herabgesetzt werden. Ist sie sehr gering, kann sie hochgesetzt werden.

Alle Statistikauswertungen sowie die Sprachdaten der kleinen Anamnese mit ihren Nasalitätsauswertungen werden in der Datenbank gespeichert.

6.2 Modul „Spracherkennung“

Da am Lehrstuhl keine Erfahrungen mit einem Spracherkennungssystem vorhanden sind, sollte für die Spracherkennung ein kommerzielles Produkt zum Einsatz kommen. Folgende Anforderungen wurden dabei an das Produkt gestellt:

1. Erkennung deutscher Sprache
2. hohe Spracherkennungsgüte
3. Integration in die Trainingsumgebung. Dies bedeutet vor allem: geeignete Schnittstellen zum Nasalitäts- und Trainingsmodul.

Die erste Anforderung führte zum Ausschluss des Betriebssystems MacOS. Für deutsche Sprache existierten damals nur Spracherkennungsprodukte unter dem Betriebssystem Windows. Alle kommerziellen Produkte basierten letztendlich auf folgenden Sprach-erkennern: „ViaVoice98 Executive“ von IBM, „Naturally Speaking Preferred (Vers. 3.5)“ von Dragon Systems, „VoiceXpress Professional (Vers. 1.1)“ von Lernout&Hauspie und „FreeSpeech98“ von Philips. Alle Systeme für kontinuierliche Sprache erreichen nach einem Training Erkennungsraten größer 90% [Kög00]. Der besseren Lesbarkeit wegen werden im weiteren Verlauf der Arbeit die Spracherkenner über ihren Firmennamen angesprochen.

Die besten Erkennungsraten erzielten IBM und Dragon. Erst bei der Frage der Integration trennte sich die Spreu vom Weizen. Lediglich IBM und Dragon bieten ein Development Kit zur Integration an. Da die Spracheingaben bzgl. ihrer Nasalität ausgewertet werden müssen und in der Datenbank gespeichert werden müssen, war es notwendig auf die gesprochene

Eingabe zugreifen zu können. Diese Möglichkeit bietet aber keines der Systeme an, da der kontinuierliche Mikrophonstrom lediglich fensterweise betrachtet wird. Glücklicherweise bestand aber die Möglichkeit der IBM Engine eine Audiodatei zur Erkennung vorzulegen. Dadurch konnte man eine Spracheingabe als Datei speichern und diese Datei sowohl dem Spracherkennungs- als auch dem Nasalitätsmodul zur Bewertung vorlegen. Das Speichern in der Datenbank war somit ebenfalls realisierbar. Zusätzlich besteht die Möglichkeit diese Audiodateien dem Benutzer noch einmal als Feedback vorzuspielen.

In einer Diplomarbeit wurde dann ein Spracherkennungsmodul auf der Basis der Spracherkennungs-Engine vom „ViaVoice98“ der Firma IBM entwickelt und bzgl. seiner Eignung für die Trainingsumgebung getestet [Kög00]. Um die Eignung und das Handling des Moduls zu testen, wurden weiterhin erste Versionen des Anamnese- und Steuermoduls entwickelt. Während das Spracherkennungsmodul in „Visual C++“ entwickelt ist, wurden das Anamnese- und Steuermodul mit dem Autorensystem „Director“ der Firma „Macromedia“ implementiert. Damit war es nun möglich umfangreiche Untersuchungen der Spracherkennungsgüte zu tätigen.

Eine Verbesserung der Spracherkennungsrate um bis zu 20% kann durch eine Anpassung des Systems an den Benutzer erreicht werden [Mal98, Kuh99]. Bei ViaVoice kann diese Adaption beim ersten Start des Systems in einer ca. 90-minütigen Trainingssitzung erfolgen[62]. Im Anschluss an die Sitzung, wird dann ein persönliches Sprachprofil angelegt. Dieses Training ist für jeden Anwender des Systems erforderlich. Bei einem Wechsel des Benutzers muss dann das entsprechende Profil ausgewählt werden. Weiterhin kann eine Verbesserung durch die Benutzung kleiner Vokabulare mit „einfachen“ Grammatiken erreicht werden.

Über das Speech Development Kit von IBM ist es möglich sich eine eigene Grammatik zu definieren. Unter Grammatik wird in diesem Zusammenhang eine strukturierte Sammlung von Wörtern und Phrasen verstanden, welche in Verbindung mit bestimmten Regeln alle zu einem gegebenen Zeitpunkt gültigen Ausdrücke angibt. Bei der Entwicklung von Grammatiken sollten einige Grundsätze beachtet werden. So sollten in einer Grammatik so wenige Phrasen wie möglich verwendet und die Syntax einfach gehalten werden. Dadurch ist eine schnellere und genauere Spracherkennung möglich. Weiterhin sollten sich die gültigen Ausdrücke in ihrem Klangbild nicht zu stark ähneln.

---

[62] In der aktuellen Version ViaVoice Millenium Pro von IBM wurde das minimale Training auf ca. 15 Minuten verkürzt. Nach einem Test in [Kuh00] war die Erkennungsleistung nach dem Minimaltraining insgesamt aber nur mäßig. Als Testsieger bzgl. der Erkennungsleistung ging Naturally Speaking Preferred 4.0 von Dragon Systems hervor. Weiterhin lässt sich Naturally Speaking nun auch mit einer importierten Textdatei trainieren. Damit ließe sich aus den Anamnesedaten der Patienten ein automatisches Sprachbenutzerprofil erstellen. Es wäre in diesem Zusammenhang die Integration der Spracherkennung in die Trainingsumgebung zu überprüfen.

Bevor eine Grammatik für die Spracherkennung mittels der ViaVoice-Engine eingesetzt werden kann, muss aus ihr ein Vokabular erstellt werden. Im Vokabular wird jedem in der Grammatik verwendetem Wort, seine Phonemfolge zugeordnet. Hier wird somit festgelegt, wie sich jedes gesprochene Wort anhört. Da für ein Wort durch Variation der Betonung oder Dehnung oft mehrere Aussprachevarianten existieren, sind diese Alternativen ebenfalls zu beachten. Zur Erstellung eines Vokabulars zu einer Grammatik bietet IBM das Tool „Dictionary Builder“ an. Dabei wird jedes in der Grammatik verwendete Wort von seiner alphanumerischen Darstellung in eine Phonemfolge übersetzt.

Während einer Spracherkennungssitzung besteht die Möglichkeit dynamisch zwischen den Vokabularen zu wechseln und sogar mehrere gleichzeitig zu verwenden. Diese Eigenschaft erlaubt die kontextsensitive Aktivierung von Vokabularen im Spielmodul, wodurch sich die Erkennungsleistung der Spracherkennung steigern lässt.

6.2.1 Güte des Spracherkennungsmoduls

Die Zielgruppe bei kommerziellen Spracherkennungsprodukten bilden „normal“ sprechende Erwachsene. Auf sie ist das Produkt optimiert. Es stellte sich daher die Frage, wie gut die Spracherkennung mit sprechgestörten Aussprachen, insbesondere von Kindern, zurechtkommen würde. Nachteilig für das Spracherkennungssystem ist weiterhin, dass es nicht an die Sprecher der nasalen Daten über eine Trainingssitzung angepasst werden kann, da auf die ursprünglichen Sprecher nicht mehr zugegriffen werden konnte. Leider kann man das Spracherkennungssystem auch nicht mit den Passagen des Heidelberger Rhinophoniebogens trainieren. ViaVoice lässt bezüglich des Trainings keine Möglichkeiten für Änderungen am Umfang oder Inhalt des Trainings zu.

Zur Beantwortung dieser Frage wurden mehrere Tests mit dem Spracherkennungsmodul durchgeführt. Die Daten der gesamten Datenbank NASAL wurden in Audiodateien extrahiert und in drei Gruppen eingeteilt:

- Logopäden: normal sprechend
- Logopäden. Nasalität simulierend
- Kinder mit nasaler Aussprache

Die Erkennungsrate bei den normal sprechenden Logopäden war recht gut. Sobald der Spracherkenner aber mit nasalen Daten konfrontiert wurde, sank die Erkennungsrate drastisch unter 30%. Bei den Kindern mit nasaler Aussprache waren die Raten am kleinsten. Da die Raten für die nasale Sprache sehr gering ausgefallen sind, muss die Eignung des IBM-Systems ohne vorherige Trainingssitzung für die Erkennung sprechgestörter Ausdrücke in Frage gestellt werden. Hier muss eine neue Untersuchung zeigen, wie die Erkennungs-

raten bei Training des Systems mit der nasalen Aussprache eines Patienten ausfallen. Wird dann eine gute Erkennungsrate erreicht, bedeutet dies, dass für jeden Benutzer des Systems eine komplette Trainingssitzung durchgeführt werden muss. Sollten die Raten allerdings zu gering ausfallen, muss vom Einsatz der IBM Sprach-Engine Abstand genommen werden und eine eigene Spracherkennung implementiert werden. Diese hätte den großen Vorteil, dass sie mit nasalen Sprachdateien trainiert werden könnte und man sich beim Vokabular größtenteils auf den Heidelberger Rhinophoniebogen beschränken könnte. Das Training ließe sich dann implizit im Rahmen der Anamnese innerhalb der Trainingsumgebung durchführen.

Leider fehlt bei ViaVoice die Möglichkeit, das Training der Spracherkennung in eine externe Applikation einzubinden. Zwar kann man ein eigenes Vokabular erstellen und nur dieses während der Erkennung aktivieren. Das Training muss aber mit den von IBM vorgeschlagenen Passagen geschehen. Wünschenswert wäre in diesem Zusammenhang ein Training mit den Sprachdaten des Heidelberger Rhinophoniebogens und die Einbettung des Trainings in die Anamnese und Spielmodule.

6.3 Modul „Training“

Das Trainingsmodul setzt sich aktuell aus jeweils einem Steuer-, Anamnese-, Präferenz- und Spielmodul zusammen. Es wurde mit der Autorensoftware „Director“ der Firma „Macromedia“ entwickelt. Dabei erfolgte die Implementierung des Spielmoduls im Rahmen einer Studienarbeit auf dem Betriebssystem MacOS [Kög98]. Die anderen drei Module wurden zusammen mit dem Spracherkennungsmodul im Rahmen einer Diplomarbeit auf dem Betriebssystem „Win95“ entwickelt [Kög00]. Der Betriebssystembruch war wegen des Einsatzes der kommerziellen Spracherkennung notwendig, da für das Betriebssystem MacOS keine kommerzielle Spracherkennung basierend auf deutscher Sprache erhältlich war. Um eine räumliche und betriebssystemabhängige Kommunikation zwischen den Modulen zu ermöglichen, werden die Daten über Text- und Sounddateien ausgetauscht. Als Trennzeichen kommen in den Textdateien der Tabulator und ein Carriage Return (Zeilenvorschub) zum Einsatz. Bei den Sounddateien wurde das WAV-Format gewählt, was sowohl auf dem Betriebssystem Windows als auch auf MacOS bekannt ist.

Die Aufgabe des Steuermoduls liegt in der Kontrolle der Submodule, d. h. im Aufrufen dieser Module nach Nachrichten von den anderen Modulen und dem Betriebssystem, sowie nach Benutzeraktionen. Insbesondere erfolgt hier auch eine Überprüfung der Modulschnittstellen vor dem Aufruf mit einer entsprechenden Fehlerbehandlung.

In dem Modul „Präferenzen“ werden drei Parameter eingestellt. Mit der gewünschten Anzahl der Wiederholversuche wird festgelegt, wie viel Versuche der Patient hat, um eine gültige Eingabe zu machen. Der zweite Parameter gibt die erstrebte Verbesserungsrate der Aussprache in Prozentpunkten an. Und der dritte Parameter bestimmt, ob die Nasalitätsbewertung absolut oder relativ durchgeführt werden soll. Aus den letzten zwei Parametern werden somit die Schwellwerte für das Training bestimmt. Alle Parameter werden in einer Textdatei gespeichert, welche den Spielmodulen als Input dient.

In dem Modul „Anamnese“ wird dem Probanden eine Wortfolge vorgespielt, die er nach jedem Wort sprechen muss. Dabei werden die Spracheingaben aufgenommen und als Sprachdatei gespeichert. Diese Sprachdatei wird dann von den Modulen „Spracherkennung“ und „Nasalitätsbewertung“ verarbeitet. Das Modul „Nasalitätsbewertung“ analysiert jede Spracheingabe hinsichtlich ihrer Nasalität. Da bekannt ist welches Wort gesprochen werden sollte, werden die entsprechenden Normwerte herangezogen. Die Ergebnisse werden in einer Textdatei abgelegt. Das Modul Spracherkennung erzeugt ebenfalls eine Textdatei, welche zur Ermittlung der Spracherkennungsgüte herangezogen werden kann. Im Einzelnen steht dort, welches Wort zu erkennen gewesen wäre, ob die Spracheingabe einem Wort zugeordnet werden konnte, und ob das Wort richtig erkannt wurde. Im Fall einer Fehlerkennung wird zusätzlich das erkannte Wort gespeichert. Mit Hilfe dieser Daten lässt sich dann die „Tagesform“ des Probanden feststellen sowie der Einsatz der Spracherkennung für ein Spielmodul überprüfen. Weiterhin dienen die Daten der Spracherkennungsgüte auch der Feststellung, welche Wörter gut erkannt werden und wo eine Verwechslungsgefahr zwischen den Wörtern besteht. Damit lassen sich dann die für die Kontrolle eines Spielmoduls über Sprachsteuerung geeigneten Wörter identifizieren. Die aufgenommenen Sprachdateien werden alle in einem Verzeichnis gespeichert, von wo sie dann in die Datenbank abgelegt werden können. Damit nachfolgende Aufrufe der Anamnese vorhergehende Sitzungen nicht überschreiben, werden die Audio- und Textdateien entsprechend einer Probanden-ID und einer Sitzungsnummer aufsteigend benannt.

Wie bereits erwähnt, wurde das Spielmodul im Rahmen einer Studienarbeit implementiert [Kög98]. Die Navigation der Spielfigur im Labyrinth erfolgt hierbei über die Tastatur. Da das Spielmodul auf dem Betriebssystem MacOS implementiert, die Spracherkennung aber unter Windows realisiert wurde, ist die Steuerung über die Spracherkennung in das Spielmodul noch nicht realisiert. Aufgrund der niedrigen Erkennungsgüte des Spracherkennungsmoduls für nasale Sprache, müssen für das weitere Vorgehen noch Tests mit trainierten nasalen Rednern durchgeführt werden. Danach wird über die Notwendigkeit einer Implementierung einer eigenen Spracherkennung sowie der Entwicklungsplattform entschieden. Abhängig ob nun auf MacOS oder Windows weiter entwickelt wird, muss entweder das Spielmodul oder das Trainingsmodul portiert werden. Da beide unter „Macromedia Director“ entwickelt

wurden, sollte der Aufwand für die Portierung im Rahmen der Zeitdauer einer Studienarbeit realisierbar sein. Es folgt nun eine Beschreibung des existierenden Spielmoduls, wobei die Angaben bzgl. der Spracherkennung noch zu realisieren sind.

Im Spielmodul hat der Proband eine Spielfigur durch das Labyrinth zu steuern. Die Steuerung wird mittels Spracheingaben erfolgen, welche auf ihre Nasalität hin analysiert werden. Das Ergebnis dieser Untersuchung wird dann mit dem im Modul „Präferenz“ eingestellten Schwellwert verglichen. Nur wenn der Schwellwert unterschritten ist, wird der Befehl ausgeführt. Der Befehl jeder Spracheingabe wird zusammen mit dem Zeitpunkt seiner Eingabe und dem ermittelten Nasalitätsgrad in einer Datei gespeichert. Weiterhin wird am Ende des Spieles die Gesamtzahl der Eingaben in dieser Datei gespeichert. Die entsprechenden Daten werden ebenfalls in der Datenbank abgelegt. Aus ihnen lassen sich z. B. der Ermüdungsgrad des Benutzers oder Aussagen über die Bedienerfreundlichkeit des Spielmoduls ableiten.

6.4 Zusammenfassung

Alle Module der computergestützten Trainingsumgebung sind implementiert. Die Module „Nasalitätsbewertung“, „Training“ und „Datenbank“ sind für den Einsatz bereits geeignet. Insbesondere kann die Nasalität mit einer hohen Güte bewertet und die gewonnenen Daten in der multimedialen Datenbank gespeichert werden. Das Trainingsmodul als quasi Kontroll-Modul muss lediglich um eine Schnittstelle zum *automatischen* Speichern der Daten in die Datenbank erweitert werden[63]. Zurzeit werden die Daten als Dateien abgelegt.

Das Modul „Spracherkennung“ hingegen liefert nur gute Erkennungsraten bei normal sprechenden Personen. Die Erkennungsrate sinkt jedoch drastisch bei nasalen Daten. Hier müssen Untersuchungen zeigen, inwieweit der Einsatz bei vorherigem Training des Probanden vertretbar ist. Bei negativen Ergebnissen empfiehlt es sich, eine eigene Spracherkennung zu implementieren. Diese kann dann mittels den nasalen Daten trainiert werden. Weitere Vorteile wären ein kurzes in die Trainingsumgebung eingebundenes Training durch das Nachsprechen des Heidelberger Rhinophoniebogens.

[63] Erste Vorabuntersuchungen auf dem MacOS zeigten die prinzipielle Eignung der Skriptsprache AppleScript für diese Aufgabe. Ein entsprechendes Äquivalent auf Windows wäre das Windows Scripting.

7 Bestimmung der Parameter

Dieses Kapitel zeigt die Vorgehensweise bei der Extraktion der Parameter auf. Die generierten Parameter werden im Rahmen dieser Arbeit in zwei Klassen eingeteilt. Die Frequenzbandparameter basieren dabei direkt auf den Frequenzbandintensitäten. Die Sprachmodellparameter werden über spektrale Maxima und Minima berechnet. Die angewandten Algorithmen, wurden in der Entwicklungsumgebung MatLab5.1 implementiert[64].

Weiterhin geht es in diesem Kapitel um die Handhabung der Sprachdaten, sowie die Probleme bei der Bestimmung der Sprachmodellparameter.

Abschließend werden für die erarbeiteten Sprachmodell-Parameter (Grundfrequenz, Formanten und Antiformanten) gruppenspezifische Intensitäten, Bandbreiten und Lagen für nonnasale und nasale Sprachdaten präsentiert.

Eine erste Diskussion über mögliche nasalitätsrelevante Sprachmodell-Parameter schließt dieses Kapitel ab.

7.1 Datenhandling

Um mit der großen Menge an Sprachdaten effizient arbeiten zu können, wurden die Sprachdaten von ihrer ursprünglichen Speicherart (einzelne Sounddateien) in die für MatLab geeigneten Matrizen konvertiert. Jeder Sprachlaut wird dabei separat in einer Matrix gespeichert. Die Zeilen der Matrix entsprechen den einzelnen Sprechern, die Spalten den Sample-Werten. Durch eine Längennormierung (vgl. Kapitel 7.2.1 „Vorverarbeitung“ weiter unten), besitzt jede der Zeilen die gleiche Länge. Den Sprachdaten vorangestellt ist ein Vektor mit den Datensatz kennzeichnenden Attributen - eine Sprecher-ID, der Sprachlaut, das Alter und Geschlecht des Sprechers, die Gruppenzugehörigkeit und die logopädische Bewertung hinsichtlich der Nasalität der Aufnahme. Zusätzlich wird auch eine Sitzungsnummer gespeichert. Damit können Aufnahmen desselben Sprechers unterschieden werden. Die Sitzungsnummer fand insbesondere bei den Sprachdaten der

[64] MatLab ist ein matrix-orientiertes Mathematikprogramm. In MatLab implementiert sind Funktionen zum Bearbeiten von Matrizen, Vektoren und Skalaren, alle trigonometrischen Funktionen, Befehle zum Darstellen und Drucken von Signalen und Funktionen zum Programmieren in einer eigenen MatLab-Sprache. Spezielle (komplexe) Funktionen existieren daneben als sogenannte M-Files. Das sind in der MatLab-Sprache geschriebene Batch-Dateien, die genauso ausgeführt werden können wie implementierte Funktionen. Für spezielle Anwendungsgebiete existieren Sammlungen solcher M-Files (Toolboxes). Die für uns interessante Toolbox ist die Signalverarbeitungs-Toolbox, welche uns insbesondere die Transformationen (FFT, LPC, ARMA, Cepstrum) bereit stellt.

Logopäden Verwendung, welche in der 1. Sitzung normal, in der 2. Sitzung nasal offen und in der 3. Sitzung nasal geschlossen sprachen.

ID	Gruppe	Geschlecht	Alter	Laut	Sitzungs-Nr.	Nasalitäts-grad
< 1000	0 – Patienten 1 – Logopäden 2 – Kinder	0 – männl. 1 – weibl.	0 – Erw. 1 – Kind	01 – A 02 – E ...	< 10	0 – kein 1 – gering 2 – mittel 3 – stark

Tab. 6: Kennzeichnung der Sprachdaten

Durch diese Art der Speicherung ist ein schneller und einfacher Zugriff auf den gesamten Sprachdatenbestand möglich. Insbesondere kann man elegant einzelne Sprecher oder Gruppen selektieren[65].

7.2 Merkmalsextraktion

Dieses Kapitel behandelt die Extraktion der Parameter aus den stimmhaften Sprachdaten. Die Merkmalsgewinnung erfolgt grob eingeteilt in zwei Stufen (vgl. Abbildung 36). In der ersten Stufe, der Vorverarbeitung, finden Berechnungen auf dem Zeitsignal statt. Die dabei angewandten Algorithmen führen keine Änderungen an den Original-Sprachdaten durch. Somit enthält die Sprachdatenbank NASAL grundsätzlich das unveränderte Datenmaterial. Die zweite Stufe, die eigentliche Merkmalsgewinnung, setzt nach der Transformation der vorverarbeiteten Sprachdaten in den Frequenzbereich ein.

7.2.1 Vorverarbeitung

Eine Überprüfung der aufgenommenen Daten ergab, dass die Amplitudendichteverteilung bei vielen aufgenommenen Daten nicht symmetrisch zum Mittelwert 0 war, d. h. die Summation der Abtastwerte x_i über der Zeit ergab nicht 0. Daher wird als erstes aus jeder Aufnahme der Gleichspannungsanteil herausgerechnet. (vgl. Kapitel 3.1.2 „Eliminierung des Gleichspannungsanteils“).

Danach wird zur Eliminierung störender Geräusche sowie vor allem zur Bestimmung des Lautbeginns, an allen Sprachdaten eine Pausenerkennung mittels des Energieschwellverfah-

[65] Weiterhin öffnet diese Speicherart die Möglichkeit auch auch Algorithmen der 2-D-Bildbearbeitung auf die (Unter-) Matrizen anzuwenden.

ren durchgeführt (vgl. Kapitel 3.1.3 „Pausenerkennung“). Ist der Beginn des quasistationären Teils lokalisiert, erfolgt eine Längennormierung, indem von dem Lautbeginn an 2205 Samples ausgewählt werden. Dieser Länge entspricht bei einer Samplerate von 22050 kHz eine Frequenzauflösung von 10 Hz.

Um den Einfluss von Lautstärkeneinflüssen, welche ihre Ursache hauptsächlich in der individuellen Sprecherlautstärke und dem Abstand des Sprechers zum Mikrophon haben, zu eliminieren, wird jeder Sprachdatensatz auf die gleiche Gesamtenergie normiert. Die Energie eines Sprachsignals wird dabei berechnet nach:

$$E = \sum_{n=0}^{N-1} \left| x_n \right|^2 \qquad (51)$$

Für jedes Sprachsignal s_i wird danach eine Verstärkung $Gain_i$ berechnet, so dass gilt:

$$Gain_i \cdot E_{si} = Normenergie \qquad (52)$$

Als Normenergie wird 2^{26} gewählt, welche der Durchschnittswert einer kleinen Stichprobe war. Um das Sprachsignal auf die Normenergie zu bringen, wird jeder Datenpunkt des Samples mit der Quadratwurzel von $Gain_i$ multipliziert.

Nach der Lautstärkennormierung erfolgt ein Anheben der höheren Frequenzen durch Faltung mit einem Parameter von –0.95 (vgl. [RJ93]).

Anschließend werden die Sprachdaten in Matrizen gespeichert. Damit sich keine Schnittfehler in die Matrizen einschleichen, erfolgt eine nachträgliche Überprüfung. Jede Tonsequenz wird nacheinander auf den Lautsprecher ausgegeben, und auf Verständlichkeit überprüft. Ergänzt wird diese auditive Überprüfung visuell durch einen Waveplot[66], der auf die periodische Struktur hin untersucht wird.

Von den NASAL-Sprachdaten wurden lediglich 0.6% falsch geschnitten. Die meisten Fehler verursachte ein starkes Rauschen vor der eigentlichen Aufnahme. Dies führte zu einer falschen Bestimmung des Anfangs. Derart inkorrekt geschnittene Daten lassen sich leicht über ihre kurze Länge orten. Ein zweiter Lauf, bei dem solche Sprachdaten mit kurzer Länge sofort verworfen wurden, reduzierte die Fehlerrate auf 0.1%. Die fehlerhaft geschnittenen Daten wurden aus der Matrix entfernt und daher für die Untersuchungen nicht berücksichtigt. Wie die auditive Überprüfung ergab, hatten die Vorverarbeitungen, von der Preemphase abgesehen, keinen nennenswerten Einfluss auf das Lautempfinden. Die Preemphase

[66] In einem Waveplot werden die Schalldruckschwankungen über die Zeit wiedergegeben.

bewirkte eine Anhebung der höheren Frequenzen. Ein Einfluss der Vorverarbeitungen auf die Nasalität ließ sich perzeptuell nicht feststellen.

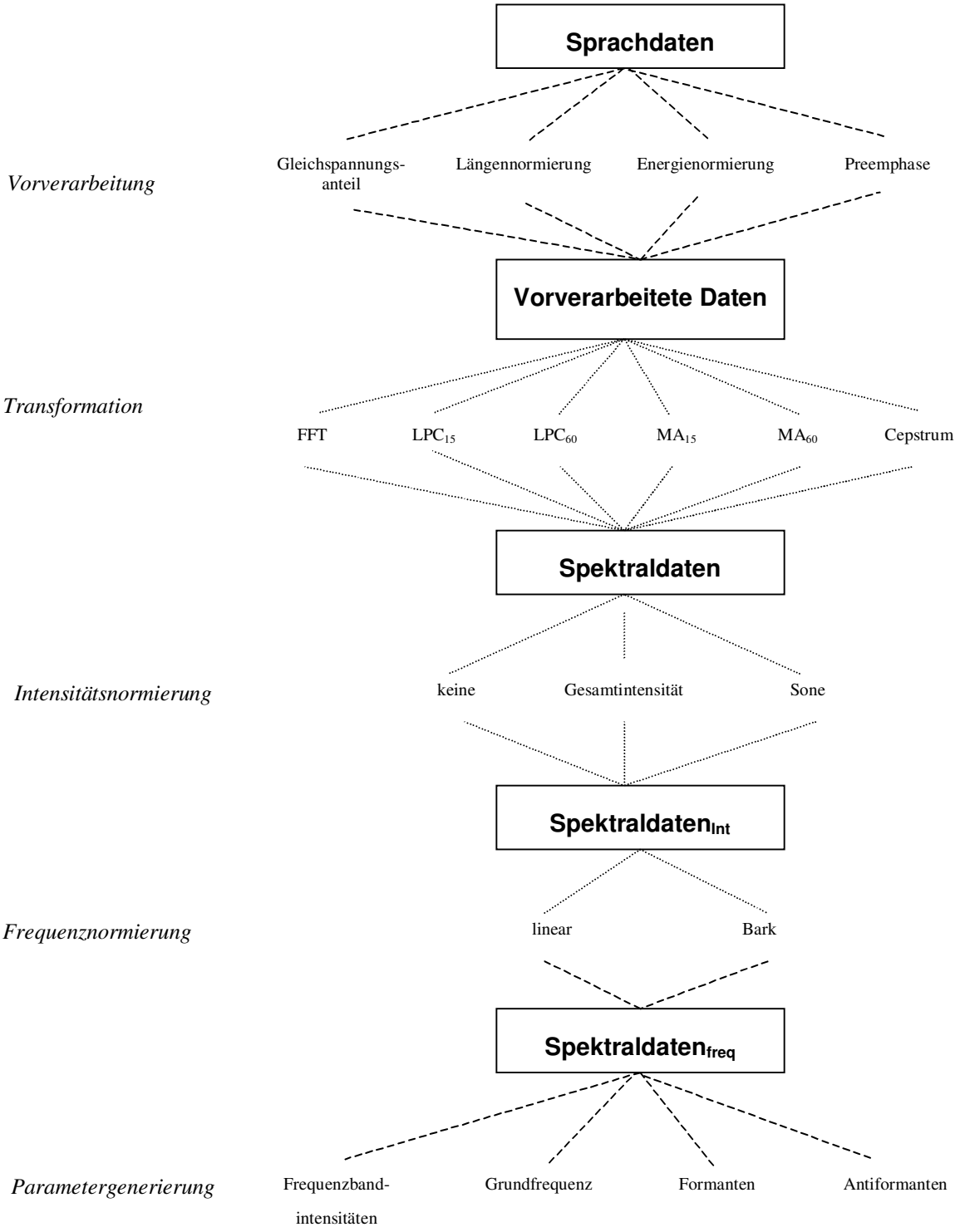

Abb. 36: Vorverarbeitung und Parametergenerierung

7.2.2 Frequenzbandparameter

Bei der Frequenzbandanalyse wird das gesamte Spektrum in Frequenzbänder unterteilt. Die Intensitäten dieser Frequenzbänder bilden die Parameter. Im Rahmen dieser Arbeit werden nur sprachrelevante Frequenzen bis 8 kHz berücksichtigt. Es wurden uniforme Frequenzbänder mit einer Bandbreite von jeweils 100 Hz und 400 Hz, sowie Bänder basierend auf der Bark-Skala untersucht. Die Bandbreiten für die uniformen Bänder wurden nach einer Voruntersuchung am Vokal /a/ bei den Logopädinnen ausgewählt. In dieser Voruntersuchung wurde das Klassifikationsergebnis in Abhängigkeit der Bandbreiten 10 Hz, 20 Hz, 50 Hz, 100 Hz, 200 Hz, 400 Hz und 800 Hz überprüft. Die besten Ergebnisse lieferten die Bandbreiten 100 Hz und 400 Hz. Bei kleineren Bandbreiten war die Parameteranzahl im Verhältnis zur Stichprobe zu groß, so dass die Anzahl der Freiheitsgrade zu groß war und die Jack-Knife-Klassifikation dadurch schlechter ausfiel. Die größeren Bandbreiten erwiesen sich als nicht detailliert genug.

Zusätzlich zur Variation der Frequenzbandbreiten, sollte auch die Auswirkung einer Intensitätsnormierung überprüft werden. So sind die echten Amplituden der Spektren für viele praktische Betrachtungen unbedeutend und werden meistens ignoriert oder einfach auf den Wert 1 normiert. Interessant ist vielfach nur der typische Verlauf der Spektren. In dieser Arbeit wurde die Intensität einmal über die Gesamtintensität des Sprachlautes normiert, das andere Mal wurde die Intensität pro Band in das psychoakustische Lautheitsmaß Sone transformiert.

Insgesamt wurden pro Sprachlaut 9 Frequenzbandparametermengen ermittelt, die im weiteren Kontext der Arbeit folgendermaßen bezeichnet werden:

U100_N : Uniform 100 Hz Bandbreite, keine Intensitätsnormierung

U100_I : Uniform 100 Hz Bandbreite, Intensitätsnormierung: Gesamtintensität

U100_S : Uniform 100 Hz Bandbreite, Intensitätsnormierung: Sone

U400_N : Uniform 400 Hz Bandbreite, keine Intensitätsnormierung

U400_I : Uniform 400 Hz Bandbreite, Intensitätsnormierung: Gesamtintensität

U400_S : Uniform 400 Hz Bandbreite, Intensitätsnormierung: Sone

BARK_N : Bark, keine Intensitätsnormierung

BARK_I : Bark, Intensitätsnormierung: Gesamtintensität

BARK_S : Bark, Intensitätsnormierung: Sone

7.2.3 Sprachmodellparameter

7.2.3.1 Grundfrequenz

Die Grundfrequenz nimmt eine Schlüsselstellung unter den Sprachsignalparametern ein. Die einer sprachlichen Äußerung unterliegende Intonation wird vornehmlich mit Hilfe dieses Parameters übertragen. Das menschliche Ohr ist gegenüber Änderungen der Grundfrequenz um eine Größenordnung empfindlicher als gegenüber Änderungen anderer Sprachsignalparameter [VHH98].

Grundsätzliche Schwierigkeiten bei der Bestimmung der Sprachgrundfrequenz in stimmhaften Lauten ergeben sich unter anderem aus folgenden Gründen:

- Bei einem beliebigen Sprachsignal von einem unbekannten Sprecher kann die Grundfrequenz über 4 Oktaven variieren (50 Hz - 800 Hz). Dadurch kann sie besonders bei Frauen und Kindern dicht bei dem 1. Formanten liegen bzw. mit diesem zusammenfallen [Fel84].
- Die Stimmbandschwingung selbst ist nicht immer regelmäßig. Schon unter normalen Bedingungen existieren gelegentliche Unregelmäßigkeiten [LCS67].
- Umgebungsgeräusche bei der Spracheingabe (Lärm, mehrere gleichzeitig Sprechende,...) können die Bestimmung in erheblichem Maße störend beeinflussen.

In der Literatur wurden zahlreiche Verfahren zur Grundfrequenzanalyse vorgestellt. Keines davon funktioniert für alle Aufgabenstellungen einwandfrei. Die Auswahl eines bestimmten Verfahrens hängt wesentlich von der jeweiligen Anwendung wie auch der Beschaffenheit der zu verarbeitenden Sprachsignale ab. Daher ist es nicht das Ziel dieses Abschnittes einen Überblick über die wesentlichen Verfahren und Ansätze zur Grundfrequenzbestimmung zu geben[67]. Vielmehr wird auf die in dieser Arbeit verwendete Grundfrequenzbestimmung über die Cepstrumsanalyse eingegangen. Diese hat sich bei zahlreichen Untersuchungen als verhältnismäßig zuverlässig erwiesen [Fel84, VHH98].

Wie im Kapitel 3.4 „Cepstral-Analyse“ beschrieben, erscheint die Grundfrequenz im Cepstrum als ein lokaler Peak. Nutzt man die Tatsache aus, dass die Grundfrequenz in der Regel größer als 50 Hz und kleiner als 400 Hz ist, braucht man nur nach einem lokalen Maximum in diesem Frequenzbereich suchen. Bei einer Samplefrequenz von 22050 Hz müssen die Cepstralwerte von $\lfloor 22050/400 \rfloor = 55$ bis $\lfloor 22050/50 \rfloor = 441$ betrachtet werden. Eine Verbesserung der Grundfrequenzbestimmung konnte durch das Vorschalten eines

[67] Einen guten Überblick über die Verfahren der Grundfrequenzbestimmung findet man in [Vary98].

Bandpassfilters mit Übertragungsbereich 50 Hz - 500 Hz vor der Cepstrum-Berechnung erreicht werden. Die ersten 50 Hz werden nicht berücksichtigt, da sie nicht sprachrelevant sind und zudem auch durch die Aufnahmestation beeinflusst werden. Frequenzen über 500 Hz werden ausgeblendet um den störenden Formanteneinfluss zu mildern. Als Grundfrequenzparameter werden die Intensität, die Frequenz und die Bandbreite extrahiert. Hier wird ein zweistufiges Verfahren verwendet. In der 1. Stufe erfolgt über das Cepstrum eine erste Annäherung der Grundfrequenz. In der 2. Stufe wird im Betragsspektrum ein 100 Hz breites, um diese erste Annäherung zentriertes Band, nach dem Maximum abgesucht. Dieses Maximum bestimmt die für die Untersuchungen verwendete Grundfrequenz. Da die exakte Lage der Grundfrequenz für die nachfolgenden Untersuchungen sehr wichtig ist, wurden die Grundfrequenzen visuell an dem Betragsspektrum nachgeprüft. Es ergab sich bei der Bestimmung eine durchschnittliche Fehlerrate von 3.45 %. Eine Untersuchung der Bestimmungsfehler in Abhängigkeit des Nasalitätsgrades brachte keine Signifikanzen hervor (vgl. Tabelle 7). Die Datensätze mit den fehlerhaft bestimmten Grundfrequenzen gingen in die Untersuchungen nicht ein.

	Anzahl			Fehler		
Laut	**Gesamt**	**Nonnasal**	**Nasal**	**Nonnasal**	**Nasal**	**Gesamt**
A	441	310	117	10	4	3.17 %
E	432	302	112	12	6	4.17 %
I	438	300	114	20	4	5.48 %
O	434	305	113	14	2	3.69 %
U	434	297	113	21	3	5.53 %
M	144	91	51	1	1	1.39 %
N	144	92	51	1	0	0.69 %

Tab. 7: Fehlerraten bei der Bestimmung der Grundfrequenz

Die Bandbreite der Grundfrequenz wird so gewählt, dass sich bei lokaler Normierung des Spektrums um die Grundfrequenz, 80% der Energie in dem Band befinden. Diese Energie wird als Intensitätsparameter für die Klassifizierung verwendet. Tabelle 8 gibt die extrahierten Grundfrequenzparameter wieder.

		Nonnasal			Nasal		
Gruppe	**Laut**	**Frequenz**	**Intensität**	**Bandbreite**	**Frequenz**	**Intensität**	**Bandbreite**
Frauen	**A**	209	59.2	21.2	221	62.0	16.6
	E	219	60.6	19.0	222	62.4	14.3
	I	219	61.8	10.8	235	63.0	10.5
	O	222	59.6	22.8	229	62.7	15.8
	U	231	62.4	12.4	234	63.3	11.0
	M	237	18.0	10.2	216	18.0	11.0
	N	235	18.0	11.0	228	18.0	10.6
Männer	**A**	124	58.2	48.8	127	59.6	44.2
	E	133	59.3	10.8	132	60.4	21.0
	I	133	59.3	12.7	156	62.1	24.3
	O	130	58.6	12.1	139	61.0	28.3
	U	145	60.2	16.0	147	61.8	26.9
	M	150	17.9	18.3	140	17.9	18.1
	N	153	17.9	13.3	146	17.9	11.9
Kinder	**A**	232	58.3	20.4	234	59.0	21.9
	E	242	59.7	18.4	200	60.6	21.7
	I	252	62.6	10.8	250	62.3	12.4
	O	242	59.3	18.6	213	60.3	16.4
	U	253	62.3	11.5	239	62.0	11.0
	M	243	18.0	11.8	221	18.0	11.7
	N	256	18.0	12.2	235	18.0	12.3

Tab. 8: Grundfrequenzparameter

7.2.3.2 Formanten

In Abbildung 37 ist das mittels einer Fouriertransformation gewonnene Spektrum als untere Frequenzreihe dargestellt. Um eine bessere Vergleichbarkeit zu ermöglichen, ist das Spektrum auf eine maximale Amplitude von 0.9 normiert. Die Feinstruktur in dem Spektrum rührt von den Harmonischen der Grundfrequenz. Für die Bestimmung der Formanten ist eine „Glättung“ vorteilhaft. Eine solche „Glättung“ leistet die LPC. Das über die LPC gewonnene Spektrum ist darüber dargestellt. Das Spektrum ist ebenfalls auf eine maximale Amplitude von 0.9 normiert und zwecks der Darstellung zusätzlich um den konstanten Betrag von 0.5 erhöht. Deutlich kann man sehen, wie sich die Grobstruktur des Fourierspektrums nachbilden lässt.

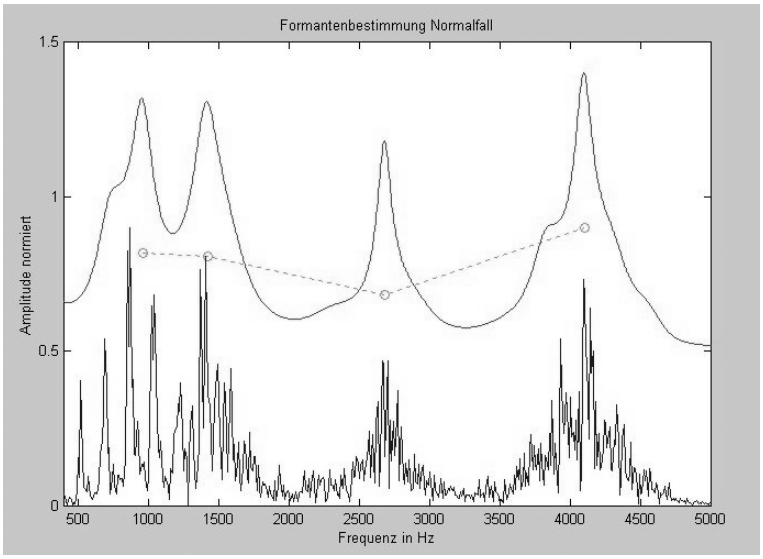

Abb. 37: Formantenbestimmung Normalfall (Vokal /a/, Frau)

Entscheidend beim LPC-Modell ist die gewählte Ordnung, welche die Auflösung der LPC bestimmt. Bei einer niedrigen Ordnung besitzt die entsprechende Funktion wenige Maxima. In obigem Beispiel wurde das Spektrum aus einem LPC-Modell der Ordnung 15 berechnet. Es besitzt 4 Maxima, welche die Formantenlage kennzeichnen. Die Kreise markieren diese Formanten. Auch hier ist die Intensität auf 0.9 normiert, jedoch wurde kein konstanter Betrag dazu addiert.

Der zur Formantenbestimmung in dieser Arbeit eingesetzte Algorithmus arbeitet auf Basis einer selbst erstellten Formantentabelle, welche sich bei der Formantenlage an die Tabelle von Peterson und Barney anlehnt (vgl. Tabelle 1). Es erfolgte ein visueller Abgleich der Formantenmittelwerte an einer kleinen Stichprobe, sowie die Ermittlung des 4. Formanten bei den Vokalen und aller 4 Formanten bei den Nasalen /m/ und /n/. Zusätzlich wurde pro Laut und Formant eine Bandbreite bestimmt, in der nach dem Formanten zu suchen ist.

Der 2-stufige Algorithmus versucht zuerst über eine LPC niedriger Ordnung (p = 15) die hohen Formanten F_3 und F_4 zu bestimmen. Hierbei fängt er mit der Berechnung von F_4 an. Aus der Formantentabelle werden die laut- und gruppenspezifischen Mittenfrequenz $\overline{F_{4_{Laut\_Gruppe}}}$ und Bandbreite $\overline{BB_{4_{Laut\_Gruppe}}}$ ausgelesen. Als Lage des Formanten F_4 wird die Frequenz F_{4Freq} aus dem Bereich $\overline{F_{4_{Laut\_Gruppe}}} \pm \overline{BB_{4_{Laut\_Gruppe}}}$ mit dem lokalen Intensitätsmaximum gewählt. Danach wird die Frequenz von F_3 nach dem gleichen Prinzip bestimmt. Es erfolgt lediglich eine Anpassung der rechten Grenze auf das Minimum zwischen $\{F_{3Laut\_Gruppe} + BB_{3Laut\_gruppe} ; F_{4Freq} - 1\}$. Die gefundene Lage der Formanten F_3 und F_4 wird in einem zweiten Schritt mit einer LPC höherer Ordnung (p = 60) genauer bestimmt. Zusätzlich werden in diesem 2. Schritt auch die niedrigen Formanten nach dem gleichen Prinzip

bestimmt. Die Bestimmung der Bandbreite der Formanten und ihrer Intensitäten erfolgt analog zu der Vorgehensweise bei der Grundfrequenz.

Ausschlaggebend für die Qualität der Formantenanalyse ist die Bestimmung der Formantenlage. Trotz der scheinbaren Einfachheit stellt dies in der Praxis selbst bei stimmhaften Lauten ein Problem dar. Dies rührt hauptsächlich daher, dass die Formantfrequenzen für den gleichen Laut erheblich differenzieren können, und die Tabellen für die Formantenlagen daher nur erste grobe Anhaltspunkte sein können.

Es sind verschiedene Probleme bei der Formantenbestimmung zu bewältigen. Ein erstes Problem stellen nahe beieinander liegende Formanten dar, welche bei kleinen Ordnungen von der LPC zusammengefasst werden, sowie tiefe Formanten, welche nicht mehr aufgelöst werden. Um diese ebenfalls zu ermitteln, muss man die Auflösung erhöhen. Dies geschieht über die Wahl einer höheren Ordnung. Im allgemeinen lassen sich alle Formanten eines Vokals im resultierenden Spektrum abbilden. Leider werden aber auch die Harmonischen stärker abgebildet, was sich im Spektrum in mehreren lokalen Maxima zeigt, welche wiederum zu stärkeren Deutungsproblemen führen. So zeigt Abbildung 38 am Beispiel des Vokals /a/ wie der 2. Formant fälschlicherweise als 1. Formant klassifiziert wird, während der 1. Formant bei ca. 450 Hz aufgrund der Nähe zur Grundfrequenz überhaupt nicht erkannt wird.

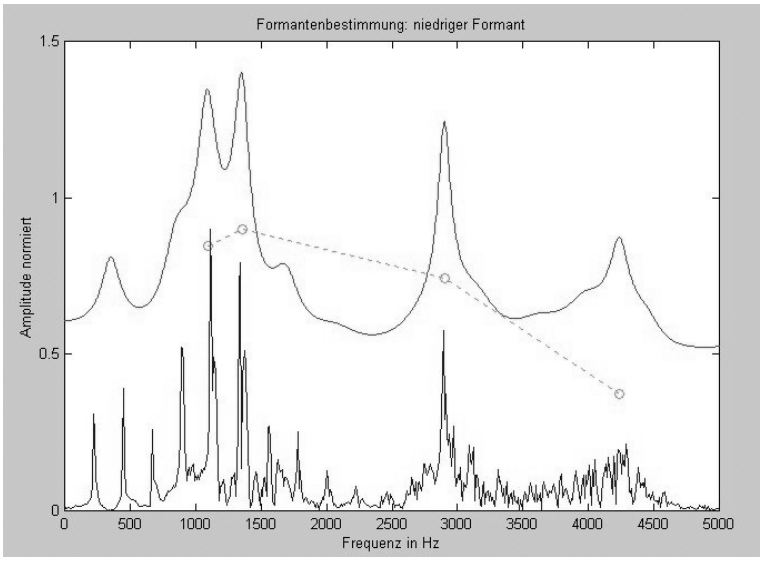

Abb. 38: Fehlertyp I: niedriger Formant (Vokal /a/, Frau)

Ein weiteres Problem stellt das Auftreten zusätzlicher Resonanzen dar. Betrachtet man das Fourierbetragsspektrum in Abbildung 39, sieht man unterhalb 2 kHz drei starke Resonanzfrequenzen. Welche die beiden tiefen Formanten des Vokals /a/ und welche die zusätzliche Resonanzfrequenz darstellt, ist nicht eindeutig bestimmbar. Nimmt man die Intensität des

LPC-Spektrums als Entscheidungsgrundlage, werden die Frequenzen um 1 kHz und 1.5 kHz als F_1 und F_2 klassifiziert. Dagegen würde beim Betragsspektrum die Wahl auf die Frequenz um 600 Hz als F_1 und 1 kHz als F_2 fallen. Je nach Wahl entsteht somit evtl. ein Fehler von 400 Hz bei der Bestimmung von F_1 und von 500 Hz bei F_2.

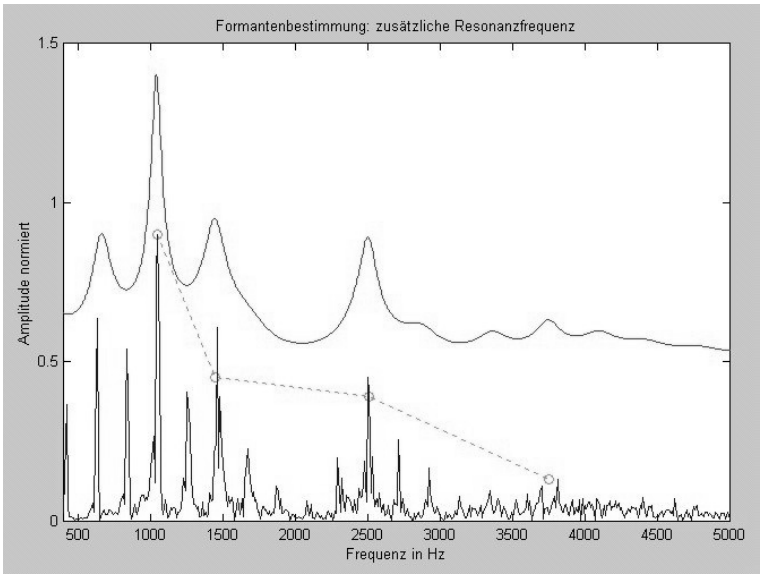

Abb. 39: Fehlertyp II: zusätzliche Resonanzfrequenz (Vokal /a/, Frau)

So wie mehrere Resonanzfrequenzen auftreten können, kann es vorkommen, dass das Fourierspektrum nur ein eindeutiges lokales Maximum im betrachteten Bereich hat. Dies kommt vor, wenn zwei Formanten im Spektrum zusammenfallen. So ist in Abbildung 40 lediglich die Frequenz um 1100 Hz als Formant erkennbar.

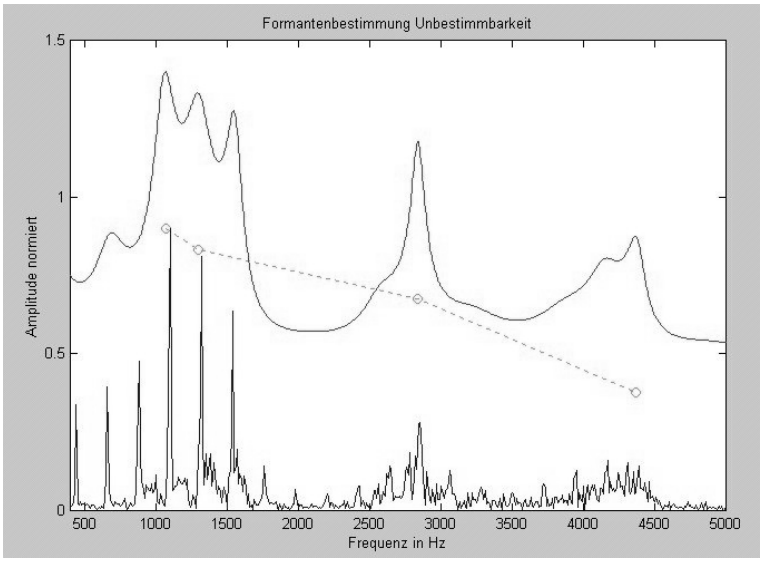

Abb. 40: Fehlertyp III: Unbestimmbarkeit (Vokal /a/, Frau)

Aufgrund der großen Varianz der Sprachdaten kann es zu Fehlzuweisungen kommen. So führt in Abbildung 41 eine zu kleine Suchbandbreite für die Suche des stark nach unten verschobenen 4. Formanten dazu, dass dieser nicht gefunden wird. Fälschlicherweise wird er als 3. Formant klassifiziert[68]. Sind die Bandbreiten dagegen zu groß gewählt, kommt es zur Überlappung der Suchbereiche. Dies führt dazu, dass die Formanten falsch zugeordnet werden. So würde in Abbildung 38 der 3. Formant irrtümlich als 4. Formant klassifiziert werden. Um eine eindeutige Zuordnung zu den einzelnen Bändern zu gewährleisten, ist es nötig, die Bänder nicht überlappen zu lassen und möglichst „schmal“ zu wählen.

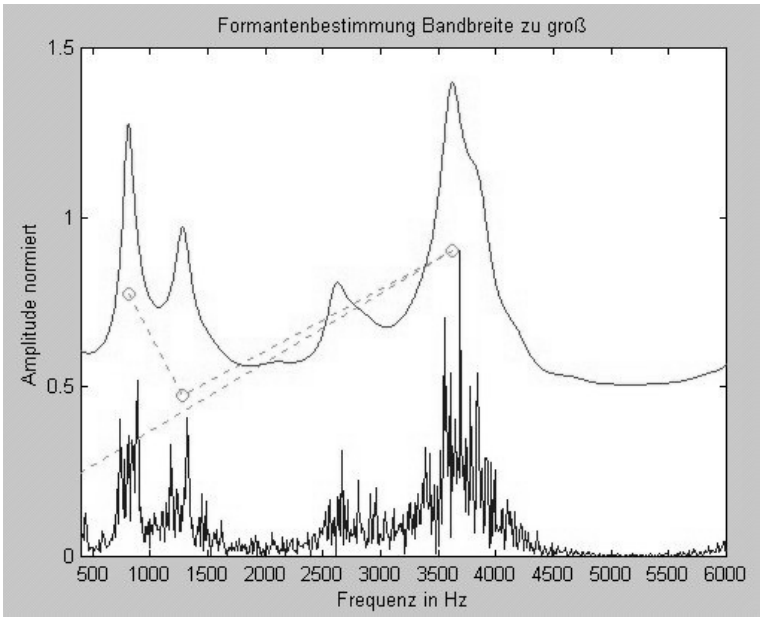

Abb. 41: Formantenbestimmung Bandbreite zu groß (Vokal /a/, Frau)

Zusammenfassend lässt sich sagen, dass eine automatische Bestimmung der Formanten nicht fehlerfrei zu realisieren ist. Neben den algorithmischen Problemen, wie der Wahl der Bandbreite und der LPC-Ordnung, gibt es auch Einschränkungen aufgrund der nicht immer eindeutigen Zuordnung. Tabelle 9 gibt die Häufigkeiten für die Fehlertypen 1-3 zusammen sowie die algorithmischen Fehler an. Auffallend hoch sind die Fehler bei Kindern. Dies liegt im wesentlichen an der großen Streuung der Sprachdaten, so dass sich die Formanten stark überschneiden. Weiterhin fällt auf, dass die beiden ersten Formanten häufig zusammenfallen und somit nicht eindeutig bestimmbar sind.

Die fehlerhaft klassifizierten Datensätze wurden visuell anhand der Überprüfung der Spektren ermittelt. Sie gingen in die Formantenanalyse nicht ein.

---

[68] Nicht erkannte Formanten werden vom Algorithmus 0 gesetzt, so dass in der Abbildung vom 3. Formanten die gestrichelte Linie auf die Ordinate führt.

Gruppe	Gruppen-größe	Fehler Typ I-III	Fehler Algorithmus	Kumulierte Fehler der Gruppe in %
Frauen	201	33	8	20.4 %
Männer	34	3	4	20.6%
Kinder	159	28	20	30.2 %

Tab. 9: Fehler bei der Formantenbestimmung beim Vokal /a/

Die Tabellen 10, 11 und 12 beschreiben die Mittelwerte der Frequenzen, Intensitäten und Bandbreiten der Formanten F_1 bis F_4 in Abhängigkeit der Personengruppe.

		Nonnasal				Nasal			
Gruppe	**Laut**	**F_1**	**F_2**	**F_3**	**F_4**	**F_1**	**F_2**	**F_3**	**F_4**
Frauen	**A**	978	1439	2855	4057	775	1390	3086	3996
	E	445	2577	3228	4165	401	2567	3148	4077
	I	397	2626	3582	4267	390	2687	3488	4270
	O	465	800	2750	3865	486	930	2653	3920
	U	434	604	2694	3968	403	621	2725	4096
	M	453	1447	2823	4256	424	1348	2753	4165
	N	1034	2244	3019	4234	1101	2276	2968	4217
Männer	**A**	751	1223	2779	3872	594	1199	3167	3732
	E	389	2231	2804	3604	334	2134	2798	3662
	I	342	2295	2998	3628	307	2120	2909	3563
	O	441	756	2542	3300	471	989	1988	3737
	U	375	727	2628	3625	291	916	2594	3468
	M	451	1222	2195	3900	428	1364	2499	3856
	N	1397	2267	3310	4082	1188	2333	3249	4018
Kinder	**A**	1001	1512	3147	4237	911	1497	3401	4420
	E	490	2564	3300	4309	439	2407	3267	4190
	I	394	2710	3571	4371	462	2651	3509	4401
	O	500	882	2978	3991	504	920	2880	4029
	U	418	821	2932	4202	469	811	2893	4178
	M	457	1469	2807	4241	459	1512	2689	4213
	N	1073	2280	3212	4254	1043	2300	3257	4259

Tab. 10: Formantfrequenzen in Hz

Gruppe	Laut	Nonnasal F_1	F_2	F_3	F_4	Nasal F_1	F_2	F_3	F_4
Frauen	**A**	41.7	40.6	38.5	37.5	37.1	38.4	34.8	36.1
	E	42.8	38.6	39.4	39.5	42.2	36.0	37.0	37.1
	I	38.2	36.0	39.0	39.0	39.2	32.7	33.9	34.3
	O	43.2	36.2	28.5	32.6	38.2	34.4	29.2	32.3
	U	42.8	38.1	25.6	30.0	43.5	34.4	29.5	29.5
	M	44.3	28.7	31.6	29.4	43.2	27.3	29.2	26.8
	N	27.9	29.3	30.2	28.7	27.4	27.0	29.3	26.8
Männer	**A**	39.7	40.6	38.5	38.4	35.9	38.3	33.8	35.6
	E	38.7	38.5	39.6	40.1	37.9	34.8	37.0	37.9
	I	39.2	35.4	38.7	39.7	39.2	30.8	35.5	36.9
	O	39.9	36.5	30.4	33.0	36.5	32.8	29.9	34.8
	U	39.3	32.8	28.4	30.2	38.8	30.2	29.7	31.0
	M	39.7	32.1	33.5	31.0	38.8	29.6	32.7	30.7
	N	30.5	33.7	30.8	28.8	28.9	30.6	29.8	28.2
Kinder	**A**	41.8	42.1	38.9	37.6	41.3	40.8	38.9	38.5
	E	43.5	39.6	40.7	39.7	40.7	35.6	39.0	39.2
	I	43.4	37.7	39.9	39.9	42.9	34.0	37.6	36.3
	O	43.7	39.1	31.4	34.4	41.9	38.1	31.2	32.9
	U	43.7	36.3	28.2	31.4	42.6	37.2	30.1	30.6
	M	43.5	30.6	31.8	31.0	42.6	30.2	31.6	30.4
	N	32.5	31.7	31.9	31.6	30.5	30.5	30.9	29.9

Tab. 11: Formantintensitäten in dB

		Nonnasal				Nasal			
Gruppe	**Laut**	F_1	F_2	F_3	F_4	F_1	F_2	F_3	F_4
Frauen	**A**	43.9	57.8	54.7	74.3	45.2	43.8	59.0	66.5
	E	13.7	51.0	75.5	66.1	15.5	48.3	64.7	68.3
	I	21.9	43.7	67.4	60.6	22.0	56.5	70.3	58.0
	O	16.2	29.2	59.8	40.1	33.5	39.6	66.4	54.4
	U	11.9	21.6	75.6	51.2	11.0	38.2	58.3	62.0
	M	10.2	59.6	36.7	70.7	11.0	60.4	59.4	90.4
	N	61.7	63.7	54.5	64.5	75.0	88.1	61.2	85.4
Männer	**A**	56.3	45.0	68.8	73.3	60.0	31.7	75.0	102.5
	E	34.6	64.0	82.1	77.5	42.0	47.0	76.0	90.0
	I	23.4	48.6	76.4	70.5	23.6	94.3	62.1	57.9
	O	32.1	36.2	71.8	51.8	42.5	60.8	82.5	63.3
	U	28.8	47.5	85.0	94.5	24.4	51.3	53.1	47.5
	M	25.0	31.7	60.0	51.7	29.4	65.6	48.8	98.1
	N	38.3	60.0	60.0	95.0	62.5	66.9	66.3	95.6
Kinder	**A**	47.6	40.2	59.4	80.2	41.2	51.3	65.4	79.2
	E	13.7	47.9	66.2	73.2	22.1	65.0	71.2	62.9
	I	13.4	40.4	55.3	54.8	19.7	73.2	59.7	80.5
	O	15.0	23.3	53.9	45.6	31.7	36.2	73.6	58.8
	U	12.9	26.1	63.8	53.4	17.0	25.8	56.8	79.8
	M	12.3	56.4	53.2	76.4	12.7	75.0	69.7	92.3
	N	41.5	55.5	60.1	60.5	61.7	69.7	79.3	90.0

Tab. 12: Formantbandbreiten in Hz

7.2.3.3 Antiformanten

Die Antiformanten sind definiert als spektrale Minima zwischen den Formanten. Daher setzt die Bestimmung der Antiformanten auf den Ergebnissen der Formantenbestimmung auf. Es werden 4 Antiformanten AF_i pro Sprachdatensatz bestimmt. Sei F_0 die Grundfrequenz, dann gilt für die Lage der Antiformanten:

$$F_{i-1} < AF_i < F_i \quad , \text{für} \quad i = 1,2,3,4 \tag{53}$$

Die hohen Antiformanten wurden über ein MA-Modell[69] mit niedriger Ordnung (p = 15) bestimmt, die niedrigen Antiformanten über ein MA-Modell mit hoher Ordnung (p = 60), wobei bei der Bestimmung nur innerhalb benachbarter Formanten gesucht wurde. Die Intensität und Bandbreite wurde analog zur Formantenbestimmung berechnet.

Die Tabellen 13, 14 und 15 beschreiben die Mittelwerte der Frequenzen, Intensitäten und Bandbreiten der Antiformanten AF_1 bis AF_4 in Abhängigkeit der Personengruppe.

		Nonnasal				**Nasal**			
Gruppe	**Laut**	**AF_1**	**AF_2**	**AF_3**	**AF_4**	**AF_1**	**AF_2**	**AF_3**	**AF_4**
Frauen	**A**	470	1188	2211	3394	477	1050	2488	3475
	E	353	1307	2818	3697	378	1572	2795	3602
	I	346	1286	2929	3932	359	1634	2983	3876
	O	351	655	1937	3146	379	729	1911	3131
	U	339	456	1898	3098	358	452	1689	3386
	M	376	972	2023	3503	348	1006	1953	3318
	N	757	1579	2578	3649	806	1675	2525	3456
Männer	**A**	374	970	2018	3322	435	818	2483	3415
	E	318	1124	2469	3179	280	1208	2362	3125
	I	257	1159	2533	3315	269	1195	2409	3142
	O	315	636	1734	2861	343	761	1412	2373
	U	283	596	1770	2962	265	718	1529	3056
	M	324	843	1540	3210	381	921	1754	3374
	N	972	1812	2990	3597	843	1688	2751	3636
Kinder	**A**	512	1228	2416	3693	498	1162	2521	3890
	E	393	1321	2840	3817	373	1436	2686	3680
	I	355	1362	2978	3955	367	1531	3022	3989
	O	388	694	2024	3375	384	758	1899	3346
	U	360	604	2007	3375	378	589	1863	3462
	M	369	911	1990	3531	371	956	2008	3549
	N	753	1605	2690	3695	735	1552	2789	3735

Tab. 13: Antiformantfrequenzen in Hz

[69] MA-Modell = Moving Average - Modell (vgl. Kapitel 3.2 „Transformationen“).

		Nonnasal				Nasal			
Gruppe	**Laut**	**AF_1**	**AF_2**	**AF_3**	**AF_4**	**AF_1**	**AF_2**	**AF_3**	**AF_4**
Frauen	**A**	0.6	0.2	1.1	0.8	0.8	0.7	2.9	1.0
	E	0.6	5.9	0.3	0.4	0.6	5.3	0.6	0.5
	I	0.7	10.4	0.7	0.3	0.5	5.7	0.9	0.7
	O	0.1	0.2	3.4	1.6	0.3	0.6	2.1	1.7
	U	0.1	0.3	3.9	2.3	0.2	0.5	2.4	1.9
	M	1.1	1.9	1.2	1.8	1.0	1.8	1.8	1.5
	N	1.3	1.9	0.7	1.7	1.4	2.3	1.1	1.3
Männer	**A**	0.6	0.3	1.4	0.8	0.4	0.4	2.0	0.8
	E	0.3	5.4	0.3	0.4	0.2	3.3	1.3	0.4
	I	0.4	9.7	0.7	0.3	0.1	3.2	1.2	0.3
	O	0.1	0.1	3.3	1.1	0.2	0.5	1.5	1.8
	U	0.1	0.3	3.8	1.3	0.0	0.7	2.0	1.0
	M	1.2	1.4	0.9	1.3	1.4	1.6	1.3	1.3
	N	1.2	1.1	0.9	1.0	1.3	1.2	1.1	1.2
Kinder	**A**	1.1	0.3	1.3	0.7	1.0	0.5	1.5	0.5
	E	0.8	4.4	0.4	0.4	0.5	3.4	0.8	0.5
	I	0.6	8.1	0.7	0.3	0.6	4.5	1.1	0.7
	O	0.2	0.3	3.4	1.1	0.2	0.3	2.5	1.4
	U	0.2	0.6	3.7	1.8	0.2	0.6	3.5	1.6
	M	1.3	1.9	1.3	1.3	1.2	2.0	1.4	1.3
	N	1.3	1.6	0.9	1.3	1.4	2.0	1.1	1.3

Tab. 14: Antiformantintensitäten in dB

		Nonnasal				Nasal			
Gruppe	**Laut**	**AF_1**	**AF_2**	**AF_3**	**AF_4**	**AF_1**	**AF_2**	**AF_3**	**AF_4**
Frauen	**A**	107.5	96.8	170.3	149.7	89.5	104.3	142.0	125.3
	E	55.1	158.5	115.3	135.3	56.9	146.3	99.4	152.5
	I	57.4	157.4	135.4	117.3	60.5	151.3	116.1	121.1
	O	58.3	70.2	147.3	139.3	70.4	88.8	139.3	138.8
	U	52.3	66.0	142.3	149.4	55.0	84.2	130.0	136.7
	M	112.4	128.1	123.1	133.3	113.5	124.2	152.5	147.9
	N	106.2	127.1	111.6	148.6	107.1	127.9	118.8	141.3
Männer	**A**	141.3	110.0	207.3	145.3	106.0	111.0	158.0	132.0
	E	66.3	193.8	121.9	138.8	65.0	200.0	145.0	195.0
	I	52.5	147.5	142.5	95.0	45.0	115.0	130.0	90.0
	O	79.4	71.3	175.6	119.4	78.3	120.8	135.0	145.0
	U	53.3	76.7	191.7	110.0	45.0	95.0	150.0	160.0
	M	123.0	145.0	161.7	126.7	118.6	161.9	136.3	134.4
	N	131.7	125.0	120.0	135.0	137.5	147.5	150.0	135.6
Kinder	**A**	111.6	94.9	156.2	150.0	109.8	104.0	159.4	140.2
	E	61.4	144.1	116.7	130.4	63.1	159.2	114.6	128.8
	I	59.2	155.8	127.9	130.0	61.4	166.4	147.1	131.4
	O	64.4	73.4	154.3	127.7	71.4	81.8	154.3	128.9
	U	56.0	75.2	134.0	146.8	59.0	86.0	154.5	157.0
	M	119.4	125.1	126.8	133.8	122.5	125.7	123.0	158.3
	N	103.1	125.3	125.8	124.8	100.0	128.0	128.3	136.0

Tab. 15: Antiformantbandbreiten in Hz

7.3 Diskussion

Nachdem die wesentlichen Sprachmodellparameter generiert sind, lassen sich die Laute anhand von 27 Werten darstellen. Und zwar jeweils die Lage, Intensität und Bandbreite der Grundfrequenz sowie der 4 Formanten und 4 Antiformanten. Beispielhaft ist dies für die Mittelwerte des Vokals /a/ in Abbildung 42 skizziert. Dabei sind die Kurven zwischen den Formanten und Antiformanten interpoliert. Schaut man sich die Abbildung an, stellt sich die Frage nach den über Laut- und Gruppengrenzen hinweg nasalitätsrelevanten Parameter.

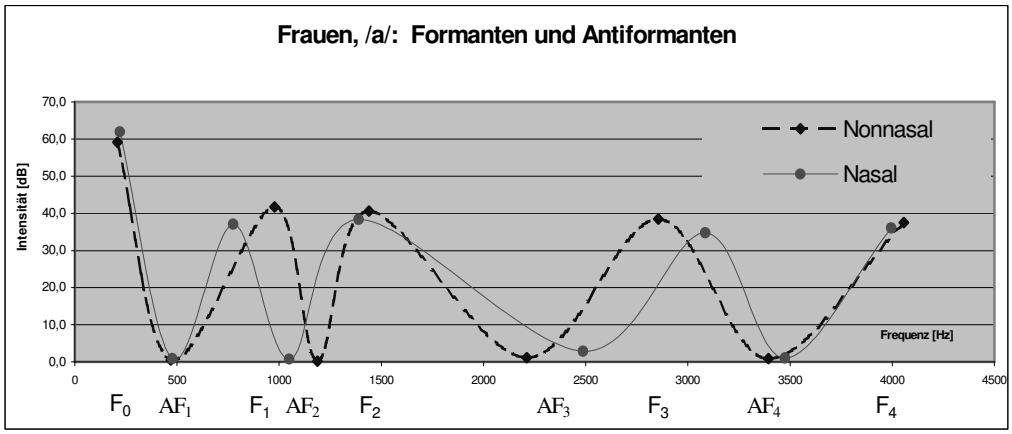

Abb. 42: Frauen, /a/: Skizze der nonnasalen und nasalen Waveplots anhand der Mittelwerte von Frequenzen und Intensitäten der Formanten und Antiformanten

Es sollen nun die generierten Sprachmodellparameter bezüglich ihres möglichen Einflusses auf die Nasalität analysiert werden. In Tabelle 16 bzw. Tabelle 17 ist die Differenz der Mittelwerte bei der Grundfrequenz und den Formanten bzw. den Antiformanten berechnet. In den letzten Zeilen ist die Anzahl der positiven und negativen Differenzen aufsummiert. An ihnen lassen sich die laut- und gruppenunabhängigen Parameter identifizieren. Es wird betont, dass über die Güte der Trennfähigkeit der in diesem Abschnitt herausgearbeiteten Parameter keine Aussage gemacht wird.

Betrachten wir zunächst das offene Näseln. Dann lassen sich folgende differenzierbare laut- und gruppenunabhängige Aussagen über die Formanten und Antiformaten tätigen. Die Aussagen beziehen sich dabei auf die Einstufung der nasalen Parameter im Vergleich zu den nonnasalen Parameter.

Offenes Näseln - Formanten

- Frequenz: Die Lage der Grundfrequenz ist erhöht,
  die Lage der Formanten F_1, F_2 und F_3 ist erniedrigt.
- Intensität: Die Intensität der Grundfrequenz ist erhöht,
  die Intensität aller 4 Formanten F_1, F_2, F_3 und F_4 erniedrigt.
- Bandbreite: die Bandbreite ist bei allen 4 Formanten F_1, F_2, F_3 und F_4 erhöht

Offenes Näseln - Antiformanten

- Frequenz: Die Lage der Antiformanten AF_1 und AF_2 ist erhöht,
  die Lage des Antiformanten AF_3 erniedrigt.

- Intensität: Die Intensität der Antiformanten AF_2, AF_3, und AF_4 ist erhöht, die Intensität des Antiformanten AF_1 erniedrigt.
- Bandbreite: Die Bandbreite von AF_2 ist erhöht, die Bandbreite von AF_3 erniedrigt.

Zusammenfassend lässt sich für das offene Näseln aussagen, dass die Intensität aller 4 Formanten abnimmt. Dies geht einher mit einer gleichzeitigen Zunahme der Bandbreiten der Formanten. Im Vergleich zu den nonnasalen Daten, werden somit bei vorhandener offener Nasalität die Formanten nicht so klar gebildet. Im geglätteten Spektrum äußert sich dies in einem energieärmeren, breiteren Formanten. Ein Teil der Energie scheint in die Antiformanten überzugehen. Dahingegen nimmt sowohl die Lage der Grundfrequenz als auch ihre Intensität zu. Eine mögliche Erklärung hierfür wäre der Versuch einer Kompensation durch näselnde Patienten. So gelangt insbesondere bei LKG-Spaltpatienten ein Teil der Energie über die Spalte in den Nasenraum. Evtl. versuchen näselnde Personen diesen Nachteil durch eine höhere Energie bei der Stimmbildung zu kompensieren. Die daraus resultierende angestrengtere Sprechweise würde auch die Erhöhung der Lage der Grundfrequenz erklären.

Gruppe	Nasalitäts-art	Laut	F_0			F_1			F_2			F_3			F_4		
			Freq.	Int.	Band	Freq.	Int.	Band	Freq.	Int.	Band	Freq.	Int.	Band	Freq.	Int.	Band
Frauen	**offen**	**A**	12	2.8	-4.6	-203	-4.6	1.3	-49	-2.2	-14	231	-3.7	4.3	-61	-1.4	-7.8
		E	3	1.8	-4.7	-44	-0.6	1.8	-10	-2.6	-2.7	-80	-2.4	-10.8	-88	-2.4	2.2
		I	16	1.2	-0.3	-7	1	0.1	61	-3.3	12.8	-94	-5.1	2.9	3	-4.7	-2.6
		O	7	3.1	-7	21	-5	17.3	130	-1.8	10.4	-97	0.7	6.6	55	-0.3	14.3
		U	3	0.9	-1.4	-31	0.7	-0.9	17	-3.7	16.6	31	3.9	-17.3	128	-0.5	10.8
	geschl.	**M**	-21	0	0.8	-29	-1.1	0.8	-99	-1.4	0.8	-70	-2.4	22.7	-91	-2.6	19.7
		N	-7	0	-0.4	67	-0.5	13.3	32	-2.3	24.4	-51	-0.9	6.7	-17	-1.9	20.9
Männer	**offen**	**A**	3	1.4	-4.6	-157	-3.8	3.7	-24	-2.3	-	388	-4.7	6.2	-140	-2.8	29.2
		E	-1	1.1	10.2	-55	-0.8	7.4	-97	-3.7	-17	-6	-2.6	-6.1	58	-2.2	12.5
		I	23	2.8	11.6	-35	0	0.2	-175	-4.6	45.7	-89	-3.2	-14.3	-65	-2.8	-12.6
		O	9	2.4	16.2	30	-3.4	10.4	233	-3.7	24.6	-554	-0.5	10.7	437	1.8	11.5
		U	2	1.6	10.9	-84	-0.5	-4.4	189	-2.6	3.8	-34	1.3	-31.9	-157	0.8	-47
	geschl.	**M**	-10	0	-0.2	-23	-0.9	4.4	142	-2.5	33.9	304	-0.8	-11.2	-44	-0.3	46.4
		N	-7	0	-1.4	-209	-1.6	24.2	66	-3.1	6.9	-61	-1	6.3	-64	-0.6	0.6
Kinder	**offen**	**A**	2	0.7	1.5	-90	-0.5	-6.4	-15	-1.3	11.1	254	0	6	183	0.9	-1
		E	-42	0.9	3.3	-51	-2.8	8.4	-157	-4	17.1	-33	-1.7	5	-119	-0.5	-10.3
		I	-2	-0.3	1.6	68	-0.5	6.3	-59	-3.7	32.8	-62	-2.3	4.4	30	-3.6	25.7
		O	-29	1	-2.2	4	-1.8	16.7	38	-1	12.9	-98	-0.2	19.7	38	-1.5	13.2
		U	-14	-0.3	-0.5	51	-1.1	4.1	-10	0.9	-0.3	-39	1.9	-7	-24	-0.8	26.4
	geschl.	**M**	-22	0	-0.1	2	-0.9	0.4	43	-0.4	18.6	-118	-0.2	16.5	-28	-0.6	15.9
		N	-21	0	0.1	-30	-2	20.2	20	-1.2	14.2	45	-1	19.2	5	-1.7	29.5
	offen	**+**	10	13	7	5	2	12	6	1	10	4	4	9	8	3	9
		-	5	2	8	10	12	3	9	14	5	11	10	6	7	12	6
	geschl.	**+**	0	0	2	2	0	6	5	0	6	2	0	5	1	0	6
		-	6	0	4	4	6	0	1	6	0	4	6	1	5	6	0

Tab. 16: Differenz der Mittelwerte bei der Grundfrequenz und den Formanten in Hz

Gruppe	Nasalitäts-art	Laut	AF_1 Freq.	AF_1 Int.	AF_1 Band	AF_2 Freq.	AF_2 Int.	AF_2 Band	AF_3 Freq.	AF_3 Int.	AF_3 Band	AF_4 Freq.	AF_4 Int.	AF_4 Band
Frauen	**offen**	**A**	7	0.2	-18	-138	0.5	7.5	277	1.8	-28.2	81	0.2	-24.3
		E	25	0	1.8	265	-0.6	-12.2	-23	0.3	-15.9	-95	0.1	17.2
		I	13	-0.2	3.1	348	-4.7	-6.1	54	0.2	-19.3	-56	0.4	3.8
		O	28	0.2	12.1	74	0.4	18.6	-26	-1.3	-8	-15	0.1	-0.5
		U	19	0.1	2.7	-4	0.2	18.2	-209	-1.5	-12.3	288	-0.4	-12.7
	geschl.	**M**	-28	-0.1	1.1	34	-0.1	-3.9	-70	0.6	29.4	-185	-0.3	14.6
		N	49	0.1	0.9	96	0.4	0.8	-53	0.4	7.2	-193	-0.4	-7.3
Männer	**offen**	**A**	61	-0.2	-35.3	-152	0.1	1	465	0.6	-49.3	93	0	-13.3
		E	-38	-0.1	-1.3	84	-2.1	6.2	-107	1	23.1	-54	0	56.2
		I	12	-0.3	-7.5	36	-6.5	-32.5	-124	0.5	-12.5	-173	0	-5
		O	28	0.1	-1.1	125	0.4	49.5	-322	-1.8	-40.6	-488	0.7	25.6
		U	-18	-0.1	-8.3	122	0.4	18.3	-241	-1.8	-41.7	94	-0.3	50
	geschl.	**M**	57	0.2	-4.4	78	0.2	16.9	214	0.4	-25.4	164	0	7.7
		N	-129	0.1	5.8	-124	0.1	22.5	-239	0.2	30	39	0.2	0.6
Kinder	**offen**	**A**	-14	-0.1	-1.8	-66	0.2	9.1	105	0.2	3.2	197	-0.2	-9.8
		E	-20	-0.3	1.7	115	-1	15.1	-154	0.4	-2.1	-137	0.1	-1.6
		I	12	0	2.2	169	-3.6	10.6	44	0.4	19.2	34	0.4	1.4
		O	-4	0	7	64	0	8.4	-125	-0.9	0	-29	0.3	1.2
		U	18	0	3	-15	0	10.8	-144	-0.2	20.5	87	-0.2	10.2
	geschl.	**M**	2	-0.1	3.1	45	0.1	0.6	18	0.1	-3.8	18	0	24.5
		N	-18	0.1	-3.1	-53	0.4	2.7	99	0.2	2.5	40	0	11.2
	offen	**+**	10	4	8	10	7	12	5	9	4	7	8	8
		-	5	7	7	5	6	3	10	6	10	8	4	7
	geschl.	**+**	3	4	4	4	5	5	3	6	4	4	1	5
		-	3	2	2	2	1	1	3	0	2	2	2	1

Tab. 17: Differenz der Mittelwerte bei den Antiformanten in Hz

Betrachten wir jetzt das geschlossene Näseln, dann lassen sich folgende Aussagen tätigen:

Geschlossenes Näseln - Formanten

- Frequenz: Die Lage des Formanten F_2 ist erhöht,
  die Lage der Grundfrequenz und der Formanten F_1, F_3, F_4 ist erniedrigt.
- Intensität: Die Intensität aller 4 Formanten F_1, F_2, F_3, F_4 ist erniedrigt.
- Bandbreite: Die Bandbreite aller 4 Formanten F_1, F_2, F_3, F_4 ist erhöht.

Geschlossenes Näseln - Antiformanten

- Frequenz: Die Lage der Antiformanten AF_2, AF_4, ist erhöht.
- Intensität: Die Intensität der Antiformanten AF_1, AF_2, AF_3 ist erhöht.
- Bandbreite: Die Bandbreite aller 4 Antiformanten AF_1, AF_2, AF_3, AF_4 ist erhöht.

Wie beim offenen Näseln ist auch beim geschlossenen Näseln die Intensität bei allen 4 Formanten erniedrigt. Dies geht ebenfalls mit einer Erhöhung der Bandbreiten der Formanten und Antiformaten einher. Auch hier ist die Zunahme der Intensität der Antiformanten zu beobachten. Diese Aussage lässt sich somit als zentrale Aussage zur Auswirkung der Nasalität auf stimmhafte Laute festhalten.

Es wird aber betont, dass die oben ausgearbeiteten Aussagen nicht universell für alle Laute gelten. Die Tabellen 18 und 19 geben eine Zusammenfassung der Parameter pro Laut wieder auf die sich die Nasalität stark auswirkt. Die Sprechergruppen sind dabei nicht berücksichtigt.

Um optimale Ergebnisse bei der Nasalitätsmessung zu erlangen, müssen natürlich für jede Sprechergruppe und jeden Laut die am besten geeigneten Parameter ermittelt werden. Weiterhin erfolgten bisher noch keine Aussagen über die Güte der gewonnen Parameter zur Messung der Nasalität. Diese Aspekte werden im nächsten Kapitel behandelt.

Laut	**Formanten - Gemeinsamkeiten**
A	Frequenz von F_0 und F_3 erhöht, Frequenz von F_1 und F_2 erniedrigt Intensität von F_0 erhöht, von F_1 und F_2 erniedrigt Bandbreite von F_3 erhöht
E	Frequenz von F_1, F_2 und F_3 erniedrigt Intensität von Grundfrequenz erhöht und von allen 4 Formanten erniedrigt Bandbreite von F_1 erhöht
I	Frequenz von F_3 erniedrigt Intensität von F_2, F_3 und F_4 erniedrigt F_1 Band und F_2 Band erhöht
O	Frequenz von F_1, F_2 und F_4 erhöht von F_3 erniedrigt Intensität von F_0 erhöht, von F_1 und F_2 erniedrigt Bandbreite von F_1, F_2, F_3 und F_4 erhöht
U	Intensität F_3 erhöht Bandbreite F_3 erniedrigt
M	Frequenz von F_0 und F_4 erniedrigt Intensität von F_1, F_2,F_3 und F_4 erniedrigt Bandbreite F_1, F_2 und F_4 erhöht
N	Frequenz F_0 erniedrigt, F_2 erhöht Intensität F_1, F_2, F_3 und F_4 erniedrigt Bandbreite F_1, F_2, F_3 und F_4 erhöht

Tab. 18: Formanten – Gruppenunabhängige Gemeinsamkeiten der Laute

Laut	Antiformanten - Gemeinsamkeiten
A	Frequenzen AF_2 erniedrigt, AF_3 und AF_4 erhöht Intensität AF_2 und AF_3 erhöht Bandbreiten AF_1 und AF_4 erniedrigt, AF_2 und AF_3 erhöht
E	Frequenz AF_2 erhöht, AF_3 und AF_4 erniedrigt Intensität AF_3 und AF_4 erhöht, AF_2 erniedrigt
I	Frequenzen AF_1 und AF_2 erhöht Intensitäten AF_2 erniedrigt, AF_3 und AF_4 erhöht
O	Frequenzen AF_2 erhöht, AF_3 und AF_4 erniedrigt Intensitäten AF_1, AF_2 und AF_4 erhöht, AF_3 erniedrigt Bandbreite AF_2 erhöht
U	Frequenzen AF_3 erniedrigt, AF_4 erhöht Intensitäten AF_2 erhöht, AF_3 und AF_4 erniedrigt Bandbreite AF_2 erhöht
M	Frequenzen AF_2 erhöht Intensitäten AF_3 erniedrigt Bandbreiten AF_4 erhöht
N	Intensitäten AF_1, AF_2 und AF_3 erhöht Bandbreiten AF_2 und AF_3 erhöht

Tab. 19: Antiformanten – Gruppenunabhängige Gemeinsamkeiten der Laute

8 Klassifikationsergebnisse

In diesem Kapitel wird die Eignung der extrahierten Parameter zur Klassifikation der Nasalität diskutiert. Konkret handelt es sich um die Sprachsignalparameter (Grundfrequenz, Formanten, Antiformanten) und die Frequenzbandparameter. Hauptziel der Untersuchungen war die Ausarbeitung der relevanten, d. h. die Nasalität beeinflussenden Parameter. Über eine hohe Klassifizierungsgüte lassen sich dann Aussagen über die Ursachen der Nasalität treffen. Sind die Faktoren identifiziert, lassen sich die Algorithmen optimieren, indem man sich nur auf diese Parameter konzentriert.

Bei der Nasalität wird zwischen offenem und geschlossenen Näseln differenziert. Das offene Näseln wird an den Vokalen, das geschlossene Näseln an den Nasalen /n/ und /m/ untersucht. Um auch Alters- und Geschlechtseffekte bewerten zu können, wird der Datenbestand in 3 Gruppen eingeteilt, in *FRAU* (Frauen ab 16 Jahren), *MANN* (Männer ab 16 Jahren) und *KIND* (sowohl Jungen, als auch Mädchen bis 16 Jahren).

Es wurden 3 Experimentreihen durchgeführt.

- Klassifikationsaufgabe „*N01*“: Zuordnung eines Sprachdatensatzes zu einer Klasse „0 - nonnasal“ oder „1 - nasal“. Bei den nasalen Daten bleibt die Ausprägung unberücksichtigt.
- Klassifikationsaufgabe „*N123*“: Zuordnung eines bereits als nasal klassifizierten Sprachdatensatzes zu einer Klasse „1 – leicht“, „2 – mittelgradig“ oder „3 – ausgeprägt“.
- Klassifikationsaufgabe „*N0123*“: Zuordnung eines Sprachdatensatzes zu einer Klasse „0 - nonnasal“, „1 – leicht“, „2 – mittelgradig“ oder „3 – ausgeprägt“.

Die erste Versuchsreihe sollte klären, inwieweit sich nasale Laute von nonnasalen algorithmisch unterscheiden lassen. Hierfür wurden alle nasalen Laute unabhängig von ihrer Ausprägung in einer Klasse „nasal“ zusammengefasst. Die Klassifikationsaufgabe bestand somit in der Feststellung, ob ein Laut nasal gesprochen war oder nicht, d. h. in der Zuordnung zu einer Klasse „nasal“ oder „nonnasal“. In der zweiten Versuchsreihe wurden alle von den Logopäden vorklassifizierten nasalen Datensätze hinsichtlich ihrer Ausprägungen „leichtes Näseln“, „mittelgradiges Näseln“ und „stark ausgeprägtes Näseln“ klassifiziert. Die dritte Versuchsreihe erweiterte die ersten beiden Klassifikationsaufgaben um die Betrachtung aller Ausprägungen.

Hatte die erste Versuchsreihe das Ziel, die nasalitätsrelevanten Parameter zu finden, sollte die zweite Versuchsreihe klären, inwieweit sich die Nasalität quantifizieren lässt. Aus Anwendungssicht dienen die Ergebnisse der 1. Reihe dazu, die Nasalität bei einer Person zu

diagnostizieren. Ist eine Nasalität bereits diagnostiziert (z. B. auch über einen Logopäden), bräuchte man für die Quantifizierung die Klasse „nonnasal“ nicht mehr berücksichtigen. Die dritte Versuchsreihe sollte klären, inwieweit sich die Nasalität bewerten lässt, wenn der Nasalitätsstatus des Datensatzes unbekannt ist. Insbesondere sollte die Frage geklärt werden, ob bei unbekanntem Nasalitätsstatus ein zweistufiges Vorgehen, erst die Nasalität feststellen und danach quantifizieren, gegenüber einer Klassifizierung über alle Ausprägungen vorteilhafter ist.

8.1 Eignung der Sprachdatenbank NASAL

Überprüfen wir die Sprachdatenbank NASAL bezüglich der beiden Forderungen an die Gleichverteilung der Klassen sowie der Stichprobenmindestanzahl und der damit verbundenen Eignung der Klassifikationsaufgaben.

Wie man in Tabelle 5 „Verteilung der Sprachdaten je Sprachlaut, Sprechergruppe und Nasalitätsausprägung“ sehen kann, sind von den ca. 200 Datensätzen pro Vokal bei *FRAU* 59% nonnasal und 41% nasal[70]. Von den 59 Nasallauten sind 51% nonnasal und 49% nasal. Somit sind obige Anforderungen für die Klassifikationsaufgabe „*N01*“ bei *FRAU* weitgehend erfüllt. Schaut man sich die nasalen Laute separat an, verteilen sich die ca. 80 Vokale bzw. 29 Nasale mit jeweils ca. 19%, 56% und 25% bzw. 21%, 31% und 48% auf die Nasalitätsgrade 1,2 und 3. Wir haben hier sowohl bei den Vokalen als auch bei den Nasalen eine dominante Klasse, so dass das Klassifikationsergebnis durch die korrekte Klassifikation dieser Gruppen beeinflusst wird. Dieser Umstand findet sich auch bei der Klassifikationsaufgabe „*N0123*“, in der die nonnasalen Daten ca. 60% bei den Vokalen und 51% bei den Nasalen ausmachen.

Betrachten wir die Sprechergruppe *MANN*. Dort sind von den ca. 33 Datensätzen bei den Vokalen 76% nonnasal und 24% nasal. Die nasalen Sprachdaten besitzen zudem alle die Ausprägung 2. Somit reduzieren sich die Klassifikationsaufgaben bei den Vokalen auf die Klassifikationsaufgabe „*N01*“. Bei den 11 Nasallauten sind 27% nonnasal und 73% nasal. Es existieren zu allen 3 Nasalitätsausprägungen Datensätze, so dass alle Klassifikationsaufgaben durchgeführt werden können. Aufgrund der kleinen Stichprobengröße sind diese jedoch nur mit Vorbehalt interpretierbar.

Von der Sprechergruppe *KIND* existieren je Vokal ca. 200 Datensätze, davon sind 85% nonnasal und 15% nasal. Von den Nasalen existieren 74 Datensätze, davon sind 80%

[70] Die Prozentangaben werden der besseren Leserlichkeit wegen im Text auf Ganzzahlen gerundet angegeben. Die genauen Angaben können den Tabellen entnommen bzw. aus dieser berechnet werden.

nonnasal und 20% nasal. Um hier „gleich große“ Klassen zu erhalten, wurde in der Klassifikationsaufgabe „*N01*“ die Anzahl der Datensätze aus der nonnasalen Gruppe auf die Größe der nasalen Gruppe beschränkt. In der Klassifikationsaufgabe „*N0123*“ wurde die Größe der nonnasalen Gruppe gleich der Größe der größten nasalen Gruppe gewählt.

Zusammenfassend lässt sich sagen, dass die Aussagen bei den Sprechergruppen *FRAU* und *KIND* als signifikant und repräsentativ angesehen werden können. Bei *MANN* hingegen sind aufgrund der geringen Sprachdatenanzahl nur die Ergebnisse bzgl. des offenen Näselns aussagefähig. Um hier die Aussagen verbessern zu können sowie auch Aussagen für das geschlossene Näseln tätigen zu können, wird die Sprachdatenbank gegenwärtig weiter ausgebaut.

8.2 Formantenanalyse

Unter (Anti-) Formantenanalyse wird in dieser Arbeit die Untersuchung des Einflusses der Nasalität auf die (Anti-) Formanten verstanden. Einheitliche Aussagen diesbezüglich existieren in der Literatur nicht.

Bei den Klassifizierungen wurden fünf Untersuchungsreihen mit unterschiedlichen Parametergruppen durchgeführt, welche wie folgt bezeichnet werden:

- *GF* : Lage, Intensität und Bandbreite der Grundfrequenz
- *FREQ* : Lage der Formanten F_1 bis F_4 bzw. Antiformanten AF_1 bis AF_4
- *INT* : Energie der Formanten F_1 bis F_4 bzw. Antiformanten AF_1 bis AF_4
- *BAND* : Bandbreite der Formanten F_1 bis F_4 bzw. Antiformanten AF_1 bis AF_4
- *ALLE* : beinhaltet die obigen 4 Parametergruppen sowie die Gesamtintensität der Aufnahme

Dabei gilt die Namenskonvention, dass der Parameter durch einen Unterstrich '_' und ein Kürzel F, I oder B, stellvertretend für die Frequenz, Intensität oder Bandbreite, ergänzt wird. Ist von der Bandbreite des 1. Antiformanten die Rede, wird dieser durch AF_1_B bezeichnet. Auf die Gesamtintensität der Aufnahme wird durch GES_I verwiesen.

Die erste Versuchsreihe sollte überprüfen, ob die Grundfrequenz tatsächlich eine so dominante Stellung unter den nasalitätsrelevanten Parametern besitzt. Die nächsten 3 Reihen sollten klären, ob sich die Nasalität evtl. nur in Frequenzverschiebungen bzw. in der Abnahme/Zunahme der Intensität und/oder Bandbreiten ausreichend aufzeigen lässt. In der letzten Versuchsreihe wurden alle 15 Parameter betrachtet. Über die schrittweise lineare Diskriminanzanalyse wurden von allen möglichen Parametern die zur Klassifizierung geeig-

neten ausgewählt. Wie bereits erwähnt, begeht die schrittweise Diskriminanzanalyse bei jedem Schritt statistische Fehler erster und zweiter Art, die nicht korrigiert werden. Die Ergebnisse sind dadurch ein wenig mit Unsicherheit belastet. Es kann sogar vorkommen, dass eine Klassifikation mit einer Untermenge von Parametern ein besseres Ergebnis erbringt, als die Obermenge[71]. In der Regel fielen jedoch die Klassifizierungsgüten mit der Parametergruppe *ALLE* am besten aus. Die bei dieser Versuchsreihe extrahierten Parameter werden in zwei Tabellen, je eine für die Formantenanalyse und eine für die Antiformantenanalyse, wiedergegeben.

8.2.1 Klassifizierungsgüte

Tabelle 20 gibt den Gesamtprozentsatz der richtigen Klassifizierungen der Nasalität pro Klassifikationsaufgabe für jede Sprecher- und Parametergruppe mittels der LDA wieder. Anhand dieser Tabelle sollen für jede Klassifikationsaufgabe, Nasalitätsart und Sprechergruppe folgende Punkte beantwortet werden:

- Wie hoch ist die durchschnittliche Erkennungsrate über alle Laute bei der Parametergruppe *ALLE*?

  Diese Parametergruppe interessiert besonders, weil sie fast immer die besten Klassifizierungsergebnisse liefert. Da beim Durchschnitt auch nicht erfolgreich vorgenommene Klassifizierungen mitberücksichtigt werden, gibt dieser Wert weiterhin Aufschluss über die Robustheit der Parameter beim Einsatz in der entsprechenden Klassifikationsaufgabe.

- Wie hoch und bei welchem bzw. welchen Laut(en) liegt die maximale Erkennungsrate?

  Hier geht es um das Ausarbeiten der zur Nasalitätsmessung besonders geeigneten Laute. Insbesondere lassen sich Algorithmen durch eine stärkere Gewichtung dieser Laute optimieren.

- Welche Parametergruppe, von *ALLE* abgesehen, liefert die besten Ergebnisse?

  Anhand dieses Punktes soll überprüft werden, ob sich die Nasalität besonders an der Grundfrequenz bzw. jeweils an den Lagen, Intensitäten oder Bandbreiten der Formanten messen lässt. Detaillierte Aussagen über die konkreten Parameter können dann den Parametertabellen entnommen werden.

---

[71] So ist die Klassifizierungsgüte von 67.8% in Tabelle 20 für *KIND* in der Klassifikationsaufgabe *N01* bei *BAND* geringfügig höher als die Klassifikationsgüte von 67.1% für ALLE.

Nasal	Gruppe	Parameter	A	E	I	O	U	M	N	Ø offen	Ø gschl.
N01	**Frauen**	*GF*	83.4%	74.4%	57.2%	87.4%	60.4%	75.5%	-	72.6%	37.8%
		FREQ	83.4%	89.0%	64.8%	71.3%	61.5%	73.6%	-	74.0%	36.8%
		INT	76.9%	66.5%	73.8%	71.9%	80.2%	71.7%	-	73.9%	35.9%
		BAND	56.8%	52.4%	65.5%	77.8%	70.3%	64.2%	62.5%	64.6%	63.4%
		ALLE	92.3%	89.0%	76.6%	89.2%	81.8%	73.6%	62.5%	85.8%	68.1%
	Männer	*GF*	76.9%	93.1%	86.2%	87.0%	71.4%	-	-	82.9%	-
		FREQ	88.5%	65.5%	62.1%	100.0%	85.7%	-	-	80.4%	-
		INT	92.3%	79.3%	69.0%	73.9%	64.3%	-	-	75.8%	-
		BAND	69.2%	-	82.8%	65.2%	60.7%	-	81.8%	55.6%	40.9%
		ALLE	100.0%	93.1%	89.7%	100.0%	89.3%	-	81.8%	94.4%	40.9%
	Kinder	*GF*	-	69.6%	-	65.2%	-	-	70.0%	27.0%	35.0%
		FREQ	67.1%	65.8%	79.5%	-	71.4%	-	-	56.8%	-
		INT	67.1%	75.3%	-	-	65.1%	-	-	41.5%	-
		BAND	67.8%	80.4%	75.6%	80.7%	72.4%	70.3%	71.4%	75.4%	70.9%
		ALLE	67.1%	82.9%	76.9%	85.1%	70.3%	70.3%	70.0%	76.5%	70.2%
N123	**Frauen**	*GF*	41.0%	-	53.6%	64.4%	-	-	53.8%	31.8%	26.9%
		FREQ	50.8%	26.2%	41.1%	-	-	-	-	23.6%	-
		INT	-	63.9%	57.1%	62.7%	52.1%	-	50.0%	47.2%	25.0%
		BAND	-	-	-	55.9%	-	-	26.9%	11.2%	13.5%
		ALLE	47.5%	59.0%	57.1%	64.4%	52.1%	-	61.5%	56.0%	30.8%
	Männer	*GF*	-	-	-	-	-	-	-	-	-
		FREQ	-	-	-	-	-	-	-	-	-
		INT	-	-	-	-	-	-	-	-	-
		BAND	-	-	-	-	-	-	-	-	-
		ALLE	-	-	-	-	-	-	-	-	-
	Kinder	*GF*	-	-	-	-	-	-	-	-	-
		FREQ	61.5%	90.5%	84.2%	81.0%	-	-	-	63.4%	-
		INT	-	-	-	-	-	-	-	-	-
		BAND	-	-	-	-	-	-	-	-	-
		ALLE	61.5%	90.5%	84.2%	81.0%	-	-	-	63.4%	-
N0123	**Frauen**	*GF*	60.9%	54.9%	37.2%	64.7%	-	56.6%	25.0%	43.5%	40.8%
		FREQ	65.7%	67.1%	44.8%	53.9%	-	52.8%	-	46.3%	26.4%
		INT	50.3%	37.2%	46.9%	56.3%	63.0%	45.3%	32.1%	50.7%	38.7%
		BAND	-	-	-	66.5%	54.7%	47.2%	-	24.2%	23.6%
		ALLE	65.1%	71.3%	53.1%	67.7%	62.5%	45.3%	32.1%	63.9%	38.7%
	Männer	*GF*	-	-	-	-	-	-	-	-	-
		FREQ	-	-	-	-	-	54.5%	-	-	27.3%
		INT	-	-	-	-	-	-	-	-	-
		BAND	-	-	-	-	-	-	-	-	-
		ALLE	-	-	-	-	-	54.5%	-	-	27.3%
	Kinder	*GF*	-	63.9%	-	-	-	-	-	12.8%	-
		FREQ	59.2%	59.5%	70.5%	41.6%	66.7%	-	-	59.5%	-
		INT	37.5%	69.0%	-	-	29.2%	-	-	27.1%	-
		BAND	-	75.3%	72.4%	82.0%	-	-	61.4%	45.9%	30.7%
		ALLE	60.5%	78.5%	73.7%	80.1%	64.6%	-	61.4%	71.5%	30.7%

Tab. 20: Klassifikation Formantenanalyse: Erkennungsraten in %

Konnte eine Klassifizierung in Tabelle 20 nicht durchgeführt werden, ist diese durch einen Bindestrich gekennzeichnet. Gründe für eine fehlende Klassifizierung sind einerseits eine zu starke Ähnlichkeit der Nasalitätsgruppen, so dass die LDA keine signifikante(n) Diskriminanzfunktion(en) erzeugen kann, andererseits zu wenig unterschiedliche Nasalitätsgruppen. Beispielsweise besitzen die Vokale bei *MANN* alle die gleiche Nasalitätsausprägung, so dass die Klassifikationsaufgabe *N123* bei Ihnen nicht durchgeführt werden kann.

Klassifikation *N01*

Betrachten wir die Klassifikationsaufgabe *N01*. In dieser wird ein Sprachdatensatz zu einer Klasse „nonnasal“ oder „nasal“ zugeordnet. Bei den nasalen Daten wird die Ausprägung nicht berücksichtigt. Unter Verwendung der über die Formantenanalyse ermittelten Parameter, soll in dieser Klassifikationsaufgabe geklärt werden, ob und wie gut sich die nasalen Laute von den nonnasalen Lauten unterscheiden lassen.

Untersucht man für das offene Näseln die durchschnittliche Klassifizierungsgüte, erkennt man, dass die Nasalität sehr gut feststellbar ist. Die durchschnittliche Erkennungsrate über sämtliche stimmhafte Laute beträgt für die Parametergruppe *ALLE* bei *FRAU* 85.8%, bei *KIND* 76.5% und bei *MANN* 94.4%. Als Laute eignen sich vor allem die Vokale /a/ und /o/. So erzielt bei *FRAU* das beste Ergebnis der Laut /a/ mit 92.3%, bei *KIND* der Laut /o/ mit 85.1% und bei *MANN* die Laute /a/ und /o/ mit 100%. Fragt man sich nach der besten Parametergruppe erhält man ein uneinheitliches Bild. So stechen bei *FRAU* die Gruppe *FREQ* und bei *MANN* die Gruppe *GF* hervor, während bei *KIND* mit *BAND* die besten Klassifizierungsgüten erreicht werden.

Im Vergleich zum offenen Näseln ist die Klassifizierungsgüte für das geschlossene Näseln geringer. Die durchschnittlichen Klassifizierungsraten für die Parametergruppe *ALLE* betragen bei *FRAU* 68.1%, bei *KIND* 70.2% und bei *MANN* 40.9%. Die geringe durchschnittliche Erkennungsrate bei *MANN* ist darauf zurückzuführen, dass lediglich bei dem Laut /n/ ein Klassifizierungsergebnis von 81.8% ermittelt werden konnte. Die besten Lautergebnisse ergeben bei *FRAU* /m/ mit 75.5% und bei *KIND* /n/ mit 71.4%. Bzgl. der Parametergruppen konnte lediglich mit *BAND* eine erfolgreiche Klassifizierung mit allen Gruppen durchgeführt werden. Auffallend ist, dass die LDA beim geschlossenen Näseln deutlich weniger Klassifizierungsergebnisse im Vergleich zum offenen Näseln liefert. Lediglich bei *FRAU* können beim /m/ über alle Parametergruppen hinweg Klassifizierungen beobachtet werden.

Klassifikation *N123*

In der Klassifikationsaufgabe *N123* werden lediglich von Logopäden vorklassifizierte nasale Datensätze betrachtet. Die nonnasalen Daten sind nicht mit berücksichtigt. Dabei soll überprüft werden, inwieweit sich die Nasalität auf einer 3-stufigen Skala quantifizieren lässt.

Die Untersuchungen an der offenen Nasalität beschränken sich auf die Sprechergruppen *FRAU* und *KIND*, da von *MANN* nur Sprachdaten mit einer Nasalitätsausprägung vorhanden sind. Die durchschnittlichen Erkennungsraten betragen bei *FRAU* 56% und bei *KIND* 63.4% und sind deutlich geringer als bei der Klassifikationsaufgabe *N01*. Die besten Erkennungsraten sind bei *FRAU* 64.4% mit /o/ und bei *KIND* 90.5% mit /e/ zu finden. Die höchsten Erkennungsraten wurden bei *FRAU* mit der Parametergruppe *INT*, bei *KIND* mit *FREQ* erzielt. Bei *KIND* war dies auch die einzige relevante Parametergruppe.

Zum geschlossenen Näseln können lediglich für die Sprechergruppe *FRAU* Ergebnisse präsentiert werden. Da diese nur für den Laut /n/ erarbeitet werden konnten, beträgt die durchschnittliche Erkennungsrate lediglich 30.8%. Die Nasalität des Lautes /n/ wurde dabei mit 61.5% korrekt klassifiziert. Mit der Parametergruppe *GF* wurden die besten Ergebnisse erzielt.

Zusammenfassend lässt sich sagen, dass eine 3-stufige Quantifizierung der Nasalität mittels Formanten bei *FRAU* ein nur befriedigendes Ergebnis liefert. Es beträgt im jeweils besten Fall 64.4% beim offenen Näseln und 61.5% beim geschlossenen Näseln. Bei *KIND* dagegen erhalten wir für das offene Näseln eine maximale Erkennungsrate von 90.5%. Hier spielt vor allem die Parametergruppe *FREQ* eine entscheidende Rolle. Zur Quantifizierung des geschlossenen Näselns bei Kindern können auf Grund der fehlenden Werte keine Aussagen gemacht werden.

Klassifikation *N0123*

In der Klassifikationsaufgabe *N0123* werden alle Nasalitätsausprägungen, also auch die nonnasalen, betrachtet. Damit soll untersucht werden, inwieweit sich die Nasalität bewerten lässt, wenn der Nasalitätsstatus des Datensatzes unbekannt ist. Dabei wird bei der Behandlung der offenen Nasalität nicht auf die Sprechergruppe *MANN* eingegangen, da diese aufgrund der einen Nasalitätsausprägung bereits mit der Klassifikationsaufgabe „*N01*“ abgehandelt wird.

Betrachten wir als erstes wieder die höchsten durchschnittlichen Klassifizierungsraten bei offener Nasalität. Diese betragen bei *FRAU* 63.9% und bei *KIND* 71.5%. Die besten Lautklassifizierungen wurden bei *FRAU* mit 71.3% beim /e/ und bei *KIND* mit 82% beim /o/

erreicht. Bei den Parametergruppen liefert *INT* bei *FRAU* und *FREQ* bei *KIND* die maximalen durchschnittlichen Klassifizierungsgüten.

Die höchsten durchschnittlichen Klassifizierungsraten betragen für das geschlossene Näseln 38.7% bei *FRAU*, 27.3% bei *MANN* und 30.7% bei *KIND*. Die geringen Güten sind über die Mittelung zu erklären, da vor allem bei *MANN* und *KIND* oft keine Ergebnisse ermittelt werden konnten. Die besten Ergebnisse für die Laute betragen bei *FRAU* 56.6% für /m/, bei *MANN* 54.5% für /m/ und bei *KIND* 61.4% für /n/. Bzgl. der Parametergruppen ergibt sich das gleiche uneinheitliche Bild wie beim offenen Näseln. So liefert *GF* die beste Klassifizierung bei *FRAU*, *FREQ* bei *MANN* und *BAND* bei *KIND*.

Zusammenfassend liefert *N0123* recht brauchbare Ergebnisse. So betragen die maximalen Klassifizierungsgüten für das offene Näseln bei *FRAU* 71.3% und bei *KIND* 82%. Die maximalen Klassifizierungsgüten für das geschlossene Näseln sind bei *FRAU* 56.6%, bei *MANN* 54.5% und bei *KIND* 61.4%.

Zweistufiges Vorgehen

Hier soll die Frage geklärt werden, ob bei unbekanntem Nasalitätsstatus ein zweistufiges Vorgehen, erst die Nasalität feststellen (entspricht Aufgabe *N01*) und danach quantifizieren (entspricht Aufgabe *N123*), gegenüber einer Klassifizierung über alle Ausprägungen (entspricht Aufgabe *N0123*) vorteilhafter ist. Das zweistufige Vorgehen wird im Kontext dieser Arbeit als *N01*N0123* bezeichnet. Die Klassifikationsgüte des zweistufigen Vorgehens wird dabei nach folgender Formel ermittelt:

$$\frac{p(N0) * Anz_{N0} + p(N1) * Anz_{N1} * p(N123)}{Anz_{N0} + Anz_{N1}} \qquad (54)$$

Dabei stehen p(N0) und p(N1) für die Wahrscheinlichkeiten, dass ein Datensatz in der Klassifikationsaufgabe *N01* als nonnasal bzw. nasal klassifiziert wird. Diese Wahrscheinlichkeiten werden mit der Gruppengröße Anz_{N0} respektive Anz_{N1} gewichtet. Ist ein Datensatz als nasaler erkannt, wird die Klassifikationsgüte der Klassifikationsaufgabe *N123* über p(*N123*) mitberücksichtigt.

Tabelle 21 fasst für die Parametergruppe *ALLE* und die Klassifikationsaufgabe *N0123* noch mal die durchschnittlichen Klassifikationsgüten über alle Laute sowie den Laut mit der besten Klassifizierung zusammen. Zum Vergleich werden für *ALLE* auch die nach obiger Formel berechneten zweistufigen Klassifikationsergebnisse wiedergegeben.

		Offen - *ALLE*			Geschlossen - *ALLE*		
	Gruppe	**ϕGüte**	**Laut**	**Güte**	**ϕGüte**	**Laut**	**Güte**
N0123	*FRAU*	63.9%	E	71.3%	38.7%	M	56.6%
	KIND	71.5%	O	82.0%	30.7%	N	61.4%
	MANN	-	-	-	27.3%	M	54.5%
N01 * N123	*FRAU*	72.1%	O	78.1%	46.2%	N	62.9%
	KIND	73.8%	O	83.8%	58.3%	M	69.5%
	MANN	-	-	-	-	-	-

Tab. 21: Formantenanalyse: zweistufige Klassifikation

Deutlich erkennbar ist die Überlegenheit des zweistufigen Ansatzes. In den Fällen, in denen eine Klassifizierung erfolgreich durchgeführt werden konnte, wurde auch eine bessere Güte erreicht. Hinsichtlich der Robustheit zeigt sich *N0123* überlegen. So konnten beim geschlossenen Näseln bei *MANN* mittels *N0123* signifikante Ergebnisse ermittelt werden, während dies mit *N123* und somit für das zweistufige Vorgehen nicht möglich war.

8.2.2 Parameter

Nasal	Gruppe	Art	Laut	1. Par	2. Par	3. Par	4. Par	5. Par	6. Par	7. Par	8. Par
N01	***FRAU***	Offen	A	F_0_I	F_1_F	F_3_F	F_1_I	F_0_F			
			E	F_1_F	F_0_F	F_3_F	F_4_I	F_1_I	F_2_I		
			I	GES_I	F_0_F	F_3_F					
			O	F_0_I	F_2_F	F_1_B					
			U	F_3_I	F_2_I	F_1_F	F_1_B	F_4_F			
		Geschl.	M	F_1_F	F_0_I						
			N	F_4_B							
	KIND	Offen	A	F_3_F	F_1_F						
			E	F_1_B	GES_I	F_1_I					
			I	F_2_B	F_1_F	F_4_B					
			O	F_1_B	GES_I	F_2_F					
			U	F_1_F	F_4_B	F_2_I	F_1_B				
		Geschl.	M	F_3_B							
			N	F_4_B							
	MANN	Offen	A	GES_I	F_4_F	F_3_F					
			E	F_0_B	F_1_F	F_2_F					
			I	F_2_B	F_2_F	F_4_B					
			O	F_2_F	F_3_F	GES_I	F_2_B	F_4_I			
			U	F_2_F	GES_I	F_2_I					
		Geschl.	M	-							
			N	F_1_B							
N123	***FRAU***	Offen	A	F_0_F	F_1_B						
			E	F_1_I	GES_I						
			I	F_1_I							
			O	F_0_I	F_0_F	F_2_I					
			U	F_1_I							
		Geschl.	M	-							
			N	F_0_I	F_1_B						
	KIND	Offen	A	F_4_F	F_3_F						
			E	F_2_F							
			I	F_2_F	F_1_F						
			O	F_2_F	F_3_F						
			U	--							
		Geschl.	M	--							
			N	--							
	MANN	Geschl.	M	--							
			N	--							
N0123	***FRAU***	Offen	A	F_1_F	F_0_I	F_3_F					
			E	F_1_F	F_0_F	F_3_F	F_4_I	F_1_I			
			I	GES_I	F_0_F	F_3_F					
			O	F_0_I	F_1_B	F_2_F					
			U	F_3_I	F_1_I	F_1_F	F_2_I				
		Geschl.	M	F_1_I							
			N	F_0_I	F_1_I						
	KIND	Offen	A	F_3_F	F_4_F	F_2_I	F_4_I				
			E	F_2_F	F_1_B	F_0_I					
			I	F_2_F	F_1_F	F_2_B					
			O	F_1_B	F_2_F	GES_I					
			U	F_2_I	F_4_B	F_1_F					
		Geschl.	M	-							
			N	F_4_B							
	MANN	Geschl.	M	F_3_F							
			N	-							

Tab. 22: Parameter Formantenanalyse

(F = Frequenz, I = Intensität, B = Bandbreite)

Tabelle 22 liefert für jede Sprechergruppe und jeden Laut die mittels der LDA generierten Parameter optimalen Parameter zur Bestimmung der Nasalität im Kontext der Klassifikationsaufgabe. Die Parameter sind nach ihrer Aufnahme durch die schrittweise LDA sortiert. Parameter mit einem kleinen Index tragen stärker zur Diskriminierung bei als Parameter mit höherem Index.

Eine klassifikations-, gruppen- und lautunabhängige Sichtweise liefert die folgende Tabelle 23. In der letzten Spalte sind die Parameter über alle Klassifikationsaufgaben je für das offene und das geschlossene Näseln aufsummiert. In den letzten 3 Zeilen wird über die Gruppen *FREQ*, *INT* und *BAND* kumuliert.

Betrachten wir zuerst die Summe über die Parameter für das offene Näseln. Deutlich ist der Einfluss der offenen Nasalität auf den 1. Formanten sichtbar. So ist insbesondere die Frequenz des 1. Formanten für die Feststellung der Nasalität, seine Intensität zur Quantifizierung geeignet. Diese Aussagen kann nach Betrachtung der Summe über die Gruppen derart verallgemeinert werden, dass sich die Präsenz der offenen Nasalität über Frequenzverschiebungen der Formanten F_1, F_2, F_3 feststellen lässt; die Stärke des Näselns sich in der Intensität der Grundfrequenz und der Formanten F_1 und F_2 widerspiegelt.

Auch beim geschlossenen Näseln hat F_1 neben der Grundfrequenz die stärkste Bedeutung. Eine so deutliche Aussage wie bei der offenen Nasalität bzgl. des Einflusses auf die Parameterausprägungen kann jedoch nicht abgegeben werden.

Fragt man sich nach der Art des Einflusses der Nasalität auf diese Parameter, liefert Tabelle 16 „Differenz der Mittelwerte bei der Grundfrequenz und den Formanten in Hz“ erste Antworten. Wie man dort sehen kann, sind bei den nasal offenen Lauten überwiegend die Frequenzen bei F_1, F_2 und F_3 erniedrigt. Die Frequenz von F_0 ist bei den offenen Lauten erhöht und bei den geschlossenen erniedrigt. Bei den Intensitäten ist bei den nasal offenen eine Erhöhung bei der Grundfrequenz sowie eine Erniedrigung bei dem 1. Formanten zu beobachten. Diese Erniedrigung der Intensität von F_1 ist auch bei den nasal geschlossenen gegeben. Die Bandbreite des 1. Formanten ist bei beiden Nasalitätsarten erhöht.

Parameter	*N01* offen	*N01* geschl.	*N123* offen	*N123* geschl.	*N0123* offen	*N0123* geschl.	Σ offen	Σ geschl.
F_0_F	3		4		2		9	0
F_0_I	2	1	2	2	3	1	7	4
F_0_B	1						1	0
F_1_F	7	1			5		12	1
F_1_I	3		6		2	2	11	2
F_1_B	5	1	2	2	3		10	3
F_2_F	6				4		10	0
F_2_I	4		2		3		9	0
F_2_B	3				1		4	0
F_3_F	6				4	1	10	1
F_3_I	1				1		2	0
F_3_B		1					0	1
F_4_F	2				1		3	0
F_4_I	2				2		4	0
F_4_B	3	2			1	1	4	3
Σ *FREQ*	24	1	4		16	1		
Σ *INT*	12	1	10	2	11	3		
Σ *BAND*	12	4	2	2	5	1		

Tab. 23: Formantenanalyse: absolute Häufigkeit der Parameter (laut- und gruppenunabhängig)
(F = Frequenz, I = Intensität, B = Bandbreite)

8.2.3 Zusammenfassung

Die untenstehende Tabellen 24 und 25 fassen die Ergebnisse der Formantenanalayse für das offene und geschlossene Näseln zusammen.

Pro Klassifikationsaufgabe wird für jede Sprechergruppe angegeben:

- die durchschnittliche Klassifizierungsgüte der Parametergruppe *ALLE.* Dieser Wert kann als Robustheit des Verfahrens für die Klassifikationsaufgabe interpretiert werden.
- die maximale Klassifizierungsgüte
- der Laut und die Parameter, mit denen die max. Klassifizierung erreicht wurde

	Gruppe	Güte *ALLE*	Bester Laut	Güte max.	1. Par	2. Par	3. Par	4. Par	5. Par	6. Par
N01	*FRAU*	85.8%	A	92.3%	F_0_I	F_1_F	F_3_F	F_1_I	F_0_F	
	KIND	76.5%	O	85.1	F_1_B	GES_I	F_2_F			
	MANN	94.4%	A O	100%	GES_I F_2_F	F4_F F_3_F	F_3_F GES_I	 F_2_B	 F_4_I	
N123	*FRAU*	56.0%	O	64.4%	F_0_I	F_0_F	F_2_I			
	KIND	63.4%	E	90.5%	F_2_F					
	MANN	-	-	-	-					
N0123	*FRAU*	63.9%	E	71.3%	F_1_F	F_0_F	F_3_F	F_4_I	F_1_I	
	KIND	71.5%	O	80.1%	F_1_B	F_2_F	GES_I			
	MANN	-	-	-	-					

Tab. 24: Zusammenfassung offenes Näseln Formantenanalyse

	Gruppe	Güte *ALLE*	Bester Laut	Güte	1. Par	2. Par	3. Par	4. Par	5. Par	6. Par
N01	*FRAU*	68.1%	M	75.5%	F_1_F	F_0_I				
	KIND	70.2%	N	71.4%	F_4_B					
	MANN	40.9%	N	81.8%	F_1_B					
N123	*FRAU*	30.8%	N	61.5%	F_0_I	F_1_B				
	KIND	-	-	-	-					
	MANN	-	-	-	-					
N0123	*FRAU*	38.7%	M	56.6%	F_1_I					
	KIND	30.7%	N	61.4%	F_4_B					
	MANN	27.3%	M	54.5%	F_3_F					

Tab. 25: Zusammenfassung geschlossenes Näseln Formantenanalyse

8.3 Antiformantenanalyse

8.3.1 Klassifikationsgüte

Die Tabelle 26 gibt den Gesamtprozentsatz der richtigen Klassifizierungen der Nasalität pro Klassifikationsaufgabe für jede Sprecher- und Parametergruppe. Die Parametergruppe *GF* ist dabei nicht mehr berücksichtigt, da diese bei der Formantenanalyse abgehandelt wurde.

N01

Wie die Formanten eignen sich auch die Antiformanten gut zur Feststellung der Nasalität. Die durchschnittliche Erkennungsrate für das offene Näseln beträgt für *ALLE* bei *FRAU* 88.2%, bei *MANN* 95.0% und bei *KIND* 77.6%. Die besten Klassifizierungsraten werden bei *FRAU* mit 94.1% bei /a/, bei *MANN* mit 100% bei /o/ und bei *KIND* mit 85.8% bei /i/ erzielt. Als Parametergruppe liefert *INT* die besten Ergebnisse. Lediglich bei *KIND* liefert *FREQ* eine um 0.8% bessere Klassifizierung.

Das geschlossene Näseln ist wie bei den Formanten im Vergleich zum offenen Näseln schlechter klassifizierbar. Die durchschnittlichen Erkennungsraten für *ALLE* betragen 74.5% bei *FRAU* und 72.7% bei *KIND*. Zu *MANN* konnten keine Ergebnisse gewonnen werden. Die beste Klassifizierung wurde mit /m/ sowohl bei *FRAU* als auch bei *KIND* mit 83.0% bzw. 81.1% erreicht. Auch hier liefert die Parametergruppe *INT* die besten Ergebnisse.

N123

Wie bei den Formanten bereits erwähnt, existiert von *MANN* nur eine Nasalitätsausprägung, so dass diese Sprechergruppe in dieser Klassifikationsaufgabe nicht berücksichtigt wurde. Allgemein lässt sich sagen, dass in dieser Klassifikationsaufgabe über die Antiformanten für einige Laute keine Diskriminanzparameter ermittelt werden konnten. Dies schlägt sich auf den Durchschnitt der Klassifizierungsraten nieder. So beträgt dieser bei *FRAU* beim offenen Näseln lediglich 47.1%. Bei *KIND* konnten in der Parametergruppe *ALLE* für jeden Laut eine Diskriminierung durchgeführt werden. Die durchschnittliche Erkennungsrate beträgt für diese Gruppe 80.0%. Die beste Erkennungsrate wurde bei *FRAU* mit /i/ mit 69.6% und bei Kind mit /a/ mit 100% erreicht. Bzgl. der Parametergruppe wurden die besten Ergebnisse mit *FREQ* erzielt.

Zur Quantifizierung des geschlossenen Näselns eignen sich die Antiformanten in dieser Klassifikationsaufgabe wenig. Für *KIND* konnten keine Ergebnisse ermittelt werden, für *FRAU* beträgt die beste Erkennungsrate für /n/ lediglich 50%. Da für /m/ keine Resultate ermittelt werden konnten, beträgt die durchschnittliche Erkennungsrate nur 25%. Relevante Ergebnisse konnten auch nur mit der Parametergruppe *INT* erzielt werden.

Nasal	Gruppe	Parameter	A	E	I	O	U	M	N	Ø offen	Ø gschl.
N01	***FRAU***	***FREQ***	85.2%	86.0%	75.2%	71.3%	73.0%	58.5%	64.2%	78.1%	61.4%
		INT	84.0%	79.9%	79.3%	79.6%	82.5%	67.9%	64.2%	81.1%	66.1%
		BAND	75.7%	86.6%	55.2%	71.9%	47.6%	67.9%	-	67.4%	34.0%
		ALLE	94.1%	87.8%	88.3%	87.4%	83.6%	83.0%	66.0%	88.2%	74.5%
	MANN	***FREQ***	80.8%	-	-	95.7%	82.1%	-	-	51.7%	-
		INT	73.1%	89.7%	82.8%	91.3%	92.9%	-	-	86.0%	-
		BAND	-	-	-	100.0%	-	-	-	20.0%	-
		ALLE	92.3%	96.6%	89.7%	100.0%	96.4%	-	-	95.0%	-
	KIND	***FREQ***	69.1%	67.5%	84.5%	71.3%	78.9%	81.1%	-	74.3%	40.6%
		INT	66.4%	77.7%	83.9%	58.8%	78.9%	81.1%	64.3%	73.1%	72.7%
		BAND	-	-	84.5%	-	78.9%	68.9%	-	32.7%	34.5%
		ALLE	68.4%	82.2%	85.8%	76.9%	74.7%	79.7%	64.3%	77.6%	72.0%
N123	***FRAU***	***FREQ***	-	47.5%	57.1%	-	-	-	-	20.9%	-
		INT	-	27.9%	48.2%	-	-	-	50.0%	15.2%	25.0%
		BAND	-	27.9%	42.9%	-	-	-	-	14.2%	-
		ALLE	-	47.5%	69.6%	-	-	-	50.0%	47.1%	25.0%
	MANN	***FREQ***	-	-	-	-	-	-	-	-	-
		INT	-	-	-	-	-	-	-	-	-
		BAND	-	-	-	-	-	-	-	-	-
		ALLE	-	-	-	-	-	-	-	-	-
	KIND	***FREQ***	-	60.0%	63.2%	85.7%	-	-	-	41.8%	-
		INT	84.6%	75.0%	-	-	-	-	-	31.9%	-
		BAND	-	-	-	-	72.0%	-	-	14.4%	-
		ALLE	100.0%	75.0%	63.2%	85.7%	76.0%	-	-	80.0%	-
N0123	***FRAU***	***FREQ***	60.9%	68.3%	52.4%	49.7%	54.5%	-	-	57.2%	-
		INT	65.7%	49.4%	47.6%	56.9%	66.7%	-	-	57.3%	-
		BAND	53.8%	61.6%	41.4%	62.9%	-	-	-	43.9%	-
		ALLE	72.2%	72.6%	70.3%	69.5%	61.4%	-	-	69.2%	-
	MANN	***FREQ***	-	-	-	-	-	-	-	-	-
		INT	-	-	-	-	-	-	-	-	-
		BAND	-	-	-	-	-	-	-	-	-
		ALLE	-	-	-	-	-	-	-	-	-
	KIND	***FREQ***	61.8%	58.0%	69.7%	57.5%	76.3%	-	-	64.7%	-
		INT	21.7%	72.6%	77.4%	-	76.3%	-	-	49.6%	-
		BAND	-	-	83.9%	-	80.0%	-	-	32.8%	-
		ALLE	53.9%	77.7%	83.2%	63.8%	80.0%	-	-	71.7%	-

Tab. 26: Klassifikation Antiformantenanalyse: Erkennungsraten in %

N0123

In der Klassifikationsaufgabe *N0123* beträgt die durchschnittliche Erkennungsrate für das offene Näseln 69.2% bei *FRAU* und 71.7% bei *KIND*. Die besten Erkennungsraten wurden

bei *FRAU* mit 72.6% bei /e/ und bei *KIND* mit 83.9% bei /i/ erzielt. Die Parametergruppe *INT* liefert bei *FRAU* die besten Erkennungen, bei *KIND* ist es *FREQ*.

Zum geschlossenen Näseln konnte für keine der 3 Gruppen ein Ergebnis ermittelt werden. Dies unterstreicht noch mal die weniger gute Eignung der Antiformantenanalyse zur Quantifizierung der Nasalität.

Zweistufiges Vorgehen

Wie bei der Formantenanalyse zeigt sich der zweistufige Ansatz auch bei der Antiformantenanalse der Klassifikationsaufgabe *N0123* überlegen. Die größere Robustheit des einstufigen Ansatzes kann hier aber nicht bestätigt werden. So liefert der zweistufige Ansatz beim geschlossenen Näseln für *FRAU* Ergebnisse, der einstufige jedoch nicht.

		Offen - *ALLE*			**Geschlossen - *ALLE***		
	Gruppe	**ϕGüte**	**Laut**	**Güte**	**ϕGüte**	**Laut**	**Güte**
N0123	*FRAU*	69.2%	E	72.6%	-	-	-
	KIND	71.7%	I	83.2%	-	-	-
	MANN	-	-	-	-	-	-
N01 * N123	*FRAU*	71.6%	A	79.1%	25.8%	N	51.5%
	KIND	76.0%	I	83.0%	-	-	-
	MANN	-	-	-	-	-	-

Tab. 27: Antiformantenanalyse: zweistufige Klassifikation

8.3.2 Parameter

Nasal	Gruppe	Art	Laut	1. Par	2. Par	3. Par	4. Par	5. Par	6. Par	7. Par	8. Par
N01	***FRAU***	Offen	A	AF_3_F	AF_2_F	AF_3_B	AF_4_I	AF_1_B	AF_1_F	AF_2_B	
			E	AF_1_B	AF_2_F	AF_2_B					
			I	AF_2_I	AF_2_F	AF_4_I	AF_3_B	AF_4_F			
			O	AF_3_I	AF_2_I						
			U	AF_3_I	AF_2_I	AF_4_F	AF_3_F	AF_2_F			
		Geschl.	M	AF_3_B	AF_4_F	AF_3_I	AF_4_I	AF_2_F	AF_2_I		
			N	AF_4_F	AF_3_I	AF_2_I	AF_4_I				
	KIND	Offen	A	AF_4_F	AF_2_F	AF_4_I					
			E	AF_3_I	AF_2_I						
			I	AF_2_I	AF_1_B	AF_3_I	AF_3_F	AF_4_I			
			O	AF_2_F	AF_3_I	AF_3_F					
			U	AF_3_B	AF_3_I						
		Geschl.	M	AF_1_I	AF_4_B						
			N	AF_3_I							
	MANN	Offen	A	AF_3_F	AF_1_I						
			E	AF_3_I	AF_3_F						
			I	AF_2_I							
			O	AF_2_I	AF_4_F	AF_3_F	AF_2_B	AF_3_B			
			U	AF_3_I	AF_2_I	AF_4_I					
		Geschl.	M	-							
			N	-							
N123	***FRAU***	Offen	A	-							
			E	AF_1_I	AF_2_F						
			I	AF_2_I	AF_2_F	AF_3_B					
			O	-							
			U	-							
		Geschl.	M	-							
			N	AF_2_I							
	KIND	Offen	A	AF_4_I	AF_3_I	AF_4_F	AF_1_B	AF_2_B			
			E	AF_2_I							
			I	AF_3_F							
			O	AF_4_F							
			U	AF_4_B	AF_2_I						
		Geschl.	M	-							
			N	-							
	MANN	Geschl.	M	-							
			N	-							
N0123	***FRAU***	Offen	A	AF_3_F	AF_2_F	AF_3_B	AF_2_I				
			E	AF_1_F	AF_2_F						
			I	AF_2_I	AF_2_F	AF_3_B					
			O	AF_3_I	AF_4_I						
			U	AF_3_I	AF_2_I	AF_3_F	AF_4_F				
		Geschl.	M	-							
			N	-							
	KIND	Offen	A	AF_4_I	AF_2_I	AF_4_F					
			E	AF_3_I	AF_2_I						
			I	AF_4_I	AF_3_I	AF_2_I	AF_1_F	AF_3_F			
			O	AF_2_F	AF_4_F	AF_3_I					
			U	AF_3_B	AF_2_B	AF_4_B	AF_1_B				
		Geschl.	M	AF_1_I							
			N	-							
	MANN	Geschl.	M	-							
			N	-							

Tab. 28: Parameter Antiformantenanalyse

(F = Frequenz, I = Intensität, B = Bandbreite)

Tabelle 28 liefert pro Klassifikationsaufgabe für jede Sprechergruppe und jeden Laut die optimalen Parameter zur Quantifizierung der Nasalität. Die Parameter sind nach ihrer Aufnahme durch die schrittweise LDA sortiert. Parameter mit einem kleinen Index tragen stärker zur Diskriminierung bei als Parameter mit höheren Indizes.

Eine klassifikations-, gruppen- und lautunabhängige Sichtweite liefert Tabelle 29. Die letzen zwei Spalten summieren die Parameter in Abhängigkeit der Nasalitätsart. In den letzten drei Zeilen wird über die Gruppen *FREQ*, *INT* und *BAND* aufsummiert.

Die stärkste Bedeutung beim offenen Näseln haben die Intensitäten und Frequenzen der Antiformanten AF_2 und AF_3. Ein Vergleich mit Tabelle 17 „Differenz der Mittelwerte bei den Antiformanten in Hz“ zeigt, dass die Intensitäten bei beiden Antiformanten erhöht sind. Die Frequenz ist bei AF_2 erhöht, bei AF_3 erniedrigt.

Beim geschlossenen Näseln sind die Intensitäten von AF_2 und AF_3 ebenfalls erhöht, die Frequenz von AF_4 ebenfalls erhöht.

Auffallend gering ist die Anzahl der Parameter für das geschlossene Näseln. Für diese Nasalitätsart erweist sich die Antiformantenanalyse als ein wenig robustes Verfahren.

	N01		***N123***		***N0123***		**Σ**	**Σ**
Parameter	**offen**	**geschl.**	**offen**	**geschl.**	**offen**	**geschl.**	**offen**	**geschl.**
AF_1_F	1				2		3	0
AF_1_I	1	1	1			1	2	2
AF_1_B	3		1		1		5	0
AF_2_F	6	1	2		4		12	1
AF_2_I	8	2	3	1	6		17	3
AF_2_B	3		1		1		5	0
AF_3_F	7		1		3		11	0
AF_3_I	8	3	1		5		14	3
AF_3_B	4	1	1		3		8	1
AF_4_F	4	2	2		3		9	2
AF_4_I	5	2	1		3		9	2
AF_4_B		1	1		1		2	1
Σ *FREQ*	18	3	5	0	12	0		
Σ *INT*	22	8	6	1	14	1		
Σ *BAND*	10	2	4	0	6	0		

Tab. 29: Antiformantenanalyse: absolute Häufigkeit der Parameter (laut- und gruppenunabhängig) (F = Frequenz, I = Intensität, B = Bandbreite)

8.3.3 Zusammenfassung

Die untenstehende Tabellen 30 und 31 fassen die Ergebnisse der Antiformantenanalyse für das offene und geschlossene Näseln zusammen.

Pro Klassifikationsaufgabe wird für jede Sprechergruppe angegeben:

- die durchschnittliche Klassifizierungsgüte der Parametergruppe *ALLE*
- die maximale Klassifizierungsgüte
- der Laut und die Parameter, mit denen die max. Klassifizierung erreicht wurde

Vergleicht man die Erkennungsraten mit denen der Formantenanalyse, sieht man bei der Antiformantenanalyse höhere Ergebnisse. Dafür liefert die Antiformantenanalyse aber weniger robuste Klassifizierungen; dies insbesondere für das geschlossene Näseln.

	Gruppe	**Güte *ALLE***	**Bester Laut**	**Güte max.**	1. Par	2. Par	3. Par	4. Par	5. Par	6. Par	7. Par
N01	*FRAU*	88.2%	A	94.1%	AF_3_F	AF_2_F	AF_3_B	AF_4_I	AF_1_B	AF_1_F	AF_2_B
	KIND	77.6%	I	85.8%	AF_2_I	AF_1_B	AF_3_I	AF_3_F	AF_4_I		
	MANN	95.0%	O	100%	AF_2_I	AF_4_F	AF_3_F	AF_2_B	AF_3_B		
N123	*FRAU*	47.1%	I	69.6%	AF_2_I	AF_2_F	AF_3_B				
	KIND	80.0%	A	100%	AF_4_I	AF_3_I	AF_4_F	AF_1_B	AF_2_B		
	MANN	-	-	-	-						
N0123	*FRAU*	69.2%	E	72.6%	AF_3_F	AF_2_F	AF_3_B	AF_2_I			
	KIND	71.7%	I	83.9%	AF_4_I	AF_3_I	AF_2_I	AF_1_F	AF_3_F		
	MANN	-	-	-	-						

Tab. 30: Zusammenfassung offenes Näseln Antiformantenanalyse

	Gruppe	**Güte *ALLE***	**Bester Laut**	**Güte max.**	1. Par	2. Par	3. Par	4. Par	5. Par	6. Par	7. Par
N01	*FRAU*	74.5%	M	83%	AF_3_B	AF_4_F	AF_3_I	AF_4_I	AF_2_F	AF_2_I	
	KIND	72.0%	M	81.1%	AF_1_I	AF_4_B					
	MANN	-	-	-	-						
N123	*FRAU*	25.0%	N	50.0%	AF_2_I						
	KIND	-	-	-	-						
	MANN	-	-	-	-						
N0123	*FRAU*	-	-	-	-						
	KIND	-	-	-	-						
	MANN	-	-	-	-						

Tab. 31: Zusammenfassung geschlossenes Näseln Antiformantenanalyse

8.4 Frequenzbandanalyse

Die Frequenzbandanalyse untersucht den gesamten Frequenzbereich eines Lautes. Bei der Klassifikation werden die Intensitäten gleicher Bänder über die Nasalitätsgruppen hinweg verglichen. Veränderungen der Lage, Intensität oder Bandbreite eines Formanten durch die Nasalität, spiegeln sich in Intensitätsunterschieden gleicher Bänder bei den nonnasalen Daten wider. Die Parameter des Sprachmodells, d. h. Grundfrequenz, Formanten und Antiformanten werden in der Frequenzbandanalyse nur indirekt beobachtet. So umfasst bei der Bandbreite von 400 Hz der erste Bandpass die Grundwelle der Sprachgrundfrequenz, ohne sie genau zu lokalisieren. Insbesondere würde bei einem tiefen 1. Formanten ein Teil seiner Energie in dieses Band mit einfließen. Eine Abnahme der Frequenz des 1. Formanten würde als Erhöhung der Grundfrequenzintensität sowie als Intensitätsabnahme des 1. Formanten interpretiert werden.

Ziel der Untersuchungen war die Beantwortung folgender Teilprobleme zur Nasalitätsbewertung für das offene und das geschlossene Näseln:

- Welches Verfahren liefert die besten Klassifizierungen?
- Bringen die Normierungen Verbesserungen?
- Welcher Laut bringt sprechgruppenunabhängig die besten Güten? Allgemeiner formuliert, stellt sich hier die Frage nach der Eignung der verschiedenen Laute.
- In welchen Frequenzbändern sind die Folgen der Nasalität besonders deutlich?

8.4.1 Klassifikationsgüte

Die Tabellen geben den Gesamtprozentsatz der richtigen Klassifizierungen pro Klassifikationsaufgabe für jede Sprechergruppe und Verfahren an.

N01

Die beste Klassifizierungsrate für das offene Näseln liefert für alle Sprechergruppen *U100_N*. Sie betragen bei *FRAU* 92.6%, bei *KIND* 82.7% und bei *MANN* 100%. Eine Frequenz- und Intensitätsnormierung erbrachte keine Verbesserungen. Die besten Erkennungsraten pro Laut werden bei *FRAU* mit 97.5% bei /a/, bei *KIND* mit 90.7% bei /i/ und bei *MANN* mit 100% bei allen Lauten erreicht.

Auch beim geschlossenen Näseln konnte mit *U100_N* bei *FRAU* und *MANN* die besten Klassifizierungsgüten mit jeweils 92.4% und 95.5% erreicht werden. Bei *KIND* brachte eine Normierung der 100 Hz Bandbreiten auf die Gesamtintensität das beste Ergebnis mit 80.4%.

Die besten Erkennungsraten pro Laut wurden bei *FRAU* mit 93.2% bei /m/, bei *KIND* mit 82.4% bei /n/ und bei *MANN* mit 100% bei /m/ erreicht.

Zusammenfassend lässt sich sagen, dass mit *U100_N* die besten Klassifizierungsgüten erreicht wurden. Eine Normierung auf die Gesamtintensität brachte nur beim geschlossenen Näseln bei den Kindern eine Verbesserung.

Gruppe	Verfahren	A	E	I	O	U	M	N	Ø offen	Ø gschl.
FRAU	*U100_N*	97.5%	88.4%	89.5%	94.4%	93.4%	93.2%	91.5%	92.6%	92.4%
	U100_I	94.0%	86.4%	87.5%	94.4%	90.4%	83.1%	81.4%	90.5%	82.3%
	U100_S	97.0%	89.9%	90.5%	92.4%	90.4%	89.8%	81.4%	92.0%	85.6%
	U400_N	94.5%	81.3%	89.0%	83.8%	89.9%	78.0%	76.3%	87.7%	77.2%
	U400_I	85.6%	77.8%	88.0%	82.7%	90.4%	74.6%	72.9%	84.9%	73.8%
	U400_S	96.0%	82.8%	90.5%	86.3%	88.4%	71.2%	81.4%	88.8%	76.3%
	BARK_N	97.0%	86.4%	87.0%	90.9%	90.4%	79.7%	79.7%	90.3%	79.7%
	BARK_I	89.6%	82.3%	84.5%	89.3%	89.9%	74.6%	66.1%	87.1%	70.4%
	BARK_S	96.0%	86.4%	88.0%	90.9%	90.4%	81.4%	74.6%	90.3%	78.0%
MANN	*U100_N*	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	90.9%	100.0%	95.5%
	U100_I	97.1%	97.0%	100.0%	100.0%	100.0%	100.0%	72.7%	98.8%	86.4%
	U100_S	100.0%	100.0%	97.0%	100.0%	100.0%	100.0%	90.9%	99.4%	95.5%
	U400_N	94.1%	90.9%	81.8%	97.0%	93.9%	-	90.9%	91.5%	45.5%
	U400_I	88.2%	-	81.8%	100.0%	87.9%	-	72.7%	71.6%	36.4%
	U400_S	100.0%	97.0%	84.8%	93.9%	100.0%	-	81.8%	95.1%	40.9%
	BARK_N	100.0%	100.0%	97.0%	100.0%	100.0%	-	90.9%	99.4%	45.5%
	BARK_I	100.0%	97.0%	97.0%	100.0%	100.0%	100.0%	90.9%	98.8%	95.5%
	BARK_S	100.0%	100.0%	90.9%	100.0%	100.0%	-	90.9%	98.2%	45.5%
KIND	*U100_N*	79.1%	87.6%	86.8%	77.9%	82.3%	78.4%	78.4%	82.7%	78.4%
	U100_I	70.4%	78.7%	85.9%	77.9%	86.2%	78.4%	82.4%	79.8%	80.4%
	U100_S	79.1%	86.6%	81.5%	76.5%	85.2%	77.0%	70.3%	81.8%	73.7%
	U400_N	71.4%	86.6%	83.9%	82.8%	79.8%	74.3%	68.9%	80.9%	71.6%
	U400_I	66.5%	67.8%	82.0%	77.5%	83.3%	77.0%	75.7%	75.4%	76.4%
	U400_S	72.3%	86.6%	77.1%	70.1%	75.9%	60.8%	-	76.4%	30.4%
	BARK_N	76.2%	84.2%	88.3%	77.9%	78.8%	74.3%	73.0%	81.1%	73.7%
	BARK_I	68.4%	75.2%	90.7%	73.0%	81.3%	74.3%	78.4%	77.7%	76.4%
	BARK_S	76.2%	86.6%	84.9%	73.0%	72.4%	70.3%	67.6%	78.6%	69.0%

Tab. 32: Frequenzbandanalyse: Klassifizierung *N01*

N123

Bei *FRAU* liefert beim offenen Näseln *U100_S* die beste Klassifizierung mit 72.6%. Bei den Kindern wird mit *U100_N* eine Klassifizierungsrate von 86.7% erzielt. Die besten Erkennungsraten pro Laut betragen bei *FRAU* mit /u/ 78.2% und bei *KIND* mit /o/ 100%. Da von den Männern beim offenen Näseln nur eine Nasalitätsausprägung existiert, können zum offenen Näseln für diese Klassifikationsaufgabe keine Angaben gemacht werden.

Gruppe	Verfahren	A	E	I	O	U	M	N	Ø offen	Ø gschl.
FRAU	*U100_N*	63.4%	55.7%	63.8%	75.6%	57.7%	96.6%	58.6%	63.2%	77.6%
	U100_I	63.4%	48.1%	66.3%	62.8%	73.1%	41.4%	82.8%	62.7%	62.1%
	U100_S	72.0%	72.2%	68.8%	71.8%	78.2%	51.7%	65.5%	72.6%	58.6%
	U400_N	58.5%	60.8%	55.0%	39.7%	64.1%	27.6%	58.6%	55.6%	43.1%
	U400_I	48.8%	45.6%	51.3%	56.4%	60.3%	44.8%	51.7%	52.5%	48.3%
	U400_S	37.8%	55.7%	52.5%	57.7%	67.9%	48.3%	51.7%	54.3%	50.0%
	BARK_N	65.9%	57.0%	58.8%	56.4%	57.7%	65.5%	55.2%	59.2%	60.4%
	BARK_I	59.8%	54.4%	60.0%	56.4%	57.7%	48.3%	65.5%	57.7%	56.9%
	BARK_S	65.9%	62.0%	60.0%	60.3%	61.5%	48.3%	48.3%	61.9%	48.3%
MANN	*U100_N*	-	-	-	-	-	87.5%	-	-	43.8%
	U100_I	-	-	-	-	-	50.0%	-	-	25.0%
	U100_S	-	-	-	-	-	87.5%	-	-	43.8%
	U400_N	-	-	-	-	-	100.0%	-	-	50.0%
	U400_I	-	-	-	-	-	25.0%	-	-	12.5%
	U400_S	-	-	-	-	-	100.0%	-	-	50.0%
	BARK_N	-	-	-	-	-	25.0%	75.0%	-	50.0%
	BARK_I	-	-	-	-	-	37.5%	87.5%	-	62.5%
	BARK_S	-	-	-	-	-	100.0%	-	-	50.0%
KIND	*U100_N*	80.6%	73.3%	96.6%	96.6%	86.2%	-	86.7%	86.7%	43.4%
	U100_I	77.4%	-	82.8%	100.0%	75.9%	-	-	67.2%	-
	U100_S	93.5%	73.3%	75.9%	86.2%	93.1%	-	-	84.4%	-
	U400_N	74.2%	80.0%	79.3%	75.9%	-	-	-	61.9%	-
	U400_I	-	-	-	72.4%	-	-	-	14.5%	-
	U400_S	71.0%	73.3%	-	79.3%	-	-	-	44.7%	-
	BARK_N	-	80.0%	72.4%	82.8%	-	-	-	47.0%	-
	BARK_I	-	-	69.0%	82.8%	-	66.7%	-	30.4%	33.4%
	BARK_S	61.3%	73.3%	69.0%	75.9%	82.8%	-	-	72.5%	-

Tab. 33: Frequenzbandanalyse: Klassifizierung *N123*

Beim geschlossenen Näseln bietet *U100_N* die besten Erkennungsraten bei Frauen und Kindern mit 77.6% bzw. 43.4%. Das beste Verfahren bei *MANN* ist *BARK_I* mit 62.5%. Die

besten Lautergebnisse wurden bei *FRAU* mit 96.6% bei /m/, bei *KIND* mit 86.7% bei /n/ und bei *MANN* mit 100% bei /m/ erreicht.

Zusammenfassend lässt sich sagen, dass (wie bei *N01*) *U100_N* die besten Ergebnisse liefert. Eine Sone-Intensitätsnormierung brachte beim offenen Näseln bei *FRAU* eine Verbesserung. Bei *MANN* war das beste Verfahren beim geschlossenen Näseln *BARK_I*.

Gruppe	Verfahren	A	E	I	O	U	M	N	Ø offen	Ø gschl.
FRAU	*U100_N*	84.1%	67.2%	66.0%	81.2%	76.3%	54.2%	15.3%	75.0%	34.8%
	U100_I	72.6%	57.1%	65.0%	74.1%	75.3%	49.2%	22.0%	68.8%	35.6%
	U100_S	79.6%	67.7%	75.0%	78.2%	76.3%	47.5%	45.8%	75.4%	46.7%
	U400_N	73.1%	60.6%	60.0%	69.0%	73.2%	28.8%	-	67.2%	14.4%
	U400_I	57.2%	52.5%	62.5%	60.9%	73.7%	30.5%	-	61.4%	15.3%
	U400_S	77.1%	59.6%	69.5%	71.1%	69.2%	33.9%	27.1%	69.3%	30.5%
	BARK_N	77.1%	60.6%	68.0%	74.6%	70.2%	55.9%	32.2%	70.1%	44.1%
	BARK_I	61.7%	59.1%	63.0%	65.5%	70.7%	50.8%	-	64.0%	25.4%
	BARK_S	79.1%	63.1%	72.0%	72.6%	74.7%	49.2%	-	72.3%	24.6%
MANN	*U100_N*	-	-	-	-	-	90.9%	81.8%	-	86.4%
	U100_I	-	-	-	-	-	81.8%	-	-	40.9%
	U100_S	-	-	-	-	-	100.0%	90.9%	-	95.5%
	U400_N	-	-	-	-	-	72.7%	-	-	36.4%
	U400_I	-	-	-	-	-	54.5%	-	-	27.3%
	U400_S	-	-	-	-	-	54.5%	-	-	27.3%
	BARK_N	-	-	-	-	-	54.5%	72.7%	-	63.6%
	BARK_I	-	-	-	-	-	72.7%	-	-	36.4%
	BARK_S	-	-	-	-	-	54.5%	81.8%	-	68.2%
KIND	*U100_N*	66.5%	85.1%	80.5%	70.6%	76.4%	71.6%	63.5%	75.8%	67.6%
	U100_I	39.3%	72.3%	81.0%	66.7%	81.3%	63.5%	60.8%	68.1%	62.2%
	U100_S	72.8%	83.7%	69.3%	67.2%	69.0%	-	-	72.4%	-
	U400_N	58.3%	84.2%	74.1%	60.8%	69.0%	63.5%	63.5%	69.3%	63.5%
	U400_I	-	73.8%	73.2%	44.6%	72.4%	66.2%	68.9%	52.8%	67.6%
	U400_S	57.3%	77.7%	69.8%	55.9%	68.5%	-	-	65.8%	-
	BARK_N	63.6%	82.2%	78.5%	68.6%	70.0%	59.5%	64.9%	72.6%	62.2%
	BARK_I	48.5%	76.2%	79.5%	65.2%	70.4%	70.3%	63.5%	68.0%	66.9%
	BARK_S	61.7%	79.7%	76.1%	67.6%	65.5%	58.1%	-	70.1%	29.1%

Tab. 34: Frequenzbandanalyse: Klassifizierung *N0123*

N0123

Die beste Klassifizierungsgüten sowohl beim offenen als auch beim geschlossenen Näseln liefern die Frequenzbandbreiten mit 100 Hz.

Beim offenen Näseln brachte eine Sone-Intensitätsnormierung bei den Frauen eine Verbesserung. Es wurde mit *U100_S* eine Klassifizierungsgüte von 75.4% erreicht. Bei *KIND* wurde mit *U100_N* ein Ergebnis von 75.8% erreicht. Die besten Laute waren bei *FRAU* /a/ mit 84.1% und bei *KIND* /e/ mit 85.1%. Bzgl. der Ergebnisse zu *MANN* wird aufgrund der einen Nasalitätsausprägung auf *N01* verwiesen.

Beim geschlossenen Näseln wurden bei *FRAU* und *MANN* mit *U100_S* 46.7% bzw. 95.5% erzielt. Bei *KIND* wurden mit *U100_N* 67.6% erzielt. Der beste Laut war für alle drei Sprechergruppen der Laut /m/ mit jeweils 55.9% bei *FRAU*, 100% bei *MANN* und 71.6% bei *KIND*.

Zweistufiges Vorgehen

Vergleicht man die Vorgehensweise *N0123* mit dem zweistufigen Vorgehen *N01 * N123*, so sehen wir wie bereits bei der Formanten- und Antiformantenanalyse beim offenen Näseln, deutlich bessere Ergebnisse beim zweistufigen Verfahren. Lediglich beim geschlossenen Näseln konnte mit *N0123* bei *MANN* eine bessere Erkennungsgüte erreicht werden.

		Offen - *ALLE*			**Geschlossen - *ALLE***		
	Gruppe	**ϕGüte**	**Laut**	**Güte**	**ϕGüte**	**Laut**	**Güte**
N0123	*FRAU*	75.4%	A	84.1%	46.7%	M	55.9%
	KIND	75.8%	E	85.1%	67.6%	M	71.6%
	MANN	-			95.5%	M	100%
N01 * N123	*FRAU*	82.1%	A	85.9%	82.2%	M	91.7%
	KIND	81.4%	I	87.4%	70.0%	N	76.4%
	MANN	-	-	-	68.8%	M	90.9%

Tab. 35: Frequenzbandanalyse: zweistufige Klassifikation

8.4.2 Parameter

Tabelle 36 listet pro Klassifikationsaufgabe für jede Sprechergruppe und jedes Verfahren die optimalen Parameter für jeden Laut. Da die Parameter aufgrund der unterschiedlichen Skalen nicht unmittelbar miteinander vergleichbar sind, werden sie in der Tabelle durch die entsprechende lineare Bandmittenfrequenz wiedergegeben.

Nasal	Gruppe	Art	Verfahren	Laut	1. Par	2. Par	3. Par	4. Par	5. Par	6. Par	7. Par	8. Par
N01	***FRAU***	**Offen**	*U100_N*	A	250	1050	950	4350	850	1250	2950	4050
				E	250	450	5350	650	550	1150	3550	3050
				I	3750	1050	950	4450	7150	150	4350	1150
				O	250	450	1750	550	350	7550	3250	2750
				U	2850	350	750	650	2350	450	950	3150
		Geschl.	*U100_N*	M	2850	150	50	850	7850	7650	1550	
				N	650	1950	5250	4350	5850	2650	2450	4150
	KIND	**Offen**	*U100_N*	A	550	4750	150	5850	2250	2650	650	
				E	850	2650	150	1050	5250	450	1950	2250
				I	1750	550	150	4250	1650	650	450	750
				O	150	1650	350	3150	3750	6850	5750	450
				U	2750	650	550	150	1150	5050	450	4650
		Geschl.	U100_I	M	7050	1750	1550	1050				
				N	450	3650	2250	1350	6950			
	MANN	**Offen**	*U100_N*	A	150	250	2750	7250	6050	5650	6850	1650
				E	150	1050	50	2150	4450	1850	2050	5850
				I	250	3150	1550	350	1450	1650	6050	450
				O	3950	4350	550	1150	150			
				U	1050	4650	1550	6950	2350	3250	1250	7250
		Geschl.	*U100_N*	M	6050	7650	50	450	950	5850		
				N	50							
N123	***FRAU***	**Offen**	*U100_S*	A	2550	1250	3050	7950	5550	6850	7250	1450
				E	4550	6550	450	5750	5950	7850		
				I	450	5950	4550	7250	5750	1450		
				O	450	150	2950	4450	7250	5850	1950	3950
				U	450	4450	5550	7250	3750	7450	4750	2050
		Geschl.	*U100_N*	M	1650	850	1450	1150	6150	550	950	
				N	4350							
	KIND	**Offen**	*U100_N*	A	950	850	3150					
				E	1750							
				I	2150	550	3650	2050	2750	6450	5350	6550
				O	2950	6250	950	350	4050	6450	7450	1050
				U	1050	1450	550	3850				
		Geschl.	*U100_N*	M	-							
				N	850	3450						
	MANN	**Geschl.**	BARK_I	M	3425	2925	50	4050				
				N	455	2160	350	4050				
N0123	***FRAU***	**Offen**	*U100_S*	A	2750	1050	250	1250	7950	950	1850	7750
				E	450	250	4550	1150	5250	6550	650	550
				I	4250	1050	3750	5250	450	150	5950	2250
				O	250	450	2250	550	4550	2750	3250	350
				U	2950	350	5550	2050	750	650	2350	6350
		Geschl.	*U100_S*	M	2850	150						
				N	4350	650						
	KIND	**Offen**	*U100_N*	A	550	4750	650	3250				
				E	1750	850	2650	150	1050	1950	2150	
				I	1850	550	4450	1450	1950	150	1250	1650
				O	1550	3050	3750	6850	2250	150	450	
				U	2750	1050	2050	650	4650	150	550	5050
		Geschl.	*U100_N*	M	3750	7050	1450	450	3850			
				N	450							
	MANN	**Geschl.**	*U100_S*	M	2950	2050	7150	5450	3150			
				N	50	7550	1650	6550	5550	6350		

Tab. 36: Parameter Frequenzbandanalyse

Um die Parameter der Frequenzbandanalyse mit den Ergebnissen der Formanten- und Antiformantenanalyse vergleichen zu können, löst Tabelle 37 die Parameter der Frequenzbandanalyse in Sprachmodellparameter auf.

Nasal	Gruppe	Art	Verfahren	Laut	1. Par	2. Par	3. Par	4. Par	5. Par	6. Par	7. Par	8. Par
N01	***FRAU***	**Offen**	*U100_N*	A	F_0	F_1	F_1	4350	F_1	AF_2	F_3	F_4
				E	F_0	F_1	5350	650	550	1150	AF_4	F_3
				I	3750	1050	950	4450	7150	F_0	F_4	1150
				O	F_0	F_1	1750	F_1	AF_1	7550	3250	F_3
				U	2850	AF_1	750	F_2	2350	F_1	AF_4	3150
		Geschl.	*U100_N*	M	F_3	F_0	50	850	7850	7650	F_2	
				N	AF_1	1950	5250	F_4	5850	AF_3	AF_3	F_4
	KIND	**Offen**	*U100_N*	A	AF_1	4750	F_0	5850	2250	AF_3	650	
				E	850	F_2	F_0	1050	5250	F_1	1950	2250
				I	1750	F_1	F_0	F_4	AF_2	650	F_1	750
				O	F_0	1650	AF_1	3150	3750	6850	5750	F_1
				U	F_3	AF_2	AF_2	F_0	1150	5050	F_1	4650
		Geschl.	U100_I	M	7050	1750	F_3	AF_2				
				N	450	AF_4	F_2	1350	6950			
	MANN	**Offen**	*U100_N*	A	F_0	250	F_3	7250	6050	5650	6850	1650
				E	F_0	AF_2	50	F_2	4450	1850	F_2	5850
				I	F_0	AF_4	1550	F_1	1450	1650	6050	450
				O	3950	4350	F_1	1150	F_0			
				U	1050	4650	AF_3	6950	2350	3250	1250	7250
		Geschl.	*U100_N*	M	6050	7650	50	F_1	AF_2	5850		
				N	50							
N123	***FRAU***	**Offen**	*U100_S*	A	AF_3	AF_2	F_3	7950	5550	6850	7250	F_2
				E	4550	6550	F_1	5750	5950	7850		
				I	F_1	5950	4550	7250	5750	1450		
				O	F_1	F_0	2950	4450	7250	5850	AF_3	F_4
				U	AF_2	4450	5550	7250	3750	7450	4750	2050
		Geschl.	*U100_N*	M	1650	AF_2	F_2	1150	6150	F_1	AF_2	
				N	F_4							
	KIND	**Offen**	*U100_N*	A	F_1	F_1	F_3					
				E	1750							
				I	2150	F_1	F_3	2050	F_2	6450	5350	6550
				O	F_3	6250	F_2	AF_1	F_4	6450	7450	F_2
				U	1050	1450	AF_2	3850				
		Geschl.	*U100_N*	M	-							
				N	AF_1	3450						
	MANN	**Geschl.**	BARK_I	M	AF_4	2925	50	F_4				
				N	455	F_2	350	F_4				
N0123	***FRAU***	**Offen**	*U100_S*	A	F_3	F_1	F_0	AF_2	7950	F_1	1850	7750
				E	F_1	F_0	4550	1150	5250	6550	650	550
				I	F_4	1050	AF_4	5250	F_1	F_0	5950	2250
				O	F_0	F_1	2250	F_1	4550	F_3	3250	AF_1
				U	2950	AF_1	5550	2050	750	F_2	2350	6350
		Geschl.	*U100_S*	M	F_3	F_0						
				N	F_4	AF_1						
	KIND	**Offen**	*U100_N*	A	AF_1	4750	650	F_3				
				E	1750	850	F_2	F_0	1050	1950	2150	
				I	1850	F_1	F_4	AF_2	1950	F_0	1250	1650
				O	1550	F_3	3750	6850	2250	F_0	F_1	
				U	F_3	1050	AF_3	AF_2	4650	F_0	AF_2	5050
		Geschl.	*U100_N*	M	3750	7050	F_2	F_1	3850			
				N	450							
	MANN	**Geschl.**	*U100_S*	M	2950	F_3	7150	5450	AF_4			
				N	50	7550	AF_2	6550	5550	6350		

Tab. 37: Parameter Frequenzbandanalyse aufgelöst in Sprachmodell-parameter

Zählt man nun die Sprachmodellparameter der obigen Tabelle pro Klassifikationsaufgabe und Nasalitätsart zusammen, erhält man die aggregierte Tabelle 38. Wie bereits bei der For-

manten- und Antiformantenanalyse zeigt sich die Bedeutung der Grundfrequenz und des 1. Formanten für die Feststellung der Präsenz der offenen Nasalität. In der Klassifikationsaufgabe *N123* verliert die Grundfrequenz an Bedeutung. Dafür treten der 2. und der 3. Formant stärker hervor. Somit lässt sich sagen, dass die Grundfrequenz und der 1. Formant sich gut zur Feststellung der offenen Nasalität eignen, die ersten 3 Formanten zur weiteren Unterteilung. Bzgl. der geschlossenen Nasalität hebt sich kein Parameter im nennenswerten Umfang hervor.

	N01		***N123***		***N0123***		**Σ**	**Σ**
Parameter	**offen**	**geschl.**	**offen**	**geschl.**	**offen**	**geschl.**	**offen**	**geschl.**
F_0	13	1	1	0	8	1	22	2
F_1	14	1	6	1	8	1	28	3
F_2	4	2	4	2	2	1	10	5
F_3	5	2	4	0	5	2	14	4
F_4	3	2	2	3	2	1	7	6
AF_1	4	1	1	1	3	1	8	3
AF_2	5	2	3	2	4	1	12	5
AF_3	2	2	2	0	1	0	5	2
AF_4	3	1	0	1	1	1	4	3

Tab. 38: Frequenzbandanalyse: absolute Häufigkeit der Parameter (laut- und gruppenunabhängig)

8.4.3 Zusammenfassung

Folgende Tabelle 39 fasst die Klassifizierungsergebnisse zusammen. Es wird jeweils das Verfahren angegeben, mit denen die beste Klassifizierung über alle Laute erzielt wurde. In der Spalte „Laut“ wird der Laut angegeben, mit dem die höchste Klassifizierung erzielt wurde. Die besten Parameter werden aus Übersichtlichkeitsgründen nicht mit angegeben. Sie können der Parametertabelle entnommen werden.

Festzuhalten ist, dass mit der Frequenzbandanalyse die besten Erkennungen in dieser Arbeit sowohl für das offene als auch für das geschlossene Näseln erreicht wurden. Sie zeigt sich der Analyse mittels den Sprachmodellparametern überlegen. Insgesamt lässt sich die Nasalität mit der Frequenzbandanalyse für bestimmte Laute mit sehr hohen Güten quantifizieren.

		Offen				Geschlossen			
	Gruppe	**Bestes Verfahren**	**Güte**	**Bester Laut**	**Güte**	**Bestes Verfahren**	**Güte**	**Bester Laut**	**Güte**
N01	*FRAU*	*U100_N*	92.6%	A	97.5%	*U100_N*	92.4%	M	93.2%
	KIND	*U100_N*	82.7%	I	90.7%	U100_I	80.4%	N	82.4%
	MANN	*U100_N*	100%	Alle	100%	*U100_N*	95.5%	M	100%
N123	*FRAU*	*U100_S*	72.6%	U	78.2%	*U100_N*	77.6%	M	96.6%
	KIND	*U100_N*	86.7%	O	100%	*U100_N*	86.7%	N	86.7%
	MANN	-	-	-	-	*BARK_I*	62.5%	M	100%
N0123	*FRAU*	*U100_S*	75.4%	A	84.1%	*U100_S*	46.7%	M	55.9%
	KIND	*U100_N*	75.8%	E	85.1%	*U100_N*	67.6%	M	71.6%
	MANN	-	-			*U100_S*	95.5%	M	100%
N01 * N123	*FRAU*	*U100_S*	82.1%	A	85.9%	*U100_N*	82.2%	M	91.7%
	KIND	*U100_N*	81.4%	I	87.4%	*U100_N*	70.0%	N	76.4%
	MANN	-	-	-	-	*BARK_I*	68.8%	M	90.9%

Tab. 39: Zusammenfassung Klassifikation Frequenzbandanalyse

9 Zusammenfassung und Ausblick

9.1 Zusammenfassung

In der vorliegenden Arbeit wurde ein sprachgestütztes Trainingssystem zur Evaluierung der Nasalität aus Mikrophonaufnahmen konzipiert und prototypisch implementiert. Das Trainingssystem soll bei Patienten mit stark nasaler Aussprache, insbesondere bei Lippen-Kiefer-Gaumenspaltträgern (LKG), Einsatz finden. Die wichtigste Voraussetzung für dieses Vorhaben ist somit die automatische Bestimmung der Nasalität mit einer hohen Güte. Daher konzentrierte sich die Arbeit auf die grundsätzliche Fragestellung, ob und wie aus kurzen Sprachsequenzen oder Lauten auf Sprechstörungen geschlossen werden kann.

Da für den deutschen Sprachraum keine geeigneten Sprachaufnahmen vorhanden waren, wurde die Sprachdatenbank NASAL mit verschiedenen Sprechergruppen, Passagen und Nasalitätsausprägungen aufgebaut. Die Sprachaufnahmen basieren auf dem standardisierten Heidelberger Rhinophoniebogen, der speziell auf das hier zu untersuchende Problem angepasst wurde. Sie wurden hinsichtlich der Nasalitätsausprägung auf einer 4-stufigen Skala logopädisch evaluiert. Die Datenbank enthält aktuell 2468 Sprachpassagen von 116 Sprechern.

Für die Quantifizierung der Nasalität musste die prinzipielle Vorgehensweise unter den gegebenen Rahmenbedingungen geklärt werden. Einerseits sollte die Erkennung online möglich sein, also der zu bearbeitende Datensatz klein sein, auf der anderen Seite ist bei der hier zu untersuchenden Sprechergruppe mit einer hohen sprecherspezifischen Variabilität zu rechnen. Damit stellt sich hier ein hoch komplexes System dar, das grundlegende analytische Fragestellungen aufwarf und zunächst die Erarbeitung eines geeigneten Verfahrens voraussetzte.

Demzufolge erfolgten auf der Grundlage der Sprachdatenbank umfangreiche Untersuchungen zur Quantifizierung der Nasalität. Als ein wesentliches Ergebnis dieser Untersuchungen wird ein automatisches Verfahren zur Quantifizierung der Nasalität mit einer hohen Güte vorgestellt. Dieses beruht auf den Frequenzbandintensitäten und Sprachmodellparametern stimmhafter Laute. Bei den Sprachmodellparametern handelt es sich jeweils um die Lage, Intensität und Bandbreite der Grundfrequenz der ersten 4 Formanten sowie der ersten 4 Antiformanten der Vokale und Nasale. Die offene Nasalität wird dabei aus Vokalen, die geschlossene aus den Nasalen bestimmt. Um Aussagen über die beeinflussenden Faktoren der Nasalität gewinnen zu können, wurde als Klassifizierungsverfahren die Lineare Diskriminanzanalyse gewählt. Hierbei handelt es sich um ein statistisches Verfahren, das sich sehr

gut bei Fragestellungen mit bekannter Gruppenzugehörigkeit eignet. Ein für die Praxis wichtiger Vorteil statistischer Verfahren ist, dass sich der Einfluss der einzelnen Variablen auf die abhängige Variable ermitteln lässt (im Gegensatz zu Neuronalen Netzen).

Für den praktischen Einsatz des entwickelten Verfahrens zur Nasalitätsbewertung wurde eine sprachgestützte Trainingsumgebung prototpyisch implementiert. Diese besteht aus den Komponenten „Spracherkennung“, „Sprachdatenbank“, „Eingangsuntersuchung“ und „Trainingsmodul“. In der Eingangsuntersuchung wird der Patient zum Nachsprechen des Heidelberger Rhinophoniebogens aufgefordert. Die über das Mikrophon aufgenommenen Sprachdaten werden bzgl. der Spracherkennungsgüte und Nasalitätsausprägung bewertet und zusammen mit den Bewertungen in der Sprachdatenbank NASAL gespeichert. Die Bedienung des Trainingsmoduls erfolgt über gesprochene Navigationsbefehle. Die hierbei verwendete Spracherkennungstechnologie basiert auf dem kommerziellen Produkt „ViaVoice“ der Firma IBM. Um den Einsatz einer Spracherkennung bei sprechgestörter Aussprache zu eruieren, erfolgen Untersuchungen zur Spracherkennungsgüte. Diese verdeutlichten die Komplexität bei der Erkennung. Die Güte der kommerziellen Sprach-erkennung bei sprechgestörter Aussprache war sehr niedrig und hatte eine geringe Trefferquote.

9.2 Ausblick

In der vorliegenden Arbeit wurden vielfältige Aspekte der spektralen Nasalitätsmessung betrachtet. Ein wesentliches Ergebnis ist der Aufbau der Sprachdatenbank NASAL. Diese enthält Sprachaufnahmen aus dem deutschen Sprachraum von verschiedenen Sprecher-gruppen, Passagen und Nasalitätsausprägungen. Aufbauend auf dieser Sprachdatenbank erfolgten umfangreiche Untersuchungen der spektralen Verfahren und Parameter zur Quan-tifizierung der Nasalität. Es konnte gezeigt werden, dass die Erkennung und die 3-stufige Quantifizierung der offenen und geschlossenen Nasalität mit hohen Erkennungsraten für Frauen und Kinder realisierbar sind. Für die Sprechergruppe der Männer kann aufgrund der geringen Datenbasis (lediglich eine nasalen Ausprägung der Vokale) nur das Vorhandensein einer offenen Nasalität festgestellt werden.

Insgesamt ist aber auch klar geworden, dass aufgrund der Breite der klinischen Fälle die automatische Bestimmung der Nasalität nicht bei allen Patienten erfolgreich durchgeführt werden kann. So zeigen einige Patienten neben der nasalen Aussprache noch weitere über-lagerte Sprechstörungen, die zu diesem Zeitpunkt noch nicht berücksichtigt werden konnten. Eine Computeranalyse für diese Patienten ist daher mit großer Unsicherheit behaftet und zum jetzigen Zeitpunkt nicht angebracht. Um auch für solche Gruppen charakteristische

Aussagen machen zu können, sollte ein zusätzlicher Klassifikator untersucht werden, der mit vagen Informationen umgehen kann. Interessant wäre die Kombination uniformer Frequenzbandparameter als Parameter mit einem Neuronalen Netz als Klassifikator.

Der Zweck der Instrumentalisierung der Nasalitätsmessung ist die Verbesserung der Genauigkeit und Wiederholbarkeit in schwierigen, diagnostischen Behandlungssituationen. Sollen Instrumente zur Unterstützung des behandelnden Therapeuten benutzt werden, ist die Übereinstimmung der Messergebnisse mit dem Höreindruck des Therapeuten ein wichtiges Kriterium. Daher kommt dem Test und der Optimierung der Nasalitätsbestimmung im klinischen Einsatz eine große Bedeutung zu. Zur Optimierung ist an erster Stelle der Ausbau der Sprachdatenbank NASAL mit den Daten des standardisierten Heidelberger Rhinophoniebogens zu nennen. Dies garantiert die Vergleichbarkeit der Datensätze verschiedener Patienten und verbessert im Laufe der Zeit die statistischen Aussagen über die entsprechenden Normwerte nasaler und nonnasaler Aussprache.

Ein weiterer beachtenswerter Punkt ist die Subjektivität der logopädischen Einschätzung. Insbesondere, da diese die Voraussetzung für den Klassifizierungsvergleich ist. Eine Steigerung der Signifikanz der Ergebnisse ließe sich durch eine Kreuzvalidierung der perzeptuellen Klassifizierungen mehrerer Logopädenbewertungen erreichen.

Die Nasalitätsbestimmung basiert im Rahmen dieser Arbeit auf isoliert gesprochenen stimmhaften Lauten. Diese spiegeln den alltäglichen Gebrauch der Stimme jedoch nur ungenügend wieder und lassen nur bedingt Rückschlüsse auf die letztendlich im Alltag relevante Stimmqualität zu (Spontansprache). Zudem wird die Phonation isolierter Laute von vielen Personen als anstrengender empfunden, als das freie Vorlesen eines Textes. Diese Gründe legen nahe, die Nasalitätsbestimmung auf die Sätze der Sprachdatenbank NASAL auszudehnen.

Der Einsatz in medizinischen Anwendungen erfordert für eine hohe Benutzerakzeptanz neben der sehr hohen Erkennungsrate auch eine effektive Dialoggestaltung. Inwieweit ein Einsatz einer Sprachsteuerung bei pathologischen Stimmstörungen und unter welchen Einschränkungen (z. B. Vokabulargröße) sinnvoll ist, muss genauer untersucht werden. Da sich kommerzielle Systeme aufgrund der geringen Flexibilität bei ihrer Anpassung für diese Aufgabenstellung nicht eignen, ist eine eigene Spracherkennung zu implementieren. Diese kann dann insbesondere mit den nasalen Daten der Datenbank NASAL trainiert werden und sollte daher erheblich bessere Erkennungsraten ergeben. Eine Anpassung an den Benutzer könnte dabei bei den Eingangsuntersuchungen erfolgen. Um mit den teilweise schwer verständlichen Sprachdaten umgehen zu können, sollte wie bei der Nasalitätsbewertung als zusätzlicher Klassifikator ein neuronales Netz mit den Frequenzbandintensitäten als Eingabeparameter für die Spracherkennung benutzt werden.

E Anhang - Parametertabellen

E.1 Grundfrequenz und Formanten

E.1.1 Klassifikation „*N01*“

Laut	Methode	Parameter 1	2	3	4	5	6	7	8
A	*GF*	F_0_I	F_0_B						
	FREQ	F_1_F	F_3_F						
	INT	F_1_I	F_3_I						
	BAND	F_2_B							
	ALLE	F_0_I	F_1_F	F_3_F	F_1_I	F_0_F			
E	*GF*	F_0_I	F_0_B						
	FREQ	F_1_F	F_3_F						
	INT	F_4_I	F_1_I						
	BAND	F_3_B							
	ALLE	F_1_F	F_0_F	F_3_F	F_4_I	F_1_I	F_2_I		
I	*GF*	F_0_I							
	FREQ	F_3_F	F_2_F						
	INT	F_3_I	F_1_I						
	BAND	F_2_B							
	ALLE	GES_I	F_0_F	F_3_F					
O	*GF*	F_0_I							
	FREQ	F_2_F							
	INT	F_1_I							
	BAND	F_1_B	F_4_B						
	ALLE	F_0_I	F_2_F	F_1_B					
U	*GF*	F_0_I	F_0_F						
	FREQ	F_4_F	F_1_F	F_2_F					
	INT	F_3_I	F_2_I	F_1_I					
	BAND	F_2_B	F_3_B	F_1_B					
	ALLE	F_3_I	F_2_I	F_1_F	F_1_B	F_4_F			
M	*GF*	F_0_F	F_0_I						
	FREQ	F_1_F							
	INT	F_1_I	F_4_I						
	BAND	F_3_B							
	ALLE	F_1_F	F_0_I						
N	*GF*	--							
	FREQ	--							
	INT	--							
	BAND	F_4_B							
	ALLE	F_4_B							

Tab. 40: Formantenanalyse: *FRAU*, *N01*

Laut	Methode	Parameter							
		1	2	3	4	5	6	7	8
A	*GF*	--							
	FREQ	F_3_F	F_1_F						
	INT	F_2_I	F_4_I						
	BAND	F_2_B							
	ALLE	F_3_F	F_1_F						
E	*GF*	F_0_F	F_0_I						
	FREQ	F_2_F	F_1_F						
	INT	F_1_I	F_2_I						
	BAND	F_1_B							
	ALLE	F_1_B	GES_I	F_1_I					
I	*GF*	--							
	FREQ	F_1_F							
	INT	--							
	BAND	F_2_B	F_4_B						
	ALLE	F_2_B	F_1_F	F_4_B					
O	*GF*	F_0_F	F_0_I						
	FREQ	--							
	INT	--							
	BAND	F_1_B	F_3_B						
	ALLE	F_1_B	GES_I	F_2_F					
U	*GF*	--							
	FREQ	F_1_F							
	INT	F_2_I	F_1_I						
	BAND	F_4_B							
	ALLE	F_1_F	F_4_B	F_2_I	F_1_B				
M	*GF*	--							
	FREQ	--							
	INT	--							
	BAND	F_3_B							
	ALLE	F_3_B							
N	*GF*	F_0_I							
	FREQ	--							
	INT	--							
	BAND	F_4_B							
	ALLE	F_4_B	F_3_I						

Tab. 41: Formantenanalyse: KIND, N01

Laut	Methode	Parameter 1	2	3	4	5	6	7	8
A	*GF*	F_0_I							
	FREQ	F_3_F	F_4_F	F_1_F					
	INT	F_1_I	F_3_I						
	BAND	F_2_B							
	ALLE	GES_I	F_4_F	F_3_F					
E	*GF*	F_0_B							
	FREQ	F_1_F							
	INT	F_3_I							
	BAND	--							
	ALLE	F_0_B	F_1_F	F_2_F					
I	*GF*	F_0_I							
	FREQ	F_2_F							
	INT	F_3_I							
	BAND	F_2_B							
	ALLE	F_2_B	F_2_F	F_4_B					
O	*GF*	F_0_I	F_0_B						
	FREQ	F_2_F	F_3_F						
	INT	F_1_I							
	BAND	F_2_B							
	ALLE	F_2_F	F_3_F	GES_I	F_2_B	F_4_I			
U	*GF*	F_0_I							
	FREQ	F_2_F	F_1_F						
	INT	F_2_I							
	BAND	F_4_B							
	ALLE	F_2_F	GES_I	F_2_I					
M	*GF*	--							
	FREQ	--							
	INT	--							
	BAND	--							
	ALLE	--							
N	*GF*	--							
	FREQ	--							
	INT	--							
	BAND	F_1_B							
	ALLE	F_1_B							

Tab. 42: Formantenanalyse: *MANN, N01*

E.1.2 Klassifikation „*N123*“

Laut	Methode	Parameter							
		1	2	3	4	5	6	7	8
A	*GF*	F_0_F							
	FREQ	F_1_F							
	INT	--							
	BAND	--							
	ALLE	F_0_F	F_1_B						
E	*GF*	--							
	FREQ	F_1_F							
	INT	F_1_I	F_2_I						
	BAND	--							
	ALLE	F_1_I	GES_I						
I	*GF*	F_0_F							
	FREQ	F_1_F							
	INT	F_1_I							
	BAND	--							
	ALLE	F_1_I							
O	*GF*	F_0_I	F_0_F						
	FREQ	--							
	INT	F_1_I							
	BAND	F_1_B							
	ALLE	F_0_I	F_0_F	F_2_I					
U	*GF*	--							
	FREQ	--							
	INT	F_1_I							
	BAND	--							
	ALLE	F_1_I							
M	*GF*	--							
	FREQ	--							
	INT	--							
	BAND	--							
	ALLE	--							
N	*GF*	F_0_I							
	FREQ	--							
	INT	F_1_I							
	BAND	F_3_B							
	ALLE	F_0_I	F_1_B						

Tab. 43: Formantenanalyse: *FRAU*, *N123*

Laut	Methode	Parameter							
		1	2	3	4	5	6	7	8
A	*GF*	--							
	FREQ	F_4_F	F_3_F						
	INT	--							
	BAND	--							
	ALLE	F_4_F	F_3_F						
E	*GF*	--							
	FREQ	F_2_F							
	INT	--							
	BAND	--							
	ALLE	F_2_F							
I	*GF*	--							
	FREQ	F_2_F	F_1_F						
	INT	--							
	BAND	--							
	ALLE	F_2_F	F_1_F						
O	*GF*	--							
	FREQ	F_2_F	F_3_F						
	INT	--							
	BAND	--							
	ALLE	F_2_F	F_3_F						
U	*GF*	--							
	FREQ	--							
	INT	--							
	BAND	--							
	ALLE	--							
M	*GF*	--							
	FREQ	--							
	INT	--							
	BAND	--							
	ALLE	--							
N	*GF*	--							
	FREQ	--							
	INT	--							
	BAND	--							
	ALLE	--							

Tab. 44: Formantenanalyse: *KIND*, *N123*

Laut	Methode	Parameter							
		1	2	3	4	5	6	7	8
A	*GF*	Es gibt nur eine nicht-leere Gruppe und 6.000 (6 ungewichtete) Fälle, die gültig sind.							
E	*FREQ*								
I	*INT*	Nicht genügend nicht-leere Gruppen.							
O	*BAND*								
U	*ALLE*								
M	*GF* *FREQ* *INT* *BAND* *ALLE*								
N	*GF* *FREQ* *INT* *BAND* *ALLE*								

Tab. 45: Formantenanalyse: *MANN, N123*

E.1.3 Klassifikation „*N0123*“

Laut	Methode	Parameter							
		1	2	3	4	5	6	7	8
A	*GF*	F_0_I							
	FREQ	F_1_F	F_3_F						
	INT	F_1_I							
	BAND								
	ALLE	F_1_F	F_0_I	F_3_F					
E	*GF*	F_0_I							
	FREQ	F_1_F	F_3_F						
	INT	F_1_I	F_4_I						
	BAND								
	ALLE	F_1_F	F_0_F	F_3_F	F_4_I	F_1_I			
I	*GF*	F_0_I							
	FREQ	F_3_F	F_2_F	F_1_F					
	INT	F_3_I	F_1_I						
	BAND								
	ALLE	GES_I	F_0_F	F_3_F					
O	*GF*	F_0_I							
	FREQ	F_2_F							
	INT	F_1_I							
	BAND	F_1_B							
	ALLE	F_0_I	F_1_B	F_2_F					
U	*GF*								
	FREQ								
	INT	F_3_I	F_2_I	F_1_I					
	BAND	F_2_B							
	ALLE	F_3_I	F_1_I	F_1_F	F_2_I				
M	*GF*	F_0_F							
	FREQ	F_1_F							
	INT	F_1_I							
	BAND	F_3_B							
	ALLE	F_1_I							
N	*GF*	F_0_I							
	FREQ								
	INT	F_1_I							
	BAND								
	ALLE	F_0_I	F_1_I						

Tab. 46: Formantenanalyse: *FRAU*, *N0123*

Laut	Methode	Parameter 1	2	3	4	5	6	7	8
A	*GF*								
	FREQ	F_3_F	F_4_F	F_1_F					
	INT	F_2_I	F_4_I						
	BAND								
	ALLE	F_3_F	F_4_F	F_2_I	F_4_I				
E	*GF*	F_0_F	F_0_I						
	FREQ	F_2_F							
	INT	F_1_I	F_2_I						
	BAND	F_1_B							
	ALLE	F_2_F	F_1_B	F_0_I					
I	*GF*								
	FREQ	F_2_F	F_1_F						
	INT								
	BAND	F_2_B							
	ALLE	F_2_F	F_1_F	F_2_B					
O	*GF*								
	FREQ	F_2_F							
	INT								
	BAND	F_1_B							
	ALLE	F_1_B	F_2_F	GES_I					
U	*GF*								
	FREQ	F_1_F							
	INT	F_2_I							
	BAND								
	ALLE	F_2_I	F_4_B	F_1_F					
M	*GF*								
	FREQ								
	INT								
	BAND								
	ALLE								
N	*GF*								
	FREQ								
	INT								
	BAND	F_4_B							
	ALLE	F_4_B							

Tab. 47: Formantenanalyse: *KIND, N0123*

Laut	Methode	Parameter 1	2	3	4	5	6	7	8
A	*GF*	F_0_I							
	FREQ	F_3_F	F_4_F	F_1_F					
	INT	F_1_I	F_3_I						
	BAND	F_2_B							
	ALLE	GES_I	F_4_F	F_3_F					
E	*GF*	F_0_B							
	FREQ	F_1_F							
	INT	F_3_I							
	BAND								
	ALLE	F_0_B	F_1_F	F_2_F					
I	*GF*	F_0_I							
	FREQ	F_2_F							
	INT	F_3_I							
	BAND	F_2_B							
	ALLE	F_2_B	F_2_F	F_4_B					
O	*GF*	F_0_I	F_0_B						
	FREQ	F_2_F	F_3_F						
	INT	F_1_I							
	BAND	F_2_B							
	ALLE	F_2_F	F_3_F	GES_I	F_2_B	F_4_I			
U	*GF*	F_0_I							
	FREQ	F_2_F	F_1_F						
	INT	F_2_I							
	BAND	F_4_B							
	ALLE	F_2_F	GES_I	F_2_I					
M	*GF*								
	FREQ	F_3_F							
	INT								
	BAND								
	ALLE	F_3_F							
N	*GF*								
	FREQ								
	INT								
	BAND								
	ALLE								

Tab. 48: Formantenanalyse: *MANN, N0123*

E.2 Antiformanten

E.2.1 Klassifikation „*N01*“

Laut	Methode	Parameter							
		1	2	3	4	5	6	7	8
A	*FREQ*	AF_3_F	AF_2_F						
	INT	AF_3_I	AF_2_I						
	BAND	AF_1_B	AF_2_B	AF_3_B					
	ALLE	AF_3_F	AF_2_F	AF_3_B	AF_4_I	AF_1_B	AF_1_F	AF_2_B	
E	*FREQ*	AF_1_F	AF_2_F						
	INT	AF_1_I	AF_3_I						
	BAND	AF_1_B	AF_3_B						
	ALLE	AF_1_B	AF_2_F	AF_2_B					
I	*FREQ*	AF_2_F	AF_4_F						
	INT	AF_2_I	AF_4_I						
	BAND	AF_3_B	AF_1_B						
	ALLE	AF_2_I	AF_2_F	AF_4_I	AF_3_B	AF_4_F			
O	*FREQ*	AF_2_F							
	INT	AF_3_I	AF_2_I	AF_1_I					
	BAND	AF_2_B	AF_3_B	AF_4_B	AF_1_B				
	ALLE	AF_3_I	AF_2_I						
U	*FREQ*	AF_4_F	AF_3_F						
	INT	AF_3_I	AF_2_I						
	BAND	AF_3_B	AF_2_B						
	ALLE	AF_3_I	AF_2_I	AF_4_F	AF_3_F	AF_2_F			
M	*FREQ*	AF_4_F							
	INT	AF_3_I	AF_4_I						
	BAND	AF_3_B							
	ALLE	AF_3_B	AF_4_F	AF_3_I	AF_4_I	AF_2_F	AF_2_I		
N	*FREQ*	AF_4_F							
	INT	AF_3_I							
	BAND	--							
	ALLE	AF_4_F	AF_3_I	AF_2_I	AF_4_I				

Tab. 49: Antiformantenanalyse: *FRAU*, *N01*

Laut	Methode	Parameter							
		1	2	3	4	5	6	7	8
A	*FREQ*	AF_4_F	AF_2_F						
	INT	AF_4_I	AF_2_I	AF_1_I					
	BAND	--							
	ALLE	AF_4_F	AF_2_F	AF_4_I					
E	*FREQ*	AF_3_F	AF_2_F						
	INT	AF_3_I	AF_2_I	AF_1_I					
	BAND	--							
	ALLE	AF_3_I	AF_2_I						
I	*FREQ*	AF_1_F							
	INT	AF_2_I	AF_4_I	AF_1_I	AF_3_I				
	BAND	AF_2_B							
	ALLE	AF_2_I	AF_1_B	AF_3_I	AF_3_F	AF_4_I			
O	*FREQ*	AF_2_F	AF_3_F						
	INT	AF_3_I							
	BAND	--							
	ALLE	AF_2_F	AF_3_I	AF_3_F					
U	*FREQ*	AF_1_F							
	INT	AF_1_I							
	BAND	AF_3_B							
	ALLE	AF_3_B	AF_3_I						
M	*FREQ*	AF_1_F							
	INT	AF_1_I							
	BAND	AF_4_B							
	ALLE	AF_1_I	AF_4_B						
N	*FREQ*	--							
	INT	AF_3_I							
	BAND	--							
	ALLE	AF_3_I							

Tab. 50: Antiformantenanalyse: *KIND, N01*

Laut	Methode	Parameter							
		1	2	3	4	5	6	7	8
A	*FREQ*	AF_3_F							
	INT	AF_3_I							
	BAND	--							
	ALLE	AF_3_F	AF_1_I						
E	*FREQ*	--							
	INT	AF_3_I							
	BAND	--							
	ALLE	AF_3_I	AF_3_F						
I	*FREQ*	--							
	INT	AF_2_I	AF_3_I						
	BAND	--							
	ALLE	AF_2_I							
O	*FREQ*	AF_3_F	AF_4_F	AF_2_F					
	INT	AF_2_I	AF_3_I						
	BAND	AF_2_B	AF_3_B	AF_4_B					
	ALLE	AF_2_I	AF_4_F	AF_3_F	AF_2_B	AF_3_B			
U	*FREQ*	AF_2_F							
	INT	AF_3_I	AF_2_I						
	BAND	--							
	ALLE	AF_3_I	AF_2_I	AF_4_I					
M	*FREQ*	--							
	INT	--							
	BAND	--							
	ALLE	--							
N	*FREQ*	--							
	INT	--							
	BAND	--							
	ALLE	--							

Tab. 51: Antiformantenanalyse: *MANN, N01*

E.2.2 Klassifikation „*N123*“

Laut	Methode	Parameter							
		1	2	3	4	5	6	7	8
A	*FREQ*	--							
	INT	--							
	BAND	--							
	ALLE	--							
E	*FREQ*	AF_2_F	AF_1_F						
	INT	AF_1_I							
	BAND	AF_4_B							
	ALLE	AF_1_I	AF_2_F						
I	*FREQ*	AF_2_F	AF_1_F						
	INT	AF_2_I							
	BAND	AF_3_B							
	ALLE	AF_2_I	AF_2_F	AF_3_B					
O	*FREQ*	--							
	INT	--							
	BAND	--							
	ALLE								
U	*FREQ*	--							
	INT	--							
	BAND	--							
	ALLE	--							
M	*FREQ*	--							
	INT	--							
	BAND	--							
	ALLE	--							
N	*FREQ*	--							
	INT	AF_2_I							
	BAND	--							
	ALLE	AF_2_I							

Tab. 52: Antiformantenanalyse: *FRAU*, *N123*

Laut	Methode	Parameter							
		1	2	3	4	5	6	7	8
A	*FREQ*	--							
	INT	AF_4_I	AF_3_I						
	BAND	--							
	ALLE	AF_4_I	AF_3_I	AF_4_F	AF_1_B	AF_2_E			
E	*FREQ*	AF_2_F							
	INT	AF_2_I							
	BAND	--							
	ALLE	AF_2_I							
I	*FREQ*	AF_3_F							
	INT	--							
	BAND	--							
	ALLE	AF_3_F							
O	*FREQ*	AF_4_F	AF_2_F						
	INT	--							
	BAND	--							
	ALLE	AF_4_F	AF_2_F						
U	*FREQ*	--							
	INT	--							
	BAND	AF_4_B							
	ALLE	AF_4_B	AF_2_I						
M	*FREQ*	--							
	INT	--							
	BAND	--							
	ALLE	--							
N	*FREQ*	--							
	INT	--							
	BAND	--							
	ALLE	--							

Tab. 53: Antiformantenanalyse: *KIND*, *N123*

Laut	Methode	Parameter							
		1	2	3	4	5	6	7	8
A E I O U	*FREQ* *INT* *BAND* *ALLE*	Es gibt nur eine nicht-leere Gruppe und 6 ungewichtete Fälle, die gültig sind. Nicht genügend nicht-leere Gruppen.							
M	*FREQ*	--							
	INT	--							
	BAND	--							
	ALLE	--							
N	*FREQ*	--							
	INT	--							
	BAND	--							
	ALLE	--							

Tab. 54: Antiformantenanalyse: *MANN, N123*

E.2.3 Klassifikation „*N0123*“

Laut	Methode	Parameter							
		1	2	3	4	5	6	7	8
A	*FREQ*	AF_3_F	AF_2_F						
	INT	AF_3_I	AF_4_I						
	BAND	AF_1_B	AF_2_B						
	ALLE	AF_3_F	AF_2_F	AF_3_B	AF_2_I				
E	*FREQ*	AF_1_F	AF_2_F						
	INT	AF_1_I	AF_3_I						
	BAND	AF_1_B							
	ALLE	AF_1_F	AF_2_F						
I	*FREQ*	AF_2_F	AF_1_F						
	INT	AF_2_I	AF_4_I						
	BAND	AF_3_B							
	ALLE	AF_2_I	AF_2_F	AF_3_B					
O	*FREQ*	AF_2_F							
	INT	AF_3_I	AF_1_I						
	BAND	AF_2_B	AF_3_B						
	ALLE	AF_3_I	AF_4_I						
U	*FREQ*	AF_4_F	AF_3_F						
	INT	AF_3_I	AF_2_I						
	BAND	--							
	ALLE	AF_3_I	AF_2_I	AF_3_F	AF_4_F				
M	*FREQ*	--							
	INT	--							
	BAND	--							
	ALLE	--							
N	*FREQ*	--							
	INT	--							
	BAND	--							
	ALLE	--							

Tab. 55: Antiformantenanalyse: *FRAU*, *N0123*

Laut	Methode	Parameter 1	2	3	4	5	6	7	8
A	*FREQ*	AF_4_F							
	INT	AF_4_I	AF_2_I						
	BAND	--							
	ALLE	AF_4_I	AF_2_I	AF_4_F					
E	*FREQ*	AF_3_F	AF_2_F						
	INT	AF_3_I	AF_2_I						
	BAND	--							
	ALLE	AF_3_I	AF_2_I						
I	*FREQ*	AF_1_F	AF_3_F						
	INT	AF_4_I	AF_3_I	AF_2_I	AF_1_I				
	BAND	AF_2_B	AF_4_E						
	ALLE	AF_4_I	AF_3_I	AF_2_I	AF_1_F	AF_3_F			
O	*FREQ*	AF_2_F	AF_4_F						
	INT	--							
	BAND	--							
	ALLE	AF_2_F	AF_4_F	AF_3_I					
U	*FREQ*	AF_1_F							
	INT	AF_1_I							
	BAND	AF_3_B	AF_2_E	AF_4_B	AF_1_E				
	ALLE	AF_3_B	AF_2_E	AF_4_B	AF_1_E				
M	*FREQ*	--							
	INT	AF_1_I							
	BAND	--							
	ALLE	AF_1_I							
N	*FREQ*	--							
	INT	--							
	BAND	--							
	ALLE	--							

Tab. 56: Antiformantenanalyse: *KIND*, *N0123*

Laut	Methode	Parameter							
		1	2	3	4	5	6	7	8
A	*FREQ*	AF_3_F							
	INT	AF_3_I							
	BAND	--							
	ALLE	AF_3_F	AF_1_I						
E	*FREQ*	--							
	INT	AF_3_I							
	BAND	--							
	ALLE	AF_3_I	AF_3_F						
I	*FREQ*	--							
	INT	AF_2_I	AF_3_I						
	BAND	--							
	ALLE	AF_2_I							
O	*FREQ*	AF_3_F	AF_4_F	AF_2_F					
	INT	AF_2_I	AF_3_I						
	BAND	AF_2_E	AF_3_E	AF_4_B					
	ALLE	AF_2_I	AF_4_F	AF_3_F	AF_2_E	AF_3_E			
U	*FREQ*	AF_2_F							
	INT	AF_3_I	AF_2_I						
	BAND	--							
	ALLE	AF_3_I	AF_2_I	AF_4_I					
M	*FREQ*	--							
	INT	--							
	BAND	--							
	ALLE	--							
N	*FREQ*	--							
	INT	--							
	BAND	--							
	ALLE	--							

Tab. 57: Antiformantenanalyse: *MANN, N0123*

E.3 Frequenzbänder

E.3.1 Klassifikation „*N01*“

Laut	Methode	Parameter							
		1	2	3	4	5	6	7	8
A	*U100_N*	250	1050	950	4350	850	1250	2950	4050
	U100_I	1050	250	950	7950	7750	450	4050	850
	U100_S	2750	250	1050	1950	4350	7050	850	3850
	U400_N	1000	2600	2200	4600	3000	3800	1800	
	U400_I	1000	2200	4600	7000	2600	6600	3800	3000
	U400_S	2600	1000	2200	3000	4600	3800	1800	7000
	BARK_N	1000	250	845	2925	1375	3425	2510	2160
	BARK_I	1000	250	845	1375	2925	3425	700	4850
	BARK_S	2925	1000	250	2510	2160	4850	4050	1860
E	*U100_N*	250	450	5350	650	550	1150	3550	3050
	U100_I	450	5350	650	7450	7550	250	550	1150
	U100_S	250	450	5350	1150	3550	7450	4050	4250
	U400_N	600	1000	5400	3800	3000	3400		
	U400_I	600	1000	5400					
	U400_S	3400	1000	600	5400				
	BARK_N	250	455	700	1175	570	350	4050	2925
	BARK_I	455	250	4850	700	570	1175		
	BARK_S	250	455	1175	4850	350	570	845	
I	*U100_N*	3750	1050	950	4450	7150	150	4350	1150
	U100_I	3750	950	1050	4450	5250	150	7150	1150
	U100_S	4250	1050	3650	5250	150	950	7150	
	U400_N	3800	1000	5400	4600	4200	3000	7000	
	U400_I	3800	1000	5400	4600				
	U400_S	4200	1000	5400	3800	3000	3400		
	BARK_N	1000	4050	1175	4850	150	7050	700	
	BARK_I	1000	4050	1175	4850	250	700	3425	
	BARK_S	4050	1000	4850	150	570	700	7050	
O	*U100_N*	250	450	1750	550	350	7550	3250	2750
	U100_I	450	350	550	1750	5750	3250	2750	7550
	U100_S	250	450	2250	550	3450	2750	350	1750
	U400_N	600	1400	3400	6600	7400			
	U400_I	600	1400	7400	2600	2200	3400	6600	6200
	U400_S	2200	600	2600	1400				
	BARK_N	250	455	570	1860	350	3425	150	4050
	BARK_I	455	350	570	5850	3425	4050	1375	
	BARK_S	250	455	2160	570	3425	350	150	4050
U	*U100_N*	2850	350	750	650	2350	450	950	3150
	U100_I	2050	650	550	2950	350	5450	2350	6350
	U100_S	350	750	2250	650	1050	850	2850	6350
	U400_N	3000	600	2200	200	6200	5400	5000	3800
	U400_I	2200	600	3000	3800	6200	5400		
	U400_S	3000	600	2200	6200	3800	200		
	BARK_N	2160	700	570	2925	350	455	1000	845
	BARK_I	2160	700	570	2925	350	1000	455	
	BARK_S	2925	700	350	2160	570	7050	1000	845
M	*U100_N*	2850	150	50	850	7850	7650	1550	
	U100_I	2850	250	850					
	U100_S	2850	5550	7850	850	50			
	U400_N	3000	5800	200					
	U400_I	3000	5800						
	U400_S	3000	5800	200					
	BARK_N	570	150	2925	2160	1860	50		
	BARK_I	250	2925	570					
	BARK_S	150	2925	3425					
N	*U100_N*	650	1950	5250	4350	5850	2650	2450	4150
	U100_I	1950	650	2750	2950	2350			
	U100_S	650	1950	2350	550	2650			
	U400_N	600	2200						
	U400_I	2200	600						
	U400_S	2200	600						
	BARK_N	1860	700	570	1000				
	BARK_I	1860	700						
	BARK_S	700	2160						

Tab. 58: Frequenzbandanalyse: *FRAU*, *N01*

Laut	Methode	Parameter							
		1	2	3	4	5	6	7	8
A	*U100_N*	550	4750	150	5850	2250	2650	650	
	U100_I	150	550	1150					
	U100_S	550	4750	150	2650	2350	5750	650	
	U400_N	4600	200	600					
	U400_I	1000	4600	5800					
	U400_S	4600	600	1000	5800				
	BARK_N	570	4850	5850	150				
	BARK_I	1175	570						
	BARK_S	570	4850	5850	150				
E	*U100_N*	850	2650	150	1050	5250	450	1950	2250
	U100_I	150	850	450	2650	1950	350		
	U100_S	2650	850	1050	5250	450	550	150	7750
	U400_N	1000	2600	1800	600	2200			
	U400_I	1000	600						
	U400_S	1000	2600	600	1800	7800	2200	5400	
	BARK_N	845	150	2510	1860	455	350	570	
	BARK_I	150	845	455	250				
	BARK_S	845	2510	150	1860	455	570	1000	
I	*U100_N*	1750	550	150	4250	1650	650	450	750
	U100_I	1750	4350	550	150	6650	450	750	650
	U100_S	1750	4250	550	1650	5850	150	250	2750
	U400_N	600	1800	2600	6600	4200	3400	3000	
	U400_I	600	6600	4200	1800	5800			
	U400_S	600	4200	1800	6600				
	BARK_N	1375	4050	570	150	455	2925	3425	2510
	BARK_I	570	150	1375	4050	2925	1860	3425	5850
	BARK_S	570	4050	1600	150	50	2925	3425	845
O	*U100_N*	150	1650	350	3150	3750	6850	5750	450
	U100_I	150	1650	3750	3050	250	450		
	U100_S	150	1350	3750	3050	5250	450		
	U400_N	3000	3800	7000	1400	5400			
	U400_I	3800	3000						
	U400_S	3000	3800	7000					
	BARK_N	150	350	455	2925	4050	1375	1000	
	BARK_I	150	1600	455	250				
	BARK_S	150	1375	4050	455	2925	1000		
U	*U100_N*	2750	650	550	150	1150	5050	450	4650
	U100_I	3050	150	7250	7050	4450	650	1150	3950
	U100_S	3050	450	550	5050	1150	7250	7350	3350
	U400_N	2600	600	3000	3800	4200			
	U400_I	3000	600	3800	4600	5000			
	U400_S	3000	600	5000					
	BARK_N	2925	570	700	150	455	1175	845	
	BARK_I	2925	150	570	1175	845	700		
	BARK_S	2925	455	570					
M	*U100_N*	7050	3750	3850	450	650	1350		
	U100_I	7050	1750	1550	1050				
	U100_S	7050	450	650	1550	1750	1050	3350	
	U400_N	600							
	U400_I	600	7000						
	U400_S	7400							
	BARK_N	455							
	BARK_I	455	7050						
	BARK_S	455	7050	1375					
N	*U100_N*	450	3650	2250	1350	6950			
	U100_I	450	3650	1350	2250	6950			
	U100_S	450							
	U400_N	600	1400						
	U400_I	600	1400						
	U400_S								
	BARK_N	455							
	BARK_I	455	570	1375	50	1000			
	BARK_S	455							

Tab. 59: Frequenzbandanalyse: *KIND, N01*

Laut	Methode	Parameter							
		1	2	3	4	5	6	7	8
A	*U100_N*	150	250	2750	7250	6050	5650	6850	1650
	U100_I	150	650						
	U100_S	650	2850	7250	550	1050	2050	7050	2350
	U400_N	600	2600	1000					
	U400_I	600							
	U400_S	2600	600	7400	7800	1800			
	BARK_N	150	570	1000	1860	1600			
	BARK_I	150	570	1860	1000				
	BARK_S	700	2925	570	4850	1000			
E	*U100_N*	150	1050	50	2150	4450	1850	2050	5850
	U100_I	150	6650	2150	50				
	U100_S	1050	2150	150	50	6550			
	U400_N	1000	600	7400	2200	3400	3000		
	U400_I								
	U400_S	2200	1000	2600	600				
	BARK_N	150	1000	50					
	BARK_I	150	1000	4850	50				
	BARK_S	2160	1000	150	50				
I	*U100_N*	250	3150	1550	350	1450	1650	6050	450
	U100_I	150	3250	3350	7850	1750	2650	7450	50
	U100_S	3150	1550	6650	6350	3550	1250		
	U400_N	3400	1000	200	4600				
	U400_I	4200	600						
	U400_S	4200	1000	6600	200				
	BARK_N	250	2925	7050	1600	350	1375	455	
	BARK_I	150	1600	7050	455	1375	350		
	BARK_S	250	2925	455	4850				
O	*U100_N*	3950	4350	550	1150	150			
	U100_I	150	3950	4150	2750	950			
	U100_S	3950	6950	3150	1050	6150	3350		
	U400_N	3800	2600	6200	2200				
	U400_I	3800	6600	2600	2200	1800	1400	3400	
	U400_S	3800	2600	6200	2200	7000	5400		
	BARK_N	4050	7050	1000	150	2160	3425		
	BARK_I	150	1000	3425	7050	4050	2160		
	BARK_S	4050	7050	1000	2925	2160			
U	*U100_N*	1050	4650	1550	6950	2350	3250	1250	7250
	U100_I	150	650	1550	4850	6950	3950	7750	
	U100_S	550	2150	4650	3850	3050	6150	5350	650
	U400_N	2200	7000	200	3000				
	U400_I	2200	3000						
	U400_S	600	2200	3800	3000	6200	4600	5800	
	BARK_N	700	1600	1375					
	BARK_I	150	1600	700					
	BARK_S	700	1600	1375					
M	*U100_N*	6050	7650	50	450	950	5850		
	U100_I	50	6050	5950	2350	3850	950	3150	
	U100_S	6050	5950	7050	1050	6950			
	U400_N								
	U400_I								
	U400_S								
	BARK_N								
	BARK_I	50	1000	3425	1375	2160	250		
	BARK_S								
N	*U100_N*	50							
	U100_I	2350							
	U100_S	50							
	U400_N	200	3400						
	U400_I	2600							
	U400_S	200	4600						
	BARK_N	50							
	BARK_I	2510	150						
	BARK_S	50							

Tab. 60: Frequenzbandanalyse: *MANN, N01*

E.3.2 Klassifikation „*N123*“

Laut	Methode	Parameter							
		1	2	3	4	5	6	7	8
A	*U100_N*	2550	1250	3050	7850	5550	4850		
	U100_I	1850	1450	1250	7950	150			
	U100_S	2550	1250	3050	7950	5550	6850	7250	1450
	U400_N	2600	1000	3400					
	U400_I	1000	3400						
	U400_S	2600							
	BARK_N	2510	1175	1600	250				
	BARK_I	1175	3425	150	2160				
	BARK_S	1175	2925	2510					
E	*U100_N*	4550	6550						
	U100_I	4550							
	U100_S	4550	6550	450	5750	5950	7850		
	U400_N	4600	6600						
	U400_I	4600							
	U400_S	4600	6600						
	BARK_N	4850	455	7050					
	BARK_I	4850	845						
	BARK_S	4850	455	7050	50				
I	*U100_N*	450	4650	5950	7250				
	U100_I	5950	450	2250	5750				
	U100_S	450	5950	4550	7250	5750	1450		
	U400_N	4600	600						
	U400_I	4600	600						
	U400_S	6200	4600						
	BARK_N	455	4850						
	BARK_I	455	4850	700					
	BARK_S	455	4850						
O	*U100_N*	450	250	7250	2750	4650	7050	5850	650
	U100_I	450	2750	4550	150	2350			
	U100_S	450	150	2950	4450	7250	5850	1950	3950
	U400_N	600							
	U400_I	600	4600	2600					
	U400_S	600	4600	2600	1000				
	BARK_N	455	250	2925					
	BARK_I	455	150						
	BARK_S	455	250	150	2925	4850			
U	*U100_N*	450	5550	6950	4650	150			
	U100_I	5550	6950	450	3250	4750	6450	250	2850
	U100_S	450	4450	5550	7250	3750	7450	4750	2050
	U400_N	600	6600	200	5800	4600			
	U400_I	5400	4600	600	6600				
	U400_S	200	5400	4600	600	6600	3400	3800	
	BARK_N	455	4050	2925	150				
	BARK_I	455	5850	4050	250				
	BARK_S	455	4050	150	3425				
M	*U100_N*	1650	850	1450	1150	6150	550	950	
	U100_I	1650	3150						
	U100_S	1650							
	U400_N	4600							
	U400_I	1800							
	U400_S	4200							
	BARK_N	1600	845	455					
	BARK_I	1600	845						
	BARK_S	845							
N	*U100_N*	4350							
	U100_I	4350	350	2350	5750	7150	250	5350	
	U100_S	4350							
	U400_N	4200							
	U400_I	4200							
	U400_S	4200							
	BARK_N	4050							
	BARK_I	4050	350	150					
	BARK_S	4050							

Tab. 61: Frequenzbandanalyse: *FRAU, N123*

Laut	Methode	Parameter							
		1	2	3	4	5	6	7	8
A	*U100_N*	950	850	3150					
	U100_I	5550	950						
	U100_S	1250	4150	750	3750	3550	850		
	U400_N	1400	4200						
	U400_I								
	U400_S	1000							
	BARK_N								
	BARK_I								
	BARK_S	1000							
E	*U100_N*	1750							
	U100_I								
	U100_S	1750							
	U400_N	1800							
	U400_I								
	U400_S	1800							
	BARK_N	1860							
	BARK_I								
	BARK_S	1860							
I	*U100_N*	2150	550	3650	2050	2750	6450	5350	6550
	U100_I	1950	550	850					
	U100_S	1950	7250						
	U400_N	1800	3800	600	5400				
	U400_I								
	U400_S								
	BARK_N	1860	250						
	BARK_I	1860							
	BARK_S	1860	2510						
O	*U100_N*	2950	6250	950	350	4050	6450	7450	1050
	U100_I	2950	450	3150	950	6850	6450	1050	1650
	U100_S	2950	5050						
	U400_N	1400							
	U400_I	6600							
	U400_S	3000	1400	3800					
	BARK_N	2925	5850	3425	455				
	BARK_I	455	1375						
	BARK_S	2925	1375						
U	*U100_N*	1050	1450	550	3850				
	U100_I	1050							
	U100_S	1050	1750	7250	3850	6950	850		
	U400_N								
	U400_I								
	U400_S								
	BARK_N								
	BARK_I								
	BARK_S	1000	1860						
M	*U100_N*								
	U100_I								
	U100_S								
	U400_N								
	U400_I								
	U400_S								
	BARK_N								
	BARK_I	700							
	BARK_S								
N	*U100_N*	850	3450						
	U100_I								
	U100_S								
	U400_N								
	U400_I								
	U400_S								
	BARK_N								
	BARK_I								
	BARK_S								

Tab. 62: Frequenzbandanalyse: *KIND*, *N123*

Laut	Methode	Parameter							
		1	2	3	4	5	6	7	8
M	*U100_N*	3450	3050	6150	2550				
	U100_I	3450	6450	4050	650				
	U100_S	3450	3050	6150	2550				
	U400_N	3400	200	3000	600				
	U400_I	3400	1400	3000	600				
	U400_S	3400	3000						
	BARK_N	2925	455	3425	2160				
	BARK_I	3425	2925	50	4050				
	BARK_S	3425	2925	50	1860				
N	*U100_N*								
	U100_I								
	U100_S								
	U400_N								
	U400_I								
	U400_S								
	BARK_N	455	2160	350	1600	4850			
	BARK_I	455	2160	350	4050				
	BARK_S								

Tab. 63: Frequenzbandanalyse: *MANN, N123*

E.3.3 Klassifikation „*N0123*“

Laut	Methode	Parameter							
		1	2	3	4	5	6	7	8
A	*U100_N*	250	1050	950	1250	7950	7750	4350	850
	U100_I	2950	250	1050	450	7950	7750	1250	3350
	U100_S	2750	1050	250	1250	7950	950	1850	7750
	U400_N	1000	2600	2200	3000	4600	1800	3800	
	U400_I	1000	2200						
	U400_S	1000	2600	2200	3000	4600	1800	3800	
	BARK_N	1000	250	845	1175	2925	3425	1375	
	BARK_I	1000	250	845	1375				
	BARK_S	1000	250	2510	2160	2925	1175	1600	
E	*U100_N*	250	450	4450	650	5350	6250		
	U100_I	450	350	5350	650				
	U100_S	450	250	4550	1150	5250	6550	650	550
	U400_N	600	1000	5400	4600	3800			
	U400_I	600	1000	5400					
	U400_S	4600	1000	600	1800	6600			
	BARK_N	250	455	4850	700				
	BARK_I	455	250	4850	700				
	BARK_S	455	250	4850	1175	845			
I	*U100_N*	3750	1050	950	4450	450	7150	150	
	U100_I	3750	950	1050	4450	450	150	5250	
	U100_S	4250	1050	3750	5250	450	150	5950	2250
	U400_N	1000	5400	600	3800				
	U400_I	3800	1000	5400	600	4600			
	U400_S	4200	1000	5400	3800	600	6200		
	BARK_N	4050	1000	455	150	4850	1175		
	BARK_I	1000	4050	455	1175	4850			
	BARK_S	4050	1000	455	4850	150			
O	*U100_N*	250	450	550	1750	350	3250	2750	4550
	U100_I	450	350	550	5750	3250	4650		
	U100_S	250	450	2250	550	4550	2750	3250	350
	U400_N	600	1400	4600	2200	2600	5400		
	U400_I	600	1400	4600	2600	5800			
	U400_S	2200	600	4600	2600	1000	5400		
	BARK_N	250	455	570	1860	350	845		
	BARK_I	455	350	570	5850				
	BARK_S	250	455	2160	570	845	4850		
U	*U100_N*	2850	2050	350	750	5550	650	3250	6450
	U100_I	2850	2050	5550	350	750	650	3250	2350
	U100_S	2950	350	5550	2050	750	650	2350	6350
	U400_N	2200	600	5400	6600	3000	200	3400	5000
	U400_I	2200	600	5400	3000	6200	5000	3400	
	U400_S	3000	600	2200	6600	5400	4600	200	3400
	BARK_N	2160	700	570	2925	350	455	5850	7050
	BARK_I	2160	700	570	2925	350	5850	7050	455
	BARK_S	2925	700	350	2160	7050	1375	250	455
M	*U100_N*	850	2850	150	1550				
	U100_I	2850	4550	250					
	U100_S	2850	150						
	U400_N	4600							
	U400_I	1800							
	U400_S	4200							
	BARK_N	845	150						
	BARK_I	570	1860						
	BARK_S	570	1860						
N	*U100_N*	1650							
	U100_I	1650							
	U100_S	4350	650						
	U400_N								
	U400_I								
	U400_S	4200							
	BARK_N	4050							
	BARK_I								
	BARK_S								

Tab. 64: Frequenzbandanalyse: *FRAU*, *N0123*

Laut	Methode	Parameter							
		1	2	3	4	5	6	7	8
A	*U100_N*	550	4750	650	3250				
	U100_I	550	1250						
	U100_S	550	4750	5750	3250	950	650		
	U400_N	4600	200						
	U400_I								
	U400_S	4600	600	1000					
	BARK_N	570	4850	5850					
	BARK_I	1175	570						
	BARK_S	570	4850	1175					
E	*U100_N*	1750	850	2650	150	1050	1950	2150	
	U100_I	150	850	1950	2650	450	350		
	U100_S	2650	850	1950	1050	150	450	550	7750
	U400_N	1800	2600	1000	600	2200			
	U400_I	1000	600	1800	2600				
	U400_S	1000	2600	1800	600	7800			
	BARK_N	1860	2510	845	150	2160			
	BARK_I	845	150	455	1860	2510			
	BARK_S	845	2510	1860	150	455	570		
I	*U100_N*	1850	550	4450	1450	1950	150	1250	1650
	U100_I	550	2150	4450	1450	150	1950	6450	2250
	U100_S	4250	550	2750	2150	1650	1950	1250	1450
	U400_N	1800	600	2600	5400				
	U400_I	600	6600	4200	1800	3000	1400		
	U400_S	1800	4200	600	2600	5400	3400		
	BARK_N	1860	570	2925	1375	150	455	845	50
	BARK_I	570	150	1860	2925	1375	2510	4850	
	BARK_S	1860	4050	1600	570	2925	3425		
O	*U100_N*	1550	3050	3750	6850	2250	150	450	
	U100_I	1550	150	450					
	U100_S	1550	150	450	2950	3750	2250	6850	
	U400_N	1400	3000	3800	7000	2200			
	U400_I	3000	3800	6600	2200				
	U400_S	3000	3800	1400	2200	7000			
	BARK_N	1600	2925	4050	150	455			
	BARK_I	1600	150	455	250				
	BARK_S	2925	4050	1600	2160	150	455		
U	*U100_N*	2750	1050	2050	650	4650	150	550	5050
	U100_I	3050	2050	1050	150	4650	3450	650	3750
	U100_S	3050	1050	450	2050	5650	5250		
	U400_N	600	3000	1800					
	U400_I	3000	1800	600					
	U400_S	3000	600	5000					
	BARK_N	2925	570	700	1600				
	BARK_I	2925	1000	1600	150				
	BARK_S	2925	1000	455	570				
M	*U100_N*	3750	7050	1450	450	3850			
	U100_I	6750	1250						
	U100_S								
	U400_N	600	1400						
	U400_I	600	1400						
	U400_S								
	BARK_N	455	1375						
	BARK_I	455	1375	1860					
	BARK_S	455	1375						
N	*U100_N*	450							
	U100_I	450							
	U100_S								
	U400_N	600							
	U400_I	600							
	U400_S								
	BARK_N	455							
	BARK_I	455							
	BARK_S								

Tab. 65: Frequenzbandanalyse: *KIND, N0123*

Laut	Methode	Parameter							
		1	2	3	4	5	6	7	8
A	*U100_N*	150	250	2750	7250	6050	5650	6850	1650
	U100_I	150	650						
	U100_S	650	2850	7250	550	1050	2050	7050	2350
	U400_N	600	2600	1000					
	U400_I	600							
	U400_S	2600	600	7400	7800	1800			
	BARK_N	150	570	1000	1860	1600			
	BARK_I	150	570	1860	1000				
	BARK_S	700	2925	570	4850	1000			
E	*U100_N*	150	1050	50	2150	4450	1850	2050	5850
	U100_I	150	6650	2150	50				
	U100_S	1050	2150	150	50	6550			
	U400_N	1000	600	7400	2200	3400	3000		
	U400_I								
	U400_S	2200	1000	2600	600				
	BARK_N	150	1000	50					
	BARK_I	150	1000	4850	50				
	BARK_S	2160	1000	150	50				
I	*U100_N*	250	3150	1550	350	1450	1650	6050	450
	U100_I	150	3250	3350	7850	1750	2650	7450	50
	U100_S	3150	1550	6650	6350	3550	1250		
	U400_N	3400	1000	200	4600				
	U400_I	4200	600						
	U400_S	4200	1000	6600	200				
	BARK_N	250	2925	7050	1600	350	1375	455	
	BARK_I	150	1600	7050	455	1375	350		
	BARK_S	250	2925	455	4850				
O	*U100_N*	3950	4350	550	1150	150			
	U100_I	150	3950	4150	2750	950			
	U100_S	3950	6950	3150	1050	6150	3350		
	U400_N	3800	2600	6200	2200				
	U400_I	3800	6600	2600	2200	1800	1400	3400	
	U400_S	3800	2600	6200	2200	7000	5400		
	BARK_N	4050	7050	1000	150	2160	3425		
	BARK_I	150	1000	3425	7050	4050	2160		
	BARK_S	4050	7050	1000	2925	2160			
U	*U100_N*	1050	4650	1550	6950	2350	3250	1250	7250
	U100_I	150	650	1550	4850	6950	3950	7750	
	U100_S	550	2150	4650	3850	3050	6150	5350	650
	U400_N	2200	7000	200	3000				
	U400_I	2200	3000						
	U400_S	600	2200	3800	3000	6200	4600	5800	
	BARK_N	700	1600	1375					
	BARK_I	150	1600	700					
	BARK_S	700	1600	1375					
M	*U100_N*	3050	3450	450	4550				
	U100_I	2950	50	3650	4150	150	7150		
	U100_S	2950	2050	7150	5450	3150			
	U400_N	3000							
	U400_I	3000							
	U400_S	3000	3400						
	BARK_N	2925							
	BARK_I	50	4050	1860	1000				
	BARK_S	2925							
N	*U100_N*	50	7550	6050	6650	650	350		
	U100_I								
	U100_S	50	7550	1650	6550	5550	6350		
	U400_N								
	U400_I								
	U400_S								
	BARK_N	50	5850						
	BARK_I								
	BARK_S	50	5850	1860					

Tab. 66: Frequenzbandanalyse: *MANN, N0123*

F Literaturverzeichnis

[AS91] Acero, A. / Stern, R.: Robust Speech Recognition by Normalization of the Acoustic Space; Proc. Int. Conf. On Acoustics, Speech and Signal Processing, Toronto; S. 893-896;1991

[Bau63] Bauer,H.: Klanganalytische Untersuchungen der deutschen Sprache bei offenem, geschlossenem und gemischtem Näseln; Habilitationsschrift, Heidelberg; 1963

[BdW99] Bild der Wissenschaft; http://www.bdw.de/bdw/ticker/index.html; 3.5.99

[Ber80] Bergs, S.: Optimalität bei Clusteranalysen – Experimente zur Bewertung numerischer Klassifikationsverfahren; Dissertation, Universität Münster; 1980

[BEP96] Backhaus, K./ Erichson, B. / Plinke, W. / Weiber, R.: Multivariate Analysemethoden - Eine anwendungsorientierte Einführung; Springer-Verlag; 1996

[BHT63] Bogert, B. / Healy, M. / Tukey, J.: The Quefrency Analysis of Time Series for Echoes; Proc. Symp. On Time Series Analysis; S.209-243; 1963

[Bri87] Brigham, E. Oran: FFT – Schnelle Fourier-Transformation; R. Oldenbourg Verlag München Wien, 3. Auflage; 1987

[BS89] Bronstein, I. N./Semendjajew, K. A.: Taschenbuch der Mathematik; Verlag Harri Deutsch Frankfurt/Main, 24. Auflage; 1989

[BWM97] Barton, Siegmund / Wesseling, Sven / Maier, Torsten: Spracherkennung; http://monet.fh-friedberg.de/users/secunet/sprache/sp00001.htm; FH Friedberg; 1997

[CC70] Counihan, D. / Cullinan W.: Reliability and Dispersion of Nasality Ratings; Cleft Palate Journal 7; S. 216-270; Januar 1970

[CC72] Counihan, D. / Cullinan W.: Some Relationships between Vocal Intensity and Rated Nasality; Cleft Palate Journal 9; S.101-108; April 1972

[Cou86] Coulon, F. d.: Signal Theory and Processing; Artech House, Dedham, Massachusetts; 1986

[CT65] Cooley, J. / Tukey, J.: An Algorithm for the Machine Computation of Complex Fourier Series; Math. Computation Bd.19, S. 297-381; 1965

[DGH93] Dobler, S. / Geller, D. / Haeb-Umbach, R. / Meyer, P. / Ney, H. / Ruehl, H.W.: Design and Use of Speech Recognition Algorithms for a Mobile Radio Telepohne; Speech Communication, Bd. December, S. 221-229; 1993

[DPH93] Deller, J.R. / Proakis, J.G. / Hansen, J.H.L.: Discrete-Time Processing of Speech Signals; Maxwell Macmillan International, New York, Oxford, Singapore, Sydney; 1993

[DJ88] Duncan, G. / Jack, M.: Speech Formant Trajectory Pattern Recognition Using Multiple-order Pole-focused LPC Analysis; Proc. Int. Conf. on Acoustics, Speech and Signal Processing, New York; S. 484-486; 1988

[DM80] Davis, S. / Mermelstein, P.: Comparison of Parametric Representation for Monosyllabic Word Recognition in Continuously Spoken Sentences; IEEE Tran. on Acoustics, Speech and Signal Processing, Bd. 28, Nr. 4, S. 357-366; 1980

[Dou92] Doubrava, C.: A New Algorithm for DSP Applications; DSP Applications, Bd.1, Nr. 1, S. 45-53; 1992

[Efr82] Efron, B.: The Jackknife, the Bootstrap, and Other Resampling Plans; Pa. Society for Industrial and Applied Mathematics; Philadelphia; 1982

[Fan60] Fant, G.: The Acoustic Theory of Speech Production; Mouton & Co., The Hague; 1960

[Fel84] Fellbaum, K.: Sprachverarbeitung und Sprachübertragung; Springer-Verlag Heidelber, New York, Tokyo; 1984

[Fle76] Fletcher, S. G.: Nasalance vs. Listener Judgements of Nasality; Cleft Palate J. 13, S. 31-44; 1976

[Fle89] Fletcher, S. G.: Palatometric Specification of Stop, Affricate and Sibilant Sounds; Journal of Speech and Hearing Research 32; S. 736-748; December 1989

[Fur86] Furui, S.: Speaker-independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum; IEEE Trans. on Acoustics, Speech and Signal Processing; Bd. 34, Nr. 1, S. 52-59; 1986

[Fur89] Furui, S.: Digital Speech Processing, Synthesis and Recognition; Marcel Dekker, New York; 1989

[GK98] Gediga, G. / Kuhnt, T.: Praktische Methodenlehre; Universität Osnabrück; Skript zur Vorlesung; http://www.psycho.uni-osnabrueck.de/ggediga/www/pm98 ;1998

[GGC96] Garnier, S. / Gallego, S. / Collet, L. / Berger-Vachon, C.: Spectral and Cepstral Properties of Vowels as a Means for Characterizing Velopharyngeal Impairment in Children; Cleft Palate-Craniofacial Journal, Vol.33 No.6 S. 507-512; Nov. 1996

[Haa91] Haapanen, M.: Nasalance Scores in Normal Finnish Speech; Folia Phoniatrica 43; S. 197-203; 1991

[Hab86] Habermann, G.: Stimme und Sprache – Eine Einführung in ihre Physiologie und Hygiene; Georg Thieme Verlag Stuttgart, New York, 2. Auflage, 1986

[Hag97] Hagen, C.: Neuronale Netze zur statistischen Datenanalyse; Shaker Verlag, 1997

[Hau93] Hauenstein, A.: Optimierung von Algorithmen und Entwurf eines Prozessors für die automatische Spracherkennung; Dissertation Technische Universität München; 1993

[HBA93] Huang, X. / Belin, M. / Alleva, F. / Hwang, M.: Unified Stoschastic Engine (USE) for Speech Recognition; Proc. Int. Conf. on Acoustics, Speech and Signal Processing, Minneapolis; Bd. 2, S. 636-639; 1993

[Her90] Hermansky, H.: Perceptual Linear Predictive (PLP) Analysis of Speech; Journal of the Acoustical Society of America, 87(2), S. 1738-1752; 1990

[HM60] Hess, D. / McDonald, E.: Consonatal Nasal Pressure in Cleft Palate Speakers; Journal Speech Hearing 3; S.201-211; September 1960

[HMB91] Hermansky, H. / Morgan, N. / Bayya, A. / Kohn, P.: Compensation for the Effect of the Communication Channel in Auditory-like Analysis of Speech (RASTA-PLP); Proceedings of European Conference on Speech Technologies, Genova, Italy, S. 1367-1370; 1991

[HP91] Haddad, R. A./ Parsons T. W.: Digital Signal Processing – Theory, Applications and Hardware; Computer Science Press; 1991

[HVM92] Hardin, M. / Van Demark, D. / Morris, H. / Payne, M.: Correspondence between Nasalance Scores and Listener Judgements of Hypernasality and Hyponasality; Cleft-Palate-Craniofacial Journal 29; S.346-351; July 1992

[HWS91] Heppt, W. / Westrich, M. / Strate, B. / Möhring, L.: Nasalanz – Ein neuer Begriff der objektiven Nasalitätsanalyse; Laryngo-Rhino-Otologie 70, S. 169-228; 1991

[IU87] Itakura, F. / Umezaki, T.: Distance Measure for Speech Recognition Based on the Smoothed Group Delay Spectrum; Proc. Int. Conf. on Acoustics, Speech and Signal Processing, Dallas; S. 1257-1260; 1987

[Jäh95] Jähne, B.: Digital Image Processing; Springer-Verlag, Berlin;Heidelberg;1995

[JW89] Junqua, J.-C. / Wakita, H.: An Comparative Study of Cepstral Lifters and Distance Measures for All Pole Models of Speech in Noise; Proceedings of the International Conference on Acoustics, Signal and Speech Processing, Glasgow, Scottland, S. 476-479; 1989

[KK96] Kammeyer, K.D. / Kroschel, K.: Digitale Signalverarbeitung – Filterung und Spektralanalyse; B. G. Teubner Stuttgart; 1996

[KMO96] Kataoka, R. / Michi, K.-I. / Okabe, K. / Miura T. / Yoshida, H.: Spectral Properties and Quantitative Evaluation of Hypernasility in Vowels; Cleft Palate-Craniofacial Journal, Vol.33 No.1 S.43 – 50; Jan. 1996

[Koh77] Kohler, K.: Einführung in die Phonetik des Deutschen; Erich Schmidt Verlag Berlin; 1977

[Kög98] Kögel, H.: Entwicklung einer sprachgesteuerten Entwicklungsumgebung für sprechgestörte Kinder; Studienarbeit am Lehrstuhl für Wirtschaftsinformatik II; Universität Mannheim; Sommersemester 1998

[Kög00] Kögel, H.: Entwicklung eines automatischen Spracherkennungsmoduls für den Einsatz in einer sprachgesteuerten Trainingsumgebung für sprechgestörte Kinder; Diplomarbeit am Lehrstuhl für Informatik V; Mannheim; Aug. 2000

[Kra90] Kratzer, K. P.: Neuronale Netze – Grundlagen und Anwendungen; Hanser Verlag München, Wien; 1990

[Kyt69] Kyttä, J.: The Influence of the Nose on the Acoustic Pattern of Nasal Sounds; Actaoto-Laryngologica Supplement 263; S. 95-98; 1969

[Kuh99] Kuhlmann, U.: Wie bitte ? – vier Diktiersysteme im Test; C't Magazin für Computertechnik; Heise Verlag; S. 124-132; Feb. 1999

[Kuh00] Kuhlmann, U.: Hörmaschine - Fünf Diktiersysteme für Windows und das Mac OS; C't Magazin für Computertechnik; Heise Verlag; S. 118-125; Aug. 2000

[LCS67] Liberman, A. / Cooper, F. / Shankweiler, D. / Studdert-Kennedy, M.: Perception of the Speech Code; Psychological Review, Bd. 74, S. 431-461; 1967

[LD92] Litzaw, L. / Dalston, R.: The Effect of Gender upon Nasalance Scores among Normal Adult Speakers; Journal of Communication Disorders 25; S.55-64; March 1992

[LRP90] Lee, C. / Rabiner, L. / Pieraccini, R. / Wilpon, J.: Acoustics Modeling for Large Vocabulary Speech Recognition; Computer Speech & Language, Bd. 4, Nr. 2, S. 127-165; 1990

[LRR81] Lamel, F.L. / Rabiner, L.R. / Rosenberg, A.E. / Wilpon, J.G.: An Improved Endpoint Detector for Isolated Word Recognition; IEEE Trans. On Acoustics, Signal and Speech Processing, 29(4), S, 777-785; 1981

[Kal98] Malaske, U.: Sprechen statt Schreiben - fünf Diktiersysteme im Test; C't Magazin für Computertechnik; Heise Verlag; S. 110-119; März 1998

[Mar72] Markel, J.: Digital Inverse Filtering – A new Tool for Formant Trajectory Estimation; IEEE Trans. on Audio Electroacoustics, Bd. 20, Nr.2, S. 129-137; 1972

[McC74] McCandless, S.: An Algorithm for Automatic Formant Extraction using Linear Prediction Spectra; IEEE Trans. on Acoustics, Speech and Signal Processing, Bd. 22, S. 135-140; 1974

[MG76] Markel, J./ Gray Jr., A.: Linear Prediction of Speech; Communications and Cybernetics, Bd. 12; Springer Verlag, Berlin, Heidelberg, New York; 1976

[MK93] Mitra, S. K./Kaiser, J. F.: Handbook for Digital Signal Processing; John Wiley & Sons ;1993

[Moo89] Moore, B.: An Introduction to the Psychology of Hearing; Academic Press, London; 1989

[MY91] Murthy, H. / Yegnanarayana, B.: Speech Processing using Group Delay Functions; Signal Processing, Bd. 22, Nr. 3, S. 259-267; 1991

[Ney80] Ney, H.: Automatic 'Voiceprint' Comparison by Computer; International Conference: Security Through Science and Engineering, Berlin; 1980

[Nol67] Nol, A.: Cepstrum Pitch Determination; Journal Acoustic Soc. Amer., Bd. 41, Nr.2, S. 293-309; 1967

[Nor96] Nordbruch, S.: Automatische Klassifikation von Objekten für die vision-basierte Roboterregelung; Diplomarbeit, Universität Bremen; Institut für Automatisierungstechnik;1996

[PB52] Peterson, G.E.; Barney, H.L.: Control Methods Used in a Study of the Vowels; J. Acoust. Soc. Am. 24(2), S. 175-194, March 1952

[PEJ91] Paynter, E. / Edmonson, T. / Jordan, W.: Accuracy of Information Reported by Parents and Children Evaluated by a Cleft Palate Team; Cleft Palate-Craniofacial Journal 28; S. 329-337; October 1991

[PM96] Power Macintosh 7100, Handbuch; Apple Computer; 1996

[Pfe99] Pfeiffer, Silvia: Information Retrieval aus digitalisierten Audiospuren von Filmen; Doktorarbeit, Universität Mannheim; Aachen, Shaker Verlag; 1999

[Rab75] Rabiner, L.R.: An Algorithm for Determining the Endpoints of Isolated Utterances; Bell System Technical Journal, 54(2), S. 297-315; 1975

[Reg88] Regel, P.: Akustisch-phonetische Transkription für die automatische Spracherkennung; Fortschrittsberichte, Reihe: Informatik/Kommunikationstechnik, Bd. 10; VDI Verlag, Düsseldorf; 1988

[RHH90] Reitemeier, G. / Heidelbach, J. / Hloucal U. / Reitemeier, B.: Prüfmöglichkeiten der Stimm- und Sprachfunktion aus stomatologischer Sicht; Zahn-, Mund- und Kieferheilkunde mit Zentralblatt 78;5; S.417-423; 1990

[RJ93] Rabiner, L. / Juang, B.-H.: Fundamentals of Speech Recognition; Prentice Hall, Englewood Cliffs, New Jersey; 1993

[Rob98] Robinson, Tony: ABBOT - Large Vocabulary Connectionist Speech Recognition; http://svr-www.eng.cam.ac.uk/~ajr/abbot.html; Cambridge University Engineering Dept./Speech Vision Robotics group; 1998

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

[Rus82] Ruske, G.: Auditory Perception and its Application to Computer Analysis of Speech; Computer Analysis and Perception, CRC Press, Boca Raton, FL; 1982

[Rus94] Ruske, G.: Automatische Spracherkennung – Methoden der Klassifikation und Merkmalsextraktion; Oldenbourg München, 2. Auflage; 1994

[RV91] Rioul, O. / Vetterli, M.: Wavelets and Signal Processing; IEEE Signal Processing Magazine, Bd. October, S. 14-38; 1991

[Sav89] Savoji, M.H.: A Robust Algorithm for Accurate Endpointing of Speech Signals; Speech Communication, 8(1), S. 45-60; 1989

[Sch77] Schürmann, J.: Polynomklassifikatoren für die Zeichenerkennung; Oldenbourg, München; 1977

[Sch92] Schrüffer, E.: Signalverarbeitung – Numerische Verarbeitung digitaler Signale; Hanser Verlag München, Wien, 2. Auflage; 1992

[Sch95] Schürer, Tilo: Sprecherunabhängige Ziffern- und Ziffernkettenerkennung über Telefonkanäle; Dissertation D83, Technische Universität Berlin; 1995

[SDL91] Seaver, E. / Dalston, R. / Leeper, H.:A study of Nasometric Values for Normal Nasal Resonance; Journal of Speech and Hearing Research 34; S.715-721; August 1991

[SH90] Stearns, S. D./ Hush D. R.: Digital Signal Analysis; Prentice Hall, Englewood Cliffs, New Jersey, Second Edition; 1990

[Söh95] Söhngen, D.: Nasalanzbestimmung mit NASOBEM 2.0; Studienarbeit am Lehrstuhl für Praktische Informatik V, Universität Mannheim; Wintersemester 1994/95

[St95] Schukat-Talamazzini, E. G.: Automatische Spracherkennung – Statistische Verfahren der Musteranalyse; Vieweg Verlag; 1995

[SHK94] Stellzig, A./ Heppt, W. / Komposch, G.: Das Nasometer. Ein Instrument zur Objektivierung der Hyperrhinophonie bei LKG-Patienten; Fortschritte der Kieferorthopädie, 55, S. 176-180;1994

[Ste57] Stevens, S.: On the Psychophysical Law; Psychological Review, Bd. 64, S. 153-181, 1957

[Ste98] Stellzig, A.: Gesichtsschädelwachstum von Patienten mit Lippen-Kiefer-Gaumenspalten unter interdisziplinären Gesichtspunkten; Habilitationsschrift, Heidelberg; 1998

[SV93] Smolders, J. / Van Compernolle, D.: In Search for the Relevant Parameters for Speaker Independent Speech Recognition; Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Minneapolis, USA, Vol. II, S.684-687;1993

[Ter74] Terhardt, E.: On the Perception of Periodic Sound Fluctuations (Roughness); Acustica, Bd. 30, S. 201-213; 1974

[TKM91] Tsopanaglou, A. / Kyriakis-Bitzaros / Mourjopoulos, J. / Kokkinakis, G.: A Real Time Speech Decoder Using Instantaneous Frequency and Energy; Proc. European Conf. on Speech Technology, Bd. 3, S. 1349; 1991

[TS99] Tillmann, H. G. / Schiel, F.: Akustische Phonetik–Kapitel II: Was ist Sprachschall?; Universität München; Vorlesungsskript; www.Phonetik.uni-muenchen.de/AP/APKap2.html; 1999

[VHH98] Vary, P./Heute, U./Hess, W.: Digitale Sprachsignalverarbeitung; B. G. Teubner Stuttgart; 1998

[VM99] Deutsches Forschungszentrum für Künstliche Intelligenz: Verbmobil; www.dfki.de./verbmobil; 1999

[Vog75] Vogel, A.: Ein gemeinsames Funktionsschema zur Beschreibung der Lautheit und der Rauhigkeit; Biol. Cybernetics, Bd. 18, S. 31-40; 1975

[Wai99] Waibel, A: The Janus Project. http://werner.ira.uka.de/ISL.speech.janus.html, Interactive System Labs; Universität Karlsruhe; Juni 1999

[Wak73] Wakia, H.: Direct Estimation of the Vocal Tract Shape by Inverse Filtering of Acoustic Speech Waveform; IEEE Trans. on Audio Electroacoustics, Bd. 21, S. 417-427; 1973

[WD64] Warren, D. / Dubois, A.: A Pressure-flow Technique for Measuring Velopharnygeal Orifice Area during Continuous Speech; Cleft Palate Journal 16, S.52-71; January 1964

[WDM88] Wilpon, J.G. /DeMarco, D.M. / Mikkilineni, R.P.: Isolated Word Recognition over the DDD Telephone Network – Results of Two Extensive Field Studies; Proceedings of the International Conference on Acoustics, Signal and Speech Processing, New York City, USA, S. 55-58; 1988

[Wir94] Wirth, G.: Sprachstörungen, Sprechstörungen und kindliche Hörstörungen; Deutscher Ärzte-Verlag; Köln,4.Auflage,1994

[WMG90] Wilpon, J.G: / Mikkilineni, R. / Gokcen, S.: Speech Recognition: From the Laboratory to the Real World; AT&T Technical Journal, S.14-24; Sep./Oct. 1990

[WMW93] Watterson, T. / McFarlane, S. / Wright, D.: The Relationship between Nasalance and Nasality in Children with Cleft Palate; J-Commun-Disord.,S. 13-28; April 1993

[WN76] White, G. M. / Neely, R. B.: Speech Recognition Experiments with Linear Prediction, Bandpass Filtering and Dynamic Programming; IEEE Trans. Vol. ASSP-24, No. 2, 183-188; 1976

[WRG89] Wolfgang Rosenthal Gesellschaft e.V.: Lippen-, Kiefer-, Gaumen-, Segel-Spalten: Informationen zur Sprachentwicklung und –behandlung; Informationsreihe der Wolfgang Rosenthal Gesellschaft e.V., Heft 5; 1989

[WRM84] Wilpon, J.G. / Rabiner, L.R. / Martin, T.: An Improved Word-Detection Algorithm for Telefone-Quality Speech Incorporating Both Syntactic and Semantic Constraints; AT&T Bell Laboratories Technical Journal, 63(3), S. 479-498; 1984

[Zec92] Zecevic, A.: Implementierung einer benutzerfreundlichen Oberfläche zur Unterstützung der Diagnose und Behandlung von offener und geschlossener Rhinophonie anhand des Nasalanzmaßes; Studienarbeit am Lehrstuhl Wirtschaftsinformatik II, Universtität Mannheim; 1992

[ZF67] Zwicker, E. / Feldtkeller, R.: Das Ohr als Nachrichtenempfänger; Hirzel Verlag, Stuttgart; 1967

[ZF90] Zwicker, E./ Fastl, H.: Psychoacoustics: Facts and Models; Number 22 in Springer Series in Information Sciences, Springer Verlag Berlin; 1990

[ZLM96] Zajac, D. / Lutz, R. / Mayo, R.: Microphone Sensitivity as a Source of Variation in Nasalance Scores; -Speech-Hear-Res., S. 1228-31; 1996

[Zwi80] Zwicker, E.: Analytical Expressions for Critical-Band Rate and Critical Bandwith as a Function of Frequency; Journal of the Acoustical Society of America, 68(5); Nov. 1980

[Zwi82] Zwicker, E.: Psychoakustik; Springer Verlag Berlin, Heidelberg; 1982

G Index

H Eidesstattliche Erklärung

Hiermit erkläre ich an Eides Statt, dass ich die hier vorgelegte Dissertation selbständig und ausschließlich unter Verwendung der angegebenen Quellen und sonstigen Hilfsmitteln verfasst habe. Die Arbeit wurde in gleicher oder ähnlicher Form keiner anderen Prüfungsbehörde zur Erlangung eines akademischen Grades vorgelegt.

Sinsheim, den 05.08.2002