Simulation emotionaler Sprechweise mit Sprachsyntheseverfahren

vorgelegt von
Felix Burkhardt

Vom Fachbereich Kommunikations- und Geschichtswissenschaften
der Technischen Universität Berlin
genehmigte Dissertation zur Erlangung des akademischen Grades
Doktor der Philosophie

Berichter: Prof. Dr. Walter F. Sendlmeier
Berichter: Prof. Dr. Klaus Fellbaum

Tag der wissenchaftlichen Aussprache: 23.10.2000

Berlin 2000

D 83

Im Gedenken an D. H. Klatt

Vorwort

Die vorliegende Arbeit wurde in den Jahren 1999-2000 verfasst. Sie stellt die Erweiterung der Magisterarbeit (Burkhardt (1999)) dar und entstand als Teil eines DFG-Projektes am Institut für Kommunikationswissenschaft der TU-Berlin. Dieses Projekt stand unter der Leitung von Prof. Dr. Walter F. Sendlmeier und trägt den Titel ”Reduktion und Elaboration bei emotionaler Sprechweise” (Kennziffer Se 462/3-1).

Ich danke vor allem meinem Doktorvater, Prof. Dr. Walter F. Sendlmeier, für die Bereitstellung des Themas und die zahllosen hilfreichen Hinweise und Gespräche. Ohne ihn wäre diese Arbeit nicht entstanden. Ein weiterer Dank gilt Prof. Dr. Klaus Fellbaum für die Übernahme des Zweitreferats.

Weiterhin danke ich meinen Kollegen für die gute Zusammenarbeit und die vielen Diskussionen. Vor allem Miriam Kienast und Astrid Paeschke, die zeitgleich an Dissertationen mit verwandter Thematik arbeiteten, bin ich für die große Unterstützung in vielerlei Hinsicht dankbar.

Ich danke der Deutschen Forschungsgemeinschaft für die Finanzierung dieser Arbeit.

Ein anonymer Dank gilt allen Testhörern, die ich aus dem Institutsflur und meinem Freundeskreis rekrutiert habe. Sie haben mir durch ihre geduldige Teilnahme an den Perzeptionsexperimenten und viele wertvolle Hinweise sehr geholfen.

Weiterhin bin ich der weltweiten Gemeinde der Entwickler freier Software sehr verpflichtet. Programme wie **Linux**, **Latex**, **R**, **xv**, **sox**, **MBROLA** oder die Implementation des Klatt-Synthesizers von J. Iles und N. Simmons haben mir die Arbeit sehr erleichtert.

Letztendlich danke ich meiner Familie für ihre Geduld und Unterstützung.

NB: Im Rahmen dieser Arbeit ist eine Dokumentationsseite im Internet entstanden, die einige der besprochenen Systeme zur Simulation emotionaler Sprechweise mit Sprachsyntheseverfahen durch Audiobeispiele demonstriert. Die Adresse lautet:
http://www.kgw.tu-berlin.de/∼felixbur/speechEmotions.html
Die meisten der entstandenen Hilfsprogramme sind weiterhin im Quelltext unter
http://www.kgw.tu-berlin.de/∼felixbur/software.html
herunterzuladen.

Berlin im August 2000
Felix Burkhardt

Inhaltsverzeichnis

Einleitung

> The movements of expression give vividness and energy to our spoken words.
> (Charles Darwin, 1872)

> It would be a considerable invention indeed, that of a machine able to mimic speech, with its sounds and articulations. I think it's not impossible.
> (Leonhard Euler, 1761)

Bei lautsprachlicher Kommunikation wird weitaus mehr Information übertragen als der semantische Gehalt. Stets wird auch Information über den emotionalen Zustand des Sprechers unwillkürlich dekodiert. Beim Fehlen visueller Informationen wird der emotionale Sprecherzustand aus dem Sprachsignal extrahiert. Die Frage nach den akustischen Korrelaten emotionaler Sprechweise ist seit gut 70 Jahren Gegenstand zahlreicher wissenschaftlicher Untersuchungen. Trotz dieser recht langen Forschungstradition ist eine Reihe von Fragen bis heute offen geblieben. Dies liegt zum einen an der hohen Komplexität des akustischen Sprachsignals und zum anderen an der unklaren Begrifflichkeit des Untersuchungsgegenstandes 'Emotion'.

Parallel zu der Erforschung des stimmlichen und sprecherischen emotionalen Ausdrucks haben sich mit der Entwicklung des Digitalcomputers große Fortschritte auf dem Gebiet der künstlichen Spracherzeugung ergeben. Die zunehmende Nutzung von Sprachsynthese im Alltag lässt eine Bereicherung solcher Systeme um emotionalen Sprechausdruck wünschenswert erscheinen. Diese Arbeit schlägt nun die Brücke zwischen der Erforschung emotionaler Sprechweise und der Entwicklung von Sprachsynthesizern. Sie steht damit einerseits in der Tradition von Resynthese-Experimenten, bei denen ein sprachähnliches Signal systematisch in seinen akustischen Eigenschaften variiert und auf emotionale Eindruckswirkung hin überprüft wird und andererseits gliedert sie sich an eine Reihe von Arbeiten an, deren Ziel darin besteht, emotionale Sprechweise erkennbar zu simulieren.

Klassische Anwendungsgebiete von TTS- (*Text-to-Speech*) Systemen sind die Verwendung als Prothese (etwa als Vorlesemaschine für Sehbehinderte oder Sprachausgabegerät für Sprechbehinderte), der Einsatz als Frontend für Auskunftssysteme oder bei der automatischen Übersetzung sowie ganz allgemein als sprachliches Interface bei der Mensch-Maschine Interaktion. Nach Murray et al. (1996) lässt sich die Qualität von TTS-Systemen hinsichtlich der drei Faktoren Verständlichkeit, Variabilität und Natürlichkeit bewerten. Mit Variabilität sind allgemeine Parameter gemeint, die Änderungen z.B. der Sprechgeschwindigkeit oder die Simulation verschiedener Alterstufen ermöglichen. Natürlichkeit ließe sich weiterhin in die Faktoren Stimmqualität, Sprachstil und Stimmung unterteilen.

- Eine Verbesserung der Stimmqualität erfordert komplexe Modelle der menschlichen Stimmerzeugung, vor allem des Glottissignals. Synthetisch erzeugte Sprache tendiert dazu, nicht genug Variabilität und Unregelmäßigkeiten in der Grundfrequenz und den einzelnen spektralen Bändern aufzuweisen.

- Eine Verbesserung des Sprachstils erfordert Modelle, die die Aspekte der Sprechsituation in die Synthese einbeziehen. Menschen sprechen anders, wenn sie vorlesen, vor einem großen Auditorium sprechen müssen oder beispielsweise mit Menschen anderer Hierarchiestufen kommunizieren.

- Es ist davon auszugehen, dass es keine vollkommen emotional neutrale Sprache gibt; der emotionale Gehalt von Sprachsignalen wird ständig durch generelle Einstellungen, momentane Stimmungen und weitere Einflüsse verändert.

Eine Optimierung dieser Aspekte stellt keine Schönheitskorrektur von im Prinzip funktionierenden Systemen dar, sondern ist eine notwendige Voraussetzung dafür, die Verständlichkeit und Hörerakzeptanz synthetischer Sprache auch unter erschwerten Bedingungen wie etwa bei Umgebungslärm oder bei langer Dauer zu gewährleisten.

Im Zentrum der vorliegenden Arbeit steht die Simulation von emotionalem Ausdruck bei der Sprachsynthese[1]. Um emotionalen Ausdruck im Sprachsignal zu erzeugen, sind mehrere Herangehensweisen möglich:

- Analytische Verfahren untersuchen die Ursachen emotionalen Sprechausdrucks auf verschiedenen Ebenen und bilden sie dann nach. Hierbei sind je nach Ebene zwei Ansätze vorstellbar (Cahn (1990)). Zum einen könnte man von der Sprecherseite ausgehen und seine innere Stimmung bzw. seinen derzeitigen Gefühlszustand und die Auswirkungen auf physiologische Zustände modellieren. Ansätze dazu werden bei der artikulatorischen bzw. neuromotor-command-Synthese verfolgt.

  Eine andere Vorgehensweise setzt sozusagen bei der Hörerseite an. Die zunächst für neutrale Sprechweise berechneten Merkmalsbeschreibungen für die Synthese werden den akustischen Korrelaten der zu simulierenden Emotion angepasst. Dieser Ansatz soll in der vorliegenden Arbeit verfolgt werden.

- Eine gänzlich andere Herangehensweise stellen Verfahren maschinellen Lernens dar, bei denen aus großen Datenbasen mit emotionaler Sprache Lernalgorithmen wie z.B. neuronale Netze optimiert werden, die dann gefundene Regelhaftigkeiten anwenden können (vgl. etwa Morton (1992)).

Im 1. Kapitel wird der Versuch unternommen, eine Arbeitsdefinition des Begriffs 'Emotion' zu finden. Dafür ist ein Exkurs in Gebiete der kognitiven Psychologie notwendig. Dabei wird vor allem das Konzept der 'Basisemotionen' dem der 'Emotionsdimensionen' gegenübergestellt.

Das 2. Kapitel liefert die notwendigen Grundlagen zur Beschreibung von Sprachsignalen. Es ist gemäß der lautsprachlichen Kommunikationskette Sprecher-Kanal-Hörer in drei Abschnitte

[1] Die Tatsache, dass Sprecher-Stimmungen eher durch akustische Merkmale als durch inhaltliche ausgedrückt werden, wird unter anderem in Scherer et al. (1984) nachgewiesen.

gegliedert. Im ersten werden physiologische Gegebenheiten bzgl. Respiration, Phonation und Artikulation besprochen. Neben verschiedenen Phonationsarten und artikulatorischen Settings und deren Relevanz bei emotionaler Sprechweise werden auch akustische Modelle für das Resonanzverhalten des Ansatzrohres besprochen. Der zweite Abschnitt besteht aus der Beschreibung möglicher Messverfahren für unterschiedliche Parameter von Sprachsignalen. Der dritte Abschnitt schließlich hat den menschlichen Hörvorgang zum Gegenstand. Darauf basierend, werden Einschränkungen akustischer Merkmale von Sprachsignalen in Bezug auf die Wahrnehmbarkeit behandelt.

Das 3. Kapitel gibt einen Überblick über den aktuellen Stand der Sprachsynthesetechnologie. Dabei wird zunächst der Aufbau von Text-to-Speech Systemen besprochen. Einzelne Synthese/Resyntheseverfahren werden dann näher erläutert. Es schließt mit einer relativ ausführlichen Besprechung der Formantsynthese, die in der vorliegenden Arbeit Verwendung findet.

Im 4. Kapitel wird ein Überblick über den bisherigen Stand der Forschung gegeben, der die Grundlage für die folgenden eigenen Experimente liefert. Zunächst werden Betrachtungen zur Universalität emotionaler Sprechweise angestellt und die Bedeutung einzelner Parameterklassen für emotionalen Sprechausdruck besprochen. Es folgt ein Abschnitt über die physiologischen Prozesse und ihre Auswirkungen auf das Sprachsignal bei emotionalen Sprecherzuständen. Danach werden verschiedene Betrachtungen zur Methodik von Perzeptionsexperimenten angestellt, und es werden einige Untersuchungen, die für diese Arbeit von speziellem Interesse sind, gesondert besprochen. Ein Überblick der Ergebnisse für die einzelnen betrachteten Emotionen führt abschließend zu einer tabellarischen Darstellung der wesentlichen Ergebnisse und zur Formulierung von Thesen, die im experimentellen Teil der Arbeit überprüft werden.

Mit dem 5. Kapitel beginnt der experimentelle Teil. Es wird ein erstes Perzeptionsexperiment beschrieben, das die Möglichkeiten der Simulation emotionaler Sprechweise mit einem Zeitbereichs-Syntheseverfahren betrachtet. Dabei wird ein datenbasierter einem regelbasierten Ansatz gegenübergestellt.

Ein eigener Sprachsynthesizer, mit dem die folgenden Experimente durchgeführt wurden, ist in Kapitel 6 beschrieben. Dieses System stellt einerseits ein Hilfsmittel dar, mit dem Stimuli vielseitig und regelbasiert manipuliert werden können, andererseits gibt es einen Rahmen für die Anwendungsmöglichkeit der gefundenen Resultate innerhalb eines konkreten Sprachsynthesizers.

Im 7. Kapitel wird ein Hörexperiment vorgestellt, in dem die Wirkung verschiedener akustischer Merkmale auf den emotionalen Sprecherausdruck untersucht wird. Dies geschieht durch die systematische Variation unterschiedlicher Sprachparameter.

Das 8. Kapitel greift einige im 7. Kapitel unklar gebliebenen Wirkungsweisen auf, indem ein Perzeptionsexperiment zur spezifischen Merkmalsvariation im Hinblick auf konkrete Hypothesen vorgestellt wird.

Die Dissertation schließt mit dem 9. Kapitel, in dem noch einmal die Resultate reflektiert werden und ein Ausblick gegeben wird.

NB: Sofern es sich bei der in der vorliegenden Arbeit angegebenen Literatur um Bücher oder sehr umfangreiche Artikel handelt, sind die Seitenzahlen vermerkt.

Teil I

Theoretischer Teil

Kapitel 1

Klärung des Begriffs 'Emotion'

1.1 Das Definitionsproblem

> "... almost everyone except the psychologist knows what an emotion is ..."
> Young 1973, zitiert in Kleinginna & Kleinginna (1981)

Die Definition von Emotionen ist in der Literatur sehr umstritten (vgl. z.B. Tischer (1993, S. 7)). Da die emotionsbeschreibenden Begrifflichkeiten (synonym dazu Affekte, Gefühle) subjektives Erleben beschreiben, sind sie wohl nicht eindeutig zu definieren. Abzugrenzen ist der Begriff der Stimmung, der sich auf emotionale Zustände bezieht, die für längere Zeiträume gelten. Emotionen haben die Eigenschaft, sich auf bestimmte Handlungen, Personen oder Objekte zu beziehen, während Stimmungen einen eher unstrukturierten Hintergrund haben. Als weiteres Unterscheidungskriterium wird des öfteren in der Literatur vorgeschlagen, dass Stimmungen (*motivations*) durch interne Stimuli entstehen, während Emotionen durch externe Reize ausgelöst werden (vgl. Kleinginna & Kleinginna (1981)). Dieser Punkt ist allerdings stark umstritten.

Zur Klassifizierung von Emotionen werden zwei grundsätzlich verschiedene Ansätze vorgeschlagen. Einmal die Einteilung in diskrete Begriffe, eventuell unterteilt in ein System von Basis- und Nebenemotionen und andererseits die Beschreibung als Punkt in einem mehrdimensionalen Raum. Diese beiden Beschreibungsansätze sind dabei nicht widersprüchlich, sondern durchaus auch kombinatorisch zu benutzen (z.B. in der Aussage: "*Er war freudig erregt.*"). So beschreiben Kleinginna & Kleinginna (1981) den Begriff Emotion in einer tautologischen Arbeitsdefinition:

> "Emotion is a complex set of interactions among subjective and objective factors, mediated by neural/hormonal systems, which can (a) give rise to to affective experiences such as feelings of arousal, pleasure/displeasure; (b) generate cognitive processes such as emotionally relevant perceptual effects, appraisals, labeling processes; (c) activate widespread physiological adjustments to the arousing conditions; and (d) lead to behaviour that is often, but not always, expressive, goaldirected, and adaptive."

Diese merkmalsorientierte Beschreibung soll auch der vorliegenden Arbeit genügen, da es hierbei lediglich um die Vokalisation emotionaler Zustände geht. Von daher kann als ausreichend gelten, dass Emotionen durch Gefühlswörter beschreibbar sind, die im Alltag ausreichend objektivierbar sind, um verstanden zu werden. Weiterhin gehen sie mit physiologischen Veränderungen einher und wirken von daher auf das Sprechverhalten. In den folgenden Abschnitten werden der dimensionale und der kategoriale Ansatz kurz erläutert. Abschließend wird die 'Component-Process' Theorie von Scherer (1984) vorgestellt und eine Begründung für die Auswahl der in dieser Arbeit berücksichtigten Emotionskategorien gegeben.

1.2 Dimensionaler Ansatz

Der dimensionale Ansatz legt unterschiedlich viele 'Emotionsdimensionen' zugrunde. Einzelne Emotionen sind dann in einem von diesen Dimensionen aufgespannten Raum als Punkte darstellbar (siehe Abb. 1.1). Das Dimensionsmodell setzt in der Regel drei Dimensionen an (Tischer (1993)), von denen zwei relativ unumstritten sind: Valenz (angenehm/unangenehm) und Aktivität (stark/schwach). Über die Art der dritten Dimension herrscht weitgehende Uneinigkeit. Dennoch ist zur weiteren Klassifizierung die Dimension der Potenz/Kontrolle/Dominanz (stark/schwach, überheblich/unterwerfend) bei vielen Autoren zu finden (vgl. Tischer (1993, S. 31)), die sich auch zur Differenzierung der in dieser Arbeit untersuchten Basisemotionen eignet.

Die Verwendung des Dimensionsansatzes hat im Hinblick auf das Thema den Vorteil, dass einige akustische Parameter direkt mit einigen der Dimensionen korrelieren (vgl. Murray et al. (1996); Bergmann et al. (1988) oder Frick (1985)). So sind in der Regel Sprechrate und mittlerer F_0-Wert stark mit der Erregtheit (Aktivität) eines Sprechers korreliert. Andererseits eignet sich das Dimensionsmodell nicht sehr gut dazu, Hörerurteile abzufragen, da die Begrifflichkeit zu abstrakt ist. Weiterhin ist es äußerst schwierig, Art und Anzahl der notwendigen Dimensionen festzulegen.

1.3 Kategorialer Ansatz

Im Gegensatz zum dimensionalen Ansatz werden hier die Emotionen nach nominalen Klassen unterteilt (vgl. z.B. Kleinginna & Kleinginna (1981)). Die Theorie der Basisemotion geht davon aus, dass es eine Menge von grundsätzlichen Emotionen gibt, die sich nicht auf andere zurückführen lassen, die entwicklungsgeschichtlich früh entwickelt wurden und die als Prototypen von 'Emotionsfamilien' gesehen werden können (vgl. z.B. Tischer (1993)). Wenn auch Art und Anzahl der Basisemotionen bei den verschiedenen Vertretern dieses Ansatzes stark umstritten sind, sind doch die Emotionen Ärger, Furcht, Freude und Trauer in fast allen Modellen enthalten[1].

Die Verwendung diskreter Begrifflichkeit hat den Vorteil, dass zumindest die Bedeutung der Begriffe für die Basisemotionen in der Alltagssprache gefestigt ist. Sie birgt allerdings die

[1] Eventuell unter verschiedenen Namen, z.B. Wut statt Ärger.

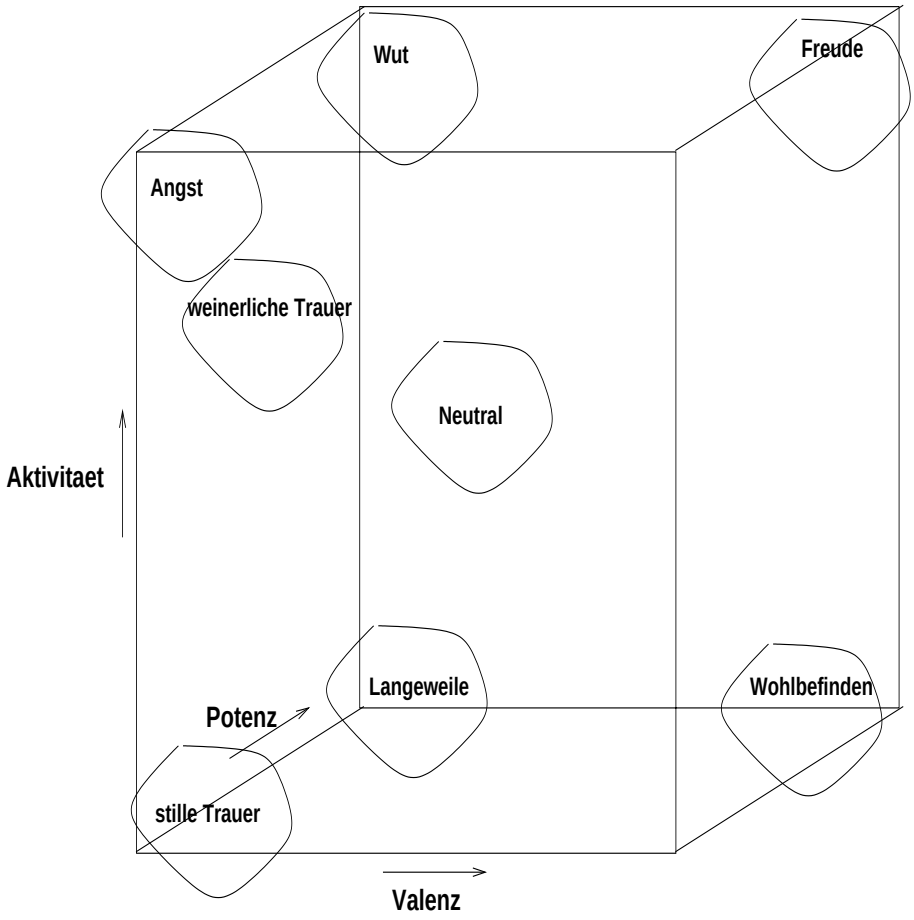

Abbildung 1.1: Einordnung der Emotionskategorien in ein Dimensionsmodell

Gefahr, dass die Begriffe nicht genügend spezifiziert werden. So lassen sich z.B. einige Widersprüchlichkeiten in der Literatur dadurch aufklären, dass nicht zwischen *hot anger* und *cold anger* unterschieden wurde (vgl. Scherer (1995, S. 93)).

1.4 Die 'Component Process' Theorie

Laut Scherer (1986) sind Emotionen nicht als statische Zustände zu sehen, sondern als Ausdruck dynamischer Prozesse, an denen mehrere Subsysteme beteiligt sind. Er entwickelt aus diesem Ansatz die *Component Process Theory* (Scherer (1984)). Diese beschreibt mehrere Emotions-Stadien, während derer sonst relativ unabhängige Subsysteme der Körpersteuerung synchron zusammenarbeiten. Diese Stadien werden durch die Abschätzung eines externen oder internen Stimulus als für den Organismus relevant ausgelöst. Die Subsysteme werden dabei durch eine Reihe von *Stimulus Evaluation Checks* (SECs) gesteuert. Dabei schlägt Scherer fünf SECs vor, die in der angegebenen Reihenfolge evaluiert werden:

1. *Novelty check*: handelt es sich um eine neue Situation?
2. *Intrinsic Pleasantness check*: ist das Ereignis angenehm oder eher unangenehm?
3. *Goal/need significance check*: ist das Ereignis relevant für aktuelle Ziele oder Bedürfnisse?
4. *Coping potential check*: ist das Ereignis handhabbar oder stellt es eine Bedrohung dar?
5. *Norm/self compability check*: entspricht das Ereignis, bzw. entsprechen die dadurch ausgelösten Aktionen der kulturellen/sozialen Norm? Steht es im Einklang mit dem gewünschten Selbstbild?

Im Rahmen dieser Theorie lassen sich Hypothesen darüber ableiten, welche prototypischen Zustände etwa die Lautproduktion als Ergebnis der SECs annimmt. Dabei wird ein funktionalistischer Standpunkt eingenommen; die Frage lautet: Welche physiologische Änderung kann als Adaption auf das Ergebnis eines SECs angesehen werden? Beispielsweise besteht die Adaption auf ein unangenehmes Objekt darin, den Rachen und die faucalen Stränge zu kontrahieren, wodurch der erste Formant steigt und die Rachenwände weniger akustische Energie absorbieren.

1.5 Auswahl der untersuchten Emotionen

Die Ausgangsbasis dieser Untersuchung stellt ein Korpus emotional gesprochener Sprache dar, der am Institut für Kommunikationswissenschaft an der TU-Berlin im Rahmen eines DFG-Projektes (Kennziffer Se462/3-1) erstellt wurde. Es handelt sich um emotional gesprochene Sätze, die von Schauspielern simuliert wurden. Dabei wurden die Basisemotionen 'Ärger', 'Trauer', 'Freude', 'Angst', 'Ekel' und 'Langeweile' berücksichtigt. Die Analysen, die auf dieser Datenbank basieren (z.B. Kienast & Sendlmeier (2000); Paeschke & Sendlmeier (2000) oder Burkhardt (1999)), stellen die Grundlage der in dieser Arbeit überprüften akustischen Merkmale emotionaler Sprechweise dar. Diese Basisemotionen wurden auch in den Perzeptionstests berücksichtigt. Die Emotionen 'Ekel' und 'Langeweile' sind in der Literatur eher selten untersucht worden. Zum Teil wurden sie auch wegen zu schlechter Erkennungsraten wieder verworfen (vgl. Scherer et al. (2000)). Bei dem im 7. Kap. beschriebenen Perzeptionsexperiment zeigt sich allerdings, dass einige Emotionen unterspezifiziert sind, weshalb bei dem nachfolgenden Experiment (Kap. 8) eine um mehrere Erregungsvarianten erweiterte Menge an Emotionskategorien berücksichtigt wird.

Kapitel 2

Parameter der lautsprachlichen Kommunikationskette

Dieses Kapitel dient der Einführung später verwendeter Begriffe. Es gibt einen Überblick über die menschliche Sprachverarbeitung und erläutert diverse Beschreibungsmodelle. Es gliedert sich, analog zur lautsprachlichen Kommunikationskette Sender-Übertragung-Empfänger in drei Bereiche. Im ersten Abschnitt wird die Generierung des Sprachsignals sowohl unter physiologischen Aspekten als auch unter Gesichtspunkten der Modellbildung betrachtet. Darauf folgt eine Besprechung unterschiedlicher Beschreibungsparameter für das Sprachsignal selber. Im dritten Teil wird letztlich auf die Perzeption eingegangen, wobei der Schwerpunkt auf den Grenzen der Wahrnehmbarkeit akustischer Merkmale liegt.

2.1 Der Erzeugungsmechanismus

Die menschliche Spracherzeugung entsteht grundsätzlich dadurch, dass zunächst ein Luftstrom erzeugt wird. Dieser wird bei stimmhaften Lauten durch die Glottis moduliert. Bei stimmlosen Lauten entsteht durch eine Verengung im Ansatzrohr eine hörbare Verwirbelung. Schließlich wird das so entstehende Anregungssignal durch die Form des Artikulationstrakts einer Filterung unterzogen. Diese drei Bereiche werden in den folgenden Abschnitten einzeln erläutert.

2.1.1 Luftstrommechanismen

Für die Erzeugung des Luftstroms gibt es zwei Richtungsmöglichkeiten und drei mögliche Luftstrommechanismen (Laver (1994, S. 161)). Die beiden Richtungen sind in den Körper hinein (ingressiv) und aus dem Körper heraus (egressiv). Für die Luftstrommechanismen kommen pulmonaler, glottaler und oraler Mechanismus in Frage. Der pulmonale Luftstrom wird durch die Atmungsorgane erzeugt. Der glottale Luftstrom wird durch ein abruptes Anheben oder Absenken des Larynx bei geschlossener Glottis erzeugt. Der orale Luftstrom entsteht innerhalb des Ansatzrohrs durch totalen Verschluss an einer bestimmten Stelle und der plötzlichen Lösung dieses Verschlusses, nachdem innerhalb ein Unter- oder Überdruck entstanden ist. Derart realisierte Laute werden *Clicks* genannt. Am weitaus häufigsten bei den Lauten der Sprachen der

Welt ist der egressive pulmonale Luftstrom. Dabei wird durch ein Anheben des Zwerchfells oder ein Absenken des Brustkorbs Luft aus den Lungen durch die Luftröhre in den Larynx gepresst.

2.1.2 Das Anregungssignal

Als Anregungssignal für die Lautbildung kommen periodische, aperiodische und transiente Signale in Frage. Aperiodische Signale kommen durch Friktion an einer Engestelle zustande und transiente durch plötzliche Öffnung eines geschlossenen Bereiches (wie etwa bei Plosiven). Periodische Signale werden durch eine Modulation des subglottalen Luftstroms durch Schwingungen der Stimmlippen erzeugt. Das anerkannteste Modell für die Stimmerzeugung durch die Glottis ist das myoelastisch-aerodynamische Modell. Die myoelastische Komponente ist hierbei durch die Elastizität der Stimmlippen und die Spannung der beteiligten Muskelgruppen gegeben, während die aerodynamische Komponente das Luftstromverhalten beim Durchströmen von Engpässen beschreibt (Laver (1980, S. 97)).

Bei stimmhafter Anregung baut sich unter der zunächst geschlossenen Glottis ein pulmonaler Druck auf. Ist der Druck groß genug geworden, werden die Stimmlippen auseinandergesprengt, und zwar vertikal gesehen nicht phasengleich sondern von unten nach oben. Nun entsteht zwischen den Stimmlippen eine sehr schmale Öffnung, durch die der Luftstrom in den Rachen strömt. Dadurch wirkt die Glottis wie ein Venturi-Rohr und durch die Bernoulli-Kräfte tritt ein lokales Druckminimum an der Glottis auf. Verstärkt durch das Bestreben der Stimmlippen, wieder in Ruhelage zu gelangen, schlagen diese wieder zusammen. Das Öffnen geschieht damit gradueller als das Schließen. Im Moment des Schließens ist die akustische Anregung des Artikulationstrakts normalerweise am stärksten. Nachdem sich wieder genug Druck unter der Glottis aufgebaut hat, wiederholt sich der Vorgang.

Dieser Vorgang lässt sich weitgehend entkoppelt durch die supralaryngale Region betrachten, wenn auch die Form des Artikulationstrakts eventuell eine Auswirkung auf das Vibrationsverhalten der Stimmlippen hat, z.B. durch stehende Wellen im Pharynx (vgl. Klatt & Klatt (1990)).

Exkurs: Der Aufbau des Kehlkopfs

Der Kehlkopf besteht aus gelenkig miteinander verbundenen Knorpeln, Muskulatur, Bändergewebe und Schleimhäuten (vgl. Laver (1980, S. 99)). Er ist durch Muskulatur mit dem Brustbein und dem Zungenbein verbunden, so dass seine Lage vertikal veränderbar ist. Der feste Teil des Larynx besteht aus drei Knorpeln. Der Cricoid (Ringknorpel) stellt die Basis für die anderen dar, die durch Muskeln mit ihm beweglich verbunden sind. Der Thyroid (Schildknorpel) liegt auf dem Cricoid auf und ist nach hinten geöffnet. Er kann nach vorne gekippt werden. An der offenen Seite des Thyroid liegen die beiden pyramidenförmigen Arytenoiden (Stellknorpel). Sie sind sehr beweglich und können sowohl gekippt und gedreht als auch seitlich bewegt werden.

An der inneren Vorderkante des Thyroid sitzt die Epiglottis auf (Kehlkopfdeckel). Die Stimmlippen sind zwischen dem mittleren Teil des Thyroid und den beiden Arytenoiden gespannt. Sie bestehen aus den inneren und den äußeren Thyroarytenoidmuskeln. Der Zwischen-

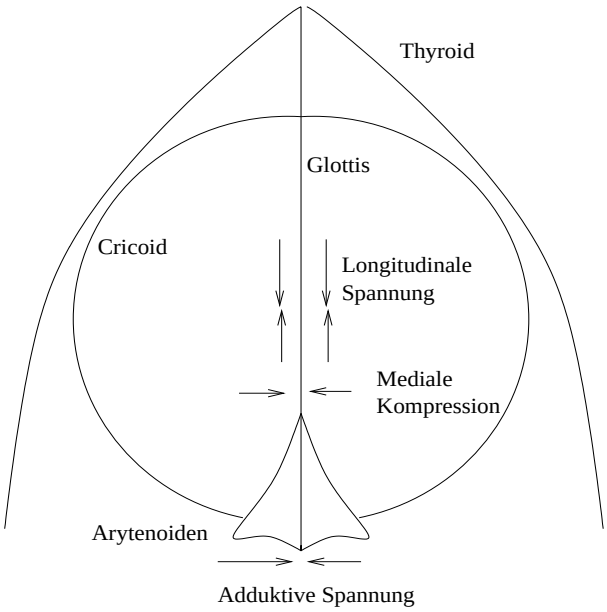

Abbildung 2.1: Glottismodell mit Spannungskräften, nach Laver (1980)

raum zwischen ihnen heißt *Glottis*. Dabei unterscheidet man zwischen dem Bereich an den Muskeln und der knorpeligen Glottis, die zwischen den Arytenoiden liegt.

Eine longitudinale Spannung der Stimmlippen kann entweder durch Kontraktion des Cricothyroid oder durch eine Anspannung der Thyroarytenoiden selber erfolgen. Der paarige hintere Cricoarytenoid dient zum Öffnen der Glottis (*Abduktion*), während sie durch Kontraktion des Cricoarytenoidus lateralis geschlossen wird (*Adduktion*). Dies geschieht jeweils durch Drehung der Arytenoiden. Weiterhin wird ein Schließen der Glottis durch Anspannung des internen Arytenoidmuskels gefördert, da hierbei die Arytenoiden zusammengeführt werden. Dies wird auch durch eine Anspannung des obliquen Arytenoidmuskels erreicht, da hierbei die Spitzen der Arytenoiden aufeinander zu bewegt werden. Zusätzlich wird das Schließen durch Kontraktion der Thyroarytenoiden selber begünstigt.

Zur Beschreibung der unterschiedlichen Phonationsarten ist es sinnvoll, die an der Glottis wirkenden Kräfteverhältnisse durch drei Kräfte zu modellieren: adduktive Spannung, mediale Kompression und longitudinale Spannung (vgl. Abb. 2.1). Die mediale Kompression unterscheidet sich von der adduktiven Spannung dadurch, dass sie auf den gewebeartigen Teil der Glottis wirkt, wodurch ein Öffnen des knorpeligen Teils auch bei medialer Kompression möglich ist.

Die Phonationsarten

Als Phonationsarten bezeichnet man verschiedene Formen der Anregung der Stimmlippen, die durch bestimmte Konfigurationen des Stimmapparates charakterisiert werden. Da die Phonationsart (zumindest im Deutschen) keine linguistische Funktion hat, wird sie stark durch den

emotionalen Sprecherzustand beeinflusst (vgl. z.B. Klasmeyer & Sendlmeier (2000)). Deshalb werden hier die zugrunde liegenden physiologischen Mechanismen und ihre akustischen Auswirkungen etwas näher erläutert.

Werden die Stimmlippen nicht angeregt, weil die Glottis weit abduziert ist, spricht man auch von 'Nullphonation'. Die Luft strömt dann relativ ungehindert durch die Glottis und erzeugt eventuell an höher liegenden Verengungen hörbare Verwirbelungen. Laver (1980) unterscheidet sechs grundlegende Phonationsarten, die zum Teil miteinander kombiniert werden können:

- **Modale Phonation** (*modal voice*)
  Die modale Phonation entspricht der 'normalen' Anregung. Dabei sind sämtliche Spannungen moderat gesetzt. Die Stimmlippen sind adduziert, mit moderater medialer Kompression und longitudinaler Spannung. Dadurch ähnelt das entstehende Zeitsignal dem Sägezahnsignal. Das Spektrum eines Sägezahns besteht aus spektralen Peaks bei der Grundfrequenz und den Vielfachen, wobei die Amplitude mit etwa 12 dB pro Oktave abfällt. Der Rauschanteil ist bei der modalen Phonation gering.

- **Behauchte Anregung** (*breathy voice*)
  Behauchte Anregung tritt ebenso wie die rauhe Stimme als Modifikator für modale Phonation auf. Die behauchte Phonation wird durch entspannte Kehlkopfmuskulatur realisiert. Dadurch ist sie nach Laver (1980, S. 133) mit keiner anderen Anregungsart als der modalen vereinbar. Eine mögliche Kombination mit der Knarrstimme wird hier allerdings unter Abschnitt 6.3.4 motiviert und begründet. Die Stimmlippen sind dick und oszillieren leicht, aber es wird kein vollständiger Verschluss erreicht, da die adduktive Spannung gering ist. Das entstehende Signal weist sowohl Periodizität als auch Rauschanteile auf, beides mit geringer Amplitude. Die höheren Frequenzbereiche sind stark gedämpft.

- **Flüsterstimme (***whispery voice*)
  Charakteristisch für die Flüsterstimme ist die Kombination von geringer adduktiver Spannung mit starker medialer Kompression. Dadurch entsteht an der knorpeligen Glottis das sogenannte 'Flüsterdreieck'. Die Luft strömt hauptsächlich durch die kleine Öffnung. Je nach longitudinaler Spannung und subglottalem Druck kann der gewebeartige Teil der Stimmlippen etwas zum Schwingen gebracht werden, aber die rauschhaften Komponenten überwiegen im akustischen Signal. In Kombination mit modaler Anregung oder Falsett erhöhen sich die Rauschanteile der resultierenden Stimme.

- **Falsettstimme** (*falsetto voice*)
  Bei der Falsettstimme ist die Glottis durch die Stellung der Arytenoiden adduziert. Der Cricothyroid und die äußeren Thyroarytenoiden sind angespannt, wodurch die Stimmlippen dünn und steif werden. Die inneren Thyroarytenoiden allerdings bleiben entspannt. Dadurch ist die adduktive Spannung groß, die mediale Kompression hoch und die passive longitudinale Spannung ebenfalls groß, während die aktive gering ist.

  In der Regel schwingen dadurch nur die innen liegenden Gewebeschichten der Stimmlippen mit sehr hoher Grundfrequenz. Die Glottis bleibt dabei oft leicht geöffnet, wodurch der subglottale Druck geringer als bei der modalen Phonation ist. Falls es zu keiner

Verschlussphase mehr kommt, entstehen starke Rauschanteile. Durch die hohe Grundfrequenz gibt es weniger Harmonische, die resultierende Sprache klingt 'dünner'. Die höheren Harmonischen sind mit etwa 20 dB/Oktave stark gedämpft.

- **Knarrstimme** (*creaky voice*)
  Die Knarrstimme ist durch starke adduktive Spannung und große mediale Kompression aber geringe longitudinale Spannung gekennzeichnet (vgl. Laver (1980, S. 123)). Eine Vibration der Stimmlippen ist dabei nur am gewebeartigen Teil in der Nähe des Thyroid möglich. Da zu dem der subglottale Druck eher gering ist, schwingen die dicken, trägen Stimmlippen mit geringer Frequenz und unregelmäßig. Die Perioden werden zum Teil als einzelne Pulse wahrgenommen. Die Steuerung der Grundfrequenz erfolgt hierbei hauptsächlich durch die aerodynamische Komponente der Stimmerzeugung, also durch Variation des subglottalen Drucks, der allerdings geringer als bei modaler Phonation ist.

- **Rauhe Anregung** (*harsh voice*)
  Die rauhe Anregung (harsh voice) ist keine eigenständige Phonationsart sondern tritt als Modifikator anderer Anregungsarten auf. Sie ist durch eine starke Anspannung des gesamten Stimmapparates gekennzeichnet. Akustisch macht sie sich vor allem durch Jitter und Shimmer bemerkbar. Sie ist der Knarrstimme ähnlich, liegt aber im Grundfrequenzbereich höher, nämlich etwas unter dem der modalen Anregung.

Ein mechanisches Modell für die Erzeugung des Glottissignals stellt das Zwei-Massen Modell von Flanagan & Ishizaka (1978) dar. Dabei werden die inneren und äußeren Stimmlippen durch zwei Massen modelliert, die mit Federn untereinander verbunden sind.

Generelle Muskelspannungszustände

Im Gegensatz zu Settings, die sich auf lokale Konfigurationen im Ansatzrohr oder am Larynx beziehen, ist auch die Unterscheidung zwischen allgemeinen muskulären Spannungszuständen sinnvoll. Bei Laver (1980) wird hier neben der normalen Spannung zwischen gespannter und ungespannter (*tense* und *lax*) Stimme unterschieden. Diese Unterscheidung ist im Hinblick auf emotionale Sprechweise günstig, da sich hieraus Vorhersagen für Emotionen ableiten lassen, die bzgl. der Aktivitätsdimension divergieren.

- **Gespannte Stimme** (*tense voice*)
  Die Stimme mit tendenziell angespannter Muskulatur ist eher laut und die Grundfrequenz ist eher hoch, bzw. auch bei tieferer F_0 liegt die wahrgenommene Grundfrequenz durch die verstärkte Energie der Frequenzen zwischen 1 und 4 kHz höher. Dies gilt auch für die wahrgenommene Lautstärke. Durch die allgemeine Anspannung wird oft eine rauhe Stimme erzeugt. Weiterhin wird faucale Spannung und eine Anhebung des Larynx wahrscheinlicher.

  Die Artikulatoren tendieren hierbei dazu, sich über ihre Zielpositionen hinauszubewegen, was zu Überartikulation führt. Die segmentale Dauer ist dabei eher länger als bei neutraler Sprechweise.

- **Entspannte Stimme** (*lax voice*)
  Bei der entspannten Stimme sind sowohl die adduktive Spannung als auch die mediale Kompression gering, wodurch als Anregungsart hauptsächlich eventuell eine behauchte modale Phonation auftreten kann. Weiterhin resultiert aus der geringen laryngalen Muskelspannung, dass die Stimme hier eher leiser und von niedriger Grundfrequenz ist.

  Die Entspannung im supralaryngalen Bereich kann in einer leichten Nasalität resultieren. Weiterhin ist der Larynx eher abgesenkt. Die Artikulatorbewegungen sind verringert, was zu einer zentralisierten Sprechweise führt. Dies resultiert in einer allgemeineren Verringerung der Formantbereiche.

2.1.3 Die Filterwirkung des Ansatzrohrs

Die Lautbildung entsteht durch Filterung des glottalen und/oder (im Falle von Frikativen oder Plosiven) supraglottalen Anregungssignals im Artikulationstrakt. Dessen Resonanzeigenschaften werden durch die Stellung der Artikulatoren (insbesondere der Zunge) bestimmt. In den folgenden beiden Abschnitten wird die Konfiguration des Ansatzrohrs unter zwei verschiedenen Gesichtspunkten betrachtet. Der erste Abschnitt befasst sich mit einer phonetischen Beschreibung länger andauernder (suprasegmentaler) Artikulationstraktkonfigurationen, den sogenannten supralaryngalen Settings. Der zweite Teil behandelt Möglichkeiten der Modellierung des Artikulationstrakts durch akustische Modelle.

Phonetische Beschreibung des Artikulationstrakts: die Settings

Der Begriff *Setting* bezieht sich auf eine bestimmte Konfiguration der spracherzeugenden Organe, die auf die segmentellen Konfigurationen Einfluss hat (vgl. Laver (1980, S. 2)). Insofern stellt es eine Einschränkung für die Artikulation dar. Ein Beispiel wäre etwa Koartikulation, aber auch länger andauernde Phänomene wie z.B. permanente Nasalierung fallen unter diesen Begriff. Die Settings sind in ihren akustischen Auswirkungen nicht notwendigerweise unabhängig voneinander, sondern beeinflussen sich gegenseitig. Zum Beispiel haben gerundete Lippen einen anderen Effekt, wenn gleichzeitig Velarisierung gegeben ist. Wirken sich verschiedene Settings auf dieselben Organe aus, so schließen sie sich entweder aus oder wirken additiv wie z.B. stimmhaftes Flüstern. Zudem unterscheiden sich Settings dadurch, inwieweit sie durch andere Settings auditiv maskiert werden. Nasalität ist z.B. ein Setting, das unter anderen Settings weniger prominent wirkt. Zudem gibt es auch einen Zusammenhang zwischen den Auswirkungen von Settings und der sprecherindividuellen Physiognomie.

Man unterscheidet zwischen phonatorischen (laryngalen) und supralaryngalen Settings. Die laryngalen Settings sind bereits oben mit den Phonationsarten besprochen worden. Da auch die supralaryngalen Settings weitgehend keine linguistische Funktion haben[1], stellen sie einen wichtigen Aspekt hinsichtlich emotionaler Sprechweise dar. Die supralaryngalen Settings können in drei Gruppen eingeteilt werden, je nachdem ob sie die longitudinale Form, die latitudinale Form des Ansatzrohrs verändern oder die velopharyngale Region modifizieren.

[1] Einige Ausnahmen hiervon werden in Abschnitt 4.2 besprochen.

Longitudinale Settings. Laver (1980) beschreibt vier Arten, die longitudinale Ausdehnung des Ansatzrohrs zu verändern. Der Larynx kann angehoben bzw. abgesenkt werden. Weiterhin können die Lippen vorgestülpt werden und viertens kann die Unterlippe angehoben und nach hinten gezogen werden, was zu labiodentalisierter Sprechweise führt. Dem Autor erscheint allerdings zusätzlich eine spezielle Berücksichtigung des Spreizens der Lippen vor allem im Hinblick auf lächelnde Sprechweise als sinnvoll.

- **Anheben des Larynx**
  Die Anhebung des Larynx wird auf zwei Arten ermöglicht. Einerseits können die Muskeln zwischen Larynx und Zungenbein kontrahieren, andererseits kann das Zungenbein selber angehoben werden. Dies führt zu einer Verkürzung des Ansatzrohrs um bis zu 1,5 cm, was eine Erhöhung vor allem des dritten und vierten Formanten zur Folge hat. Der erste und zweite Formant werden vor allem bei offenen Vokalen angehoben. Da beim Heben des Larynx auch der Cricothyroidmuskel gespannt wird, ist dies in der Regel auch durch einen Anstieg der Grundfrequenz begleitet. Weiterhin klingt die Stimme bei angehobenem Larynx angestrengt, was durch die Spannung der Muskeln bei dieser unnatürlichen Haltung verursacht wird. Dieses Setting ist damit vor allem eine Folge starker Muskelspannung, wie sie etwa bei wütender Sprechweise auftritt.

- **Absenken des Larynx**
  Der Larynx kann durch Kontraktion des Sternothyroidmuskels, der den Kehlkopf mit dem Brustbein verbindet, abgesenkt werden. Dadurch wird das Ansatzrohr um bis zu 1 cm verlängert, was allgemein zu einer Absenkung vor allem der unteren Formanten führt. Da das Absenken des Larynx eine Entspannung der Kehlkopfmuskulatur mit sich bringt, klingt die resultierende Stimme eher behaucht. Die Wahrscheinlichkeit des Auftretens dieses Settings erhöht sich damit unter emotionalen Zuständen, die von muskulärer Entspannung geprägt sind.

- **Vorstülpen der Lippen**
  Das Vorstülpen der Lippen verlängert den Artikulationstrakt. Dadurch werden vor allem die höheren Formanten abgesenkt. In der Regel geht dies mit einer Rundung der Lippen einher.

- **Labiodentales Setting**
  Durch labiodentalisierte Sprechweise wird das Ansatzrohr leicht verkürzt. Da die Verengung dabei in der Regel zu groß ist, als dass dadurch ein Rauschen entsteht, wirkt sich dieses Setting hauptsächlich auf Segmente aus, deren Artikulationsort in der Nähe der Mundöffnung liegt. Oft geht es mit einer Rundung der Lippen einher, was zusammen mit der Verengung der Mundöffnung akustisch eher zur Absenkung vor allem der höheren Formanten führt.

- **Spreizen der Lippen**
  Ein Spreizen der Lippen tritt z.B. bei lächelnder Sprechweise auf und resultiert in einer Verkürzung des Sprechtrakts. Dadurch werden die Formanten grundsätzlich erhöht, wobei allerdings dieser Effekt für die höheren Formanten eventuell durch die entstehende

Verengung der Mundöffnung wieder kompensiert wird. Die akustischen Auswirkungen eines Spreizens der Lippen werden weiterhin unter Abschnitt 6.3 diskutiert.

Latitudinale Settings. Die latitudinalen Settings bedeuten eine quasipermanente Konstriktion oder Erweiterung an einer bestimmten Stelle des Artikulationstrakts. Eine Unterscheidung nach dem hauptsächlich verursachenden Organ liefert fünf Gruppen, je nachdem ob die Manipulation durch die Lippen, die Zunge, die faucalen Stränge, den Rachen oder den Kiefer verursacht wird.

- **Labiale Settings**
  Unter der Annahme, dass die Lippen immer auf der gleichen Ebene liegen, lassen sich 18 verschiedene Formen der Mundöffnung unterscheiden. Diese Zahl entsteht aus der Kombination von horizontaler und vertikaler Verengung oder Weitung der Lippen mit und ohne Vorstülpen, zuzüglich der neutralen Lage. Obwohl alle davon physiologisch möglich sind, sind einige nur unter erhöhtem Aufwand zu bewerkstelligen und treten daher selten auf. Zum Beispiel werden bei vertikaler Weitung bzw. horizontaler Verengung die Lippen in der Regel auch vorgestülpt, während dies bei horizontaler Weitung fast nie vorkommt.

  Akustisch am einfachsten zu diskriminieren sind die Unterschiede zwischen Spreizung und Rundung der Lippen, die jeweils generell zu einer Erhöhung bzw. Absenkung der Formanten führen, wobei die Art der betroffenen Formanten hauptsächlich von der Größe der Mundöffnung abhängt. Bei gerundeten Lippen lässt sich gut zwischen ’offener’ und ’geschlossener’ Rundung unterscheiden.

- **Linguale Settings**
  Linguale Settings zeichnen sich dadurch aus, dass die Zunge an einem bestimmten Ort im Ansatzrohr tendenziell eine Verengung erzeugt. Man könnte hier nach dem hauptverantwortlichen Teil der Zunge unterscheiden, also Spitze, Schwert, Körper oder Wurzel oder eine Unterscheidung aufgrund der betroffenen Stelle treffen.

  Da die Zunge als beweglichster Artikulator stark an der Artikulation der meisten Laute beteiligt ist, ist über einen längeren Zeitraum eine zu feine Unterscheidung nicht sinnvoll. Oft wird z.B. zwischen vorderer und zurückgezogener Zungenlage diskriminiert. Üblich ist auch eine Einteilung in Dentalisierung, Velarisierung oder Pharyngalisierung, wodurch die Tendenz beschrieben wird, die Zunge in der jeweiligen Region zu halten. Bei palatalisierter Stimme liegt der zweite Formant recht hoch. Je weiter die Verengung in den vorderen Bereichen des Mundraumes liegt, also bei Dentalisierung oder Alveolisierung, desto stärker steigt der zweite Formant. Bei Settings mit zurückgezogener Zunge liegt der erste Formant höher und der zweite tiefer als beim neutralen Setting. Es gibt einige Sprachen und Dialekte, bei denen bestimmte linguale Settings besonders häufig zu beobachten sind. Das Französische wird z.B. von vielen Phonetikern als eine Sprache beschrieben, bei der vordere Zungenlage vorherrscht.

  Eine tendenzielle Verschiebung der Zunge in einen bestimmten Bereich entspricht dabei einer Vergrößerung des Raumes in davon entfernten Bereichen. Ein weiteres labiales Setting besteht darin, die Zunge kaum aus ihrer Mittellage zu bewegen (zentralisierte

Sprechweise). Es tritt oft bei traurigem oder gelangweiltem Sprechausdruck auf. Zudem kann die Tendenz bestehen, vorrangig mit der Spitze anstatt dem Zungenschwert zu artikulieren oder eine permanente Tendenz zu retroflexer Zungenlage aufzuweisen. Die akustischen Folgen, die durch Settings der Zungenspitze oder des Schwertes entstehen, sind sehr komplex, da sie durch die Lage des Zungenkörpers beeinflusst werden. Das retroflexe Setting hat vor allem eine Senkung des dritten und vierten Formanten zur Folge.

- **Faucale Settings**
  Eine weitere Gruppe von Settings ist durch die faucalen Stränge gegeben. Diese bestehen aus zwei Muskelgruppen, dem Palatoglossus und dem Palatopharyngeus, die zwischen Mund- und Rachenraum liegen und den weichen Gaumen mit dem Rachen und der Zunge verbinden. Durch Kontraktion erzeugen sie eine Enge an dieser Stelle. Weiterhin kann durch sie das Velum abgesenkt und die Zungenwurzel angehoben werden. Auch hierbei wird wie bei zurückgezogener Zunge der erste Formant angehoben und der zweite abgesenkt. Eine Anspannung der faucalen Stränge hat in der Regel auch eine Spannung der laryngalen Muskeln zur Folge. Dieses Setting würde etwa mit dem Konzept der 'Rachenenge' nach Trojan (1975) korrespondieren, also bei geekelter Sprechweise auftreten.

- **Pharyngale Settings**
  Eine Verengung des Pharynx kann sowohl durch die Zungenwurzel als auch durch eine Kontraktion der Muskeln an den Rachenwänden selber herbeigeführt werden. Der akustische Effekt einer Verengung des Rachens besteht wiederum in einer Anhebung des ersten und Senkung des zweiten Formanten, während eine Erweiterung in einer Absenkung des ersten Formanten resultiert. Eine Anspannung führt zu weniger absorbierenden Rachenwänden und damit zu schmaleren Formantbandbreiten.

- **Mandibulare Settings**
  Settings, welche die Lage des Kiefers betreffen, unterscheiden sich vor allem durch dessen Öffnungsgrad. Dabei interagieren sie mit den labialen Settings, die ja dieselbe Region betreffen. Akustisch führt eine stärkere Öffnung des Kiefers zu einer Anhebung vor allem des ersten Formanten, während er durch einen eher geschlossenen Kiefer abgesenkt wird. Sendlmeier (1997) weist darauf hin, dass die von ihm beobachtete Absenkung des ersten Formanten bei Trauer, Langeweile und Angst auf die Tendenz zurückzuführen ist, den Kiefer nicht richtig zu öffnen, während bei Freude oder Wut der gegenteilige Effekt beobachtet wurde.

Velopharyngale Settings. Velopharyngale Settings beziehen sich hauptsächlich auf das Absenken bzw. Anheben des Velums. Da das Velum unter anderem durch Mithilfe der faucalen Stränge gesenkt wird, ist dies oft von einem Zurückziehen der Zunge begleitet. Mehrere Stufen der Nasalität lassen sich aus unterschiedlichen Graden der Koppelung des Nasaltrakts zum Mundraum ableiten. In der neutralen Lage ist das Velum leicht gesenkt.

Akustisch gesehen führt die Zuschaltung des Nasaltrakts zu sogenannten nasalen Formanten bei etwa 250, 1000 und 2500 Hz. Durch Auslöschungen aufgrund der beiden Resonanzkammern kommt es zu Antiresonanzen bei etwa 600 Hz und zwischen 900 und 1800 Hz. Diese hängen

beträchtlich von der sonstigen Konfiguration des Ansatzrohrs ab. Weiterhin ist ein allgemeiner Abfall der Energie vor allem der höheren Frequenzbereiche zu beobachten sowie eine allgemeine Verbreiterung der Formantbandbreiten. Die Intensität vor allem des ersten Formanten ist dabei erheblich verringert. Sendlmeier & Heile (1998) haben bei der Analyse von Sprachaufnahmen mit simuliertem Wohlbehagen vermehrt Nasalierung beobachtet, was auf die starke Entspannung zurückzuführen ist.

Akustische Modellierung des Artikulationstrakts

Zur Voraussage der spektralen Eigenschaften eines (zunächst stationären) Signals aufgrund der Filterung eines Anregungssignals mit einem Resonator gibt es verschiedene Modelle. Im einfachsten Fall wird das Ansatzrohr als Viertel-Wellenlängen Resonator (vgl. etwa Ungeheuer (1962) oder Fant (1960)) modelliert. Dabei wird der Artikulationstrakt als gleichförmiges Rohr betrachtet, das an der Glottis geschlossen und an den Lippen offen ist. Somit herrscht an der Glottis maximaler Schalldruck und an den Lippen maximale Schallschnelle. Alle Frequenzen, deren Periodenlänge nun von der Form ist, dass der Abstand zwischen einem Maximum (=maximaler Druck) und einem Nulldurchgang (=maximale Schnelle) gerade der Länge des Resonators entspricht, werden dann verstärkt. Aus diesem Modell ergeben sich beispielsweise die Frequenzen der ersten i Formanten für den neutralen Vokal [ə] nach der Formel

$$F_i = \frac{(2i-1)c}{4l} \tag{2.1}$$

zu 500, 1500, 2500, ... Hz (bei Schallgeschwindigkeit $c = 340$ m/sek und Artikulationstraktlänge $l = 17$ cm). Daraus ergibt sich direkt, dass die Formanten bei einer Verlängerung des Ansatzrohrs tiefer liegen und bei einer Verkürzung steigen (vgl. etwa Fant (1960, S. 64)).

Um weitere Vokale modellieren zu können, muss das Ansatzrohr in mehrere Röhrenabschnitte eingeteilt werden. Für das [a] etwa besteht eine einfache Approximation des Sachverhaltes darin, zwei gleichlange Röhrenstücke anzusetzen, von denen das erste deutlich schmaler ist. Dadurch können beide wieder jeweils als Viertel-Wellenlängen Resonatoren betrachtet werden (da beide an einem Ende relativ geschlossen und am anderen Ende relativ offen sind). Da die Länge der Abschnitte nun die Hälfte der Gesamtlänge des Ansatzrohrs beträgt, verdoppeln sich die Resonanzfrequenzen. Weil sich Formanten aber aufgrund akustischer Kopplung jeweils nur auf bis zu 200 Hz annähern können, ergeben sich die ersten vier Formantwerte zu 900, 1100, 2900 und 3100 Hz. Dies stellt natürlich eine sehr grobe Vereinfachung dar; diese theoretischen Werte liegen jedoch recht nah an empirisch für das [a] ermittelten Formantfrequenzen.

Das [i] lässt sich durch ein breites Segment, an das sich ein schmales anschließt, modellieren. Dadurch entstehen zwei Teilstücke, von denen eines an beiden Enden relativ geschlossen und das andere relativ offen ist. Sie stellen damit Halb-Wellenlängen-Resonatoren dar, da sie Frequenzen verstärken, deren Abstand zwischen den Maxima bzw. den Nulldurchgängen gerade der Länge des Resonators entspricht. Daraus ergeben sich analog zu Gleichung 2.1 aus

$$F_i = \frac{ic}{2l} \tag{2.2}$$

Resonanzen für den i-ten Formanten zu geradzahligen Vielfachen von 1000 Hz. Da es sich im Falle des [i] um zwei Resonatoren handelt, deren Resonanzfrequenzen sich wieder auf maximal 200 Hz annähern können, ergeben sich der zweite und dritte Formant jeweils zu 1900 respektive 2100 Hz. Da weiterhin eine Verengung am vorderen Bereich des Ansatzrohrs insgesamt gesehen eine Absenkung des ersten Formanten zur Folge hat, liegt dieser beim [i] bei etwa 250 Hz.

Das Filter für Konsonanten lässt sich als ein dreisegmentiges Rohr modellieren, bei dem das Mittelstück eine Verengung darstellt (vgl. Fant (1960, S.73)). Somit ergeben sich Resonanzen für zwei Halb-Wellenlängen-Resonatoren (das erste und mittlere Stück) und einen Viertel-Wellenlängen-Resonator (zwischen Konstriktion und Mundöffnung). Dabei ist die Länge des vorderen Abschnitts von entscheidender Bedeutung, da das Anregungssignal, welches ja an der Konstriktion entsteht, vor allem am vorderen Segment resoniert.

Die Modellierung des Ansatzrohrs durch Röhrenstücke wird im sogenannten Röhrenmodell verfeinert, welches das Artikulationstraktfilter durch 8-12 Röhrensegmente beschreibt (vgl. Markel & Gray (1976, S. 60)). Die Querschnitte werden aus realen Spektrogrammen durch ein mathematisches Optimierungsverfahren ermittelt (Fellbaum (1984, S. 62)). Viele artikulatorische Synthesizer basieren auf diesem Modell (vgl. Abschnitt 3.4).

2.2 Akustische Parameter

Bei der Beschreibung von Merkmalen der Lautsprache sind mehrere Ebenen voneinander zu unterscheiden. Grundlegend ist die akustische Ebene, in der sich Merkmale durch physikalische Maßeinheiten ausdrücken lassen (z.B.: *"Die mittlere Grundfrequenz beträgt 120 Hz.*"); davon abzugrenzen ist die Perzeptionsseite, bei der subjektiv wahrgenommene Eigenschaften beschrieben werden. So ist etwa zwischen der Grundfrequenz (*fundamental frequency*) und der wahrgenommenen Tonhöhe (*pitch*) zu unterscheiden. Diese hängt eben nicht nur von der F_0 ab, sondern auch von der Amplitude der Obertöne. Eine weitere Ebene ist durch linguistische Beschreibungsgrößen gegeben. Begriffe wie ‚Betonung' oder ‚Akzent' sind sehr abstrakt und funktional bestimmt. Akustisch sind sie oft nur durch das komplexe Zusammenspiel mehrerer Parameter zu beschreiben. Da sich letztlich alle Beschreibungsebenen auf das akustische Signal beziehen sollten, ist dieser Abschnitt mit 'akustische Parameter' betitelt.

Die akustischen Merkmale sind in den drei Hauptgruppen prosodische, stimmqualitative und artikulatorische Eigenschaften zusammengefasst. Dieser Abschnitt dient damit vor allem dem Verständnis des 4. Kapitels, in dem der Stand der Forschung besprochen wird.

2.2.1 Prosodische Merkmale

Prosodische Merkmale sind solche, die sich über mehrere Einzellaute, also suprasegmental, erstrecken und vom spezifischen Lautinventar einer Sprache weitgehend unabhängig sind, wie *Sprechrhythmus* und *Sprechmelodie*. Die Prosodie umfasst damit Dauer, Lautheit und Tonhöhenverlauf (Intonation) auf suprasegmentaler Ebene. Wird durch Prosodie Information vermittelt, die nicht syntaktischer oder semantischer Art ist, spricht man von paralinguistischer Information bzw. nonverbaler vokaler Information, ansonsten von linguistischer Information

(vgl. Scherer et al. (1984, S. 1346)).

Die prosodischen Parameter sind einfacher zu messen als etwa Parameter der Stimmqualität, was ein Grund dafür ist, dass die meisten Untersuchungen zu emotionaler Sprechweise diese Parameter beschrieben haben. Sie spielen auch fraglos eine große Rolle für paralinguistische Informationsübertragung, haben andererseits aber auch die stärkste Rolle als Träger syntaktischer und semantischer Kennzeichnung.

Betonungen

Betonungen werden meist auf Silbenebene annotiert. Die meisten Untersuchungen in der Literatur unterscheiden drei Stufen der Betonung:

- Die *Hauptbetonung* entspricht der Satz- oder Phrasenbetonung und gibt den Satzfokus an. Sie liegt damit in der Regel auf der wortbetonten Silbe des semantisch wichtigsten Wortes. Je nach Definition kann es auch mehrere Hauptbetonungen in einer Phrase bzw. einem Satz geben.

- *Wortbetonungen* liegen auf der Silbe, die bei Mehrsilbern am stärksten betont ist.

- *Unbetont* sind alle restlichen Silben.

In der Regel wirken längere Silben mit angehobener F_0 und stärkerer Intensität stärker betont als andere. Die Intensität scheint dabei allerdings perzeptiv nicht sehr relevant zu sein, weshalb sie bei den meisten Synthesesystemen zur Modellierung der Betonung nicht berücksichtigt wird.

Im Zusammenhang mit der Betonung wird auch der Begriff 'Akzent' verwendet. Mitunter wird hier Akzent im Sinne eines Fokus innerhalb einer Betonungsgruppe verstanden. Ein Akzent erstreckt sich dann über eine Gruppe von Silben, deren Intonationskontur eine geschlossene Form aufweist. Anders ausgedrückt wird damit die Intonationskontur beschrieben, die eine anfängliche Steigung und eventuell einen abschließenden Fall beinhaltet und eine betonte Silbe enthält (vgl. Paeschke & Sendlmeier (2000) und Peters (1999)).

Durchschnittliche Silbenlängen

Eine weitere für emotionale Sprechweise wichtige Parameterklasse stellen die Zeitmaße dar, die durch Parameter wie Gesamtlänge einer Äußerung, Anzahl und Dauern von Sprechpausen, Sprechrate oder durchschnittliche Lautdauern bestimmt werden. Die meisten Untersuchungen verwenden als globalen Dauerparameter die Sprechrate oder Sprechgeschwindigkeit, welche uneinheitlich in Wörtern pro Minute oder Silben pro Sekunde ausgedrückt wird. Hierbei wird die Anzahl der verwendeten Einheiten inklusive Pausen durch die gebrauchte Zeit geteilt. Zusätzlich kann die Artikulationsrate berechnet werden, bei der die Zeit ohne Pausen verwendet wird (Scherer & Scherer (1981)).

Durchschnittliche Dauern für einzelne Lautklassen

Zur Analyse emotionaler Sprachkorpora wird die Datenbank in der Regel phonetisch transkribiert. Aus den Dauern für die einzelnen Phone lassen sich dann verschiedene statistische Aussagen ableiten, wie z.B. die durchschnittliche Dauer aller stimmhaften Frikative.

Sprechpausen

Auch die Anzahl und durchschnittliche Dauer von Sprechpausen kann für unterschiedliche Emotionen signifikant variieren. Abadjieva et al. (1993); Cahn (1990) oder Williams & Stevens (1972) messen z.B. ein erhöhtes Sprechpausenvorkommen für traurige Sprechweise. Ein weiterer Parameter im Zusammenhang mit Sprechpausen ist der Laut/Stille-Quotient (vgl. Scherer (1982)), der die Gesamtdauer der Sprechpausen mit der Vokalisationszeit in Beziehung setzt.

Durchschnittliche Grundfrequenz und F_0-Range

Aus der Grundfrequenz lassen sich verschiedene globale Parameter wie etwa die durchschnittliche Tonhöhe, der F_0-Range oder die Variabilität ableiten. Die Variabilität ist ein weiterer Streuungsparameter, üblicherweise wird dabei die Standardabweichung verwendet. Um dem logarithmischen Hörempfinden des Menschen Rechnung zu tragen, sollten Grundfrequenzabstandswerte in Halbtönen oder Prozent angegeben werden. Die Distanz D zweier Frequenzen f_1 und f_2 in Halbtönen lässt sich durch folgende Formel berechnen (t'Hart et al. (1990, S. 24)):

$$D = 12 \log_2 \frac{f_1}{f_2} = \frac{12}{\log_{10} 2} \log_{10} \frac{f_1}{f_2}$$

Weiterhin lässt sich aus der Grundfrequenzmessung der Konturverlauf von Äußerungen beschreiben. Dieser wird sinnvollerweise getrennt auf silbischer Ebene und über ganze Äußerungen bzw. Phrasen hinweg betrachtet.

Intonation auf Phrasenebene

Auf Phrasenebene lässt sich eine Intonationskontur, abgesehen von den globalen Durchschnittswerten wie F_0-Range oder mittlere Grundfrequenz, durch eine Verlaufsform beschreiben. Dabei wird in der Regeln die Deklination anhand der *Baseline* (der Interpolation des ersten und letzten Minimums der F_0-Kontur) oder der *Topline* (analog zur Baseline die Interpolation zwischen den Maxima) betrachtet (vgl. z.B. Paeschke & Sendlmeier (2000)).

Intonation auf Silbenebene

Die Intonationsverläufe auf Silbenebene wurden ebenfalls in mehreren Studien untersucht (z.B. Paeschke & Sendlmeier (2000) oder Mozziconacci & Hermes (1997)). Dabei werden die F_0-Konturverläufe über den Silben anhand eines Stilisierungsalgorithmus' (z.B. d'Alessandro & Mertens (1995)) unterschiedlichen Konturtypen zugeordnet. Als Typen werden meistens bitonale Figuren verwendet, wie steigend, fallend, gerade, steigend/fallend, usw.. Alternativ dazu

kann auch eine Analyse anhand einer Tobi-Labelung (*Tones and Break Indices*, vgl. etwa Grice et al. (1996)) erfolgen. Die relative Häufigkeit des Auftretens der Konturtypen und die durchschnittlichen Steigungen können dann für unterschiedliche Emotionen verglichen werden.

Intensität

Die Intensität von Sprachsignalen wird subjektiv als *Lautheit* wahrgenommen. Sie wird durch Respiration und Phonation (vgl. Abschnitt 2.1) beeinflusst. Physikalisch entspricht sie dem Schalldruck, im Signal drückt sie sich durch die Amplitude aus. Eine Schwierigkeit bei der Bestimmung der objektiven Intensität von Sprachaufnahmen besteht darin, den Abstand und die Richtung zum Mikrophon konstant und alle Geräte kalibriert zu halten. Die Rolle der relativen Lautheit zur Wahrnehmung von Betonung (im Gegensatz zu relativer Tonhöhe und Dauer) ist so umstritten, dass sie oft ignoriert wird.

2.2.2 Stimmqualitative Merkmale

Die Stimmqualität wird subjektiv als *Timbre* wahrgenommen. Sie wird durch Phonation und Artikulation beeinflusst. Zur objektiven Messung können etwa Relationen verschiedener spektraler Bänder verwendet werden (wie etwa bei Klasmeyer & Sendlmeier (2000) oder Banse & Scherer (1996)). Eine gespannte Stimme zeichnet sich z.B. durch mehr Energie in den höheren Frequenzbereichen aus.

Ein weiterer Parameter zur Beschreibung der Stimmqualität ist der *Frequenzrange*, der den Abstand zwischen Grundfrequenz und höchster vorhandener Frequenz angibt. Weiterhin lässt sich die Qualität akustisch durch das Vorhandensein und die Stärke von Rauschanteilen in bestimmten Frequenzbereichen beschreiben.

Stimmqualität allein kann schon zur Erkennung von emotionalen Zuständen des Sprechers führen (Scherer (1995)). Die Stimmqualität ist ein sehr wichtiger Parameter für den emotionalen Gehalt, da sie einen geringen Einfluss auf die Verständlichkeit von Sprache hat, und nicht Träger linguistischer Information ist (vgl. Scherer et al. (1984)).

Beschreibung des Anregungssignals

Zur Beschreibung der Phonationsarten dient in der Regel die Betrachtung des Glottissignals. Da es keine Möglichkeit gibt, das Signal direkt aufzunehmen, werden indirekte Zugangswege verwendet. Dabei werden vor allem Elektroglottogramme (Laryngogramme) oder durch inverse Filterung gewonnene Quellsignale analysiert. Zur Parametrisierung des Glottissignals bietet sich das LF-Modell (Fant et al. (1985)) an (vgl. Abschnitt 3.5.3). Die Anpassung des Modells an das Signal ist allerdings sehr aufwendig und muss bislang von Hand durchgeführt werden, was die Untersuchung größerer Datenmengen sehr erschwert (vgl. Gobl & Ni Chasaide (1992) oder Strik & Boves (1991)).

Ein für viele Emotionen wichtiger Parameter ist der *Jitter* (äquivalent dazu: F_0-Perturbation). Er beschreibt Schwankungen der Grundfrequenz von einer Periode zur nächsten.

Ein weiterer Parameter zur Beschreibung des Anregungssignals ist der *Shimmer*, der das Vorhandensein kurzzeitiger Schwankungen der Amplitude angibt. Jitter und Shimmer sind allerdings schwierig mit automatisierten Verfahren zu messen.

Lage und Bandbreiten der Formanten

Das Resonanzverhalten des Ansatzrohrs wird ausschließlich durch die Artikulation bestimmt. Es wird durch die Lage der Formanten beschrieben, die bestimmte Peaks in der Energieverteilung darstellen und durch Mittenfrequenz und Bandbreite festgelegt sind. Diese treten vor allem bei Vokalen auf, wobei die Lage der untersten drei die Lautqualität bestimmen und die Lage der höheren vor allem zum Klangeindruck beitragen. Als solche sind sie sprechercharakteristisch, da ihre Lage unter anderem von der Länge des Ansatzrohrs abhängt. Wenn auch ihre relative Lage weitgehend durch den artikulierten Vokal festgelegt ist, gibt es doch in der absoluten Lage durchaus Spielraum. Eine Verkürzung des Sprechtrakts, wie sie etwa durch lächelnde Sprechweise entsteht, führt z.B. zu einer generellen Erhöhung der Formanten (Fant (1957)).

Eine automatische Formantextraktion ist sehr fehleranfällig[2]. Dies gilt speziell für Sprache mit hoher Grundfrequenz, da diese unter Umständen über dem ersten Formanten liegen kann. Formantlagen und Bandbreiten sind explizite Parameter von Formantsynthesizern (siehe Abschnitt 3.4.1), weshalb diese zur Simulation emotionaler Sprechweise sehr geeignet sind.

Energieverteilung der spektralen Bänder

Die Energieverteilung in spektralen Bändern kann, wie oben bemerkt, der Beschreibung der Stimmqualität dienen. Hierfür kann der durchschnittliche Amplitudenwert für bestimmte Frequenzbereiche verwendet werden. Normalerweise werden dabei Langzeitspektren ausgewertet, bei denen eventuell Abschnitte mit bestimmten Lauteigenschaften (z.B. Stimmhaftigkeit) zusammengefasst werden.

2.2.3 Artikulatorische Merkmale

Durch die artikulatorischen Merkmale werden Reduktionen und Elaborationen auf segmenteller Ebene quantifiziert. Sie dienen damit der Bestimmung des artikulatorischen Aufwands und korrelieren mit der Aktivitätsdimension, da sie durch generelle Muskelspannungszustände beeinflusst werden.

De- bzw. Zentralisierung von Vokalen

Der artikulatorische Aufwand bei der Realisierung von Vokalen kann durch Bestimmung der Formantlagen gemessen werden. Liegen der erste und zweite Formant eher zentralisiert (*Vowel-Target Undershoot*), so ist der betreffende Vokal reduziert worden. Im entgegengesetzten Fall, wenn die Formanten über den bei neutraler Sprechweise erreichten Zielfrequenzen liegen, ist die Vokalrealisation elaboriert worden (*Vowel-Target Overshoot*). Für die Vergleiche werden in

[2] Ein Algorithmus dafür ist z.B. in Mannell (1998) beschrieben.

der Regel die Formantwerte aller auftretenden Varianten eines Vokals gemittelt (vgl. Kienast & Sendlmeier (2000)).

Artikulatorischer Aufwand bei Konsonanten

Als Parameter zur Bestimmung der Reduktion bzw. Elaboration stimmloser Frikative wurde die spektrale Balance (*spectral balance* oder *center of gravity*) vorgeschlagen (Van Son & Pols (1999)). Diese berechnet sich wie folgt:

$$SB = \frac{\sum f_i E_i}{\sum E_i}$$

Dabei ist f_i die i-te Frequenz im Spektrum und E_i der Energiewert bei dieser Frequenz. Die spektrale Balance korreliert bei stimmlosen Lauten mit dem Engegrad der Konstriktion; je enger die Konstriktion desto größer die spektrale Balance. Daher kann sie als Parameter zur Bestimmung des artikulatorischen Aufwands herangezogen werden (vgl. Kienast & Sendlmeier (2000)).

Koartikulationseffekte und lautliche Elisionen und Epenthesen

Zur Bestimmung der Koartikulationseffekte bietet sich ein Vergleich zwischen neutraler und emotionaler Sprechweise bezüglich der Anzahl und des Grads der Assimilationen an.

Die Quantifizierung von Epenthesen oder Elisionen kann durch den Vergleich der Transkription mit einer Normlautung geschehen. Bei der Untersuchung emotionaler Sprechweise bietet es sich an, als Referenz nicht eine abstrakte Normlautung zugrunde zu legen, sondern die neutrale Sprechweise. Bei der qualitativen Erfassung von Elisionen und Epenthesen wird der sogenannte Lautminderungsquotient LMQ (nach Hildebrandt (1963)) wie folgt berechnet:

$$LMQ = 10\frac{10 - n_{emot}}{n_{neut}}$$

n_{neut} entspricht dabei der Anzahl der Laute bei neutraler und n_{emot} der Anzahl bei emotionaler Sprechweise (Kienast & Sendlmeier (2000)). Ein positiver LMQ zeigt damit das Auftreten von Elisionen bei den emotionalen Äußerungen an, während ein negativer auf Epenthesen hinweist.

2.3 Verarbeitung von Sprachsignalen beim Menschen

Um die Frage nach den perzeptiv relevanten akustischen Merkmalen emotionaler Sprechweise zu beantworten, muss man zunächst klären, welche akustischen Eigenschaften der Lautsprache überhaupt perzeptiv relevant sind. Deshalb werden in dem folgenden Abschnitt zunächst kurz die physiologischen Gegebenheiten der Hörorgane betrachtet. Darauf aufbauend, werden akustische Merkmale wie Grundfrequenz, spektrale Zusammensetzung, Lautintensitäten und -Dauern bzgl. ihrer Wahrnehmbarkeitsgrenzen besprochen.

2.3.1 Physiologie des Gehörs

Die menschliche akustische Signalverarbeitung ist ein Prozess, der zunächst durch die mehrfache Wandlung des Signals geprägt ist. Diese findet in den drei Bereichen des Ohres *Außenohr*, *Mittelohr* und *Innenohr* statt. Bereits durch die Form des Gehörgangs werden bestimmte Frequenzbereiche angehoben. Am Trommelfell werden die Luftdruckschwankungen auf die Gehörknöchelchen übertragen, die auf der anderen Seite mit dem ovalen Fenster verbunden sind. Dort wird die mechanische Bewegung wiederum in Druckwellen umgewandelt, diesmal im flüssigen Medium. In der Cochlea werden durch die Wellen in Abhängigkeit von der spektralen Zusammensetzung verschiedene Haarzellen erregt, die elektrische Signale aussenden. Diese werden durch Nervenbahnen zum Gehirn übertragen und dort interpretiert.

In der Cochlea befinden sich etwa 3600 Haarzellen zwischen Helicotrema und ovalem Fenster, mit einem Abstand von jeweils 9 Mikrometer. Diese werden durch unterschiedliche Frequenzen angeregt. Da sich der Bereich der wahrnehmbaren Frequenzunterschiede von 50 bis etwa 16 kHz in 640 Stufen unterteilen lässt, können also Reize auf bis zu 6 Haarzellen genau unterschieden werden (vgl. Zwicker & Fastl (1990, Abschnitt 7.2.1)).

2.3.2 Wahrnehmbarkeit von Amplitudenänderungen

Die Bandbreite der Wahrnehmung für Frequenzen liegt zwischen etwa 50 Hz (tiefer liegende Töne werden als Einzelereignisse wahrgenommen) und 20 kHz. Der Amplitudenbereich ist frequenzabhängig und liegt zwischen dem nicht mehr wahrnehmbaren Bereich und der Schmerzgrenze. Messungen für die gerade noch wahrnehmbaren Amplitudenunterschiede werden in der Regel für Sinustöne und weißes Rauschen durchgeführt. Da bei abrupter Änderung der Amplitude eines Signals ein hörbares Klicken entsteht, werden die Messungen entweder mit Amplitudenmodulation oder sukzessiv durchgeführt. Bei sukzessiv aufeinander folgenden Signalen mit unterschiedlicher Amplitude liegt das Unterscheidungsvermögen dabei etwas höher als bei Amplitudenmodulation. Bei Sinustönen ist die Pegelabhängigkeit größer als bei weißem Rauschen. Im Schnitt werden Unterschiede um etwa 1 dB gerade noch wahrgenommen (nach Zwicker & Fastl (1990, S. 1)).

2.3.3 Wahrnehmbarkeit von Frequenzänderungen

Der gerade noch wahrnehmbare Unterschied in der Frequenz hängt dagegen kaum vom Schalldruckpegel ab. Auch dieser wird entweder durch Modulation oder sukzessiv gemessen. Bei Frequenzmodulation sind dabei über 500 Hz Unterschiede von circa 0,7 % wahrnehmbar, bei 100 Hz sinkt die Unterscheidungsfähigkeit auf etwa 3,6 % (Zwicker & Fastl (1990, S. 9)). Frequenzänderungen in den Obertönen sind bei komplexen Klängen eher wahrnehmbar als Änderungen der Grundfrequenz. Der Unterschied zwischen zwei Tönen wird dabei um den Faktor 3 genauer wahrgenommen als bei Frequenzmodulation. So werden unter 500 Hz Abweichungen um 1 Hz wahrgenommen, darüber steigt das Unterscheidungsvermögen auf 0,2 % der Frequenz. Phasenunterschiede werden kaum wahrgenommen, vor allem sinkt das Unterscheidungsvermögen rapide, wenn die Klänge nicht in einem reflexionsarmen Raum bzw. über

Kopfhörer dargeboten werden.

Zur Wahrnehmbarkeit akustischer Ereignisse unterscheiden t'Hart et al. (1990) zwischen psychoakustischen und psychophonetischen Experimenten. Erstere betrachten generell die Fähigkeit der Differenzierung zwischen akustischen Signalen, während letztere speziell die Beschränkungen bei Sprechsituationen in Betracht ziehen. So ist zu bedenken, dass etwa Frequenzunterschiede bei isoliert dargebotenen Tönen deutlich eher wahrnehmbar sind als bei gesprochener Sprache.

2.3.4 Wahrnehmbarkeit von F_0-Bewegungen

Durch ihre Rolle als prosodisches Merkmal ist die Wahrnehmbarkeit der Grundfrequenz von besonderem Interesse. Die menschliche Verarbeitung kann eine F_0 auch dann detektieren, wenn sie gar nicht im Signal vorhanden ist. Da das Gehör, wie oben beschrieben, frequenzanalytisch arbeitet, existiert ein Mechanismus, der den größten gemeinsamen Teiler (GGT) einer Menge von Frequenzen berechnet, welcher als Grundfrequenz wahrgenommen wird. Dabei reichen bereits zwei nebeneinanderliegende Frequenzen aus, um die tiefere der beiden oder den GGT als Grundfrequenz wahrzunehmen.

Die zur F_0-Detektion notwendige Dauer eines Signals muss nicht mehr als 30 msek betragen (t'Hart et al. (1990, S. 26)), allerdings sinkt die Diskriminationsfähigkeit bei kürzeren Signalen. Für die Anwendung auf Sprachsignale ist die Frage von besonderen Interesse, wie stark sich die Grundfrequenz in einem bestimmten Zeitintervall ändern muss, damit diese Änderung wahrgenommen wird. t'Hart et al. (1990, S. 32) extrapolierten dazu eine lineare Beziehung aus eigenen Daten und anderen Untersuchungen, die durch die Formel

$$g_{thr} = \frac{0,16}{T^2}$$

beschrieben wird. g_{thr} ist dabei die Glissandoschranke als gerade noch bemerkbare zeitliche Veränderung der Frequenz (in Halbtönen pro Sekunde) bei Stimulusdauer T (in Sekunden). Diese Glissandoschranke wurde in Perzeptionsexperimenten im Rahmen eines automatischen Stilisierungsalgorithmus für Grundfrequenzkonturen überprüft (d'Alessandro & Mertens (1995)) mit dem Ergebnis, dass die meisten Perzipienten keinen Unterschied zwischen resynthetisierten Stimuli mit stilisierten und originalen F_0-Konturen feststellen konnten. Interessanterweise berichteten dabei mehrere Hörer, dass sie Unterschiede aufgrund der Stimmqualität und nicht durch die F_0-Kontur wahrgenommen hätten.

Eine weitere für die Intonationswahrnehmung bedeutsame Schranke besteht bei der Unterscheidungsfähigkeit zweier Glissandi. t'Hart et al. (1990) drücken diese als das gerade noch bemerkbare Verhältnis zweier Glissandi aus ($g_{diff} = g_1/g_2$) und erhalten bei Perzeptionsexperimenten einen Wert von etwa 2.

2.3.5 Wahrnehmbarkeit von Formantänderungen

Zur Wahrnehmbarkeit von diversen Vokaleigenschaften existiert eine relativ umfangreiche Untersuchung von Carlson et al. (1979). Es wurden isolierte durch Formantsynthese erzeugte vokalähnliche Stimuli in jeweils leicht variierter Form dargeboten. Dabei wurden frühere Studien

dahingehend bestätigt, dass Formantänderungen ab etwa 3 % wahrnehmbar sind. Die subjektiv wahrgenommene Unterschiedlichkeit steigt dabei mit der Anzahl der verschobenen Formanten. Unterschiede bei den Formantbandbreiten werden nur sehr grob perzipiert; erst ab etwa 40 % Änderung nahmen die Probanden einen Unterschied wahr. Die Unterscheidungsfähigkeit steigt, wenn alle Formantbandbreiten gleichermaßen verändert werden. Diese Variationen führten zu einem stärkeren Eindruck als eine reine Amplitudenänderung[3]. Spektrale Löcher, also Frequenzbereiche mit Nullamplitude, wie sie etwa bei Kopplung des Artikulationstrakts mit dem sublaryngalen Bereich oder dem Nasaltrakt entstehen können, wurden zwischen den ersten beiden Formanten erst ab einer Bandbreite von 900 Hz wahrgenommen[4]. Da hierbei bereits die Amplitude der umliegenden Formanten betroffen wird, scheinen spektrale Löcher keine große Rolle bei der Wahrnehmung zu spielen. Änderungen der spektralen Dämpfung ergaben, dass Variationen um 2 dB pro Oktave zu einer deutlich höheren Unterschiedlichkeit führten als bei 1 dB/Okt. Bei 1 dB/Okt ergab sich eine subjektive Differenz von 1 auf einer Skala von 1-10, so dass hier auch etwa die Grenze der Wahrnehmbarkeit anzusetzen ist.

[3]Die Wirkung ist dieselbe, nur dass bei einer allgemeinen Bandbreitenänderung lediglich die Formantamplituden modifiziert werden, wohingegen die Täler zwischen den Formanten gleich bleiben.

[4]Der synthetisierte Vokal hatte einen F1 von 700 und einen F2 von 1800 Hz.

Kapitel 3

Sprachsyntheseverfahren

3.1 Vorbemerkung

Das Thema der vorliegenden Arbeit hat im wesentlichen zwei Aspekte. Zum einen geht es um die Evaluation der perzeptiven Relevanz akustischer Korrelate emotionaler Sprechweise mittels Sprachsynthese. Zum anderen werden darauf aufbauend die Ergebnisse direkt umgesetzt, indem Text-to-Speech Systeme durch 'Emotionsregeln' zur Simulation emotionalen Sprechausdrucks befähigt werden.

Daher werden in diesem Kapitel nicht nur einzelne Resyntheseverfahren besprochen, sondern es wird darüberhinaus ein Überblick über die aktuelle Sprachsynthesetechnologie gegeben. Der Unterschied zwischen Resyntheseverfahren und Sprachsynthesizer soll dabei so definiert werden, dass ein Resynteseverfahren bestimmte Merkmale eines Sprachsignals durch Manipulationen im Zeit- oder Frequenzbereich modifiziert (etwa durch PSOLA- oder LPC-Techniken), wohingegen ein Sprachsynthesizer Sprache völlig neu synthetisiert. Wird mittels eines Sprachsyntheseverfahrens eine Äußerung kopiert, so wird dafür der Begriff *Kopiesynthese* verwendet. Im ersten Abschnitt dieses Kapitels erfolgt ein Überblick über die verschiedenen Module von TTS-Systemen, im zweiten werden zugrundeliegende Syntheseverfahren näher erläutert. Ein besonderes Gewicht wird dabei auf die Formant-Synthese gelegt, da der für die vorliegende Arbeit entwickelte Synthesizer auf diesem Verfahren basiert. Für eine Auswahl an Überblicksliteratur wird auf Dutoit (1997); Klatt (1987) und O'Shaughnessy (1987) verwiesen.

3.2 Text-To-Speech Verfahren

Ein Überblicksdiagramm für eine mögliche Klassifizierung von Systemen zur künstlichen Sprachcerzeugung ist in Abb. 3.1 dargestellt.

- Als *Sprachwiedergabesysteme* werden dabei solche bezeichnet, die zuvor aufgenommene Sprachsegmente wiedergeben. Diese Segmente haben dabei zumindest Wortgröße, so dass im eigentlichen Sinne keine neue Sprache synthetisiert wird, sondern vorhandene (eventuell in geänderter Abfolge) wiedergegeben wird. Im weitesten Sinne kann es sich

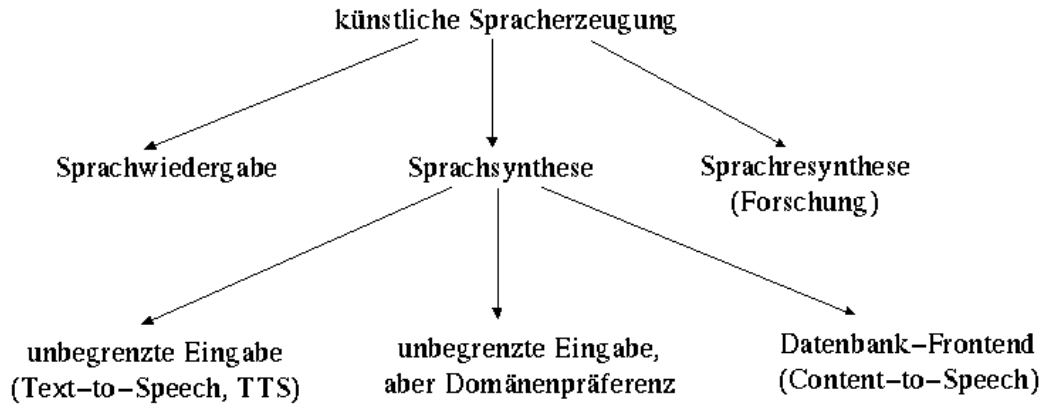

Abbildung 3.1: Überblick zur Klassifizierung von TTS-Systemen

dabei um ein Kassettengerät handeln, aber üblicherweise werden etwa Ansagesysteme in der U-Bahn als Sprachwiedergabesysteme bezeichnet.

- *Sprachsynthesesysteme* (*arbitrary speech systems*, nach Donovan (1996)) zeichnen sich im Gegensatz zu Sprachwiedergabesystemen dadurch aus, dass die Eingabe grundsätzlich nur durch die jeweilige Sprache begrenzt ist. Sie kann in schriftlicher Form (Text-to-Speech) oder durch ein Textgenerierungssystem (Concept-to-Speech) erfolgen. Als Sonderfall können Systeme mit Domänenpräferenz gelten. Sie sind zwar grundsätzlich dazu in der Lage, beliebigen Eingabetext zu verarbeiten, aber Wörter aus einem bestimmten Begriffsfeld werden besonders gut synthetisiert. Als Beispiel dafür kann das Verbmobil-Projekt gelten, welches auf Terminabsprachen beschränkt ist.

- Der Vollständigkeit halber seien hier auch *Resynthesesysteme* erwähnt. Sie werden ausschließlich in der Forschung verwendet und dienen der gezielten Manipulation ausgewählter Merkmale von Sprachsignalen.

Grundsätzlich bestehen alle Sprachsynthesizer aus mindestens zwei Komponenten: der Symbolverarbeitung, bei der eine Generierung der prosodischen Parameter sowie eine phonetische Transkription erfolgt und der akustischen Synthese, bei der das eigentliche Sprachsignal generiert wird (vgl. Abb. 3.2 und 3.3). Diese beiden großen Teilbereiche werden auch als NLP (Natural Language Processing) und DSP (Digital Speech Processing) bezeichnet.

Ein Klassifikationsschema für unterschiedliche Verfahren zur akustischen Synthese des Sprachsignals ist in Abb. 3.2 dargestellt. Die Signalgeneratoren können konzeptionell in zwei unterschiedliche Ansätze unterteilt werden.

- Artikulatorische Verfahren erzeugen das Sprachsignal durch eine Modellierung des menschlichen Sprecherzeugungsapparates. Sie sind dabei, wenn auch durch den abstrakteren Ansatz von äußerst hoher Flexibilität, derzeit noch von deutlich geringerer Sprachgüte als signalmodellierende Verfahren. Von daher ist die Nachbildung emotionaler Sprechweise aufgrund ungenügender Natürlichkeit der Ausgabe eher schwierig.

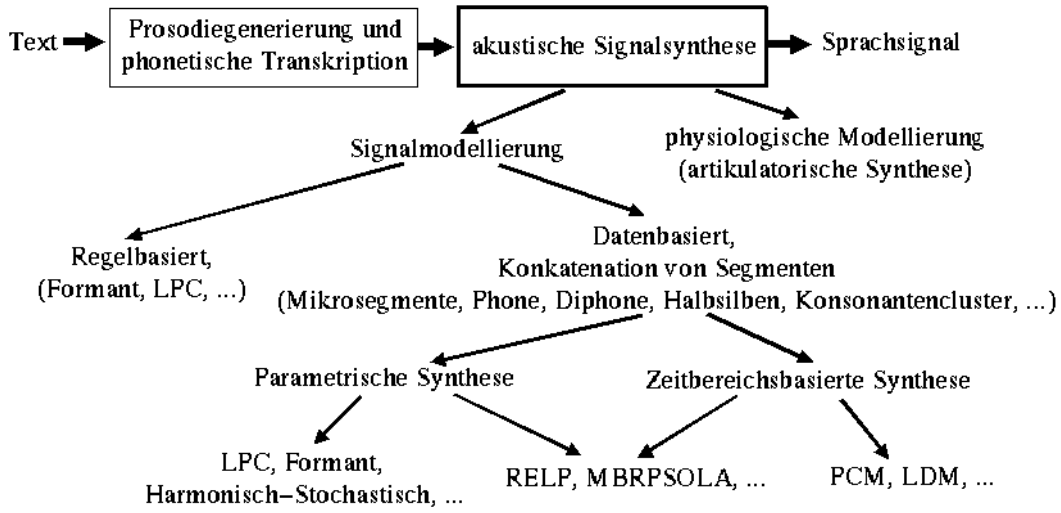

Abbildung 3.2: Überblick zur Klassifizierung von Sprachgeneratoren

- Signalmodellierende Verfahren bilden die akustischen Eigenschaften des Signals direkt nach. Dabei ist nun zu unterscheiden, ob die Synthese im wesentlichen regelbasiert erfolgt (ausgehend von Allophon-Targetwerten) oder durch Konkatenation von Basiseinheiten. Da die Konkatenationsmethode beim heutigen Forschungsstand bessere Sprachqualität als der regelbasierte Ansatz liefert, ist sie derzeit stark verbreitet. Konkatenierende Synthesizer können weiterhin dadurch klassifiziert werden, ob die zu konkatenierenden Basiseinheiten durch Parameter kodiert sind (Frequenzbereich) oder ob im wesentlichen Samples im Zeitbereich vorliegen (vgl. z.B. O'Shaughnessy et al. (1987)).

Ein Black-Box Diagramm der Komponenten eines TTS-Systems ist in Abb. 3.3 dargestellt. In den folgenden beiden Abschnitten werden das Symbolverarbeitungs- und das Signalgenerierungs-Modul einzeln besprochen.

3.3 Prosodiegenerierung und phonetische Transkription

Der zu synthetisierende Text wird zunächst durch die Symbolverarbeitung in eine phonetische Darstellung konvertiert. Zusätzlich wird eine Prosodiebeschreibung generiert. Diese beiden Vorgänge werden auch als LTS (*Letter-to-Sound*) und PG (*Prosody Generation*) bezeichnet. Um diese Aufgabe zu erfüllen, ist prinzipiell eine vollständige syntaktisch-morphologische Analyse des Eingabetextes notwendig. Darüberhinaus kann eine natürlich wirkende Prosodie im Prinzip nur durch Wissen über den semantischen und pragmatischen Gehalt des Textes generiert werden. Speziell bei der automatischen Simulation emotionaler Sprechweise ist dieses Wissen für eine natürliche Ausgabe unumgänglich. Allerdings ist bei den meisten aktuellen TTS-Systemen nicht einmal ein syntaktischer Parser implementiert. In aller Regel beschränken sich die Systeme, basierend auf Morphemlexika, auf eine Klassifizierung der einzelnen Wörter in Inhalts- und

Funktionswörter. Der Grund dafür liegt vor allem in der Echtzeitanforderung bei den meisten Anwendungen und der hohen Komplexität natürlicher Sprache.

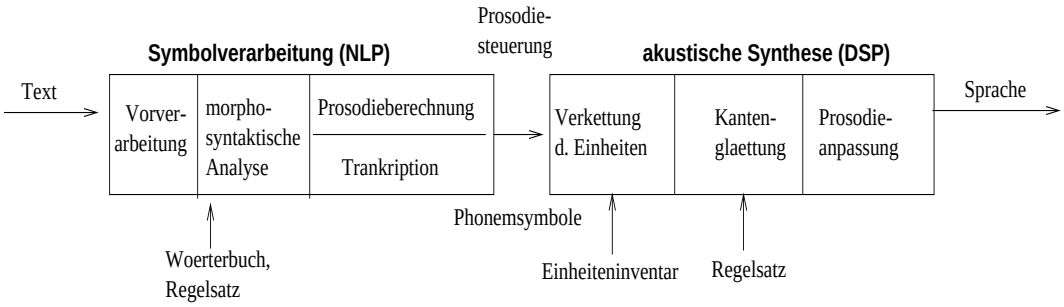

Abbildung 3.3: Allgemeines Schema für konkatenierende TTS-Systeme

Während frühere Sprachsynthesesysteme einzelne Verarbeitungsschritte sequentiell abarbeiteten, arbeiten neuere mit komplexeren Datenstrukturen. *Feature Structures* oder *Multilayer Data Structures* erlauben die Erfassung der Information aus vielen Modulen in zentral zugänglichen Strukturen (vgl. Dutoit (1997, S. 61)). Dadurch werden die Systeme unter anderem wesentlich flexibler, was z.B. die Erweiterung um Emotionsmodule betrifft. So wären Slots denkbar, die ausgehend von einem semantischen Modul den approbaten emotionalen Ausdruck für einen bestimmten Textabschnitt festlegen.

Die Verarbeitung im NLP-Modul lässt sich in mehrere Schritte untergliedern (vgl. Dutoit (1997, S. 63)).

1. In einer Vorverarbeitung werden Abkürzungen, Ziffern und Sonderzeichen in orthographische Form gebracht. Probleme gibt es hierbei vor allem mit der Vieldeutigkeit von Satzzeichen, Zahlenangaben und der Aussprache von Abkürzungen. Der Text wird in eine Kette von Wörtern in orthographischer Form umgewandelt, die zu Sätzen gruppiert sind. Zudem muss dabei eventuell eine Umformatierung der Eingabe erfolgen, um etwa Tabelleninhalte verständlich vorlesen zu können.

2. Die aus der Vorverarbeitung resultierende Wortliste wird, falls keine Analyse der Syntax erfolgt, zumindest in Funktions- und Inhaltswörter unterteilt. Da die meisten Sprachen weniger als 1000 Funktionswörter umfassen, werden diese in der Regel in einem Lexikon aufgeführt. Inhaltswörter, von denen es potentiell unendlich viele gibt, werden einer morphologischen Analyse unterzogen.

3. Da die Analyse auf Wortebene sehr viele Falschklassifikationen liefert[1], ist eine kontextuelle Analyse notwendig. Oft wird dabei auf eine komplette syntaktische Analyse verzichtet, sondern lediglich die Wahrscheinlichkeit einer Wortart bei gegebenen Wortarten der letzten N Wörter berechnet (N-Grams, vgl. Dutoit (1997, S.89)).

[1] Für das Französische 40 % laut Larreur *et. al*. (1989), zitiert in Dutoit (1997, S. 87).

4. Mit dem Wissen aus den syntaktischen und morphologischen Analysen[2] können die Wörter dann phonetisch transkribiert werden. Von besonderer Schwierigkeit ist dabei die Transkription von Eigennamen. Hierbei konkurrieren regel- mit datenbasierten Verfahren[3]. Bei regelbasierten Systemen unterscheidet man Expertensysteme, bei denen die Konvertierungsregeln von Hand erstellt werden, von trainierten Systemen, bei denen die Regeln durch statistische Lernverfahren generiert werden. Meistens werden dafür HMMs (*Hidden Markov Models*), Entscheidungsbaumverfahren (CARTs, *Classification and Regression Trees*) und seltener künstliche neuronale Netze (ANNs, *Artificial Neural Networks*, z.B. Nettalk, Sejnowski & Rosenberg (1987)) verwendet. Automatische Verfahren eignen sich besonders für multilinguale Systeme, da die manuelle Erstellung eines Regelsatzes für eine neue Sprache einen erheblichen Aufwand darstellt. Selbstverständlich ließe sich mit ihnen auch eine datenbasierte Emotionssimulation durchführen, indem das Konvertierungsmodul mit einer emotionalen Datenbank trainiert wird.

   Oft werden die transkribierten Wörter noch einer Nachverarbeitung unterzogen. Dies ist notwendig, da das LTS-Modul in der Regel nur mit isolierten Wörtern operiert. Zudem lassen sich hier phonetische Aspekte wie Reduktionen oder Assimilationen berücksichtigen. Nicht zuletzt können dabei emotionale Zustände simuliert werden.

5. Sind die Wörter nach grammatischer Funktion, Stellung im Satz und eventuell semantischem Gehalt klassifiziert, kann eine Berechnung der Prosodieparameter erfolgen.

Prosodiesteuerung

Unter dem Begriff 'Prosodie' werden im Allgemeinen die akustischen Merkmale zusammengefasst, die für den Eindruck der Sprechmelodie und des -rhythmus ausschlaggebend sind (vgl. Abschnitt 2.2.1). Insbesondere sind dies die Intonationskontur, die Lautdauern und die Lautintensitäten. Die Lautintensitäten werden allerdings bei fast allen Systemen nicht dynamisch generiert, da eine Variation perzeptiv eine wesentlich geringere Rolle spielt als bei der F_0 oder den Lautdauern. Die Funktion der Prosodie ist vielfältig. Durch sie werden Sätze in Phrasen unterteilt, Satzfokusse gesetzt, die syntaktische Struktur wird deutlich gemacht und auch der semantische Gehalt einer Äußerung kann durch die Prosodie verändert werden[4]. Nicht zuletzt drückt sich in ihr auch der emotionale Zustand des Sprechers aus. Eine natürlich wirkende Prosodie erhöht somit nicht nur die Natürlichkeit, sondern auch die Verständlichkeit synthetischer Sprachausgabe in nicht zu unterschätzendem Maße.

Gleichwohl stellt die Berechnung der Prosodie beim heutigen Stand der Technik die größte Herausforderung an TTS-Systeme dar, was vor allem an der Schwierigkeit syntaktischer und semantischer Analyse liegt (s.o.). Um eine neutrale Prosodie mit geringem Aufwand generieren

[2] Syntaktisches Wissen ist z.B. wichtig für die korrekte Aussprache heterophoner Homographen. Eine morphologische Analyse ist unverzichtbar, da die meisten Graphem/Phonem-Konvertierungsregeln nicht über Morphemgrenzen hinweg gelten.

[3] Wobei jedes System eine Mischform darstellt: Lexikonbasierte Systeme verwenden Regeln für fehlende Einträge und regelbasierte verwenden ein Lexikon für Ausnahmen.

[4] So ist z.B. die Semantik des bekannten Satzes "*Er sah den Mann mit dem Fernrohr.*" nur durch den Satzfokus ersichtlich.

zu können, arbeiten die meisten TTS-Systeme mit recht simplen Regeln. Eine einfache Methode ist der *Chinks'n Chunks*-Algorithmus (Liberman & Church (1992)), bei dem eine prosodische Phrase aus einer Kette von Chinks gefolgt von einer Kette von Chunks besteht. Als Chinks werden hierbei Funktionswörter und flektierte Verben aufgefasst, als Chunks Inhaltswörter und persönliche Pronomen. Die Gesamtprosodie für einen Satz ergibt sich dann als die Vereinigung der Prosodiebeschreibung für mehrere Phrasen und einer übergeordneten für den ganzen Satz[5].

Einen funktionalen Ansatz zur Intonationsmodellierung stellt das Modell von Fujisaki (z.B. Fujisaki (1992)) dar. Der Produktionsmechanismus (also das Glottissignal) wird dabei im Modell berücksichtigt. Eine Intonationskontur ergibt sich danach aus der Addition von Phrasenkommandos (in der Form von Impulsen) und Akzentkommandos (in der Form von Stufenfunktionen). Um mikroprosodischen Effekten Rechnung zu tragen, wird das Intonationssignal noch durch einen glottalen Oszillationsmechanismus gefiltert. Dieses Modell wurde in modifizierter Form z.B. beim Klattalk System (Klatt (1990)) verwendet.

Datenbasierte Systeme verwenden Verfahren des maschinellen Lernens wie z.B. statistische Entscheidungsbaumverfahren (CART, z.B. Malfrère et al. (1999); Fordyce & Ostendorf (1998)), oder neuronale Netze. Mittels dieser können aus großen Datenbanken jeweils approbate Prosodiebeschreibungen extrahiert werden. Ein Verfahren soll hier näher erläutert werden, da es eine Möglichkeit darstellt, die Prosodie emotionaler Sprechweise datenbasiert zu simulieren. Dazu müssten als Datenbank lediglich Äußerungen in der gewünschten Emotion verwendet werden[6]. Bei dem in Malfrère et al. (1999) beschriebenen Verfahren werden sowohl die Datenbank als auch die zu generierende Äußerung mit dem oben beschriebenen Chinks&Chunks-Algorithmus in prosodische Abschnitte unterteilt. Zur Lautdauervorhersage wird die Datenbank mittels automatischer Verfahren in eine symbolische Repräsentation konvertiert[7], bei der jedem Laut und jeder Silbe neben einer prosodischen Beschreibung eine linguistische Beschreibung ihrer Stellung zugeordnet wird. Features sind hierbei z.B. auf Lautebene die Stellung des Lautes in der Silbe, auf Silbenebene der Typ (z.B. CV oder CVC), die Art des Akzents und die Größe (Anzahl der Laute). Um also die Dauer eines Lautes vorherzusagen, wird mit dem Entscheidungsbaumverfahren der ähnlichste Laut in dieser Stellung aus der Datenbank gesucht. Die F_0-Kontur-Vorhersage funktioniert ähnlich. Jedem prosodischen Abschnitt in der Datenbank wird neben der Beschreibung der Kontur durch F_0-Targetwerte ein Schlüssel zugeordnet, der Werte für z.B. die Stellung im Satz, Anzahl der unbetonten Silben und weitere enthält. Wie bei der Dauervorhersage können dann möglichst approbate F_0-Konturen für bestimmte Abschnitte gefunden werden. Die Lautheit als prosodisches Merkmal wurde in dem erwähnten Artikel nicht beachtet. Eine Vorhersage könnte wie für die Lautdauern erfolgen, wenn auch eine Extraktion aus dem Sprachmaterial aus den in Abschnitt 2.2.1 erwähnten Gründen Schwierigkeiten bereitet.

[5] Hier wird z.B. das Anheben der Grundfrequenz am Ende bei Fragesätzen modelliert.

[6] Die Erstellung einer solchen Datenbank stellt allerdings erhebliche Probleme dar. Dazu müsste immerhin über eine Stunde Sprachmaterial, in einer Emotion von einer Person gesprochen, gewonnen werden. Daher wurde dieser Ansatz nach Kenntnis des Autors auch nur in einer nichtveröffentlichten Studie verfolgt. Untersucht wurden dabei die Stimmungen 'Schüchtern' und 'Nervös' (laut E-Mail vom Untersuchenden, F. Malfrère, Universität Mons).

[7] Spracherkennung funktioniert bei Kenntnis der Äußerung recht zuverlässig. Ein Verfahren hierzu ist in Malfrère & Dutoit (1998) beschrieben.

3.4 Akustische Signalsynthese

Die akustische Synthese generiert aus der lautlichen und prosodischen Beschreibung das Sprachsignal. Dabei wird die diskrete Kette von phonetischen Symbolen in einen kontinuierlichen Datenstrom von Signalabtastwerten oder -parametern überführt. Je nach Synthesemethode werden entweder aus Parametern Sprachsignale berechnet und/oder Glättungsregeln auf die Übergangsstellen der Einheiten angewendet. Zusätzlich wird zur Anpassung der Prosodie das Ausgabesignal manipuliert bzw. die Parameter angepasst. Wie oben ausgeführt, lassen sich hierbei signalmodellierende Systeme von solchen unterscheiden, die die menschliche Physiologie modellieren. In den folgenden Abschnitten werden die unterschiedlichen Konzepte einzeln beschrieben, wobei sich die Systematik an dem Klassifikationsschema in Abb. 3.2 orientiert.

In Bezug auf die Simulation emotionaler Sprechweise ist die Verwendung artikulatorischer Systeme äußerst attraktiv, da diese die Möglichkeit bieten, physiologische Vorgänge direkt zu modellieren. Je nach Ansatzpunkt ist zwischen *neuromotor-command-*, *articulator-* und *vocal-tract-shape*-Synthese zu unterscheiden (O'Shaughnessy (1987, S. 168)). Letztendlich wird zwischen Artikulator-Targetpositionen interpoliert, wobei kinetische Beschränkungen bzgl. der Geschwindigkeit und Freiheitsgrade der Artikulatoren berücksichtigt werden.

Das Kölner Artikulatorische Sprachsynthesesystem (Kröger et al. (1995)) z.B. besteht aus drei Subsystemen. Ausgehend von den Lautsymbolen werden zunächst mittels dynamischer Modelle Bewegungsmuster für die Artikulatoren berechnet. Das artikulatorische Subsystem berechnet aus diesen Bewegungsabläufen Querschnittsmaße für den Artikulationstrakt, aus denen mittels eines aerodynamisch-akustischen Modells das Sprachsignal generiert wird. Dieses kann zum Beispiel durch die Umrechnung der Flächenverhältnisse im Röhrenmodell in Reflexionskoeffizienten für einen Kreuzglied- (*Lattice*) Filter (vgl. z.B. Fellbaum (1984)) geschehen. Andere Systeme (z.B. das Forschungssystem HLSyn der Firma Sensymetrics, Bickley et al. (1997)) berechnen aus den Artikulatorbewegungen Formantverläufe und verwenden für die akustische Synthese einen Formantsynthesizer. Dieses Vorgehen wird auch als pseudo-artikulatorisch bezeichnet. Das Anregungssignal kann physiologisch durch das weitverbreitete Zwei-Massen Modell von Flanagan & Ishizaka (1978) beschrieben werden. Dabei werden die inneren und äußeren Stimmlippen durch zwei Massen modelliert, die mit Federn untereinander verbunden sind.

Physiologische Zustände, die für einzelne Emotionen charakteristisch sind, können hierbei auf höchster Ebene modelliert werden, was eine hohe Flexibilität ermöglicht. Weiterhin werden dabei koartikulatorische Effekte, die bei signalmodellierenden Systemen durch ein kompliziertes Regelwerk beschrieben werden müssen, direkt durch die Modellierung der Artikulatorbewegungen berücksichtigt. Langfristig gesehen scheint die artikulatorische Synthese ein vielversprechender Ansatz, natürlich klingende Sprache zu erzeugen.

Artikulatorische Ansätze leiden jedoch bislang an einem eklatanten Mangel an Daten. Vor allem der Zusammenhang zwischen Lautbildung und Bewegungsabläufen der Artikulatoren ist mangels geeigneter Messtechniken noch weitgehend unerforscht. Weiterhin sind die Zusammenhänge zwischen Glottisbewegung, Artikulatorstellung und akustischem Signal äußerst komplex (das Röhrenmodell stellt eine recht grobe Vereinfachung des Artikulationstrakts dar). Aus diesem Grund stehen heutige artikulatorische Sprachsynthesesysteme noch vor der Schwie-

rigkeit, überhaupt verständliche, geschweige denn natürliche Sprachausgabe zu generieren. Sie werden deshalb zur Zeit ausschließlich in der Forschung zur Kopiersynthese eingesetzt.

3.4.1 Signalbeschreibende Systeme: Regelbasierte Synthese

Signalmodellierende Systeme sind zunächst in regel- und datenbasierte Ansätze zu unterteilen. Bei den regelbasierten Systemen ist das phonetische Wissen explizit in Regeln enthalten, nach denen Sprache erzeugt wird. Als Grundbausteine werden meistens Allophone verwendet, deren spektrale oder artikulatorische Eigenschaften (alle artikulatorischen Synthesizer sind regelbasiert, s.o.) tabellarisch erfasst sind. Um einen kontinuierlichen Sprachfluss zu erzeugen, wird zwischen den Werten einzelner Parameter unter Berücksichtigung bestimmter Regeln interpoliert. Da die Regeln zur Lautübergangsgestaltung phonetisch motiviert sind, wird bei regelbasierten Systemen ausschließlich Formantsynthese verwendet. Noch vor wenigen Jahren waren diese Systeme sehr verbreitet (ein prominenter Vertreter ist DEC-Talk), da der Speicherbedarf im Vergleich zu datenbasierten Verfahren wesentlich geringer ist.

Die regelbasierte Synthese eignet sich insofern zur Simulation emotionaler Sprechweise, als die Regeln zur Emotionserzeugung direkt auf die Generierungsregeln angewendet werden können. Allerdings ist regelbasierte Synthese oft von schlechterer Qualität als datenbasierte. Viele Koartikulationseffekte sind zwar durch einfache Regeln beschreibbar, wie z.B. die Absenkung der Formanten bei gerundeten Lippen. Für andere jedoch, die sich etwa durch das Verkürzen der Artikulatorbewegungen zu Zielpositionen hin ergeben, gilt dies nicht. Zudem lassen Vereinfachungen wie die Modellierung des Glottissignals als Impulsfolge die resultierende Sprache unnatürlich klingen (O'Shaughnessy et al. (1987)). Bei Untersuchungen zu emotionaler Sprechweise, die mit formantbasierten Systemen (wie z.B. DEC-Talk oder GLOVE) durchgeführt wurden (vgl. Murray & Arnott (1995) und Carlson et al. (1992)), wurde bereits die neutrale Stimme von einem hohen Prozentsatz der Versuchshörern als traurig klingend bewertet, was sich eventuell durch eine zu große Monotonie im Anregungssignal erklären lässt.

3.4.2 Signalbeschreibende Systeme: Datenbasierte (konkatenierende) Synthese

Datenbasierte Systeme generieren Sprache aus Grundeinheiten, die als Inventar vorliegen und miteinander konkateniert werden. Im Gegensatz zu regelbasierten Systemen ist hier ein Großteil des phonetischen Wissens implizit in den Einheiten kodiert, weshalb die resultierende Sprachgüte zunächst höher ist. Dadurch verschiebt sich das Problem der Generierungsregeln zu dem Problem der Kantenglättung und der prosodischen Manipulation. Datenbasierte Systeme lassen sich vor allem hinsichtlich zweier Kriterien klassifizieren (vgl. O'Shaughnessy et al. (1987)): erstens welche Art von Einheiten konkateniert werden und zweitens wie diese Einheiten kodiert werden.

Wahl der Einheiten

Bei der Wahl der Einheiten ist ein Kompromiss zwischen einer möglichst geringen Anzahl der Segmente und möglichst großen Einheiten zu schließen. Je länger die Segmente sind, desto weniger Schnittstellen gibt es im resultierenden Sprachsignal. Dies ist von Bedeutung, weil die Kantenglättung eines der größten Probleme bei konkatenierenden Systemen darstellt. Allerdings wächst mit der Größe der Segmente auch deren Anzahl stark an. Dieses Problem wird durch die Tatsache verschärft, dass viele Systeme auch heute noch die Segmente nicht nur einmal, sondern bis zu dreimal (mit unterschiedlichen Betonungstufen) in der Datenbank halten[8]. Um eine Variation der Stimmqualität bei Zeitbereichsynthese zu ermöglichen, wären weiterhin mehrere Versionen für unterschiedliche Emotionen denkbar. So könnte jedes Segment z.B. zusätzlich zu einer neutralen Version einmal mit *tense* und einmal mit *lax voice* (vgl. Abschnitt 2.1.2) in der Datenbank vorliegen.

Als Einheiten werden Silben, Halbsilben, Diphone, Phone, Phonemkluster, Mischinventare oder subphonemische Einheiten verwendet (vgl. Portele (1996b, S. 11) oder O'Shaughnessy et al. (1987)). Halbsilben entstehen durch einen Schnitt im Silbenkern, bei Diphonen wird das Sprachsignal in phongroße Stücke unterteilt, mit dem Schnitt in der Mitte des Phons. Der Schnitt wird in der Regel vor der zeitlichen Mitte des Phons vorgenommen, da koartikulatorische Effekte eher früher als später auftreten. Dadurch treffen bei Konkatenation immer Einheiten mit ähnlichen spektralen Eigenschaften aufeinander, was die Glättungsregeln stark vereinfacht. Da Koartikulation sich oft über mehr als zwei Phone erstreckt, ist auch die Verwendung von Triphonen üblich. Silben sind geeignet, weil sich Koartikulation vor allem innerhalb von Silben auswirkt, wobei die Anzahl der benötigten Silben für ein vollständiges Inventar allerdings bei den meisten Sprachen zu groß ist. Ein Phonemcluster ist eine Kette von vokalischen oder konsonantischen Phonemen. Zum Teil ist die Wahl des Inventars auch sprachenabhängig. So sind Halbsilbensysteme besonders für Sprachen mit komplexen Konsonantenfolgen wie das Deutsche geeignet. Ein Halbsilbeninventar lässt sich reduzieren, indem z.B. finale Obstruenten abgetrennt werden oder Normschnittstellen zu homogenen Lauten verwendet werden (wie etwa bei Portele et al. (1990) beschrieben). Ein rein zeitbereichsbasiertes Verfahren, bei dem subphonemische Einheiten verkettet werden, ist bei Benzmüller & Barry (1996) beschrieben. Es zeichnet sich vor allem durch eine sehr kleine Datenbasis und schnelle Berechnungszeit aus, so dass es äußerst geringen Hardwareanforderungen genügt.

Die zunächst naheliegende Einheit Phon ist in der Regel zur Verkettung ungeeignet, da koartikulatorische Effekte die resultierende Ausgabe sehr unverständlich klingen lassen. Eine Ausnahme bilden Syntheseverfahren, die auf *non-uniform unit selection* basieren (vgl. z.B. Campbell (1996)). Dabei werden bei einem Sprachkorpus eines Sprechers die Phone markiert[9]. Für jedes Phon wird dann ein Merkmalsvektor berechnet, der seine segmentellen und prosodischen Eigenschaften beschreibt. Zur Synthesezeit wird für den Eingabetext ein Targetvektor von Phonen mit approbaten prosodischen Eigenschaften festgelegt. Ein statistisches Verfahren minimiert daraufhin eine doppelte Kostenfunktion. Es werden diejenigen Phone miteinander

[8] Vor der Entwicklung des PSOLA-Verfahrens (siehe Abschnitt 3.5.1) war dies für Zeitbereichssynthese unabdingbar.

[9] Bei einem zweistündigen Sprachkorpus werden dabei etwa 35000 Phone annotiert.

verkettet, die einerseits den gewünschten Targetmerkmalen und andererseits den benachbarten Phonen am ähnlichsten sind. Das System kommt so völlig ohne Kantenglättung oder Prosodiemanipulation aus, was in einer Sprachqualität von hoher Güte resultiert. Da zur Berechnung der Prosodie derselbe Sprachkorpus verwendet wird, aus dem auch die Segmente stammen, klingt die resultierende Sprachausgabe dem originalen Sprecher sehr ähnlich. Zudem ist das System leicht auf andere Sprecher oder Sprachen übertragbar. Es wurde auch zur Emotionssimulation benutzt, indem Sprachkorpora mit emotionalem Ausdruck (Freude, Wut und Trauer) für die Synthese verwendet wurden. Dabei gelang es, erkennbaren emotionalen Sprecherausdruck zu simulieren (Iida et al. (2000)).

Kodierung der Einheiten

Datenbasierte TTS-Systeme lassen sich abgesehen von der Art der Einheiten danach klassifizieren, in welcher Form die Segmente in der Datenbank kodiert sind. Die Art der Kodierung bestimmt wesentlich die notwendigen Schritte der Kantenglättung und der Prosodieanpassung. Grundsätzlich sind drei Ansätze zu unterscheiden: 1. parametrische (Frequenzbereichs-), 2. Waveform (Zeitbereichs-) und 3. halbparametrische Kodierung (hybride Verfahren).

Parametrische Kodierung. Als parametrische Kodierungsverfahren (Vocoder-Verfahren, vgl. Fellbaum (1984, S. 161)) sind vor allem LPC- und Formant-Synthese zu nennen. Beiden ist gemeinsam, dass die Spracherzeugung durch das Quelle-Filter Modell modelliert wird. Da der in dieser Arbeit entwickelte Synthesizer (siehe Kap. 6) auf der Formantsynthese-Technologie basiert, wird dieses Modell und damit verbundene Probleme im Folgenden näher besprochen.

Das Quelle-Filter Modell[10] nach Fant (1960) (siehe Abb. 3.4) basiert auf der Annahme, dass sich das Sprachsignal in ein Anregungssignal und ein Filter zerlegen lässt (analog zur Trennung des Sprechvorgangs in Phonation und Artikulation). Es bildet die Grundlage für Analyse- und Syntheseverfahren, die im Gegensatz zu allgemeineren Signalumformungsverfahren (wie etwa der Fourier-Transformation oder statistischen Verfahren) spezielle Eigenarten gesprochener Sprache berücksichtigen. So können Quelle und Filter getrennt und unter Ausnutzung gegebener Beschränkungen einzeln kodiert und verarbeitet werden. Das Quellsignal kann etwa im Extremfall lediglich durch die Grundfrequenz repräsentiert werden, was eine enorme Datenkompression ermöglicht. In Bezug auf die Sprachsynthese ist dieses Modell vor allem deshalb von Vorteil, weil prosodische Eigenschaften wie Lautdauer und Intonation explizite Parameter der Sprachgenerierung sind und ihre Modellierung somit keinen zusätzlichen Aufwand erfordert. Für den Fall der Simulation emotionaler Sprechweise ist dieser Ansatz darüberhinaus äußerst attraktiv, weil sich Eigenschaften des Glottissignals und des Artikulationstraktfilters getrennt voneinander beschreiben und manipulieren lassen. Da die Stimmquelle durch mathematische Modelle erzeugt wird, kann die Stimmqualität direkt manipuliert werden (vgl. Granström (1992) oder Karlsson (1992)). So können akustische Phänomene wie z.B. Laryngalisierung erzeugt werden, die hauptsächlich von Unregelmäßigkeiten der Glottispulsfolge herrühren.

Parametrische Kodierung spart Platz bei der Speicherung und der Übertragung eines Signals. Weiterhin sind die Übergangsglättung und die Anpassung an koartikulatorische Effekte

[10] In allgemeiner Form auch als AR- (Autoregressives) Modell bezeichnet.

sowie die Intonation einfacher zu berechnen, da auf die spektralen Eigenschaften des Signals direkt zugegriffen werden kann. Dies stellt natürlich auch einen großen Vorteil bei der Simulation emotionaler Sprache dar. Zudem kann (zumindest bei Formant-Synthesizern) die Verschiebung der Formantlagen bei einer Änderung der Grundfrequenz berücksichtigt werden.

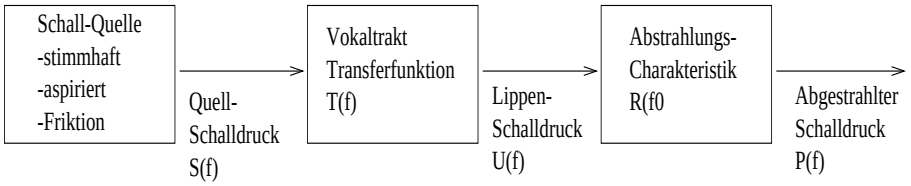

Abbildung 3.4: Quelle-Filter Modell (nach Fant (1960))

Das Quelle-Filter Modell setzt allerdings voraus, dass Quelle und Filter weitgehend unabhängig voneinander sind. Dies stellt in manchen Fällen eine ungenügende Vereinfachung dar (vgl. Klatt & Klatt (1990, S. 840)).

- Die Unabhängigkeit zwischen Quelle und Filter ist vor allem bei geschlossener Glottis gegeben. Während der geöffneten Phase bzw. bei behauchter Anregung, wenn die Glottis nie ganz geschlossen ist, kann es bei der Synthese zur Notwendigkeit weiterer Pol/Nullstellen-Paare zur Modellierung des Resonanzverhaltens der Trachea kommen.

- Die direkte Proportionalität zwischen glottaler Öffnung und glottalem Luftdruck ist durch stehende Wellen in der Pharynx nicht immer gegeben. Die Druckveränderungen dieser Wellen bewirken eine Welligkeit (*ripple*) des glottalen Luftstroms im Bereich des ersten Formanten und können sogar das mechanische Verhalten der Stimmlippen beeinflussen (siehe Fant (1985)).

  Dieser Effekt bewirkt eine Verstärkung des ersten Formanten gegenüber den anderen, da Anteile von ihm bereits im Glottissignal enthalten sind. Ein weiterer Effekt besteht darin, dass die Energie der Anregung ansteigt, wenn der erste Formant ein Vielfaches der Grundfrequenz beträgt.

- Das Filter unterliegt während eines Glottiszyklus starken Veränderungen.

- Der glottale Luftdruck ist während eines Zyklus nicht konstant, sondern proportional zur Wurzel des Druckabfalls. Dies hat einen Abfall tieffrequenter Energie während der geöffneten Glottisphase zur Folge. Die Bandbreite des ersten Formanten vergrößert sich dabei beträchtlich. Zudem kann der erste Formant um bis zu 10 % ansteigen (Klatt & Klatt (1990, S. 842)).

Eine weitere unzulängliche Vereinfachung früher Quelle-Filter Systeme stellt die Darstellung des stimmhaften Anregungssignals als Kette gleichförmiger Impulse dar. Drei bekannte Effekte widersprechen dieser Vereinfachung:

- Als **Jitter** bezeichnet man die Unregelmäßigkeit der Abstände benachbarter Perioden. Er tritt praktisch immer auf, ein Fehlen lässt das synthetische Sprachsignal hochgradig unnatürlich klingen.

- **Shimmer** bezeichnet Unregelmäßigkeiten in der Amplitude benachbarter Impulse, da dieser Effekt allerdings perzeptiv nicht sehr relevant ist, wird von einer Modellierung in der Regel Abstand genommen.

- **Diplophonisches Doppelpulsieren** bezeichnet das Auftreten benachbarter Impulse, deren Abstand zueinander jeweils paarweise verringert ist. Dabei ist in der Regel die Amplitude des ersten schwächer. Dieser Effekt tritt oft am Äußerungsende auf.

Weiterhin bedingen alle parametrischen Verfahren die Zerlegung des Signals in *Frames*, für deren Dauer (üblicherweise um die 20 ms) das Signal als stationär angenommen wird. Speziell bei Plosiven können allerdings drastische Änderungen des Signals in kürzeren Zeitabschnitten erfolgen. Zudem macht sich der durch die Parametrisierung auftretende Informationsverlust bei der Resynthese störend bemerkbar. Aus diesen Gründen sind parametrische Synthesizer gegenüber Zeitbereichsverfahren bislang noch von eher geringerer Sprachqualität.

Zeitbereichs (PCM)-Kodierung. Die Alternative zur parametrischen Kodierung ist die Kodierung im Zeitbereich, also das Operieren auf dem digitalisierten Zeitsignal. Dieses wird als PCM (*Pulse Coded Modulation*)-Signal bezeichnet. Es wurden verschiedene Ansätze zur Datenreduktion entwickelt (vgl. O'Shaughnessy (1987) oder Fellbaum (1984)), die sich zum Teil auch im Frequenzbereich interpretieren lassen (z.B. ADPCM, vgl. Abschnitt 3.5.2). Bei der LDM- (*Linear Delta Modulation*) Kodierung werden etwa im Gegensatz zur PCM-Kodierung nicht die Signalwerte selber, sondern lediglich die binäre Information, ob der jeweilige Samplewert über oder unter dem letzten liegt, abgespeichert.

Die Schwierigkeit bei zeitbereichsbasierter Kodierung liegt vor allem in der Kantenglättung und der Prosodieanpassung. Um unterschiedliche Betonungsstufen generieren zu können, speicherten frühere Systeme deshalb jedes Segment mehrmals (mit verschiedenen prosodischen Merkmalen) ab. Ähnliche Verfahren werden immer noch verwendet, da jede Manipulation der Segmente (vor allem bzgl. der Grundfrequenz) die Segmentqualität hörbar beeinträchtigt. Seit etwa 10 Jahren ist die Verwendung des TD-PSOLA (*Time Domain PSOLA*) Algorithmus (vgl. Abschnitt 3.5.1) verbreitet, durch den eine Prosodieanpassung im Zeitbereich möglich geworden ist. Einfachere Ansätze wie etwa die Mikrosegmentsynthese (Benzmüller & Barry (1996)) erzeugen eine Änderung der Grundfrequenz durch Weglassen des perzeptiv weniger relevanten hinteren Teils der Periode (beim Anheben der F_0) bzw. Einfügen von Nullen am Ende (beim Absenken der F_0). Unterschiedliche Lautdauern werden beim subphonemischen Ansatz einfach durch die Anzahl der subphonemischen Segmente generiert.

Bei der zeitbereichsbasierten Synthese ist das gesamte phonetisches Wissen fest in den Segmenten kodiert. Daraus ergibt sich eine sehr gute Sprachqualität, die allerdings durch die Unstimmigkeiten bzgl. Phase, Amplitude und Spektrum an den Segmentgrenzen verringert wird. Der Stimmeindruck dieser Systeme ähnelt sehr dem des Originalsprechers. Die zunächst gute

Qualität wird durch eine äußerst geringe Flexibilität bzgl. stimmqualitativer und artikulatorischer Merkmale erkauft, was eine Simulation emotionaler Sprechweise auf prosodische Parameter beschränkt.

Halbparametrische Kodierung. Verfahren, die sowohl im Zeit- als auch im Frequenzbereich operieren, werden als halbparametrisch bezeichnet. Es ist dabei zu unterscheiden, ob die parametrische Verarbeitung zur Synthesezeit durchgeführt wird (z.B. RELP, *Residual Excited Linear Prediction*) oder die Samples in einer Vorverarbeitungsstufe durch frequenzbasierte Verfahren modifiziert werden (z.B. MBR-PSOLA). Das RELP und ähnliche Verfahren werden unten im Zusammenhang mit der LPC-Synthese (Abschnitt 3.5.2) besprochen. MBR-PSOLA und Varianten davon werden im Zusammenhang mit PSOLA (Abschnitt 3.5.1) erläutert.

Grundsätzlich sind Systeme, die zur Laufzeit parametrische Verfahren implementieren, in der Lage, die Qualität der Signale im Zeitbereich mit der Flexibilität frequenzbasierter Verfahren zu kombinieren. Für die Generierung emotionalen Sprecherausdrucks eignen sie sich jedoch nur bedingt, da sie in der Regel auf rein mathematisch/statistischen Modellen ohne Bezug zu einer phonetisch motivierten Beschreibungsebene basieren.

3.5 Ausgewählte Sprachsyntheseverfahren

In den folgenden Abschnitten werden die wichtigsten Resynthese/Synthese Verfahren einzeln besprochen. Dabei wird zunächst das PSOLA-Verfahren erläutert. Ein weiterer Abschnitt befasst sich mit dem Prinzip der LPC-Analyse/Synthese, die wohl das wichtigste Verfahren zur Signalgenerierung, Komprimierung und Analyse darstellt. Letztendlich wird die Formantsynthese relativ ausführlich erläutert, da der in Kap. 6 beschriebene Synthesizer auf diesem Verfahren beruht.

3.5.1 PSOLA-Verfahren

Der PSOLA- (*Pitch Synchronous Overlap and Add*) Algorithmus stellt kein eigenständiges Syntheseverfahren dar, sondern ermöglicht eine Modifikation der prosodischen Eigenschaften von Sprachsignalen durch Resynthese. Bei der akustischen Synthese muss neben der Übergangsglättung an den Einheitengrenzen eine adäquate Prosodie generiert werden (s.o.). Bei parametrischer Kodierung der Segmente ist dies relativ einfach, da alle Parameter der Prosodie (Dauer, Intensität, F_0) explizit bei der Dekodierung anzugeben sind. Für die Zeitbereichskodierung ist das erst durch den PSOLA-Algorithmus möglich geworden (Charpentier & Stella (1984); Moulines & Charpentier (1990)). Hierbei wird durch überlappende Addition von gewichteten Signalabschnitten die Grundfrequenz beeinflusst und durch Eliminieren oder Hinzufügen von Signalabschnitten die Dauer verändert (siehe Abb. 3.5). Durch einen zusätzlichen Faktor bei der Gewichtung wird die Energie der Signalabschnitte korrigiert.

Dies kann beim TD-PSOLA-Verfahren (*Time-Domain-PSOLA*) zu Artefakten führen, wenn die F_0-Verschiebungen zu drastisch sind (vgl. Moulines & Charpentier (1990, S. 456)). So gibt es Probleme bei Stimmen mit hoher Grundfrequenz, da eventuell der erste Formant verfälscht

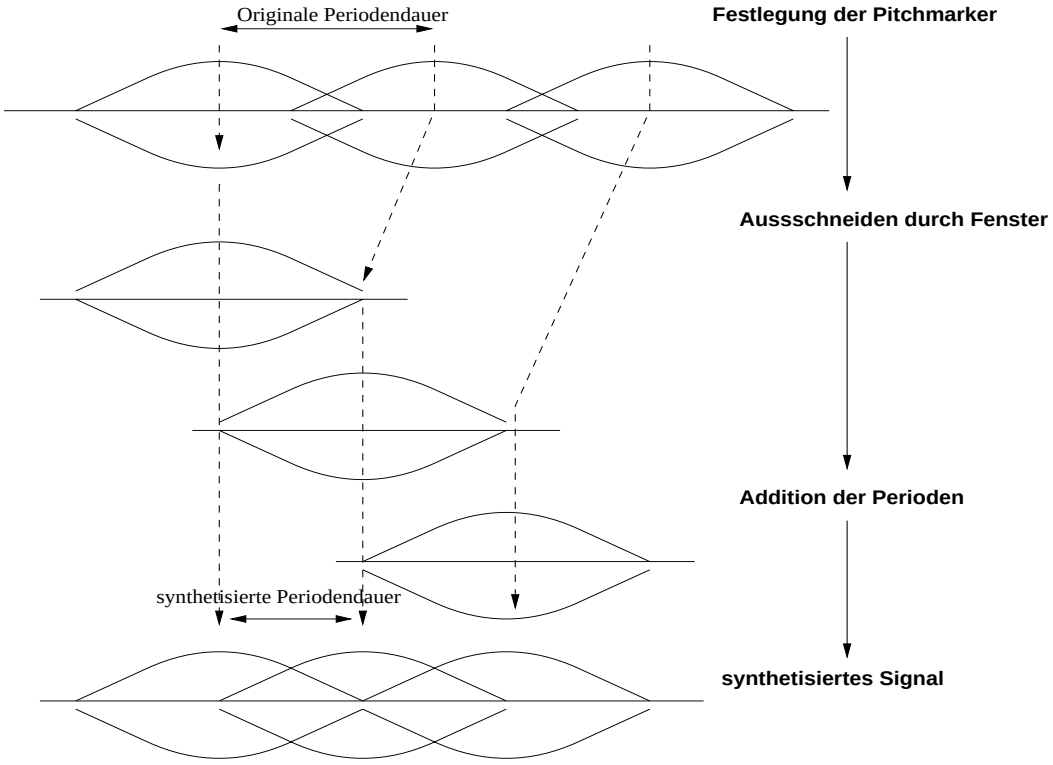

Abbildung 3.5: Erhöhung der Grundfrequenz durch das TD-PSOLA-Verfahren (nach Hirokawa & Hakoda, 1990)

wird. Zudem ist die Markierung der einzelnen Perioden (*pitch-marking*) für dieses Verfahren kritisch und muss von Hand nachgebessert werden. Weiterhin arbeitet das Verfahren unter der Prämisse, dass die für die Perzeption wichtigen Eigenschaften der Sprachsignale vor allem im vorderen Bereich der Perioden liegen.

Eine Alternative zum TD-PSOLA ist das FD-PSOLA-Verfahren (*Frequency-Domain-PSOLA*), bei dem die Signalausschnitte einer FFT (*Fast-Fourier-Transformation*) unterzogen werden. Die Syntheseergebnisse sind besser als beim TD-PSOLA, da die spektrale Glättung an den Übergängen der Fenster vereinfacht wird, das Verfahren ist jedoch äußerst aufwendig. Beim LP-PSOLA (*Linear Predictive PSOLA*) wird das Signal durch LPC-Analyse (siehe Abschnitt 3.5.2) in ein Residual und ein Signalformungsfilter getrennt und der PSOLA-Algorithmus dann auf das Residual[11] angewendet. Dabei werden die im Filter kodierten Formanten nicht beeinflusst. Viele weitere Verfahren sind vom TD-PSOLA abgeleitet, zu nennen wäre hier noch das MBR-TD-PSOLA (*Multi-Band-Resynthesis-TD-PSOLA*, Dutoit & Leich (1993 a)). Dieses Verfahren soll hier kurz erläutert werden, da der in Kapitel 5 verwendete MBROLA-Synthesizer darauf basiert. Dabei wird zur Präparierung der Segmentdatenbank eine Resynthese durchgeführt, die eine Pitch-, Phasen- und Frequenzangleichung vornimmt. Diese basiert auf dem MBE- (*Multi-Band-Excitation*) Modell, welches Sprache als Ergebnis der Addition unterschied-

[11] Bei einigen Verfahren wird auch eine komprimierte Form des Residuals verwendet

licher Frequenzbänder, die jeweils durch harmonische oder stochastische Parameter beschrieben werden, modelliert. Zunächst wird der harmonische Anteil geschätzt, der stochastische ergibt sich dann aus der Subtraktion mit dem Sprachsignal (vgl. Dutoit (1997, S. 262)). Dieses Verfahren ist zwar rechnerisch sehr aufwendig, aber es wird wie gesagt nicht zur Synthesezeit, sondern nur einmal bei der Generierung der Datenbank angewendet. Die Konkatenation der Segmente kann dann einfach durch lineare Interpolation im Zeitbereich erfolgen. Ein weiterer Vorteil dieser Vorgehensweise besteht darin, dass das für den TD-PSOLA-Algorithmus kritische Pitch-Marking automatisch und praktisch fehlerfrei durchgeführt werden kann. MBROLA (Dutoit et al. (1996)) stellt eine Erweiterung des MBR-PSOLA in Bezug auf die Komprimierung der Segmente dar. Durch die Pitch- Amplituden- und Phasengleichheit der Segmente können diese durch eine pitchsynchrone DPCM- (*Differential PCM*) Kodierung sehr effektiv komprimiert werden. Obwohl es ein halbparametrisches Verfahren darstellt, verhält sich MBROLA zur Synthesezeit wie ein reines Zeitbereichs-Verfahren, wodurch sich eine Modellierung emotionaler Sprechweise auf prosodische Parameter beschränken muss.

Zwei weitere auf PSOLA basierende Verfahren zur Veränderung der Charakteristik von Sprachsignalen werden von Valbret et al. (1992) vorgeschlagen. Das Spektrum des zu resynthetisierenden Signals wird dabei dem Spektrum eines Zielsignals alternativ durch LMR (*Linear Multivariate Regression*) oder DFW (*Dynamic Frequency Warping*) angepasst. Grundfrequenz und Dauern werden wiederum durch Anwendung von PSOLA auf das Residual verändert.

3.5.2 LPC-Verfahren

LPC (*Linear Predictive Coding*, Lineare Prädiktion) -Verfahren basieren wie die Formantsynthese auf dem Quelle-Filter Modell (vgl. Abschnitt 3.4.2). Sie wurden ursprünglich als Signalkodierungsverfahren entwickelt[12]. Die Analysekomponente berechnet für einen Frame der

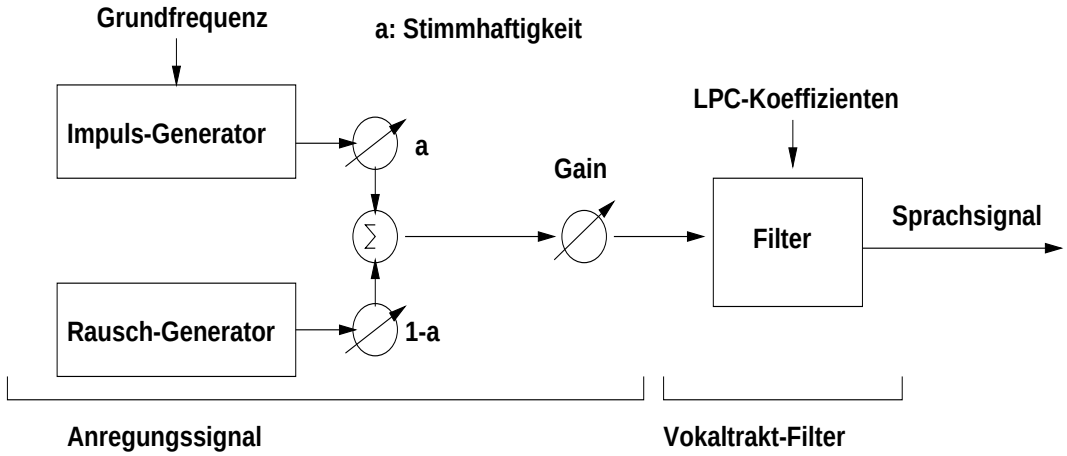

Abbildung 3.6: Einfaches LPC-Synthese Modell (nach O'Shaughnessy (1987))

[12]Eine Weiterentwicklung der PCM-Kodierung, das ADPCM (*Adaptive Differential PCM*) ist eine Form der LPC-Kodierung (vgl. Fellbaum (1984, S. 142)).

Länge N Samples die Koeffizienten für ein Filter P-ter Ordnung, der die spektrale Zusammensetzung des Frames möglichst gut beschreibt. Dabei wird jeder Signalwert $\hat{s}_n$ durch die gewichtete Summe seiner Vorgänger geschätzt (prädiziert) (vgl. Markel & Gray (1976, S. 10)):

$$\hat{s}_n = \sum_{k=1}^{P} a_k s_{n-k}, \text{ für } n \text{ von } 1..N$$

Da N in der Regel größer als P ist, bleibt dabei ein Fehlersignal (Residual) e_n der quadrierten Differenzen:

$$e_n = (\hat{s}_n - s_n)^2$$

Die Koeffizienten werden gefunden, indem die Summe der Fehlerquadrate minimiert wird.

$$\begin{aligned} E &= \sum_{n=1}^{N-P} e_n = \sum_{n=1}^{N-P} (\sum_{k=1}^{P} a_k s_{n-k} - s_n)^2 \\ \frac{\partial E}{\partial a_k} &= 0 \end{aligned}$$

Daraus ergibt sich ein lineares Gleichungssystem mit P Gleichungen zur Lösung der P Prädiktorkoeffizienten a_k. Für die Lösung dieser partiellen Differentiation werden vor allem zwei Verfahren angewendet: Autokorrelations- und Kovarianzverfahren.

Für die Synthese ist hauptsächlich die automatische Trennung in Quelle und Filter interessant. Das Residual steht dabei wiederum im Zusammenhang mit dem Glottissignal und die Prädiktionskoeffizienten mit dem Artikulationstraktfilter. Tatsächlich haben Rank & Pirker (1998) versucht, durch Manipulation des Residuals unterschiedliche Phonationsarten zu simulieren (wenn auch nicht sehr erfolgreich). Bei einfachen LPC-Synthesizern wird das Residual nicht weiter verwendet, sondern je nach Stimmhaftigkeit durch eine periodische Impulsfolge oder eine Zufallsfolge, die weißes Rauschen modelliert[13], generiert (vgl. Abb. 3.6).

Bei der Anwendung eines LPC-Verfahrens sind mehrere Parameter zu berücksichtigen. Die Anzahl der Koeffizienten (Filterordnung) hängt von der Samplerate ab. Als Faustregel gilt: Filterordnung = Samplerate (in kHz). Dies hängt damit zusammen, dass zwei Koeffizienten jeweils einen Formanten beschreiben können. Da in etwa pro kHz-Band ein Formant auftritt, und das Signal maximal Frequenzen der halben Samplerate enthalten kann, liegt die Anzahl der zur Beschreibung des Spektrums günstigen Koeffizienten ungefähr bei dieser Größenordnung.

Als Verfahren zur Koeffizientenbestimmung wird in der Regel entweder die Autokorrelations- oder die Kovarianzmethode angewendet. Die Fensterlänge für die LPC-Analyse beträgt normalerweise zwischen 5 und 20 ms. Die Verschiebungszeit des Fensters kann ebenfalls festgelegt werden oder bei stimmhaften Signalteilen pitchsynchron erfolgen, was eine vorherige Festlegung von Pitch-Markern erfordert. Die Lage der Pitch-Marker ist für eine gute Resynthesequalität kritisch und muss von Hand nachgebessert werden (vgl. KAY (1992)). Als Fensterform werden bei Sprachsignalen oft Hammingfenster verwendet.

[13] Eine Verbesserung der Qualität vor allem stimmhafter Frikative wird dadurch erreicht, dass Mischformen wie Addition von periodischen und stochastischem Signal oder periodisch geformtes Rauschen zugelassen werden.

Da die Frequenzenergie von Sprachsignalen mit etwa 6 dB pro Oktave abfällt, wird vor der LPC-Analyse Preemphase[14], auf das Signal angewendet, um auch die höheren Frequenzanteile analysieren zu können. Dies wird bei der Synthese durch Deemphase wieder ausgeglichen. Bei stimmlosen Lauten ist die Preemphase geringer zu wählen, da hier mehr Energie in den oberen Frequenzbereichen vorhanden ist.

Bei der LPC-Resynthese wird als Anregungssignal nicht (wie in Abb. 3.6 dargestellt) ein künstlich erzeugtes Gemisch aus Impulsfolge und Rauschen verwendet, sondern das Residualsignal. Die Koeffizienten des Filters können in Mittenfrequenzen und Bandbreiten der Formanten umgerechnet werden, wodurch eine Manipulation der Formantlagen durch Resynthese möglich wird. Dieses Verfahren arbeitet allerdings nicht fehlerfrei, da bei der Umrechnung der Koeffizienten und wieder zurück ein Informationsverlust auftritt. Zudem wird das Verhältnis zwischen Spektrum und Grundfrequenz bei einer Modifikation gestört. Bei der Resynthese wird diese Verfälschung des Signals deutlich hörbar. Eine weitere Fehlerquelle ist die Vereinfachung vieler LPC-Analyseverfahren, bei denen der Filter nur Pol- aber keine Nullstellen hat. Die Annahme, das Sprachspektrum hätte keine Nullstellen, stellt aber vor allem bei Nasalen eine grobe Vereinfachung dar (vgl. O'Shaughnessy (1987, S. 337) bzw. die Diskussion unter Abschnitt 3.4.2).

Die meisten Modifikationen der LPC-Synthese betreffen die Modellierung des Residuals. Multi-Pulse LP z.B. kodiert das Residual durch ein Optimierungsverfahren nach Lage und Amplitude der Impulse (vgl. Dutoit (1997, S.220)). Beim GAR- (*Glottal Autoregressive*) Modell wird das Residual noch einmal durch ein AR-Modell beschrieben, indem es in ein parametrisiertes Glottissignal-Modell und eine stochastische Komponente unterteilt wird. Die LPC-Analysegleichung wird damit folgendermaßen modifiziert:

$$s_n = \sum_{k=1}^{P} a_k s_{n-k} + u_n(\theta) + f_n$$

wobei $u_n(\theta)$ das mit θ parametrisierte Glottismodell und f_n die rauschhafte Komponente meint. Eventuell ließe sich hierdurch eine Modellierung unterschiedlicher Phonationsarten zum Zweck der Simulation emotionaler Sprechweise besser realisieren, als wenn das Residual direkt modifiziert wird. Das Verfahren ist allerdings sehr aufwendig und wohl nicht zur Synthesezeit durchzuführen.

3.5.3 Formantsynthese

Die Formantsynthese stellt das signalmodellierende Verfahren der Sprachsynthese dar, das eine phonetisch motivierte Modellierung der Sprachgenerierung am weitesten ermöglicht. Ausgangspunkt ist ebenso wie bei der LPC-Synthese das Quelle-Filter-Modell von Fant (1960) (siehe Abb. 3.4). Da der in Kap. 6 vorgestellte Synthesizer auf diesem Verfahren basiert, wird es hier näher erläutert. Dabei wird vor allem auf den von Klatt (1980) entwickelten Synthesizer Bezug genommen.

[14] Das Signal wird einer Hochpassfilterung unterzogen.

Überblick

Die Formantsynthese modelliert das Ansatzrohr durch eine Reihen- und/oder Parallelschaltung von digitalen Resonatoren, die die Frequenzen und Bandbreiten der ersten drei bis vier Formanten repräsentieren. Der Vorteil von Reihen-Formantsynthesizern besteht einerseits darin, dass die Amplituden für die einzelnen Formanten nicht explizit angegeben werden müssen und anderseits darin, dass er ein adäquateres Modell für stimmhafte Lautproduktion ohne Zuschaltung des Nasaltrakts darstellt. Allerdings wird hier zusätzlich mindestens ein parallel geschalteter Resonator benötigt, um Frikative und Plosivbursts zu modellieren. Höhere Formanten werden oft fest vorgegeben, da sie vor allem für den Klang der Stimme verantwortlich und nicht so sehr charakteristisch für einzelne Laute sind. Bei den meisten Formantsynthesizern sind auch die Formantbandbreiten nicht zeitveränderlich, da eine Variation perzeptiv kaum relevant ist. Fujimura & Lindqvist (1971) messen zwar eine von der Formantlage abhängig veränderte Bandbreite des ersten Formanten, aber laut Untersuchungen von Carlson et al. (1979) werden diese Variationen kaum wahrgenommen (vgl. Abschnitt 6.2.4).

Eine spezielle Schwierigkeit stellt hierbei die Synthese von Frauen- und Kinderstimmen dar. Einerseits ist durch die höhere Grundfrequenz eine Bestimmung der Formantlagen schwieriger als bei Männerstimmen. Zum anderen deuten informelle Untersuchungen darauf hin, dass die Vokalspektren von Frauen- und Kinderstimmen durch ein Nurpol-Modell nur unzulänglich beschrieben werden können, was an einer stärkeren Kopplung von Quelle und Filter liegt (Klatt & Klatt (1990, S. 820)).

Eine konkrete Implementation eines kombinierten Reihen/Parallel-Formantsynthesizers ist bei Klatt (1980) beschrieben. Das akustische Generierungsmodul des in Kap. 6 beschriebenen Synthesizers basiert auf dieser Implementation. Ein Schaltbild des Synthesizers ist in Abb. 3.7 dargestellt. Jeder Resonator wird dabei durch Angabe von Mittenfrequenz und Bandbreite charakterisiert. Als digitales Bandpassfilter (Polstelle, Formant) wird es durch die Formel:

$$y(nT) = Ax(nT) + By(nT - T) + Cy(nT - 2T)$$

repräsentiert. T stellt dabei die Sampledauer dar, $y(nT)$ das Ausgangs- und $x(nT)$ das Eingangssignal. A, B und C sind Konstanten, die sich nach der impulsinvarianten Transformation von Gold & Rabiner (1968) aus der Mittenfrequenz F und der Bandbreite BW wie folgt berechnen:

$$\begin{aligned} C &= -e^{-2\pi BW\ T} \\ B &= 2e^{-\pi BW\ T}\cos(2\pi F\ T) \\ A &= 1 - B - C \end{aligned} \tag{3.1}$$

Die Werte für die Resonatoren gelten wie bei der LPC-Synthese jeweils für einen Frame (in der Regel 5 msek). Weiterhin werden zur Modellierung von Sprachsignalen Antiresonatoren benötigt. Sie bilden die Nullstellen des Spektrums ab und stellen somit eine Spiegelung der Polstellen dar. Der Output eines Antiresonators verhält sich zum Input nach der Formel:

$$y(nT) = A'x(nT) + B'y(nT - T) + C'y(nT - 2T)$$

wobei A', B' und C' durch

$$A' = \frac{1}{A};\ B' = -\frac{B}{A};\ C' = -CA;$$

gegeben sind. A, B und C ergeben sich dabei aus den Gleichungen 3.1. Spektrale Nullstellen werden benötigt, um das Spektrum des stimmhaften Anregungssignals sowie nasale Antiformanten zu modellieren.

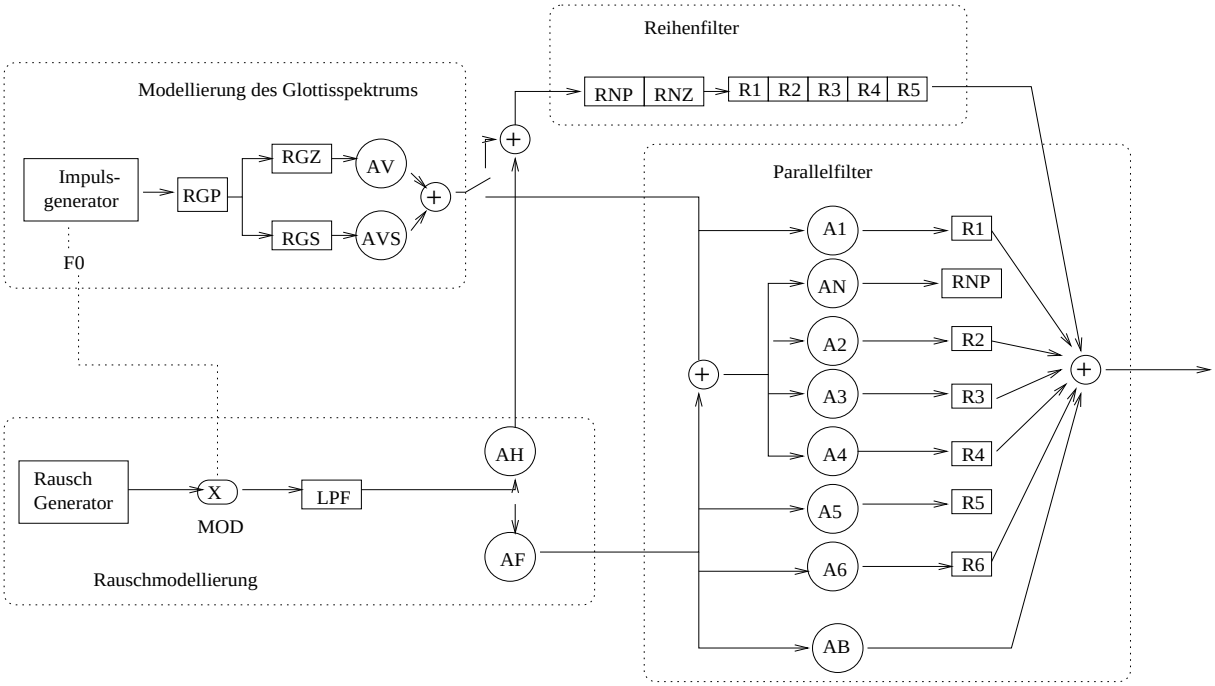

Abbildung 3.7: Blockdiagramm des KLATT80-Formantsynthesizers (nach Klatt (1980)). R=Resonator, A=Amplitudensteuerung, G=Glottissignal, N=Nasal, P=Polstelle, Z=Nullstelle

Modellierung glottaler und subglottaler Anregungssignale

Das Glottissignal $U_g(t)$ wird in der Regel durch das transglottale Luftvolumen über der Zeit dargestellt. Es ist zu unterscheiden zwischen dem oralen Luftstromsignal $U_g(t)$ und dem differenzierten Signal $U_g'(t)$, das durch die Lippenabstrahlcharakteristik ensteht (vgl. Ni Chasaide & Gobl (1997, S. 431)).

Input eines Formantsynthesizers und Output einer inversen Filterung ist die Ableitung $U_g'(t)$ dieses Signals nach der Zeit. Ursprünglich wurde dieses Signal lediglich als tiefpassgefilterte Impulskette implementiert (Ni Chasaide & Gobl (1997, S. 434)). Die zwei Hauptnachteile dieses einfachen Modells bestehen darin, dass erstens die spektrale Dämpfung des Signals nicht beeinflusst werden kann (als Parameter stehen nur F_0 und Amplitude zur Verfügung) und zweitens die Impulsantwort in Bezug auf das tatsächliche Signal zeitverdreht ist.

Deshalb gibt es viele Alternativvorschläge zur Modellierung des Glottissignals. Ananthapadmamnabha (1984) etwa schlägt ein fünf-Parameter Modell vor, welches den Signal-Verlauf aus parabolischen Funktionen approximiert.

Exkurs: Das LF-Modell

Da die Modellierung des Glottissignals eine Simulation der für emotionale Eindruckswirkung relevanten Phonationsarten ermöglicht, sollen hier einige Modelle näher erläutert werden. Sehr verbreitet ist das LF- (*Liljencrants-Fant*) Modell (Fant et al. (1985)), bei dem das Signal durch eine Kombination aus einer Sinus- und einer Exponentialfunktion beschrieben wird (siehe Abb. 3.8). Die vier beschreibenden Parameter sind:

- E_e, die maximale negative Amplitude, sie entspricht der Amplitude, mit der der Artikulationstrakt im Moment des Glottisverschlusses angeregt wird.
- t_p, der Zeitpunkt des ersten Nulldurchgangs von $U_g'(t)$, und damit der Zeitpunkt des maximalen Luftdrucks bei $U_g(t)$.
- t_e, der Zeitpunkt der maximalen negativen Amplitude
- T_a, die Dauer des Abschnitts zwischen t_e und dem Nulldurchgang einer Tangente am Signal. Dieser Parameter verhält sich dabei proportional zur spektralen Dämpfung. Er gibt die Zeit an zwischen der maximalen Erregung und dem Moment, in dem der glottale Luftstrom quasi konstant ist.

Aus diesen Modellparametern lassen sich weitere ableiten, die sich zur Beschreibung von Glottissignalen durchgesetzt haben (vgl. Ni Chasaide & Gobl (1997); Karlsson (1992)):

- Die Stärke der Anregung, **Ee**, entspricht der negativen Amplitude im differenzierten Schalldruckverlauf, typischerweise kurz vor Beginn der geschlossenen Phase. Auf der Produktionsseite hängt dieser Parameter von der Stärke des Luftstroms und der Geschwindigkeit des Schließungsvorgangs ab, auf der Perzeptionsseite entspricht er der Intensität des Signals.
- **RA**, das glottale Leck, gibt die Luftmenge an, die zwischen maximaler Anregung und kompletter (oder maximaler) Schließung der Glottis strömt. Innerhalb des LF-Modells ist RA $= T_a/t_0$.
- **OQ**, der Öffnungsquotient beeinflusst hauptsächlich die Amplitude der tieferen Komponenten im Anregungssignal.
- **FG**, die Glottale Frequenz ist definiert als $1/2T_p$. Sie wird also durch die Dauer der Öffnungsphase festgelegt. Üblicherweise wird sie in der zu F_0 normalisierten Form verwendet: **RG** = FG/F_0. Ein hoher RG-Wert stärkt die Amplitude der zweiten Harmonischen, während ein geringer (längere Öffnungsperiode) die der ersten verstärkt.

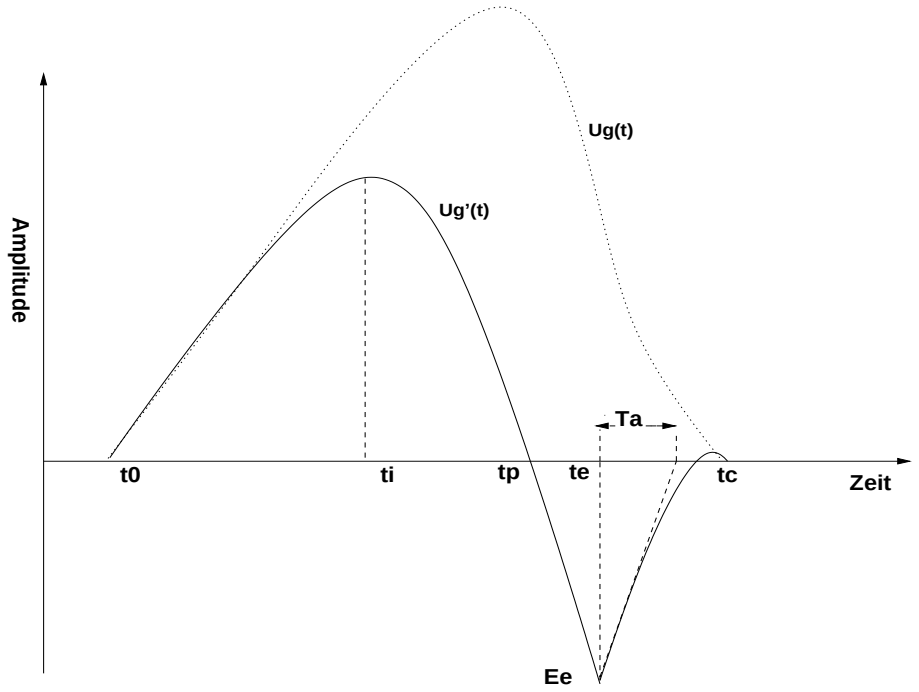

Abbildung 3.8: Parametrisierung des differenzierten Schalldruckverlaufs an der Glottis nach dem LF-Modell (nach Fant (1985))

- **RK** ist ein Parameter für die Neigung des Glottisimpulses. Bei modaler Phonation ist der Impuls etwas nach rechts geneigt, die geöffnete Phase ist dann länger als die geschlossene. Perzeptiv wirkt sich die Neigung auf die Amplitude tieferer Frequenzen sowie die Stärke spektraler Senken (fehlende oder schwache Harmonische) aus.

Diese Parameter sind im Allgemeinen nicht unabhängig voneinander. So führt z.B. in der Regel ein schwächeres EE zu einem höheren RA. Es gibt auch einen Zusammenhang zwischen RA und RK, da bei längerer Schließungsphase der Impuls eher symmetrisch ist.

Die Modellierung des Glottissignals beim KLSYN88-Synthesizer

Bei der ersten Version des Klatt-Synthesizers ist ein sehr einfaches Modell für das Glottissignal vorgesehen. Es wird durch eine tiefpassgefilterte Impulskette erzeugt, so dass das resultierende Spektrum eine Dämpfung von -12 dB/Oktave oberhalb von 50 Hz aufweist (Abb. 3.7, Resonator RGP). Aus Berechnungsgründen wird der Effekt der Lippenabstrahlcharakteristik, der als Hochpass wirkt, bereits bei der Glottissignalgenerierung berücksichtigt, so dass das Spektrum lediglich um 6 dB/Oktave gedämpft wird. Um individuelle Spektren genauer modellieren zu können, ist zudem ein Antiresonator vorgesehen (Abb. 3.7, Resonator RGZ). Zur Modellierung einer quasi-sinuidalen Anregung, wie sie als Ergebnis nicht vollständig schließender Stimmlippen auftritt, ist ein weiteres Tiefpassfilter vorgesehen (Abb. 3.7, Resonator RGS), der das Spektrum bis auf -24 dB/Oktave abdämpft (Klatt (1980, S. 976)).

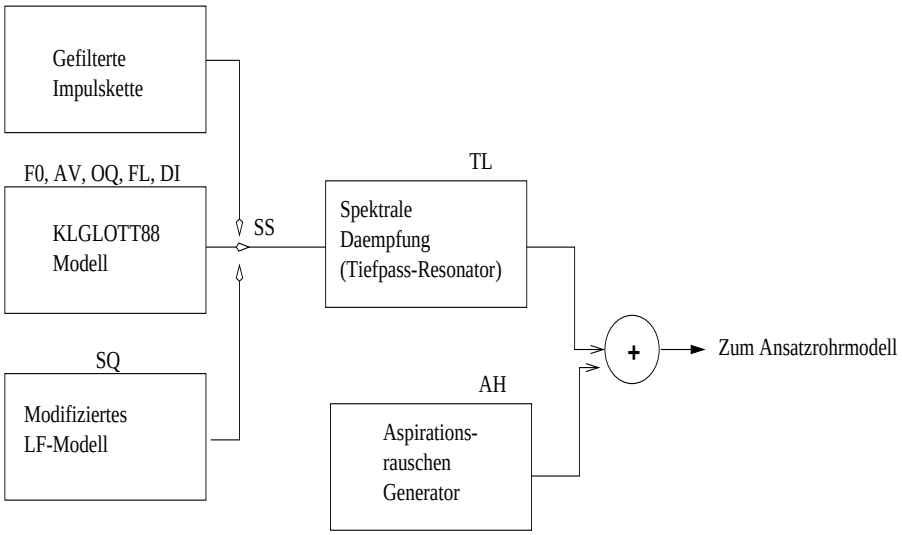

Abbildung 3.9: Blockschaltbild der modifizierten Glottis-Quellen beim KLSYN88-Synthesizer (nach Klatt & Klatt (1990))

Das Quellsignal für stimmhafte Laute kann beim später entwickelten KLGLOTT88-Modell (Klatt & Klatt (1990)) unter mehreren Modellen durch den Schalter SS (*source switch*) gewählt werden (siehe Abb. 3.9). Mit dem Ziel, ein natürlicheres Glottissignal erzeugen zu können, gibt es dabei weitere zum Teil vom LF-Modell abgeleitete Parameter, um das Glottissignal zu beschreiben. Die einzelnen Parameter zur Steuerung des Glottissignals seien hier kurz dargestellt:

- **F0**, die Grundfrequenz.
- **AV** (Amplitude of Voicing), maximale Amplitude des Signals.
- **OQ** (Open-Quotient), das Verhältnis zwischen geöffneter Phase und der Gesamtdauer des Signals.
- **AH** (Amplitude of Aspiration), die Amplitude des dem Glottissignal beigemischten Rauschens.
- **TL** (Spectral Tilt), eine Dämpfung der höheren Frequenzbereiche, die durch schwächer ausgeprägte Ecken am Signalwendepunkt erreicht wird.
- **FL** (F_0-Flutter), eine Jitter-Simulation, bei der die Periodendauern quasizufallsmäßig mit drei Sinusschwingungen verschoben werden.
- **DI** (Diplophonic Double Pulsing), resultiert in einer Verschiebung und Amplitudendämpfung jeder ersten Periode eines Paares.

Die letzten beiden Parameter beschreiben dabei nicht Eigenschaften einer Periode, sondern Irregularitäten im F_0-Verlauf. *Diplophonisches Doppelpulsieren* ist ein Effekt, der oft bei laryngalisierter Sprechweise auftritt, z.B. am Äußerungsende. Dabei rücken je zwei Perioden näher

aneinander, wobei die erste oft stark gedämpft ist. Dieser Effekt kann so stark werden, dass die beiden Perioden zu einer verschmelzen.

Alle Parameter müssen dynamisch veränderbar sein, wobei sich die Werte vor allem bei Beginn und Ende stimmhafter Perioden rapide ändern (Strik & Boves (1991)). Strik & Boves (1991) untersuchten die Änderung glottaler Parameter in fließender Sprache und stellten dabei Unterschiede zwischen Vokalen und stimmhaften Konsonanten fest. Die Konsonanten hatten ein größeres T_a und einen größeren Abstand zwischen t_p und t_e. Allerdings waren die Parameter für stimmhafte Konsonanten schwieriger abzuschätzen als für Vokale.

Modellierung glottalen Rauschens Da beim KLATT-Synthesizer die Reihenschaltung speziell zur Modellierung von Lauten mit glottaler Anregung vorgesehen ist, werden Behauchung oder glottale Frikative durch eine Filterung des Rauschens mit dem Reihentrakt modelliert (Abb. 3.7, Amplitude AH).

Quelle-Filter Interaktionen

Das ursprüngliche Konzept des Quelle-Filter-Modells, nach der Quelle und Filter als voneinander unabhängig agierend gesehen wurden, stellt eine sehr grobe Vereinfachung der Realität dar (siehe Beschreibung der Probleme des Quelle-Filter Modells, Abschnitt 3.4.2).

Einmal wird der direkte Zusammenhang zwischen Glottisfläche und transglottalem Luftstrom durch stehende Wellen in der Pharynx gestört. Diese erzeugen eine Welligkeit im Bereich des ersten Formanten. Klatt & Klatt (1990) beschreiben vier Phänomene, die bei herkömmlichen Formantsynthesizern nicht berücksichtigt werden:

- Der Einfluss stehender Wellen im subglottalen Bereich bewirkt eine Welligkeit und Verstärkung der Amplitude beim ersten Formanten. Dies wirkt sich perzeptiv erst durch eine dynamische Entwicklung aus, (Klatt & Klatt, 1990, S. 841) schlagen eine Verringerung der ersten Bandbreite und eine Vergrößerung der spektralen Dämpfung über die ersten Perioden eines Stimmeinsatzes vor.

- Die Amplitude des Glottissignals steigt, wenn der erste Formant bei einem Vielfachen der Grundfrequenz liegt. Dies ist ebenfalls eine Folge der pharyngalen stehenden Wellen.

- Die sinkende glottale Impedanz bei der Öffnungsphase verursacht eine Verbreiterung der ersten Bandbreite und ein Ansteigen der ersten Formantfrequenz. Aus diesem Grund ist beim KLSYN88-Synthesizer durch die Parameter **DF1** und **DB1** (für Delta F1 und B1) eine pitchsynchrone Änderung der Kennwerte für den ersten Formanten vorgesehen.

- Während der geöffneten Phasen bzw. bei nicht ganz schließender Glottis treten tracheale Pol- und Nullstellen auf. Diese können analog zu nasalen Pol- und Nullstellen durch ein Resonator/Antiresonator-Paar in der Reihenschaltung modelliert werden.

Modellierung supraglottaler Anregungssignale

Supraglottale Friktionsgeräusche wurden beim KLATT80-Synthesizer durch ein gaussverteiltes Rauschen modelliert. Dabei wird ein flaches Spektrum angenommen. Die Dämpfung der höheren Frequenzen wird durch die Anhebung der höheren Frequenzen durch die Lippenabstrahlcharakteristik aufgehoben (Klatt (1980, S. 977)).

Stimmhafte Frikative werden durch eine Modulation des Rauschens mit einem Rechtecksignal der Grundfrequenz modelliert.

Modellierung von Vokalen

Bis zu einer Frequenz von 5 kHz lässt sich die Artikulationstrakt-Transferfunktion durch eine Wellengleichung erster Ordnung beschreiben (Fant (1960)). Solange das Erregungssignal dabei vom Larynx herrührt und der Nasaltrakt nicht zugeschaltet ist, kann die Transferfunktion durch ein Produkt von vier bis sechs Polstellen approximiert werden.

Um die zusätzlichen Nullstellen bei leicht geöffneter Glottis zu berücksichtigen, die von der Kopplung des sub- mit dem supraglottalen Bereich herrühren, schlägt Klatt (1980) die Verbreiterung der Bandbreite des ersten Formanten auf etwa 300 Hz vor.

Zur Modellierung nasalierter Vokale, wie sie beispielsweise in der Umgebung von Nasalen vorkommen, ist für die Reihenschaltung ein weiteres Resonator/Antiresonator-Paar vorgesehen (Abb. 3.7, Resonatoren RNP und RNZ). In diesem Fall wird der Nasaltrakt als Nebenrohr des Artikulationstrakts angesehen. Für die Erzeugung von Nasalen wird dagegen der Artikulationstrakt als Nebenrohr des Nasaltrakts angesehen und der parallele Zweig benutzt (Abb. 3.7, Resonator RNP im parallelen Zweig).

Modellierung von Frikativen

Für frikative Spektren sind zusätzliche Nullstellen charakteristisch, die aus der Kopplung des Bereiches hinter der Friktionsstelle mit dem Bereich davor herrühren. Diese erzeugen Kerben im Spektrum und beeinflussen die Formantamplituden. Die Kerben sind zwar perzeptiv nicht relevant (Carlson et al. (1979)), aber die Amplitudenveränderungen lassen eine adäquate Modellierung durch eine Reihenschaltung nicht mehr zu. Deshalb werden frikative Spektren durch parallel geschaltete Resonatoren erzeugt, bei denen die Amplituden einzeln anzugeben sind. Die Amplituden hängen dann vom Artikulationsort des Frikativs ab. Für sehr weit vorn artikulierte Frikative (die ein breitbandiges Rauschen erzeugen, z.B. /f/) ist ein Bypass-Weg vorgesehen (Abb. 3.7, Amplitude AB).

Modellierung sonoranter Laute mit der Parallelschaltung

Die Reihenschaltung des Formantsynthesizers kann auch durch die Parallelschaltung ersetzt werden. Dafür sind dann die Amplituden der Formanten explizit anzugeben. Dabei sind gewisse Regeln zu beachten. Die Bandbreite und Amplitude eines Formanten hängen umgekehrt proportional miteinander zusammen (eine Verdopplung der Bandbreite verringert die Amplitude

um 6 dB). Für die Filterfunktion steigt mit der Frequenz eines Formanten auch die Amplitude, dies wird aber durch die spektrale Dämpfung der Anregungsfunktion wieder aufgehoben.

Weiterhin beeinflusst die Verschiebung eines Formanten die Amplituden der höher liegenden Formanten. Liegen zwei Formanten dicht beisammen (etwas über 200 Hz) so steigen ihre Amplituden um etwa 3 bis 6 dB an.

Zur Modellierung von Nasalen durch die Parallelschaltung ist ein weiterer Resonator vorgesehen (Abb. 3.7, Resonator RNP).

Modellierung der perzeptiven Korrelate von Phonationsarten

Da die Phonationsarten (vgl. Abschnitt 2.1) für emotionale Eindruckswirkung relevant sind, wurde eine Simulation in dem unter Kap. 6 beschriebenen Synthesizer implementiert. Daher werden an dieser Stelle die Möglichkeiten der Modellierung der akustischen Eigenarten mittels Formantsynthese der drei wichtigsten Phonationsarten für den emotionalen Stimmeindruck erörtert.

Behauchung (*breathy voice*). Die behauchte Anregung ist durch einen unvollständigen Verschluss der Stimmlippen gekennzeichnet. Dies hat zum einen ein Rauschen zur Folge, zum anderen weist das entstehende Glottissignal eine sinusartige Form auf. Die Öffnungsperiode ist dabei verlängert.

Akustisch macht sich eine behauchte Anregung durch eine leicht abgesenkte Grundfrequenz, Rauschen in den höheren Frequenzbereichen, Dämpfung der höheren Frequenzbereiche und durch eine Anhebung der Amplitude der ersten Harmonischen bemerkbar (Klatt & Klatt, 1990, S. 824).

Behauchte Stimmen wurden bei Klatt & Klatt (1990, S. 848) durch folgende Syntheseparameter modelliert:

- Verstärkung der ersten Harmonischen durch Erhöhung des Öffnungsquotienten.
- Dämpfung der höheren Frequenzbereiche durch spektralen Tilt.
- Beimischung von Rauschen bei mittleren und hohen Frequenzen durch den Parameter AH.
- Verbreiterung der ersten Bandbreite.
- Hinzufügen trachealer Pol- und Nullstellen.

Knarrstimme (*creaky voice / laryngealized voice*). Die Perioden werden bei der Knarrstimme zum Teil als einzelne Pulse wahrgenommen. Die Steuerung der Grundfrequenz erfolgt hierbei hauptsächlich durch die aerodynamische Komponente der Stimmerzeugung, also durch Variation des subglottalen Drucks. Der glottale Puls wird dabei relativ schmal, d.h. die Öffnungsphase ist verkürzt (Klatt (1990)). Die Amplitude der fundamentalen Komponente ist dabei verringert. Der Effekt entspricht dem des diplophonischen Doppelpulsierens, so dass der Parameter **DI** zur Simulation der Knarrstimme verwendet werden kann.

Falsettstimme (*falsetto*). Durch die hohe Grundfrequenz gibt es weniger Harmonische, die resultierende Sprache klingt ’dünner’. Die höheren Harmonischen sind mit etwa 20 dB/Oktave stark gedämpft (Laver (1980, S. 120)).

Gewinnung der Steuerparameter

Um verständliche Sprache zu erzeugen, benötigt ein Formantsynthesizer pro Frame einen Satz von Steuerparametern, die das Verhalten von Quelle und Filter festlegen. Die Formantlagen werden in der Regel durch Hilfsprogramme ermittelt, mittels denen sich Formantverläufe auf der Grundlage von natürlichen Sprachaufnahmen stilisieren lassen (Klatt (1980)).

Für Resynthesezwecke ist das Original zu analysieren. Dabei werden, von einer Analyse des Spektrogramms ausgehend, Werte für Grundfrequenz, Formantlagen, spektrale Zusammensetzung bei stimmlosen Lauten und weiteres festgelegt. Ein Verfahren hierfür ist zum Beispiel mit LACS (*Label Assisted Copy Synthesis*) (Scheffers & Simpson (1995)) gegeben. Im Prinzip ist es bei genügender Sorgfalt möglich, vom Original ununterscheidbare Kopien erzeugen. Ein berühmtes Beispiel einer solchen Kopie stammt von Holmes (1973).

Zur Synthese unbegrenzter Sprachausgabe kommen wiederum der regelbasierte und der datenbasierte Ansatz in Frage (vgl. Abschnitte 3.4.1 und 3.4.2). Die regelbasierte Synthese geht von einer Verkettung von Allophonen aus, für die jeweils Targetwerte in einer Datenbank abgespeichert sind. Zur Synthesezeit werden dann die Parameter durch regelbasierte Interpolation zwischen diesen Targetwerten berechnet. Der Vorteil liegt in einer sehr kleinen Datenbank und einer hohen Flexibilität des Systems. Als problematisch stellen sich vor allem Phone dar, die keinen *steady-state* besitzen, wie etwa Plosive.

Datenbasierte Systeme speichern die Steuerparameter in Segmenten. Zur Synthesezeit sind dann lediglich die Schnittstellen zwischen den Segmenten zu glätten. Die parametrisierten Segmentbeschreibungen werden zuvor aus im Zeitbereich vorliegenden Samples extrahiert. Ein Verfahren dafür ist in Mannell (1998) beschrieben. Dabei werden in einem ersten Schritt die Länge des Ansatzrohrs und die durchschnittlichen Abstände der ersten fünf Formanten durch Analyse des neutralen Vokals [ə] abgeschätzt. Ausgehend von diesen Werten wird dann in einem zweiten Schritt ein Wahrscheinlichkeitsraum für jeden Vokal und Formanten berechnet, mittels dem die Ergebnisse einer LPC-Analyse auf Glaubwürdigkeit abgeschätzt werden können. Dabei werden zwei LPC-Analysen berechnet, eine mit 14 Koeffizienten, um die Hauptpolstellen zu berechnen und eine mit 24 Koeffizienten, die ausgehend von den Ergebnissen der ersten eine genauere Lage der Formanten liefert. Vokale, die an Nasale angrenzen, werden gesondert betrachtet, da der nasale Formant ansonsten die Lage des ersten Formanten verfälschen würde.

Diese Methode ist zwar als weitgehend automatisiertes Verfahren ausgelegt, arbeitet aber leider so fehlerhaft, dass eine umfangreiche Nachkorrektur unumgänglich ist. So bleibt das Hauptproblem datenbasierter Formantsynthese die Parametrisierung des Segmentinventars.

3.6 Zusammenfassung

Für die Simulation emotionaler Sprechweise mit Sprachsynthesizern sind also mehrere Faktoren relevant. Handelt es sich um ein zeitbereichsbasiertes Verfahren, so bleibt entweder die

Beschränkung auf prosodische Modifikationen oder der Einsatz mehrerer Segmentdatenbanken. Eine Untersuchung der Simulation emotionaler Sprechweise bei Beschränkung auf prosodische Merkmalsvariation ist in Kap. 5 beschrieben. (Halb-) Parametrische Synthesizer sind wesentlich flexibler, was die Modifikation stimmqualitativer und artikulatorischer Merkmale betrifft. Allerdings klingt die Ausgabe in der Regel qualitativ schlechter, da durch Datenmangel bzw. ungenügende Modellierung bei der Signaldekodierung Artefakte auftreten.

Da sämtliche Module eines Sprachsynthesesystems im Prinzip eher regel- oder datenbasiert ausgelegt sein können, sind diese beiden Ansätze auch im Hinblick auf die Simulation emotionaler Sprechweise relevant.

Eine Schwierigkeit für die automatische Generierung emotionalen Sprechausdrucks stellt das Fehlen syntaktischer und semantischer Information dar, welches eine zuverlässige Vorhersage von Betonungs- oder Sprechpausenmöglichkeiten nicht erlaubt.

Kapitel 4

Betrachtung der Literatur

4.1 Vorbemerkung

Dieses Kapitel gibt einen Überblick über den Stand der Forschung. Es werden vor allem Untersuchungen neueren Datums berücksichtigt. Frühere Arbeiten sind u.a. in Scherer (1981); Murray & Arnott (1993); Tischer (1993) sowie Frick (1985) zusammengefasst. Zunächst werden einige grundsätzliche Fragen und Konzepte erörtert. So behandelt der erste Abschnitt die Universalität akustischer Ausdrucksformen emotionaler Sprechweise. Weitere Abschnitte betreffen die Bedeutung einzelner Parametergruppen und erläutern physiologische Korrelate emotionaler Sprecherzustände. Der darauf folgende Teil ist Überlegungen zur Methodik der Datengewinnung gewidmet. Abschließend werden ausgewählte Untersuchungen im einzelnen beschrieben und die Ergebnisse nach Emotionen geordnet zusammengefasst. Darauf aufbauend, werden einige Hypothesen formuliert, die den experimentellen Teil der vorliegenden Arbeit motivieren.

Scherer (1986) weist auf das Paradox hin, dass, obwohl emotionale Zustände anhand des akustischen Signals eher erkannt werden als bei rein visuellem Eindruck, die akustischen Eigenschaften bislang nicht eindeutig beschrieben worden sind. Dafür gibt es mehrere Gründe (vgl. Banse & Scherer (1996)):

- Meistens wurde nur ein sehr beschränktes Set aus wenigen Basisemotionen untersucht, was zur Folge hat, dass feinere Differenzierungen eher unberücksichtigt geblieben sind.

- Oft wurden relativ grobe Parameter wie durchschnittliche Grundfrequenz, F_0-Range oder spektrale Energieverteilung in Langzeitspektren untersucht, durch deren Ausprägungen sich lediglich Emotionen mit unterschiedlichem Erregungsgrad zuverlässig differenzieren lassen.

- Es gibt bislang noch keine einheitliche Theorie über den Zusammenhang von Emotionen und ihrem akustischen Ausdruck.

Schwierigkeiten bei dem Vergleich unterschiedlicher Ergebnisse ergeben sich weiterhin aus unterschiedlichen Messverfahren und unklarer Bedeutung der Emotionsbegriffe. Mitunter sind auch sprachspezifische Variationen die Ursache für abweichende Ergebnisse in der Literatur.

4.2 Betrachtung der Universalität von emotionalem Sprechausdruck

Mit Einschränkungen hat sich gezeigt, dass die Fähigkeit, emotionale Sprechweise zu diskriminieren, durchaus sprach- und kulturuniversell ist. Eine spezielle Untersuchung zu diesem Thema ist mit Scherer et al. (2000) gegeben. Weiterhin fasst Frick (1985) mehrere Studien zu diesem Thema zusammen. Als Universalia können wohl alle akustischen Korrelate gelten, die sich physiologisch begründen lassen. Scherer (1995) unterscheidet hier zwischen psychobiologischen *push*- und soziokulturellen *pull*-Faktoren (vgl. Abschnitt 4.4).

Gegenüber den physiologisch bedingten Faktoren, die sich etwa auf Hormonausschüttungen bei Angst, Freude oder Wut zurückführen lassen, gibt es andere Einflussgrößen, die einer Universalität akustischer Korrelate entgegenstehen:

- **Sprecherspezifische Unterschiede**
  Selbstverständlich spielen bei der akustischen Manifestation von emotionalen Zuständen auch sprecherspezifische Eigenheiten eine Rolle. Bei Hecker et al. (1968) stellte sich z.B. heraus, dass sich die prosodischen Merkmale unter Stress bei den Versuchspersonen zwar konsistent, aber untereinander uneinheitlich veränderten.

- **Sprachspezifische Einschränkungen**
  Die Verwendung von stimmqualitativen, prosodischen und segmentellen Merkmalen in phonologischer Funktion variiert unter den Sprachen der Welt (vgl. Scherer et al. (2000)). So wirkt etwa im Kanton-Chinesischen der Parameter Tonhöhe bedeutungsunterscheidend. Ein anderes Beispiel wäre der Phonemstatus von Klick-Lauten im südafrikanischen Zhosa, deren Gebrauch als extralinguistisches Merkmal damit eingeschränkt ist. Weiterhin dient Nasalität im Sundanesischen (auf Java) zur Kennzeichnung von Verbformen (nach Laver, 1980, S. 4).

- **Interkulturelle Eigenarten**
  Die Bedeutung bestimmter stimmlicher Eigenschaften variiert zudem zwischen einzelnen Kulturräumen. Nasalität ist zum Beispiel im Cayuvava, einer bolivianischen Sprache, ein Zeichen von Unterwürfigkeit (nach Laver, 1980, S. 4), während durch bestimmte Arten der Nasalierung im deutschen oder englischen Sprachraum eher Arroganz ausgedrückt wird. Weiterhin variiert die gesellschaftliche Akzeptanz hinsichtlich der Ausprägung des emotionalen Ausdrucks.

Es besteht ein Zusammenhang zum Konzept der Basisemotionen dahingehend, dass solche als Basisemotionen gelten, die angeboren und damit kulturell universell sind, wohingegen Unterkategorien erlernt und damit kulturabhängig sind (vgl. Murray & Arnott (1993, S. 1098)). Scherer et al. (2000) konnten zeigen, dass emotionale Äußerungen aus einem Kulturkreis in anderen unterschiedlich gut wiedererkannt wurden, obwohl dabei sinnleere Texte verwendet wurden. Die von deutschen Schauspielern realisierten Phantasiesätze, die aus Silben indogermanischen Ursprungs gebildet waren, wurden in neun Ländern Erkennungstests unterzogen. Es ergab sich

für deutsche Beurteiler die höchste durchschnittliche Erkennungsrate (74 %) für die intendierten Emotionen und für indonesische Hörer die geringste Erkennungsrate (52 %). Dabei gehörte Indonesien als einziges Land (neben den USA) nicht zum west-/südeuropäischen Kulturkreis.

4.3 Bedeutung einzelner Parameterklassen für den emotionalen Ausdruck

Bei vielen Untersuchungen werden die akustischen Merkmale in prosodische und stimmqualitative Merkmale eingeteilt. Dabei sind mit dem Begriff *Stimmqualität* in der Regel nicht nur Eigenschaften des Glottissignals gemeint, sondern auch akustische Manifestationen von supralaryngalen Settings (vgl. Abschnitt 2.1.3) wie z.B. eine Formantanhebung aufgrund einer Verkürzung des Artikulationstrakts. Zu den prosodischen Merkmalen zählen in der Regel die drei suprasegmentellen Parameter *Intonationskontur, Intensitätskontur* und *Dauerstruktur*. Segmentelle Parameter wie etwa Artikulationsgenauigkeit wurden sehr selten überprüft (z.B. Kienast et al. (1999)).

4.3.1 Prosodische Parameter

Da prosodische Parameter relativ leicht zu messen sind, wurden diese bei fast allen Untersuchungen betrachtet. Dabei wurden allerdings in der Regel nur sehr grobe Beschreibungsgrößen wie etwa die durchschnittliche F_0-Lage oder der F_0-Range berücksichtigt. Bei der Untersuchung prosodischer Parameter im Hinblick auf ihre Variation bei emotionaler Sprechweise ist zu berücksichtigen, dass sie immer auch linguistische Funktionen erfüllen. Dennoch konnte bei allen Untersuchungen ein Einfluss emotionaler Zustände auf prosodische Parameter nachgewiesen werden. Zudem kann die emotionale Eindruckswirkung allein durch Manipulation prosodischer Merkmale erzeugt werden (z.B. Schröder (1999); Heuft et al. (1998); Rank & Pirker (1998); Vroomen et al. (1993) oder Carlson et al. (1992)).

Dabei scheint der Intonationsverlauf die entscheidende Rolle zu spielen. Frick (1985) berichtet von Untersuchungen, bei denen die Synthese allein der Intensitätsverläufe zu einer durchschnittlichen Erkennungsrate von lediglich 14 % führte (das Zufallsniveau lag bei 12,5 %). Bei ausschließlicher Berücksichtigung der F_0-Kontur lag die Erkennungsrate wesentlich höher (43 %) und steigerte sich nur unwesentlich bei zusätzlicher Synthese der Intensitätsverläufe (auf 47 %). Auch der prosodische Parameter Lautdauer scheint einen geringeren Einfluss auf emotionale Erkennung zu haben als die Intonation. So zeigte sich z.B. bei Carlson et al. (1992) keine deutliche Änderung im Beurteilerverhalten, wenn lediglich die Dauerstruktur verändert wurde. Weitere Hinweise auf die Wichtigkeit prosodischer Parameter für emotionalen Sprechausdruck geben Studien, die die automatische Klassifikation emotionaler Sprechweise zum Gegenstand haben (vgl. etwa Slaney & McRoberts (1998); Dellaert et al. (1996) und Banse & Scherer (1996)). Wegen der einfachen Parametrisierbarkeit werden dabei in aller Regel lediglich prosodische Merkmale als Input für maschinelle Klassifikatoren verwendet. Dellaert et al. (1996)

z.B. erreichten damit eine durchschnittliche Fehlerrate von lediglich 20,5 %[1].

4.3.2 Stimmqualitative Parameter

Andererseits gab es auch bei fast allen Arbeiten Hinweise auf die Prominenz stimmqualitativer Merkmale. Einige Beispiele illustrieren diese Tatsache:

- Bei Bergmann et al. (1988) etwa führte die Manipulation der prosodischen Merkmale Intensität, Dauer und F_0-Kontur bei einem wütend gesprochenen Satz nicht zu einer gänzlichen Abschwächung der ursprünglichen Emotion. Dies weist auf die Bedeutung anderer Merkmale hin.

- Scherer et al. (1984) finden starke Korrelationen für Hörerurteile über vollständige und durch random-splicing bzw. Umdrehung (vgl. Abschnitt 4.5.3) maskierte Stimuli. Auch dies weist auf die Wichtigkeit stimmqualitativer Merkmale hin.

- Eine Synthese mit neutralen Diphonen aber wütenden Prosodieparametern führte lediglich zu 7 % zu einer Beurteilung der Stimuli als wütend (Montero et al. (1998)). Stimuli, die hingegen mittels Diphonen aus den wütenden Originalen, aber neutraler Prosodie generiert wurden, wurden zu 95 % als wütend klassifiziert.

- Bei Carlson et al. (1992) stieg die Erkennungsrate deutlich für Stimuli, bei denen auch das Anregungssignal eines Formantsynthesizers dem natürlichen Vorbild nachempfunden war.

- Klasmeyer (1997) stellte in einem Syntheseexperiment fest, dass Hörer stimmqualitative Manipulationen ebenso stark wahrnehmen wie etwa eine Änderung der F_0-Lage.

Obschon also die Wichtigkeit stimmqualitativer Merkmale für emotionalen Sprechausdruck als weitgehend untermauert gelten kann, fehlt es noch an Untersuchungen, die den Bezug zwischen den Phonationsarten (im Sinne von Laver (1980)) und emotionalen Zuständen zum Gegenstand haben. Fast alle Untersuchungen haben stimmqualitative Parameter lediglich indirekt durch Messung der Energie in spektralen Bändern (wie z.B. in Banse & Scherer (1996)) betrachtet. Sehr selten wurden etwa die die Öffnungsphasen der Glottis gemessen (z.B. Klasmeyer (1999) oder)). Eine Untersuchung, bei der sämtliche Parameter eines Glottissignalmodells für emotionale Sprechweise berücksichtigt wurden (wie etwa für die Phonationsarten bei Ni Chasaide & Gobl (1997)), ist dem Autor nicht bekannt. Der Grund dafür liegt sicherlich darin, dass diese Messungen nicht (oder nur teilweise) automatisiert durchzuführen sind. Sie sind deshalb äußerst zeitaufwendig und können jeweils nur für wenige Beispieldaten durchgeführt werden (vgl. Ni Chasaide & Gobl (1997)).

[1] Allerdings bei Sprecherabhängigkeit.

4.3.3 Artikulatorische Parameter

Diese Parameterklasse wurde bislang noch am seltensten untersucht. Die Ursache liegt sicherlich darin, dass akustische Korrelate artikulatorischer Settings sehr komplex sind. Weiterhin wurden viele der Analysen von Psychologen durchgeführt, denen das hierfür benötigte spezielle phonetische Wissen nicht ausreichend zur Verfügung stand. Vorgeschlagen wurde etwa das *Center of Gravity*, das die Stärke konsonantischer Reduktion angibt oder der Lautminderungsquotient (vgl. Abschnitt 2.2.3), mit dessen Hilfe ebenfalls Reduktions- oder Elaborationsphänomene parametrisierbar sind (vgl. etwa Kienast et al. (1999)). Weiterhin lässt sich der artikulatorische Aufwand bei der Vokalrealisation aufgrund der Zentralisierung oder Dezentralisierung der Formantlagen überprüfen. Da diese Phänomene stark durch generelle Muskelspannungszustände beeinflusst sind, lassen sich Vorhersagen treffen, die mit dem Erregungsgrad einer Emotion zusammenhängen.

4.4 Physiologische Zusammenhänge

Emotionale Zustände führen zu Erregungen bestimmter Teile des vegetativen Nervensystems und haben so Einfluss auf die für die Spracherzeugung wichtigen Muskeln und Gewebeteile der Lungen, des Kehlkopfs und des Artikulationstrakts. Darin liegt ja gerade zum Teil die Ursache dafür, dass sich emotionale Sprecherzustände im akustischen Sprachsignal manifestieren (vgl. die Push-Faktoren, Abschnitt 4.4, bzw. die Component Process Theorie (Scherer (1986)), Abschnitt 1.4).

Das vegetative Nervensystem unterteilt sich in die Bereiche Sympathikus und Parasympathikus, deren Erregung jeweils kontradiktorische Einflüsse auf den Körperzustand hat. Durch eine Erregung des Sympathikus werden der Herzschlag und die Atmung beschleunigt und die Sekretbildung verringert, während bei einer Erregung des Parasympathikus diese verlangsamt werden und die Sekretbildung der Speicheldrüsen verstärkt wird (vgl. Williams & Stevens (1981) oder Trojan (1975)). Da sich zudem bei sympathischer Aktivation der Muskeltonus der quergestreiften Muskulatur erhöht, ist eine ständige leichte Anspannung der Kehlkopfmuskulatur zu erwarten.

Aus diesen Überlegungen lassen sich Hypothesen bezüglich akustischer Korrelate von emotionalen Zuständen ableiten. Wird der Sympathikus erregt, spricht Trojan (1975) von einer *Kraftstimme*, wird der Parasympathikus angeregt, von der *Schonstimme*. Die Erste tritt bei Emotionen mit hoher Erregung wie Ärger oder Freude auf, die Zweite bei solchen mit niedriger Erregung wie Trauer oder Wohlbehagen. Bei freudiger Erregung wäre laut Trojan aufgrund der Kraftstimme eine Verlängerung der Konsonanten, stärkere höhere Formanten bei den Vokalen, ein breiter F_0-Range, größere Unterschiede zwischen betonten und unbetonten Silben und ein schnelleres Tempo zu erwarten (vgl. auch Williams & Stevens (1981, 1972)).

Laut Frick (1985) hat eine Erregung des Sympathikus bzw. äquivalent dazu ein emotionaler Zustand mit starker Erregung eine angehobene Grundfrequenz, höhere F_0-Maxima, einen breiteren F_0-Range und eine höhere F_0-Variabilität zur Folge. Desweiteren seien sowohl die durchschnittliche Lautheit als auch die Lautheitsmaxima erhöht sowie die Artikulationsrate beschleunigt. Allerdings gelte die Erhöhung der Sprechgeschwindigkeit nur im Vergleich zu emo-

tionaler Sprechweise bei niedrigerem Erregungsgrad. Neutrale Sprechweise ist Frick zu Folge generell schneller gesprochen als emotionale.

Einen weiteren physiologischen Zusammenhang sieht Trojan (1975, S. 70) darin, dass im allgemeinen positive Emotionen mit einem Zustand der Rachenweite verbunden sind, während bei negativen der Rachen verengt ist. Dies wird entwicklungsgeschichtlich aus den Vorgängen der Nahrungsaufnahme und dem Brechreiz hergeleitet. Rachenweite drückt sich akustisch durch eine Verminderung der Rauschanteile aus, da durch die geringere Verengung weniger Verwirbelungen der Luft entstehen. Ein anderer Aspekt physiologischer Zusammenhänge ergibt sich daraus, dass emotionale Zustände mit mimischem Ausdruck einhergehen. So wird freudige Sprache oft mit lächelndem Gesichtsausdruck erzeugt, was eine Verkürzung des Ansatzrohrs zur Folge hat.

Scherer (1995) schlägt bezüglich physiologischer Korrelate eine Unterscheidung zwischen *push*- und *pull*-Faktoren vor. Push-Faktoren sind solche, die von den oben beschriebenen Erregungen des vegetativen Nervensystems herrühren. Sie sind in der Regel vom Sprecher nicht zu kontrollieren. Pull-Faktoren entsprechen externen Vorgaben, die Ausdruck der sozialen Norm sind oder sonstige körperunabhängige Faktoren. Sie sind meist von der Intention des Sprechers motiviert, beim Empfänger einen möglichst günstigen Eindruck zu erwecken[2].

4.5 Zur Methodik der Datengewinnung

Im folgenden Abschnitt werden unterschiedliche Aspekte bzgl. der Methodik der Datenaquisition besprochen, in denen sich die berücksichtigten Studien unterscheiden. Konsequenzen daraus werden im experimentellen Teil dieser Arbeit gezogen. Dabei wird zunächst die Problematik diskutiert, emotional gesprochene Sprache aufzunehmen. Ein weiterer Abschnitt behandelt den semantischen Gehalt der Äußerungen. Weiterhin werden verschiedene Maskierungsverfahren erläutert, die bei Resyntheseexperimenten angewendet wurden. Abschließend werden Überlegungen zur Problematik des Testdesigns bei Perzeptionsexperimenten angestellt.

4.5.1 Gewinnung des Sprachmaterials: Schauspieler vs. 'echte Gefühle'

Zur Erforschung der akustischen Korrelate emotionaler Sprechweise werden Sprachäußerungen benötigt, die in ihrer Merkmalsausprägung exemplarisch für den Ausdruck bestimmter Emotionen stehen. Es ergibt sich dabei die Anforderung, dass

- die Äußerungen für alle Emotionen den gleichen Inhalt haben, da die Parameter stark durch linguistische Anforderungen festgelegt sind.
- die Anzahl der Sprecher ausreichend groß ist, um über sprecherspezifische Eigenarten hinweg generalisieren zu können.
- alle Sprecher möglichst dieselbe Emotion ausdrücken.

[2] Beispielsweise wird bei aggressiver Sprechweise die Grundfrequenz oft abgesenkt, um größer zu wirken.

- die Aufnahmen qualitativ sehr hochwertig sind, um komplexere Messungen durchführen zu können.

Aus diesen Forderungen verbietet sich der Gebrauch von Aufnahmen, die in natürlicher Umgebung entstehen. Ein weiteres Argument gegen 'echtes' Sprachmaterial ist die Tatsache, dass reine Basisemotionen nur sehr selten vorkommen, in der Regel werden gemischte Emotionen ausgedrückt. Banse & Scherer (1996) weisen weiterhin darauf hin, dass auch in echten Situationen emotionaler Ausdruck nicht ungefiltert zu beobachten ist, da jeder Mensch ständig schauspielert (in dem Sinne, dass er ständig durch soziokulturelle Konventionen beeinflusst wird und das Bedürfnis besteht, möglichst attraktiv zu wirken). Dies gilt für Kommunikationssituationen besonders stark. Zudem ist der Sprechakt ein Vorgang, der in hohem Maße feinmotorische Steuerung verlangt und daher, sieht man einmal von Interjektionen ab, großer Kontrolle unterliegt. Aus diesen Gründen werden meist Schauspieler engagiert, die den emotionalen Zustand simulieren. Dabei werden entweder lediglich Instruktionen gegeben, den Trägersatz in der gewünschten Emotion zu realisieren oder die Trägersätze werden in kurze Szenen eingebettet (z.B, Williams & Stevens (1972)). Der Vorteil der zweiten Methode besteht darin, dass die realisierten Emotionen für alle Schauspieler einheitlicher sind, da als Rahmen jeweils dieselbe Situation zugrunde liegt. Da aber andererseits jeder Mensch auf bestimmte Situationen anders reagiert, ist dieser Vorteil eventuell doch nicht gegeben und die einfache Beschreibung durch eine Begrifflichkeit vorzuziehen. Eher selten gibt es Aufnahmen mit induzierten Emotionen. Die Schwierigkeit besteht darin (vgl. Banse & Scherer (1996)), dass

- solche Experimente ethisch fragwürdig sind.
- der emotionale Ausdruck eher schwach ist.
- es keine einheitlichen Emotionen gibt, sondern verschiedene Ausdrücke von Emotionsfamilien.

Induzierter emotionaler Sprechausdruck wurde z.B. in Johnstone & Scherer (1999) untersucht. Hier wurden Studenten bei speziell vorbereiteten Computerspielen Frustrations- und Erfolgssituationen ausgesetzt.

Rückt man von der Forderung ab, gleiches Sprachmaterial für mehrere verschiedene Emotionen untersuchen zu wollen, können natürlich auch Aufnahmen aus Rundfunk oder Fernsehen analysiert werden. Eine mehrfach analysierte Aufnahme ist z.B. der Bericht des Radiosprechers zum Absturz der Hindenburg. Williams & Stevens (1972) führten einen Vergleich dieser Aufnahme mit Sprachmaterial durch, bei dem Schauspieler Angst simulierten. Dabei zeigte sich eine hohe Korrelation der akustischen Merkmale beider Aufnahmen. Weiterhin verglichen sie Feldaufnahmen mit denen der Schauspieler und kamen zu ähnlichen Ergebnissen.

In jedem Fall sollte das gewonnene Sprachmaterial vor der Analyse durch einen Perzeptionstest evaluiert werden, um die externe Validität zu gewährleisten. Dieser wurde auch bei fast allen Studien durchgeführt. Eine interessante Alternative stellt die Untersuchung von Banse & Scherer (1996) dar. Dort wurden für von Schauspielern simulierte Sprachaufnahmen die Ergebnisse des Perzeptionstests mit der Klassifizierungsleistung von statistischen Lernverfahren verglichen.

4.5.2 Wahl der Testsätze

Bei Perzeptionstests besteht grundsätzlich das Problem, dass die Hörer sich beim Beurteilen emotional gesprochener Sätze vom semantischen Gehalt leiten lassen. Es können vier Arten von Textsorten unterschieden werden (vgl. Murray & Arnott (1993)):

- Sinnleere Texte (wie etwa Alphabet oder Phantasiesilben) haben keinen semantischen Gehalt (z.B. Banse & Scherer (1996)). Sie eignen sich besonders bei sprachübergreifenden Untersuchungen (z.B. Scherer et al. (2000)).

- Emotional neutrale Texte sind solche Sätze der Alltagskommunikation, denen keine emotionale Bedeutung inhärent ist, z.B. " *Der Koffer liegt auf dem Tisch.* ". Der Satz könnte zwar emotional geäußert werden, aber aus einer Stimmung heraus; d.h. die Emotion bezieht sich nicht auf den Inhalt des Satzes.

- Als emotional unbestimmte Texte werden solche bezeichnet, bei denen zwar ein emotionaler Ausdruck im Hinblick auf die Bedeutung der Äußerung plausibel erscheint, dieser aber nicht auf eine bestimmte Emotion festgelegt ist. Oft sind dies Äußerungen (z.B. "*In sieben Stunden wird es soweit sein.*"), die auf ein Ereignis hinweisen, welches sowohl ein ärgerliches, trauriges, freudiges oder langweiliges sein kann.

- Emotionale Texte sind solche, deren Sinngehalt gerade mit einer bestimmten Zielemotion in Verbindung gebracht werden soll. Diese Art von Texten ist relativ selten zur Evaluierung von synthetisiertem Sprecherausdruck eingesetzt worden (z.B. Rank (1999) oder Murray & Arnott (1995)). Ein Beispiel wäre: "*Fahre doch nicht so schnell.*" für Angst (aus Rank (1999)).

Ein Beispiel für die Bedeutung des Sinngehalts der Testsätze wird bei Bergmann et al. (1988) erwähnt. Die Variable Akzenthöhe wurde für 'Freude' bei dem Testsatz "*Wussten Sie schon, er ist doch gekommen.*" signifikant, für den Testsatz "*Aber schriftlich habe ich das nicht bekommen.*" jedoch nicht. Eventuell wirkt der zweite Satz eben generell unfreudig. Auch bei Burkhardt (1999, S. 64) ergaben sich bei gleichen Parametermanipulationen unterschiedliche Bewertungen von Sätzen, die vermutlich auf den semantischen Gehalt zurückgeführt werden können. Murray & Arnott (1995) berücksichtigten diese Problematik, indem sie zwei Hörexperimente durchführten, einmal mit emotional neutralen Stimuli und einmal mit semantisch emotionalen Äußerungen. Es ist nicht sehr überraschend, dass die Erkennungsrate für die Äußerungen, denen bereits semantisch eine bestimme Emotion zugeordnet werden konnte, sehr viel höher lag als bei den emotional neutralen.

4.5.3 Maskierungsverfahren

Als Resyntheseverfahren werden solche angesehen, bei denen natürlichsprachige Stimuli durch technische Verfahren manipuliert oder durch Modellbildung kopiert werden. Dies geschieht jeweils mit dem Ziel, bestimmte Aspekte des Sprachsignals gesondert manipulieren zu können. Diese Verfahren sind aufgrund der Wichtigkeit für diese Arbeit in einem gesonderten Kapitel

beschrieben (siehe Abschnitt 3.5). Als Vorstufe davon kann die Maskierung angesehen werden, bei der bestimmte Signalparameter zerstört werden, um andere unabhängig davon untersuchen zu können. In der Literatur wurden diverse Maskierungsverfahren beschrieben (vgl. Scherer et al. (1984)):

- **Tiefpassfilterung** bei einer Grenzfrequenz etwas oberhalb der Grundfrequenz dient der Maskierung stimmqualitativer Merkmale. Es hat sich allerdings gezeigt, dass dies nur zu einem gewissen Grad der Fall ist (vgl. Murray & Arnott (1993)).

- **Random Splicing** Hierbei wird das Sprachsignal in Abschnitte unterteilt, die etwa drei Phone umfassen. Diese werden in zufälliger Reihenfolge neu zusammengesetzt. Dadurch wird sowohl der Inhalt als auch die F_0-Kontur zerstört. Der Nachteil besteht darin, dass zufällig neue Wörter bzw. eine andere F_0-Kontur entstehen können. zudem ist der Klang der entstehenden Signale sehr eigentümlich.

- **Rückwärts abspielen** Diese Methode dient ebenfalls der Maskierung der F_0-Kontur und der Verständlichkeit unter Beibehaltung stimmqualitativer und globaler Merkmale wie F_0-Range oder durchschnittliche Grundfrequenz. Das Problem neu entstehender F_0-Konturen wiegt hier allerdings noch schwerer als beim Random-Splicing.

Alle drei Verfahren wurden bei Scherer et al. (1984) angewendet. Andere Studien, die Maskierungsverfahren verwendet haben, sind in Murray & Arnott (1993) zusammengefasst.

4.5.4 Das Testdesign

Bei Untersuchungen zu emotionaler Sprechweise besteht oft die Notwendigkeit, emotional geäußertes Sprachmaterial von Testhörern beurteilen zu lassen. Bei reinen Analysestudien ist dies zur Sicherung der externen Validität notwendig (vgl. Abschnitt 4.5.1) und bei Synthesestudien dient es gerade der Datengewinnung. Das Design des Perzeptionstests ist also in jedem Fall von großer Bedeutung. Grundsätzlich gibt es drei Arten von Tests (vgl. Murray & Arnott (1993)), die für ein Design zur Beurteilung emotionaler Sprechweise durch Versuchshörer in Frage kommen:

- **forced-choice Tests** sind dadurch gekennzeichnet, dass sie dem Beurteiler eine vorgegebene Antwortmöglichkeit für die Stimuli präsentieren. In der Regel werden Fragebögen verwendet, bei denen bestimmte Antwortkästchen angekreuzt werden können. Oft werden hier auch Ratingskalen benutzt, bei denen die Ausprägung eines Merkmals auf einer festgelegten diskreten Skala angekreuzt werden kann. Banse & Scherer (1996) und Frick (1985) weisen daraufhin, dass der Vergleich mehrerer Emotionen innerhalb eines forced-choice Test nicht zu dem Schluss führen kann, dass die jeweiligen Emotionen erkannt worden sind, sondern lediglich, dass sie voneinander unterscheidbar sind.

- **free-response Tests** zeichnen sich durch eine unbeschränkte Antwortmöglichkeit aus. Sie erfolgt meist in schriftlicher Form. Um die Antworten miteinander vergleichbar zu machen, müssen sie nachträglich kategorisiert werden. Hierbei entsteht eine erhebliche

Schwierigkeit in der Interpretation von free-response Tests, weshalb in der Regel eine freie Antwortmöglichkeit höchstens zusätzlich angeboten wird (z.B. bei Cahn (1990) oder Murray & Arnott (1995)). Zudem erfordert diese Art der Bewertung einen wesentlich höheren Zeitaufwand der Testhörer als die anderen Methoden. Allerdings kann man bei erfolgreicher Auswertung eher davon ausgehen, dass die Emotionen tatsächlich erkannt und nicht lediglich durch Ausschlussverfahren voneinander diskriminiert wurden.

- **Tests mit Gegensatzpaaren** bieten jeweils zwei Stimuli an, zwischen denen die Beurteiler vergleichen. Auch diese Art von Testdesign erfordert einen erheblichen Zeitaufwand und wurde deshalb nur in Tests verwendet, deren Thematik eine eher binäre Fragestellung beinhaltet (z.B. Tartter (1980)).

Weiterhin spielt die Anzahl der Testhörer eine nicht zu unterschätzende Rolle. Je nach statistischem Auswertungsverfahren kann die Zahl der Teilnehmer, die für ein stabiles Ergebnis notwendig ist, variieren. Sie sollte allerdings mindestens 20 betragen. Laut einer Faustregel gilt eine Größe von $N = 30$ als sicheres Maß. Ergebnisse von Studien, bei denen deutlich weniger Perzipienten befragt wurden (wie etwa bei Rank (1999), $N = 9$), sind also in Frage zu stellen.

Von großer Bedeutung ist auch die Reihenfolge, in der die Stimuli dargeboten werden. Alle berücksichtigten Untersuchungen randomisierten die Abfolge des Tests. Wenn aber die randomisierte Reihenfolge der Stimuli für alle Testhörer dieselbe ist (wie es bei vielen Untersuchungen der Fall war), schützt dies nicht davor, dass das Hörerurteil durch den Zeitpunkt und den Kontext, in dem der Stimulus wahrgenommen wird, beeinflusst wird. So zeigte sich etwa bei Montero et al. (1998), dass die Erkennungsrate der zehn letzten Stimuli deutlich über dem Durchschnitt lag. Bei vielen neueren Untersuchungen wurden daher die Hörexperimente an einem Computerterminal durchgeführt, bei dem für jeden Hörer eine eigene Zufallsreihenfolge generiert wurde (z.B. Rank & Pirker (1998) oder Cahn (1990)). Dieses Vorgehen schränkt natürlich die Anzahl der Perzipienten ein, da die Hörversuche dadurch nicht gruppenweise durchzuführen sind. Ein Kompromiss besteht darin, mehrere Bänder mit Zufallsreihenfolgen anzufertigen und diese verschiedenen Gruppen darzubieten. Oft werden die Stimuli auch mehrmals innerhalb eines Tests angeboten, wodurch sich die Anzahl natürlich vervielfacht. Die Dauer des Tests sollte 20 min. nicht überschreiten, da die Konzentration danach stark abnimmt. Das bedeutet, pro Hörtest sollten nicht mehr als etwa 80 Stimuli dargeboten werden.

Ein weiteres Kriterium für die Erkennung dargestellter Emotion ist die Mischung der Stimuli. Im allgemeinen umfasst eine zu evaluierende Datenbank Beispiele mehrerer Sprecher, verschiedener Texte und unterschiedlicher Emotionen. Da solche Datenbanken oft zu umfangreich sind, um sie in einer Session bewerten zu können (speziell wenn die Stimuli mehrmals dargeboten werden), werden die Hörtests auf mehrere Sitzungen verteilt. Wenn nun pro Sitzung etwa alle Äußerungen eines Sprechers dargeboten werden, liegt die zu erwartende Erkennungsrate höher, als wenn eine Session alle Beispiele für einen Satz umfasst, da die Hörer die individuellen Sprechereigenschaften der Darsteller adaptieren und von diesen abstrahieren.

4.6 Besprechung ausgewählter Untersuchungen

Im folgenden Abschnitt werden verschiedene Untersuchungen im Einzelnen besprochen. Ausgewählt wurden Studien, die

- dieselbe Datenbank verwendeten, die auch der vorliegenden Arbeit zugrundeliegt (Kienast et al. (1999) und Paeschke et al. (1999)),
- 'klassische' Untersuchungen darstellen (Williams & Stevens (1972)),
- selten berücksichtigte Parameter untersuchten (Klasmeyer (1999) und Banse & Scherer (1996)),
- vergleichbare Resyntheseexperimente durchführten (Carlson et al. (1992) und Bergmann et al. (1988)),
- oder ebenfalls die Anwendung automatisierter Emotionsregeln bei Text-To-Speech Systemen zum Gegenstand hatten (Montero et al. (1999); Rank (1999); Murray & Arnott (1995) und Cahn (1990)).

Dabei werden lediglich methodische Aspekte besprochen; die Ergebnisse sind in der darauf folgenden Zusammenfassung enthalten. Der Abschnitt gliedert sich in drei Teile. Die größte Gruppe bilden Studien, die emotionales Sprachmaterial analysiert haben. Die zweite Gruppe bilden Untersuchungen, die Resyntheseverfahren eingesetzt haben, um neutrale Stimuli systematisch zu variieren (Analysis-by-Synthesis). In der dritten Gruppe werden solche Arbeiten besprochen, die durch die Anwendung von 'Emotionsregeln' den Output von Text-To-Speech Systemen emotionalisiert haben. Zu nennen wären hier auch Untersuchungen, die die automatische Emotionserkennung zum Gegenstand haben. Da diese jedoch in aller Regel mit statistischen Klassifikationsverfahren arbeiten und somit nicht zu einer regelhaft beschriebenen Darstellung emotionalen Sprechausdrucks beitragen, werden sie eher am Rande erwähnt.

4.6.1 Untersuchungen zur Analyse emotionalen Sprachmaterials

Kienast, Paeschke, Burkhardt und Sendlmeier. Die Arbeiten von Kienast et al. (1999); Paeschke et al. (1999); Burkhardt & Sendlmeier (1999b) beruhen alle auf derselben Datenbank, die im Rahmen eines DFG-Projektes zu Reduktionen und Elaborationen bei emotionaler Sprechweise (Kennziffer Se 462/3-1) an der TU-Berlin entstanden ist. Aus dieser Datenbank stammen auch die Äußerungen, die im experimentellen Teil dieser Arbeit zu Resynthesezwecken verwendet wurden. Dafür simulierten 10 Schauspieler (5 Frauen und 5 Männer) emotionale Sprechweise. Das Textmaterial umfasst fünf einphrasige und fünf zweiphrasige Sätze, die jeweils in den Emotionen Freude, Trauer, Angst, Langeweile, Ekel und Ärger realisiert wurden. Als Referenz wurde zusätzlich eine neutrale Variante erzeugt. Da einige der Sätze in mehreren Versionen geäußert wurden, ergibt sich daraus eine Sprachkorpus von etwa 800 Sätzen.

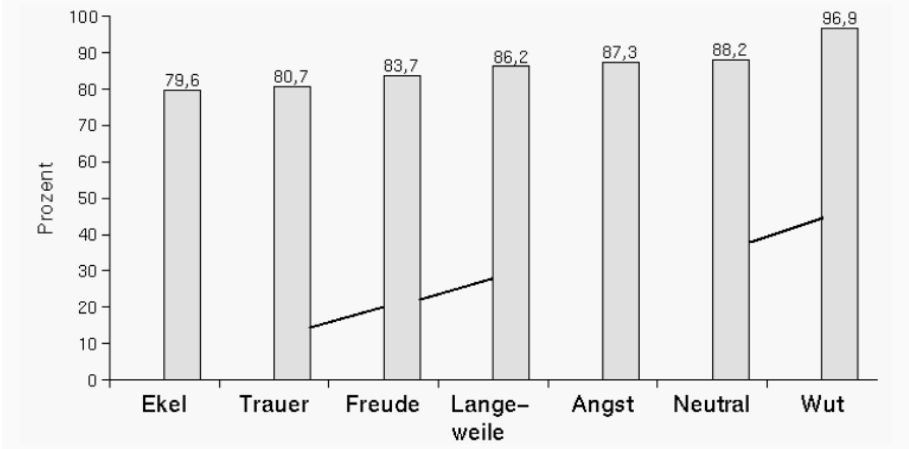

Abbildung 4.1: Durchschnittliche Erkennung der Emotionen bei der emotionalen Datenbank der TU-Berlin. Ein Verbindungsstrich zwischen den Balken steht für eine Signifikanz bzgl. des Unterschieds (siehe Text).

Die Aufnahmen wurden mit einem Sennheiser MKH 40 P48 Mikrofon und einem Tascam DA P1 portablen DAT-Rekorder in dem reflexionsarmen Raum der Technischen Akustik, TU-Berlin, durchgeführt. Darüberhinaus wurden bei den Aufnahmen Elektroglottogramme mit dem portablen Gerät von Laryngograph Ltd. angefertigt. Diese dienen einerseits als stabilere Basis für Grundfrequenzdetektoren als Oszillogramme und ermöglichen weiterhin Rückschlüsse auf die Form des Glottissignals.

Die Datenbank wurde in einem automatisierten Hörtest von 20 bis 30 Hörern bezüglich der Wiedererkennung überprüft. Zusätzlich sollten die Hörer beurteilen, wie überzeugend der emotionale Ausdruck in den Äußerungen enkodiert ist. Die durchschnittlichen Erkennungsraten für die einzelnen Emotionen sind in Abb. 4.1 dargestellt. Die Verbindungsstriche zwischen den Balken stehen dabei für einen signifikanten Unterschied (auf dem 5 % Niveau, Auswertung durch Varianzanalyse mit kompletter Messwiederholung). Die wütend intendierten Äußerungen wurden deutlich am häufigsten erkannt, obwohl die Schauspieler explizit angewiesen wurden, den Ausdruck von Wut nicht einfach durch Schreien zu simulieren.

Zur weiteren Analyse wurden jeweils nur Sätze verwendet, die von mind. 80 % der Hörer erkannt wurden. In der Dissertation von Frau Kienast (noch nicht erschienen) werden vor allem artikulatorische Phänomene wie Assimilationen, Elisionen und Elaborationen untersucht. Viele der im experimentellen Teil überprüften Hypothesen stammen aus dieser Arbeit. Die Dissertation von Frau Paeschke (noch nicht erschienen) hat vor allem intonatorische Merkmale emotionaler Sprechweise zum Gegenstand. Auch aus dieser Arbeit stammt ein Großteil der Anregungen, welche divergierenden Intonationsparameter auf perzeptive Relevanz hin zu überprüfen sind.

Klasmeyer und Sendlmeier. Die in Klasmeyer (1999); Klasmeyer & Sendlmeier (2000, 1997, 1995) beschriebenen Untersuchungen beruhen ebenfalls auf der Analyse eines einzigen Sprachkorpus. Dabei wurden die akustischen Korrelate von sechs Basisemotionen (Freude,

Angst, Wut, Trauer, Ekel und Langeweile) untersucht. Der Untersuchungsschwerpunkt lag dabei auf stimmqualitativen Merkmalen. Die Tatsache, dass diese bis dahin sehr selten berücksichtigt wurden, macht diese Arbeit besonders wertvoll. Zunächst wurde eine emotionale Datenbank erstellt. Dafür simulierten drei Schauspielschüler 10 kurze Sätze in sechs Emotionen sowie eine neutrale Version. Die so gewonnenen Sprachdaten wurden in einem Hörtest auf Natürlichkeit und Erkennbarkeit überprüft.

Die Datenbank wurde hinsichtlich prosodischer, stimmqualitativer und artikulatorischer Parameter analysiert. Zur prosodischen Analyse wurden vor allem globale F_0-, Dauer- und Intensitätsparameter verwendet. Artikulatorische Messungen wurden durch Betrachtung der Formanttransitionen vorgenommen. Rückschlüsse auf die Phonationsarten wurde vor allem durch Bestimmung des *closed-glottis* Intervalls im invers gefilterten Signal und durch Messungen der Signalintensität in verschiedenen Frequenzbändern gezogen. Bedauerlicherweise wurden die invers gefilterten Signale nicht mit üblichen Modellen wie etwa dem LF-Modell (vgl. Abschnitt 3.5.3) parametrisiert. Die Ergebnisse einer solchen Parametrisierung können direkt in Formantsynthesizern umgesetzt werden, da die Glottissignal-Modelle in der Regel (und auch bei dem in der vorliegenden Arbeit verwendeten) auf eben diesen Parametern basieren. Die Autoren weisen auf die Schwierigkeit hin, bei emotionaler Sprechweise eine inverse Filterung durchzuführen. So ist z.B. bei behauchter Phonation, wie sie oft bei trauriger Sprechweise auftritt, eine Unabhängigkeit von Quelle und Filter nicht mehr gegeben.

Banse und Scherer. Banse & Scherer (1996) stellten eine sehr umfangreiche Studie vor, bei der sowohl Analysen einer eigens erstellten emotionalen Sprachdatenbank als auch Ergebnisse früherer Untersuchungen berücksichtigt wurden (z.B. Scherer (1989)). Ungewöhnlich war dabei auch die hohe Anzahl an berücksichtigten Emotionen: es wurden (ausgehend vom dimensionalen Klassifikationsansatz) sechs Qualitäten mit jeweils hoher und niedriger Aktivierung berücksichtigt, also insgesamt 12 Emotionskategorien.

Als Ausgangsmaterial diente eine Sprachdatenbank, für die 12 Schauspieler zweimal zwei sinnleere Sätze in zwei standardisierten Situationen unter 14 verschiedenen emotionalen Zuständen[3] realisierten. Da die Schauspieler sehr unterschiedlich gut erkannt wurden, wurden für die spätere Analyse jedoch lediglich die besten zwei Äußerungen pro Situation und Emotion verwendet. Die Sätzen wurden hinsichtlich 29 akustischer Parameter analysiert. Diese wurden in fünf Gruppen unterteilt: Energieparameter, Parameter der Grundfrequenz, Dauerparameter und Energieverteilung in Frequenzbändern von Langzeitspektren, unterteilt nach stimmhaften und stimmlosen Abschnitten.

Um herauszufinden, welche der 29 Parameter für die Klassifikation der Emotionen mit statistischen Lernverfahren am besten geeignet sind, wurde einmal ein evolutionäres Verfahren und einmal eine Diskriminanzanalyse angewendet. Dabei zeigte sich, dass durch eine Untergruppe von 16 Parametern eine durchschnittliche Erkennungsrate von etwa 40 % erreicht werden kann. Erstaunlicherweise wurde Ekel durch die Diskriminanzanalyse zu 50 % korrekt klassifiziert, während menschliche Beurteiler eine Erkennungsrate von lediglich 15 % für diese Emotion aufwiesen. Ähnliche Ergebnisse gab es für starke Freude (*elation*): 38 % der Stimuli wurden

[3] Für die Sprachaufnahmen wurden zwei weitere Emotionen berücksichtigt.

von den Beurteilern erkannt, aber 63 bzw. 69 % wurden von den statistischen Verfahren korrekt klassifiziert. Dies deutet einerseits darauf hin, dass einige der akustischen Parameter perzeptiv nicht relevant sind. Ein anderer Grund liegt darin, dass menschliche Beurteiler emotionale Stimuli nach abstrakten Kategorien bewerten, während die statistischen Verfahren allein auf der Analyse dieser Datenbasis operierten. Andere Emotionen wurden wiederum deutlich besser von den menschlichen Beurteilern als durch die automatischen Verfahren erkannt, etwa Langeweile oder Verachtung. Somit scheinen bei der Auswahl der gemessenen Parameter auch nicht alle perzeptiv relevanten enthalten gewesen zu sein.

Williams und Stevens. In Williams & Stevens (1972) werden die Ergebnisse dreier vorangegangener Untersuchungen zusammengefasst. Dabei wurden Aufnahmen von vier Schauspielern analysiert, die in Szenen eingebettet waren. Zusätzlich wurden Feldaufnahmen verwendet. Im Rahmen dieser Arbeit wurde auch ein Vergleich zwischen den simulierten Emotionen und der Aufnahme des Radiosprechers beim Hindenburg-Unglück durchgeführt. Es zeigte sich, dass beide Aufnahmen sehr ähnliche Parameterausprägungen aufwiesen. Dies ist eine weitere Legitimation für die Verwendung von Daten, die von Schauspielern simuliert wurden. Untersuchungsgegenstand waren die akustischen Merkmale von Ärger, Angst und Kummer (*sorrow*). Die Auswahl der untersuchten Parameter waren vor allem durch den Einfluss physiologischer Korrelate emotionaler Zustände bestimmt. Es wurden F_0-Lage und -Range, Sprechrate, Artikulationsgenauigkeit anhand der Formantlagen und Burstintensität gemessen. Ruckschlüsse auf die Phonationsart wurden durch die Interpretation von Spektrogrammen gezogen.

4.6.2 Resyntheseexperimente

Methodisch kommen für die Überprüfung der Relevanz akustischer Merkmale emotionaler Sprechweise durch Resyntheseverfahren zwei Möglichkeiten in Frage, einmal die Manipulation emotional neutraler Stimuli mit dem Ziel, den Eindruck einer bestimmten Emotion zu verstärken und zum anderen die Manipulation emotional gesprochener Stimuli mit dem Ziel, die ursprünglich intendierte Emotion abzuschwächen. Der zweite Ansatz ist sehr viel seltener durchgeführt worden (z.B. Bergmann et al. (1988)).

Bergmann, Goldbeck und Scherer. Bergmann et al. (1988) führten vier Experimente zur Wirkung akustischer Merkmalsvariation auf emotionale Sprechwirkung durch. Im ersten Experiment wurde ein varianzanalytisches Design mit den unabhängigen Variablen F_0-Range, Intensität, F_0-Kontur und F_0-Perturbation angewendet. Ein neutraler Trägersatz wurde systematisch für die genannten Parameter variiert. Dieses geschah durch ein automatisiertes LPC-Resyntheseverfahren. Die Versuchspersonen ordneten die Stimuli jeweils einer von sechs Emotionen zu[4]. Zusätzlich standen zur Beurteilung drei eher auf die Einstellung des Sprechers bezogene Skalen zur Verfügung[5]. Es wurden nur Ergebnisse angegeben, die unter einem Signifikanzniveau von $\alpha = .01$ lagen.

[4] freudig, ärgerlich, ängstlich, verächtlich, gelangweilt und traurig

[5] nachdrücklich, vorwurfsvoll und entgegenkommend

Der F_0-Range wurde nach einem Modell von Ladd (1983) in zwei Stufen (erweitert und reduziert) verändert. Die Intensität wurde ebenfalls einmal angehoben und einmal reduziert. Als Vergleichswerte für laute und leise Sprechweise dienten hierbei Energiemittelwerte von Äußerungen, die vorher unter der Anweisung, laut bzw. leise zu sprechen, aufgenommen wurden. Die F_0-Kontur wurde dadurch manipuliert, dass einmal ein satzinitialer Hauptakzent durch Anheben der jeweils betonten Silbe modelliert wurde und einmal ein satzfinaler. Als vierter Parameter wurde die F_0-Perturbation dahingehend manipuliert, dass alle Trägersätze mit 5 % Perturbation überlagert wurden. Auf diese Weise wurden 36 Stimuli generiert und von 22 Versuchspersonen bewertet.

In einem zweiten Experiment wurde zusätzlich die Akzenthöhe in fünf Varianten verändert, Intensität und F_0-Perturbation wurden hingegen nicht verändert. Zudem wurde ein weiterer Trägersatz von einem anderen Enkodierer hinzugenommen, um sprecherindividuelle Unterschiede abschätzen zu können. Dabei gab es tatsächlich für mehrere Skalen signifikante Unterschiede. Ein Sprecher wurde von den Hörern durchgängig als ärgerlicher und erregter beurteilt als der andere. Als Konsequenz daraus wurden die Stimuli dieses Sprechers nicht weiter verwendet.

Der dritte Versuch bestand darin, dass die Variablen Akzenthöhe, Akzentdauer sowie wiederum F_0-Perturbation variiert wurden. Zudem wurde das Textmaterial variiert. Die Tonhöhe des Hauptakzents wurde reduziert und angehoben, die Dauer des Akzentsilbenkerns wurde in zwei Stufen angehoben und es wurden 2,5 bzw. 5 % Perturbation überlagert. Dies ergab ein varianzanalytisches Design von 27 (3^3) Stimuli.

Im vierten Experiment wurde eine ärgerlich realisierte Äußerung hinsichtlich der Parameter F_0-Range, Dauern der betonten Silben und Intensität in je drei Stufen variiert. Ein interessantes Ergebnis dieses letzten Versuches war, dass es nicht gelungen ist, den Eindruck von Ärger aus dem Sprachsignal zu entfernen. Offensichtlich spielen Parameter der Stimmqualität (die hier abgesehen vom Jitter nicht berücksichtigt wurden) eine wesentliche Rolle zur Attribution emotionaler Sprechweise.

Carlson, Granström und Nord. Carlson et al. (1992) führten Untersuchungen zum Einfluss der Prosodie auf emotionale Eindruckswirkung durch. Ausgangspunkt waren Stimuli aus einem früheren Experiment (Öster & Riesberg (1986)). Zwei Schauspieler hatten hierbei die Emotionen Angst, Traurigkeit, Freude und Neutral in mehreren Sätzen simuliert. Die Sätze waren in einem Hörexperiment mit 29 Teilnehmern auf Erkennbarkeit des emotionalen Ausdrucks hin überprüft worden. Dieses Material wurde zunächst hinsichtlich prosodischer Parameter wie Sprechgeschwindigkeit und durchschnittlicher F_0-Merkmale sowie Artikulationsgenauigkeit analysiert.

Im Anschluss an die Analyse wurden mittels eines Sprachsynthesesystems auf Formantbasis ein Stimulusinventar von 15 Sätzen erstellt. Hierbei wurden zunächst vier Sätze eines Sprechers für die drei Emotionen und Neutral nach der natürlichen Vorlage synthetisiert. Berücksichtigt wurden dabei die Intonationskonturen, die Dauerstruktur, die Formantverläufe und Parameter zur Beschreibung des Glottissignals. Dann wurden den drei emotionalen Sätzen die F_0-Kontur und Dauerwerte aller anderen Emotionen zugeordnet. Die neutrale Version wurde nur entsprechend der Dauerwerte der emotionalen Varianten manipuliert.

Für alle Emotionen wurden diese Stimuli immer noch von 80 % der Versuchshörer als neutral bewertet. Dies weist daraufhin, dass der Parameter Dauer keinen Haupteffekt auf die gewählten Emotionen hat. Wurden alle verfügbaren Parameter kopiert, ergab sich eine sehr unterschiedliche Wiedererkennungsrate für die einzelnen Emotionen. Trauer und Wut wurden von 95 bzw. 80 % der Hörer erkannt, die freudige Version erreichte dagegen eine Wiedererkennungsrate von lediglich 42 %. Freude wurde oft mit Ärger verwechselt, was sich durch einen ähnlichen Erregtheitsgrad der beiden Emotionen erklären lässt. Die Stimuli, die durch Kombination des ärgerlichen Originals mit freudigen, traurigen oder neutralen Prosodiewerten erzeugt waren, wurden oft als traurig klingend beurteilt. Wie auch bei Murray & Arnott (1995) besteht der Verdacht, dass eine Formantsynthese Artefakte erzeugt, die zu trauriger Eindruckswirkung führen.

4.6.3 Erzeugung von Emotionen in Text-to-Speech-Systemen

Cahn. Cahn (1990) stellte ein System zur Generierung von Steuerinstruktionen für Sprachsynthesizer vor, die sie befähigen soll, Emotionen zu simulieren. Die Autorin unterteilt ihr Modell in die vier Kategorien Grundfrequenz, Timing, Stimmqualität und Artikulation. Zu jeder dieser Kategorien gibt es bis zu sieben Parameter, die je nach gewünschter Emotion verschiedene Ausprägungen annehmen können. Die Grenzen zwischen diesen Kategorien sind allerdings nicht absolut. So hat z.B. der Parameter 'Betonungsfrequenz' sowohl Einfluss auf das Timing als auch auf die Grundfrequenzkontur. Der einzige artikulatorische Parameter ist 'Präzision', der den Grad der Elaboration bzw. Reduktion der Lautklassen angibt. Durch den Gebrauch eines kommerziellen Systems konnte dieser Parameter allerdings nicht direkt manipuliert werden. Von daher führt eine Erhöhung der Artikulationspräzision zu einer Ersetzung stimmhafter Konsonanten durch stimmlose. Zudem wird der neutrale Vokal [ə] dort, wo er eine Reduktion darstellt, durch dezentralere Vokale ersetzt. Eine Verringerung der Artikulationspräzision wird durch gegenteiliges Vorgehen erreicht. Die Parameter werden durch Werte zwischen -10 und 10 quantifiziert, wobei der Wert 0 der neutralen Sprechweise entspricht.

Die Auswahl der Parameter und ihrer Ausprägungen erfolgte nach einer Literaturrecherche. Input des sogenannten *Affect Editors* ist eine Äußerung, die in syntaktische und semantische Abschnitte unterteilt ist. Dieses Format wird in der Regel von textgenerierenden Systemen erzeugt. Er stellt damit allerdings nicht eine Erweiterung eines bestehenden Systems dar, sondern lediglich eines hypothetisch möglichen. Der Affect Editor analysiert diese Äußerung nach möglichen Betonungs- und Pausenorten, erzeugt einen Prosodieverlauf und belegt dann die oben genannten Parameter mit approbaten Werten für die gewünschte Emotion. Der Output besteht einerseits aus einem Gemisch von Text und Phonemen und andererseits aus Steueranweisungen, die für den in der Arbeit durchgeführten Evaluierungstest von einem DecTalk-Synthesizer verarbeitet wurden. Auch hier (vgl. Murray & Arnott (1995)) war es nicht möglich, alle Parametermanipulationen mit dem gewählten Synthesesystem zu realisieren.

Zur Evaluierung des Systems wurde ein Hörtest durchgeführt, bei dem Sätze mit den sechs Emotionen Wut, Ekel, Freude, Trauer, Furcht und Überraschung generiert und bewertet wurden. Die Versuchsteilnehmer wurden gebeten, jeden der 30 dargebotenen Stimuli (eine Kombination aus fünf Sätzen mit den sechs Emotionen) mit einer der Emotionen zu klassifizieren.

Zusätzlich konnten sie die Stärke der Emotion, den Grad der Sicherheit der Beurteilung und einen Kommentar angeben. Die Äußerungen waren einphrasige kurze Sätze, die unter allen Emotionen plausibel wirken sollten. Alle Emotionen wurden mit etwa 50 % erkannt, Trauer sogar mit 91 %. Die häufigsten Verwechslungen gab es zwischen Emotionen mit ähnlichem Erregtheitsgrad. Freude wurde am häufigsten mit Überraschung verwechselt. Hier ist allerdings anzumerken, dass die Verwechslungen so häufig waren, dass man eigentlich nicht mehr von einer korrekten Erkennung sprechen kann. Ekel wurde z.B. fast genauso häufig als Ärger klassifiziert wie Ekel. Die Autorin benutzt diesen Umstand allerdings, um ihre Ergebnisse noch zu verbessern, indem sie als 'justierte' Ergebnisse die Summen der Erkennungen mit den am häufigsten verwechselten Urteilen angibt. Weiterhin fällt auf, dass unter den angebotenen Klassifizierungen Neutral nicht vorkam, während die Kategorie 'Überraschung'[6] als Nebenemotion gegenüber den anderen Basisemotionen herausfällt.

Murray und Arnott. Murray & Arnott (1996, 1995) sowie Abadjieva et al. (1993) stellten ein Softwaresystem namens HAMLET[7] vor. Dieses dient zur automatischen Anwendung von 'Gefühlsregeln', um den Output von TTS-Systemen zu emotionalisieren und stellt eine Ergänzung zum kommerziellen TTS-System DecTalk dar. Der eingegebene Text wird von DecTalk phonemisiert, in HAMLET bearbeitet und das Ergebnis dann wieder DecTalk übergeben.

Es werden drei Parameter berücksichtigt: Stimmqualität, F_0-Kontur und Sprechrate. Für diese Parameter wurden acht Variablen definiert und ihnen je nach Emotion unterschiedliche Werte zugewiesen. Die Variablen sind Sprechrate, mittlere Grundfrequenz und Range, Lautheit, Laryngalisierung, Frequenz des 4. Formanten, spektrale Dämpfung und F_0-Endkontur. Zusätzlich zu den Variablen gibt es eine Menge von 11 Prosodieregeln, von denen jeweils eine Untermenge jeder Emotion zugeordnet ist. Die erste Regel besagt zum Beispiel: „Erhöhe die Grundfrequenz der betonten Vokale“[8]. Weitere Regeln betreffen z.B. die Artikulation. Auf welche Art die Variablen modifiziert und welche Regeln angewendet wurden, um bestimmte Emotionen zu simulieren, wurde der Literatur entnommen und in Vorversuchen nachgebessert. Einige Beschränkungen werden HAMLET durch den benutzten Synthesizer auferlegt. So ist es nicht möglich, Intensitätsveränderungen für einzelne Phoneme vorzunehmen oder die Stimmqualitätsparameter innerhalb einer Äußerung zu verändern, ohne Pausen zu provozieren.

Um das System zu evaluieren, wurde ein Test mit 35 Versuchspersonen durchgeführt. Dabei wurden sechs Emotionen simuliert: Wut, Freude, Trauer, Angst, Ekel und Kummer[9]. Es wurden Testsätze mit emotional unbestimmtem Inhalt und solche mit emotionalem Inhalt verwendet, die einmal neutral und einmal mit Simulation emotionaler Erregung synthetisiert wurden. Die Befragung wurde durch einen free-response- und einen forced-choice-Test durchgeführt, bei dem neben den sechs intendierten Emotionen noch 'neutral', 'andere' und sieben weitere Distraktoremotionen zur Wahl gestellt wurden.

Zu den Ergebnissen ist zunächst zu bemerken, dass bereits den emotional unbestimmten, neutral synthetisierten Sätzen von etwa zwei Dritteln der Versuchshörer Emotionen zugeordnet

[6] original: *surprise*

[7] HAMLET: Helpful Automatic Machine for Language and Emotional Talk

[8] Es werden drei Arten von Betonung unterschieden: nebenbetont, hauptbetont und emphatisch.

[9] original: *grief*

wurden, und zwar fast ausschließlich negative[10]. Offensichtlich war bereits die neutrale Syntheseleistung des verwendeten Systems nicht sehr befriedigend. Ein ähnliches Ergebnis gab es auch für die Sätze mit emotionalem Inhalt. Die emotional synthetisierten Sätze mit neutraler Semantik zeigten eine sehr schwache Erkennungsrate. Lediglich Wut und Trauer wurden mehrheitlich erkannt, alle anderen Emotionen wurden verwechselt, am stärksten mit Trauer. Allerdings ist anzumerken, dass bei diesen Versuchen aufgrund der zahlreichen Distraktoremotionen von wirklicher Erkennung (im Gegensatz zu bloßer Diskriminierung) gesprochen werden kann. Freude wurde am häufigsten mit Furcht verwechselt, allerdings streuten die Urteile sehr. Bei den Sätzen mit emotionalem Inhalt war das Ergebnis erwartungsgemäß deutlich besser.

Später (Murray et al. (1996)) wurde das System interessanterweise um die Möglichkeit erweitert, Emotionen als Punkt in einem dreidimensionalen Raum darzustellen (basierend auf Schlosberg (1954)). Dadurch ergibt sich die Möglichkeit, Hypothesen über den Einfluss bestimmter Emotionsdimensionen auf akustische Charakteristika zu überprüfen. Weiterhin können verschieden stark ausgeprägte Emotionen derselben Qualität, bzw. 'Zwischenemotionen' erzeugt werden. Die Ergebnisse scheinen aber nicht sehr erfolgreich gewesen zu sein, da in dem Artikel nicht weiter auf diese Erweiterung eingegangen wird.

Rank und Pirker. Rank (1999) bzw. Rank & Pirker (1998) beschreiben ein System zur Erweiterung eines bestehenden LPC-Halbsilben Konkatenationssynthesizers (VieCtoS, *Vienna Concept-to-Speech Synthesizer*) um emotionalen Sprechausdruck. Die Parametrisierung der Segmente durch LPC-Koeffizienten und ein Residualsignal bietet neben Formantsynthesizern eine alternative Möglichkeit, stimqualitative Merkmale sowie suprasegmentelle Merkmale wie Artikulationsgenauigkeit gezielt zu beeinflussen. Bei reiner Synthese klingen LPC-Synthesizer im Allgemeinen qualitativ besser als Formantsynthesizer, da die Modellierung des Signals weniger radikal ist und von daher mehr Information über das Originalsignal vorliegt[11]. Sollen jedoch die Eigenschaften des Sprachsignals gegenüber den Originalsegmenten umfassend verändert werden, zeigen sich die Schwächen dieser Parametrisierung, die nicht phonetisch sondern vielmehr statistisch motiviert ist. Der Zusammenhang zwischen Residual- und Glottissignal ist sehr indirekt, so dass eine Beeinflussung des Residualsignals mit dem Ziel, der Stimmqualität etwa Behauchung zuzufügen, hier nicht zum Erfolg geführt hat (vgl. Rank (1999)).

Für die Gewinnung von Emotionalisierungsregeln stützen sich die Autoren auf Ergebnisse aus Klasmeyer & Sendlmeier (1995) sowie Cahn (1990). Dabei berücksichtigten die Autoren die Basisemotionen Wut, Trauer, Angst und Ekel. Weshalb Freude, die doch bei fast allen Untersuchungen die einzige qualitativ positive Emotion darstellt, nicht berücksichtigt wurde, geht aus den Artikeln nicht hervor. Unklar bleibt auch, ob mit der englischen Bezeichnung *disgust* tatsächlich 'Ekel' gemeint war, oder 'Empörung', wie an anderen Stellen des Artikels erwähnt. Zur Evaluierung des realisierten Synthesesystems wurden zwei Hörtests durchgeführt. Der erste war ein Test mit Gegensatzpaaren und sollte die grundsätzliche Unterscheidbarkeit der generierten Emotionen überprüfen. Ein Ergebnis dieses Versuchs besteht darin, dass Trauer öfter von den Hörern als unterscheidbar eingestuft wurde als Wut. Der eigentliche Evaluierungsversuch

[10] Angst, Trauer und Ekel

[11] LPC-Verfahren waren ursprünglich lediglich zur Datenkompression entwickelt worden.

bestand darin, fünf jeweils semantisch für die Zielemotionen approbate Sätze mit den Parametern für die Zielemotionen zu generieren und diese von neun Hörern beurteilen zu lassen. Es waren also $5 * 4 = 20$ Stimuli zu bewerten. Dabei wurde eigentlich nur Trauer erkannt (zu rund 80 %). Wut wurde fast ebenso oft mit Ekel verwechselt wie erkannt. Die anderen beiden Emotionen wurden eher verwechselt (Ekel mit Trauer, Angst mit Wut oder Ekel), als erkannt.

Montero et al. Montero et al. (1998, 1999) beschreiben zwei Systeme zur Erweiterung spanischer Text-to-Speech Synthesizer um emotionalen Ausdruck. Beim ersten handelt es sich um das Infovox System (Formantsynthese mit Diphonsegmenten), welches um das GLOVE-Anregungsmodell erweitert wurde (vgl. Granström (1992)). Um Regeln für den emotionalen Ausdruck zu gewinnen, wurde eine emotionale Sprachdatenbank mit einem Sprecher aufgenommen. Diese Datenbank wurde analysiert und Unterschiede zwischen den Emotionen wurden in bis zu 123 Parametern zusammengefasst. Dabei wurden nicht nur prosodische Parameter, sondern auch solche, die vom Anregungssignal herrühren, wie etwa spektrale Dämpfung oder Rauschanteile in höheren Frequenzbereichen, berücksichtigt. Aus der Parameteranalyse wurden dann Regeln extrahiert, die zur Steuerung des Synthesizers benutzt wurden. Insgesamt gesehen, liegen die Ergebnisse der Hörererkennung durchaus im Rahmen anderer Untersuchungen (Neutral : 59 %, Trauer : 83 %, Freude : 47 %, Wut : 42 %).

Die relativ geringe Erkennung der neutralen Stimuli mag darauf zurückzuführen sein, dass zur Beurteilung weiterhin die Kategorie 'andere' zur Verfügung stand. Bemerkenswerterweise lag die Erkennungsrate der letzten 10 Stimuli deutlich über diesem Durchschnitt, bei Freude mit 87 % sogar über der Erkennungsrate der originalen Aufnahmen (75 %). Offensichtlich war für diesen Test eine lange Adaptionszeit für die Erkennung günstig. Dies unterstreicht die Wichtigkeit einer zufällig gemischten Reihenfolge für jeden Hörer.

Ein weiteres Ergebnis aus dieser Testserie besteht in der Prominenz stimmqualitativer Merkmale vor allem für wütenden Sprecherausdruck. Eine Synthese mit neutralen Diphonen aber wütenden Prosodieparametern führte zu einer Beurteilung der Stimuli als wütend von lediglich 7 %. Stimuli, die hingegen mittels Diphonen aus den wütenden Originalen, aber neutraler Prosodie generiert wurden, wurden zu 95 % als wütend klassifiziert. Es geht aus dem Artikel allerdings nicht hervor, inwieweit stimmqualitative Merkmale in den formantparametrisierten Diphonen kodiert waren.

Bei dem zweiten System, beschrieben in Montero et al. (1999), handelt es sich um Zeitbereichssynthese mit Diphonkonkatenation. Es wurde zunächst ein Resyntheseexperiment unternommen, bei dem die Diphone aus Sätzen stammen, die mit der entsprechenden Emotion geäußert wurden. Die prosodischen Merkmale wurden dabei stilisiert. Die Erkennungsraten liegen im Vergleich mit anderen Untersuchungen sehr hoch. Bei Experimenten mit unbeschränkter Synthese zeigte sich jedoch die Schwierigkeit, Diphone aus emotionalen und neutralen Samples miteinander zu vermischen. Die Untersuchung ist noch nicht abgeschlossen.

4.7 Zusammenfassung der Ergebnisse

Im folgenden werden die Ergebnisse aus der Literatur hinsichtlich der akustischen Korrelate emotionaler Sprecherzustände zusammengefasst. Dabei werden zunächst Resultate für die Emotionsdimensionen (vgl. Kap. 1) beschrieben. In den weiteren Abschnitten werden die Ergebnisse für die einzelnen Emotionskategorien zusammengefasst. Diese bilden die Grundlage für die im folgenden Kapitel dargestellten Hypothesen, die der Arbeit zugrunde liegen. Bei den Ergebnissen für die einzelnen Emotionen beziehen sich Vergleiche immer auf die neutrale Sprechweise. Soweit Ergebnisse widersprüchlich oder unklar sind, ist die originale Emotionsbezeichnung (in der Regel in Englisch) vermerkt.

4.7.1 Akustische Korrelate der Emotionsdimensionen

Aktivität

Aktivität geht mit einer Erregung des sympathischen Nervensystems einher (vgl. Abschnitt 4.4), von daher lassen sich alle Vorhersagen, die sich aus dieser Erregung ergeben, auf die Aktivitätsdimension übertragen. Dazu gehören eine angehobene F_0-Lage, ein breiterer F_0-Range, stärkere Intensität, schnellere Sprechweise, eine eher gespannte Phonation und elaborierte Artikulation. Die daraus entstehenden Korrelate machen sich perzeptiv sehr stark bemerkbar, was sich daran zeigt, dass Emotionen mit ähnlichem Aktivitätsgrad eher verwechselt werden als solche mit ähnlicher Valenz oder Potenz (Murray & Arnott (1993)).

Valenz

Davitz (1964) zufolge korrelieren Valenzdimension und Rhythmik miteinander. Positive Emotionen werden danach mit einer regelmäßigeren Rhythmik ausgedrückt als negative. Bergmann et al. (1988) äußern die Vermutung, dass Grundfrequenzperturbationen (Jitter) eher auf negative Emotionen hinweisen. Akustische Korrelate, die sich auf diese Dimension zurückführen lassen, scheinen aber nicht sehr ausgeprägt zu sein.

Potenz

Es gibt widersprüchliche Aussagen, was den Zusammenhang zwischen Potenzdimension und durchschnittlicher Grundfrequenz betrifft (Frick (1985)). Einerseits gibt es die Tendenz bei Säugetieren und Vögeln, die Stimme in aggressiver Stimmung abzusenken, um größer zu wirken, andererseits zeigen viele Untersuchungen ein Korrelation mit höherer Stimmlage. Dies hängt unter anderem mit der erhöhten Erregung des symphatischen Nervensystems zusammen. Weiterhin wurde bei Emotionen mit geringer Potenz wie Depression ein Trend zu tieferer Stimmlage festgestellt (Sedlácek & Sychra (1963)). Eventuell kommen hier auch interkulturelle Unterschiede zum tragen, die von Frick (1985) berücksichtigten Untersuchungen (Ohala (1982) und Scherer & Oshinsky (1977)) stammen aus unterschiedlichen Kulturkreisen.

Weiterhin führt Frick aus, dass hohe Stimmen eher unagressiv wirken. Dies wird durch den Zusammenhang zwischen Kinderstimmen, die durch die kürzeren Stimmlippen höher liegen, und einem geringem Aggressionsapell begründet (nach Morton (1977)).

4.7.2 Ärger, Wut, Zorn

Allgemein Ärger wird in seiner erregten Form in der Regel gut erkannt (vgl. Banse & Scherer (1996, S. 622)).

Dimensionale Einordnung Bei Ärger wird generell zwischen *hot* und *cold anger* unterschieden. In der vorliegenden Untersuchung wurde *hot anger* gewählt (Anweisung an die Schauspieler), wobei bei dem in Kap. 8 beschriebenen Perzeptionstest der Versuch unternommen wird, zwischen diesen beiden Formen zu diskriminieren. Die Valenzdimension ist wohl eher als unangenehm einzuschätzen und bzgl. der Potenzdimension ist Ärger als dominant einzuordnen.

Prosodische Merkmale

- Betonungen:
  Bei Paeschke et al. (1999) hat Ärger (von fünf untersuchten Basisemotionen und neutral) die meisten stark betonten Silben. Sendlmeier (1997) misst einen gegenüber dem F_0-Gipfel vorgezogenen Amplitudengipfel bei den Akzenten.

- Durchschnittliche Silbenlängen:
  Die durchschnittliche Sprechgeschwindigkeit ist eher erhöht (Kienast (1998); Banse & Scherer (1996); Montero et al. (1998); Scherer (1995); Murray & Arnott (1995, 1993); Abadjieva et al. (1993) und Cahn (1990)). Williams & Stevens (1972) dagegen messen eher eine Verlangsamung der Äußerungen, allerdings nicht für alle Sprecher.

- Durchschnittliche Dauern für einzelne Lautklassen:
  Kienast (1998, S. 87) und Sendlmeier (1997) zu Folge sind Vokale eher verlängert, was für wortbetonte[12] Vokale besonders stark der Fall ist. Weiterhin misst Kienast (1998) eine Verkürzung von Frikativen, Nasalen und Plosiven. Bei stimmlosen Plosiven gilt dies vor allem für die Okklusionsphase. Sendlmeier (1997) misst beim [t] eine gegenüber der Okklusionsphase längere Voice-Onset-Time in wort- und satzfinaler Position.

- Sprechpausen:
  Laut Kienast (1998, S. 87) sind Sprechpausen bei ärgerlicher Sprechweise "*selten, sehr kurz und befinden sich meist vor wortbetonten Silben*".

- Durchschnittliche Grundfrequenz und F_0-Range:
  Die durchschnittliche Grundfrequenz steigt stark an (Paeschke & Sendlmeier (2000); Banse & Scherer (1996); Scherer (1995); Murray & Arnott (1995, 1993) und Williams & Stevens (1972)). Die Messungen von Williams & Stevens (1972) ergaben eine Steigerung

[12] Im Gegensatz zu un- und satzbetonten Vokalen.

der F_0 um mindestens 50 %. Abadjieva et al. (1993) messen eine moderate durchschnittliche Grundfrequenz. Cahn (1990) geht allerdings von einer tieferen Grundfrequenz aus.

Der F_0-Range ist breiter (Paeschke & Sendlmeier (2000); Montero et al. (1998); Banse & Scherer (1996); Scherer (1995); Murray & Arnott (1995, 1993); Abadjieva et al. (1993); Cahn (1990) und Williams & Stevens (1972)).

- Grundfrequenzhöhe, -Steigung und Verlaufstyp auf Silbenebene:
  Banse & Scherer (1996) und Scherer (1995) zufolge ist die Variabilität grundsätzlich erhöht. Es soll ein Trend zu fallender Intonation auf Silbenebene geben. Laut Williams & Stevens (1972) sind einige Silben stark betont, die Vokale haben dabei F_0-Maxima. Bergmann et al. (1988) stellen fest, dass eine Anhebung sowie ein Abwärtstrend auf den betonten Silben eher ärgerlich wirkt.

  Nach Paeschke et al. (1999) ist Ärger generell durch ein erhöhtes Vorkommen von Silben mit steigendem und fallendem Verlauf gekennzeichnet. Die Steilheit der Konturen war dabei von allen fünf untersuchten Basisemotionen am größten. Der Unterschied zwischen betonten und unbetonten Silben ist bezüglich der Steilheit nicht sehr hoch. Paeschke & Sendlmeier (2000) messen eine generell größere Steilheit des F_0-Verlaufs bei akzentuierten Silben (sowohl fallend als auch steigend). Murray & Arnott (1995, 1993) und Cahn (1990) gehen von abrupten F_0-Änderungen auf betonten Silben aus.

- Grundfrequenzhöhe, -Steigung und Verlaufstyp auf Phrasenebene:
  Abadjieva et al. (1993) und Cahn (1990) zufolge führt eine stärkere Absenkung der F_0-Kontur am Phrasenende zu stärkerer Eindruckswirkung von Stimuli als ärgerlich. Auch Bergmann et al. (1988) beobachtet eine häufigere Attribuierung von Stimuli mit fallender Kontur als ärgerlich. Dies geht konform mit Montero et al. (1998), die ihren Syntheseexperimenten eine fallende Kontur zugrundelegen.

  Paeschke & Sendlmeier (2000) unterteilen Zorn in zwei Subkategorien, solche mit steigender und solche mit fallender Deklination. In beiden Fällen ist der Unterschied zwischen dem ersten und letzten F_0-Minimum recht groß.

Stimmqualitative Merkmale

- Beschreibung des Anregungssignals:
  Laut Bergmann et al. (1988) führt 5 % Jitter zu steigender Eindruckswirkung von Ärger. Auch Williams & Stevens (1972) stellen (bei betonten Silben) Irregularitäten im F_0-Verlauf fest. Murray & Arnott (1995, 1993) und Abadjieva et al. (1993) legen ihren Syntheseexperimenten eine behauchte Anregung zugrunde. Dies scheint sich allerdings nicht an Lavers Terminologie anzulehnen, da sie es mit einer gespannten Artikulation kombinieren[13]. Nach Laver (1980, S. 132) hingegen ist rauhe Stimme (*harsh voice*) ein Zeichen von Ärger. Diese Art der Anregung ist durch Jitter, Shimmer und eine generelle Anspannung der laryngalen Muskulatur gekennzeichnet.

[13] Laut Laver (1980) kann behauchte Stimme nur bei geringer Muskelspannung auftreten, was einer gespannten Artikulation widerspricht.

- Lage der Formanten:
  Williams & Stevens (1972) messen eine Tendenz zu einer Abschwächung des ersten Formanten bei den betonten Silben. Weiterhin liegt der erste Formant eher höher, was durch eine größere Öffnung des Ansatzrohrs begründet wird. Auch nach Sendlmeier (1997) und Abadjieva et al. (1993) ist die Frequenz des ersten Formanten angehoben.

- Energieverteilung der spektralen Bänder:
  Ärger ist durch erhöhte Energie in den höheren Frequenzbändern gekennzeichnet (Klasmeyer & Sendlmeier (2000); Banse & Scherer (1996); Scherer (1995); Abadjieva et al. (1993) und Cahn (1990)).

Artikulatorische Merkmale

- Allgemein
  Laut Sendlmeier (1997) und Williams & Stevens (1972) ist die gesamte Artikulationsgestik eher elaboriert, wodurch die Konsonanten mit deutlichen Friktionsstellen realisiert werden. Auch Cahn (1990) geht von einer deutlicheren Artikulation aus. Murray & Arnott (1995, 1993) und Carlson et al. (1992) weisen auf eine gespannte Artikulation mit abrupteren Vokaleinsätzen hin.

- De- bzw. Zentralisierung von Vokalen
  Kienast & Sendlmeier (2000) stellen bei den zornigen Sätzen eine Dezentralisierung der Formanten fest (Vowel-Target-Overshoot).

- Artikulatorischer Aufwand bei Konsonanten
  Kienast & Sendlmeier (2000) stellen weiterhin eine höhere spektrale Balance[14] bei stimmlosen Frikativen fest, was auf einen hohen artikulatorischen Aufwand hindeutet.

- Koartikulationseffekte und lautliche Elisionen und Epenthesen
  Laut Kienast & Sendlmeier (2000); Kienast (1998) und Sendlmeier (1997) gibt es bei ärgerlicher Sprechweise wenig Silben- und Lautelisionen. Kienast & Sendlmeier (2000) messen bei vier untersuchten Basisemotionen nur bei Ärger Elaborationen. Auch der Grad und die Anzahl der Assimilationen ist geringer.

4.7.3 Trauer

Allgemein Trauer ist eine äußerst grundlegende Emotion, die bei Syntheseversuchen sehr hohe Erkennungsraten aufweist (z.B. Schröder (1999); Rank & Pirker (1998); Murray & Arnott (1995); Granström (1992) oder Cahn (1990)). Aufgrund der Charakteristik der Kraftlosigkeit, mit der einige Formen von Trauer einhergehen, besteht allerdings eine Gefahr der Verwechslung mit Langeweile.

[14] Die spektrale Balance (*spectral balance*) gibt das Verhältnis der Energie in bestimmten Frequenzbändern zur Gesamtenergie an und dient der Beschreibung lautlicher Reduktion bzw. Elaboration bei stimmlosen Frikativen (vgl. Abschnitt 2.2.2).

Dimensionale Einordnung Murray & Arnott (1995) unterscheiden zwischen *sadness* und *grief*, wobei sich die Emotionen aber (den akustischen Parametern nach zu urteilen) vor allem in ihrer Intensität unterscheiden. Die meisten Arbeiten haben die Form mit niedriger Erregung untersucht. Aufgrund von späteren Ergebnissen wird in der vorliegenden Arbeit zwischen *stiller* und *weinerlicher* Trauer zu unterscheiden sein. Diese Formen unterscheiden sich hauptsächlich dadurch, dass weinerliche Trauer einen starken Erregungsgrad hat. Bezüglich der Valenzdimension ist Trauer in der Regel als eher unangenehm einzuordnen.

Prosodische Merkmale

- Betonungen:
  Laut Paeschke et al. (1999) sinkt die Anzahl der betonten Silben.

- Durchschnittliche Silbenlängen:
  Die Sprechgeschwindigkeit ist verringert (Kienast et al. (1999); Montero et al. (1998); Sendlmeier (1997); Banse & Scherer (1996); Scherer (1995); Murray & Arnott (1995, 1993); Abadjieva et al. (1993) und Cahn (1990)). Williams & Stevens (1972) (*sorrow*) messen eine im Durchschnitt halb so langsame Sprechgeschwindigkeit. Bei Bergmann et al. (1988) führt die Dehnung akzentuierter Silben zu einer Steigerung der Eindruckswirkung als traurig.

- Durchschnittliche Dauern für einzelne Lautklassen:
  Montero et al. (1998) und Carlson et al. (1992) weisen auf eine Verlängerung der Konsonanten hin. Auch Sendlmeier (1997) misst ein stärkere Dehnung der Konsonanten als der Vokale. Besonders die satzbetonten Vokale sowie stimmlose Frikative sind stark verlängert (Kienast (1998, S. 88)). Kienast et al. (1999) messen weiterhin eine Verlängerung der Voice-Onset-Time stimmloser Plosive.

- Sprechpausen:
  Abadjieva et al. (1993); Cahn (1990) und Williams & Stevens (1972) weisen auf ein erhöhtes Sprechpausenvorkommen hin. Sprechpausen treten nicht nur häufiger auf, sondern sind auch von längerer Dauer (Kienast (1998, S. 88)).

- Durchschnittliche Grundfrequenz und F_0-Range:
  Die durchschnittliche Grundfrequenz ist tiefer (Banse & Scherer (1996); Scherer (1995); Murray & Arnott (1995, 1993); Bergmann et al. (1988) und Williams & Stevens (1972)). Abadjieva et al. (1993) allerdings messen keine Verschiebung der Grundfrequenzlage gegenüber neutraler Sprechweise. Der F_0-Range ist schmaler (Paeschke & Sendlmeier (2000); Montero et al. (1998); Banse & Scherer (1996); Scherer (1995); Murray & Arnott (1995, 1993); Abadjieva et al. (1993); Carlson et al. (1992); Cahn (1990) und Williams & Stevens (1972)).

- Grundfrequenzhöhe, -Steigung und Verlaufstyp auf Silbenebene:
  Trauer ist eher durch fallende als durch steigende F_0-Konturen gekennzeichnet (Paeschke et al. (1999); Banse & Scherer (1996); Scherer (1995); Murray & Arnott (1995, 1993)

und Abadjieva et al. (1993)). Die Steilheiten dieser Verläufe sind dabei laut Paeschke et al. (1999) verglichen mit weiteren fünf Basisemotionen nicht sehr groß und es treten überwiegend gerade Silbenverläufe auf, steigende Verläufe sind sehr selten. Cahn (1990) dagegen geht eher von steileren Konturen auf akzentuierten Silben aus. Bergmann et al. (1988) stellen eine erhöhte Eindruckswirkung für Trauer bei abgesenkter akzenttragender Silbe fest.

- Grundfrequenzhöhe, -Steigung und Verlaufstyp auf Phrasenebene:
  Laut Cahn (1990) steigt die F_0-Kontur am Phrasenende eher an. Paeschke & Sendlmeier (2000) stellen fest, dass die typische Absenkung am Äußerungsende geringer ist. Die gesamte Deklination fällt kaum.

Stimmqualitative Merkmale

- Beschreibung des Anregungssignals:
  Williams & Stevens (1972) fällt bei ihren Messungen starker Jitter auf. Cahn (1990) legt für traurige Sprechweise starke Behauchung zugrunde. Auch Kienast (1998, S. 89) stellt eine starke Häufung behauchter Vokale fest. Klasmeyer & Sendlmeier (2000) messen für Trauer eine modal bis behauchte Sprechweise mit häufigen Laryngalisierungen.

- Lage der Formanten:
  Die Lage des ersten Formanten sinkt eher, während die des zweiten steigt (Sendlmeier (1997)), was sich durch eine seltenere Lippenrundung gepaart mit geringerem Öffnungsgrad des Kiefers aufgrund einer generell eingeschränkten Motorik erklären lässt.

- Energieverteilung der spektralen Bänder:
  Laut Banse & Scherer (1996); Scherer (1995) und Cahn (1990) sind die höheren Frequenzbereiche gedämpft. Dies folgert auch aus dem eher behauchten Anregungssignal.

Artikulatorische Merkmale

- Allgemein
  Murray & Arnott (1995, 1993) und Abadjieva et al. (1993) beschreiben die Artikulation als verschleift (*slurring*).

- De- bzw. Zentralisierung von Vokalen
  Kienast & Sendlmeier (2000) und Sendlmeier (1997) stellen eine Zentralisierung der Formanten (Target-Undershoot) fest.

- Artikulatorischer Aufwand bei Konsonanten
  Montero et al. (1998) und Carlson et al. (1992) weisen auf eine Verlängerung der Konsonanten hin. Kienast & Sendlmeier (2000) messen eine geringere spektrale Balance bei stimmlosen Frikativen.

- Koartikulationseffekte und lautliche Elisionen und Epenthesen
  Die Anzahl der segmentellen Reduktionen ist bei trauriger Sprechweise sehr hoch, Assimilationen und Elisionen treten sehr häufig auf (Kienast & Sendlmeier (2000) und Sendlmeier (1997)).

4.7.4 Freude

Allgemein Freude ist in der Regel die einzige positive Emotion bei den Untersuchungen. Sie wird in der Regel eher schlecht erkannt (vgl. Burkhardt & Sendlmeier (1999a); Scherer et al. (2000) und Carlson et al. (1992)) und oft mit Ärger verwechselt. Kienast (1998, S. 88) stellt bei Freude starke sprecherspezifische Unterschiede fest.

Es entsteht der Eindruck, dass Freude durch sehr unterschiedliche Parameterkonstellationen ausgedrückt werden kann (vgl. Burkhardt & Sendlmeier (1999a) und Kienast (1998, S. 88)).

Dimensionale Einordnung Für Freude wird in den meisten Untersuchungen (wie auch in der vorliegenden) von einer deutlichen Emotion mit hohem Erregungsgrad ausgegangen. Die Valenzdimension kann als angenehm bezeichnet werden. Bezüglich der Potenzdimension ist Freude eher dominant einzuschätzen.

Prosodische Merkmale

- Betonungen:
  Freudige Sprechweise ist durch eine erhöhte Frequenz von stark betonten Silben gekennzeichnet (Burkhardt (1999) und Paeschke et al. (1999)). Laut Trojan (1975) vergrößern sich die Unterschiede zwischen unbetonten und betonten Silben. Abadjieva et al. (1993) messen bei freudiger Sprechweise vor allem eine regelmäßige Betonungsstruktur, so dass die Sprache sehr rhythmisch wirkt. Nach Sendlmeier (1997) geht bei freudiger Sprechweise bei den Betonungen der Amplitudengipfel dem F_0-Gipfel voraus.

- Durchschnittliche Silbenlängen:
  Laut Carlson et al. (1992) und Öster & Riesberg (1986) ist freudige Sprechweise durch ein langsames Tempo gekennzeichnet. Banse & Scherer (1996); Scherer (1995); Fonagy (1981) und Davitz (1964) allerdings beschreiben das Tempo als lebendig bzw. weisen auf eine Erhöhung der Artikulationsrate hin. Bei Abadjieva et al. (1993) wies Freude sogar unter fünf berücksichtigten Emotionen die schnellste Sprechgeschwindigkeit auf. Murray & Arnott (1995, 1993) dagegen beschreiben das Tempo als schneller oder langsamer. Offensichtlich ist das Sprechtempo kein zuverlässiger Indikator für freudige Sprechweise, wie auch aus den Untersuchungen in Burkhardt (1999) und Kienast (1998, S. 88) hervorgeht. Da aber die Sprechgeschwindigkeit stark mit der Aktivitätsdimension korreliert, liegt der Verdacht nahe, dass sich Freude mit sehr unterschiedlichen Erregungstufen darstellen lässt. Nach Bergmann et al. (1988) sollten phrasenbetonende Silben von kurzer Dauer sein.

- Durchschnittliche Dauern für einzelne Lautklassen:
  Burkhardt (1999) und Kienast (1998, S. 88) haben bei freudiger Sprechweise eine Verlängerung der stimmhaften Frikative um etwa 20 % gemessen. Nach Sendlmeier (1997) werden die Vokale gedehnt, während die Konsonanten verkürzt werden. Kienast et al. (1999) messen eine Verlängerung der Vokale in wortbetonten Silben. Laut Sendlmeier (1997) sind sowohl Okklusionsphase als auch Voice-Onset-Time beim [t] verkürzt. Darüberhinaus wird der Verschluss bei satzfinaler Position seltener gelöst.

- Durchschnittliche Grundfrequenz und F_0-Range:
  Freude führt zu einem starken Anstieg der Grundfrequenz (Paeschke & Sendlmeier (2000); Burkhardt (1999); Johnstone & Scherer (1999); Banse & Scherer (1996); Murray & Arnott (1995, 1993); Abadjieva et al. (1993); Öster & Riesberg (1986); Tartter (1980) und Trojan (1975)). Cahn (1990) dagegen legt ihren Syntheseexperimenten eine leichte Absenkung zugrunde. Da die durchschnittliche F_0-Lage ebenso wie die Sprechgeschwindigkeit stark mit der Aktivitätsdimension korreliert, wurde Freude hier mit eher niedrigem Erregungsgrad simuliert.

  Freudige Äußerungen sind durch einen breiteren F_0-Range gekennzeichnet (Paeschke & Sendlmeier (2000); Burkhardt (1999); Banse & Scherer (1996); Murray & Arnott (1995, 1993); Abadjieva et al. (1993); Cahn (1990) und Bergmann et al. (1988)).

- Grundfrequenzhöhe, -Steigung und Verlaufstyp auf Silbenebene:
  Laut Scherer (1981) führt Freude zu höherer Variabilität. Der Anteil von Silben mit geradem Verlauf ist laut Burkhardt (1999) und Paeschke et al. (1999) geringer. Burkhardt (1999); Paeschke et al. (1999); Mozziconacci & Hermes (1997) und Murray & Arnott (1995, 1993) haben für Freude mehr ansteigende Konturen auf Silbenebene festgestellt als bei neutraler Sprechweise. Der Anteil fallender Konturen ist auch höher, aber steigende überwiegen. Abadjieva et al. (1993) messen allerdings eher fallende Konturen.

  Die Steilheit der Konturen ist ebenfalls größer (Burkhardt (1999)), wobei nach Paeschke et al. (1999) stark betonte Silben einen weniger steilen Verlauf haben als schwach betonte. Weiterhin fällt auf, dass hier verglichen mit fünf weiteren Basisemotionen die meisten polytonalen Verläufe auftreten.

  Nach Paeschke & Sendlmeier (2000) ist die Steilheit der Akzente (vgl. 2.2.1) (sowohl fallend als auch steigend) generell höher. Murray & Arnott (1995, 1993) zufolge sind die Steigungen eher sanft. Auch Frick (1985) fasst mehrere Untersuchungen dahingehend zusammen, dass sanfte F_0-Konturen vermehrt auftreten. Laut Cahn (1990) sind die Konturen auf akzentuierten Silben sehr steil. Bergmann et al. (1988) und Burkhardt & Sendlmeier (1999a) stellen stärkere freudige Eindruckswirkung bei der Anhebung phrasenbetonender Silben fest.

- Grundfrequenzhöhe, -Steigung und Verlaufstyp auf Phrasenebene:
  Laut Cahn (1990) führt eine leichte Anhebung der F_0 am Phrasenende zu einer freudigeren Eindruckswirkung. Dieser Einschätzung (die Autorin nennt keine expliziten Quellen) steht der Autor aber eher skeptisch gegenüber, da eine phrasenfinale Anhebung der F_0

eher unsicher wirkt[15] (vgl. Paeschke & Sendlmeier (2000)), wohingegen Freude ein emotionaler Zustand ist, der mit Selbstsicherheit einhergeht. Paeschke & Sendlmeier (2000) stellen fest, dass freudige Äußerungen entweder eine stark fallende oder stark steigende Deklination aufweisen.

Stimmqualitative Merkmale

- Beschreibung des Anregungssignals:
  Nach Murray & Arnott (1995, 1993); Abadjieva et al. (1993) und Fonagy & Magdics (1963) weist freudige Sprechweise eine behauchte Phonation auf. Dies steht allerdings im Widerspruch zu physiologischer Erregung. Vermutlich ist dabei eher Wohlbefinden gemeint. Klasmeyer & Sendlmeier (2000) dagegen beschreiben die Phonationsart als modal mit leichter Anspannung, woraus stärkere Energie in den höheren Frequenzbereichen resultiert. Die Autoren bezeichnen diese Anregungsart als 'Modal II'.
  Johnstone & Scherer (1999) messen für Freude im Vergleich mit fünf anderen Emotionen den deutlich höchsten Jitterwert. Allerdings ist der Wert absolut (in Hz) angegeben und somit auch Folge der hohen Grundfrequenz. Laut Johnstone & Scherer (1999) ist die Glottis-Schließungsphase kürzer.

- Lage der Formanten:
  Freudige Sprechweise führt zu einer Anhebung der Formanten, was sich durch ein häufigeres Auftreten gespreizter Lippen bei lächelnder Sprechweise erklären lässt (Kienast & Sendlmeier (2000); Klasmeyer & Sendlmeier (2000); Burkhardt & Sendlmeier (1999a); Schröder et al. (1998); Sendlmeier (1997) und Tartter (1980)). Davon ist besonders der erste Formant betroffen, was auch von der größeren Öffnung des Ansatzrohrs herrührt (Sendlmeier (1997)).

- Energieverteilung der spektralen Bänder:
  Die höheren Frequenzbereiche sind weniger gedämpft (Klasmeyer & Sendlmeier (2000); Banse & Scherer (1996) und Scherer (1995)). Klasmeyer & Sendlmeier (2000) messen bei freudigen Äußerungen eine erhöhte Energie des Frequenzbereichs zwischen drei und fünf kHz. Johnstone & Scherer (1999) messen weniger Energie unter 1 kHz. Auch Abadjieva et al. (1993) fassen ihre Messungen unter dem Begriff *helle Stimme* zusammen, was auf einen verstärkten Anteil höherer Frequenzbereiche hinweist.

[15] Sie tritt ja auch vor allem bei Fragesätzen auf.

Artikulatorische Merkmale

- Allgemein
  Laut Sendlmeier (1997) und Abadjieva et al. (1993) ist die Artikulationsgenauigkeit der Inhaltswörter erhöht.

- Artikulatorischer Aufwand bei Konsonanten
  Kienast & Sendlmeier (2000) messen eine erhöhte spektrale Balance bei stimmlosen Frikativen.

- Koartikulationseffekte und lautliche Elisionen und Epenthesen
  Kienast & Sendlmeier (2000) messen bei freudiger Sprechweise weniger Assimilationen.

4.7.5 Angst

Allgemein Williams & Stevens (1972) hatten bei dieser Emotion (*fear*) Probleme, klare und konsistente Korrelate zu finden. Angst wird in der Regel eher schlecht erkannt (vgl. Banse & Scherer (1996, S. 622)).

Dimensionale Einordnung Angst ist eine Emotion mit starker Erregung, die eindeutig negativ und subdominant einzuordnen ist.

Prosodische Merkmale

- Betonungen:
  Laut Cahn (1990) steigt die Anzahl der betonten Silben. Sendlmeier (1997) misst eine zeitliche Verschiebung von F_0- und Amplitudengipfel bei den Betonungen, der F_0-Gipfel liegt früher.

- Durchschnittliche Silbenlängen:
  Die Sprechgeschwindigkeit ist höher (Kienast (1998); Sendlmeier (1997); Banse & Scherer (1996); Scherer (1995); Murray & Arnott (1995, 1993) und Cahn (1990)). Abadjieva et al. (1993) messen ebenfalls eine schnellere Sprechgeschwindigkeit als bei neutraler Sprechweise, aber langsamer als bei Freude oder Wut. Williams & Stevens (1972) (*fear*) dagegen berichten von einer eher langsameren durchschnittlichen Sprechgeschwindigkeit. Diese Ergebnis mag sich allerdings durch das vermehrte Auftreten von Sprechpausen erklären lassen. Bergmann et al. (1988) beobachten eine Steigerung ängstlicher Eindruckswirkung bei Dehnung der phrasenbetonenden Silbe.

  Nach Paeschke & Sendlmeier (2000) ist die Dauer der Akzente[16] verkürzt.

- Durchschnittliche Dauern für einzelne Lautklassen:
  Laut Kienast et al. (1999) und Sendlmeier (1997) verlängert sich die Dauer der Voice-Onset-Time für stimmlose Plosive.

[16] Ein Akzent ist laut Paeschke & Sendlmeier (2000) eine zusammenhängende Betonungsgruppe, die sich über mehrere Silben erstrecken kann (vgl. Abschnitt 2.2.1).

- Sprechpausen:
  Abadjieva et al. (1993) messen ein erhöhtes Pausenvorkommen.

- Durchschnittliche Grundfrequenz und F_0-Range:
  Die durchschnittliche Grundfrequenz ist sehr stark erhöht (Banse & Scherer (1996); Scherer (1995); Murray & Arnott (1995) und Cahn (1990)). Paeschke & Sendlmeier (2000) messen ebenfalls eine deutlich höhere F_0-Lage, die allerdings unter der von zorniger und freudiger Sprechweise liegt. Abadjieva et al. (1993) messen dagegen eine lediglich leicht angehobene F_0-Lage. Der F_0-Range ist ebenfalls größer (Banse & Scherer (1996); Scherer (1995); Murray & Arnott (1995, 1993); Abadjieva et al. (1993) und Cahn (1990). Williams & Stevens (1972) (*fear*) messen eine durchschnittliche Grundfrequenz, die in etwa der der neutralen Äußerungen entspricht. Allerdings berichten sie von gelegentlichen Peaks in der F_0-Kontur.

- Grundfrequenzhöhe, -Steigung und Verlaufstyp auf Silbenebene:
  Bergmann et al. (1988) stellen fest, dass eine steigende Kontur auf den betonten Silben eher ängstlich wirkt. Paeschke et al. (1999) zufolge gibt es überwiegend gerade Silbenverläufe, der Anteil an fallenden Verläufen ist dabei unter fünf betrachteten Basisemotionen am geringsten. Bezüglich der Steilheit der F_0-Konturen sind Unterschiede zwischen betonten und unbetonten Silben festzustellen. Abadjieva et al. (1993) messen irreguläre fallende Konturen auf der Lautebene.

- Grundfrequenzhöhe, -Steigung und Verlaufstyp auf Phrasenebene:
  Paeschke et al. (1999) stellen bei ängstlicher Sprechweise ein Verbleiben auf dem sehr hohen F_0-Niveau fest, der typische Abfall am Satzende bleibt aus. Die Deklination ist dabei steigend. Cahn (1990) geht sogar von einer Steigung der F_0 am Phrasenende aus. Auch bei Bergmann et al. (1988) führt eine steigende Phrasenkontur zu einer Steigerung ängstlicher Eindruckswirkung.

Stimmqualitative Merkmale

- Beschreibung des Anregungssignals:
  Klasmeyer & Sendlmeier (2000); Abadjieva et al. (1993) und Williams & Stevens (1972) haben bei einigen ihrer Sprecher Jitter festgestellt. Auch Murray & Arnott (1995, 1993) legen bei ihren Syntheseexperimenten eine irreguläre Anregung zugrunde. Kienast (1998, S. 88) stellt bei ängstlicher Sprechweise eine häufige Aspiration bei Plosiven fest.
  Klasmeyer & Sendlmeier (2000) zufolge ist die Phonationsart modal bis flüsternd, mit möglichen Registerwechseln zu Falsett hin.

- Lage der Formanten:
  Sendlmeier (1997) misst eine Vergrößerung des Abstands der ersten beiden Formanten, was vor allem auf eine Anhebung des zweiten Formanten zurückzuführen ist. Er führt diese Effekte auf eine verminderte Motorik bei ängstlicher Sprechweise zurück.

- Energieverteilung der spektralen Bänder:
  Es gibt laut Banse & Scherer (1996); Scherer (1995) und Cahn (1990) mehr Energie in den höheren Frequenzbändern. Abadjieva et al. (1993) messen eine geringere Energie in den tieferen Frequenzbereichen.

Artikulatorische Merkmale

- Allgemein
  Murray & Arnott (1995, 1993) legen bei ihren Syntheseexperimenten eine präzise Artikulation zugrunde, dies basiert vermutlich auf Messungen von Abadjieva et al. (1993).

- De- bzw. Zentralisierung von Vokalen
  Die Formanten sind dagegen laut Kienast & Sendlmeier (2000) eher zentralisiert.

- Artikulatorischer Aufwand bei Konsonanten
  Kienast & Sendlmeier (2000) messen eine größere spektrale Balance für stimmlose Frikative.

- Koartikulationseffekte und lautliche Elisionen und Epenthesen
  Sendlmeier (1997) misst einen negativen Lautminderungsquotienten[17] bei ängstlicher Sprechweise, die Anzahl der artikulierten Segmente gegenüber einer idealen Aussprache hat sich dabei um etwa 4 % erhöht. Kienast & Sendlmeier (2000) dagegen messen eine segmentelle Reduktion, Laut- und Silbenelisionen sowie Assimilationen treten bei ängstlicher Sprechweise sehr häufig auf. Eventuell läßt sich dieser Widerspruch im Hinblick auf die Verlängerung betonter Silben dadurch auflösen, dass für betonte Silben Elaboration zugrunde gelegt wird und ansonsten eher Reduktion.

4.7.6 Langeweile

Allgemein Langeweile zählt nicht zu den k̈lassischen v̈ier Basisemotionen. Daher ist diese Emotion in der Literatur seltener betrachtet worden und es gibt weniger Ergebnisse. Da Langeweile keine Basisemotion ist, kann sie Elemente von Trauer wie von leichtem Ärger beinhalten, wodurch sie leicht verwechselbar ist. Revers (1949) unterscheidet zwischen gegenständlicher (etwas langweilt mich) und zuständlicher (ich bin gelangweilt) Langeweile.

Dimensionale Einordnung Langeweile ist eine Emotion mit niedrigem Erregungsgrad. Die Valenz ist dabei negativ. Bezüglich der Potenzdimension lässt sich Langeweile nicht eindeutig zuordnen, was sicherlich damit zusammenhängt, dass Langeweile nicht zu den Basisemotionen zählt.

[17] Der Lautminderungsquotient berechnet sich durch den Vergleich einer idealen Aussprache mit der Anzahl der tatsächlich geäußerten Lautsegmente (vgl. Abschnitt 2.2.3).

Prosodische Merkmale Entsprechend der eher geringen Erregung, die mit gelangweilter Stimmung einher geht, ist von einer langsameren Sprechweise, niedrigerer Intensität, eher abgesenkter F_0-Lage und schmaleren F_0-Range auszugehen. Dies mag sich ändern, wenn sie in Richtung Ärger geht (genervte Langeweile).

- Betonungen:
  Paeschke et al. (1999) misst bei Langeweile keine Veränderung hinsichtlich der Verteilung betonter, schwach- und unbetonter Silben gegenüber neutraler Sprechweise. Nach Sendlmeier (1997) liegt bei betonten Silben der F_0-Gipfel vor dem Amplitudengipfel.

- Durchschnittliche Silbenlängen:
  Laut Paeschke et al. (1999) unterscheidet sich Langeweile von Trauer vor allem dadurch, dass eine Häufung verlängerter Silben auftritt. Bergmann et al. (1988) beobachten eine Steigerung gelangweilter Eindruckswirkung bei Dehnung der phrasenbetonenden Silbe. Auch Paeschke & Sendlmeier (2000) messen eine längere Dauer der Akzente. Sendlmeier (1997) misst unter fünf betrachteten Emotionen bei den gelangweilten Äußerungen die niedrigste Sprechgeschwindigkeit. Die Gesamtdauer lag dabei ein Viertel über der der neutralen Versionen.

- Durchschnittliche Grundfrequenz und F_0-Range:
  Paeschke et al. (1999) zufolge gibt es bei Langeweile ähnlich wie bei Trauer eine durchschnittlich tiefere Grundfrequenz. Der F_0-Range war ebenfalls geringer, wenn auch nicht ganz so stark wie bei Trauer. Auch nach Bergmann et al. (1988) ist der Range schmaler.

- Grundfrequenzhöhe, -Steigung und Verlaufstyp auf Silbenebene:
  Paeschke et al. (1999) messen bei Langeweile 5 % mehr Silben mit geradem Verlauf und deutlich weniger mit steigendem. Die durchschnittliche Steilheit der Konturverläufe war dabei die geringste von allen betrachteten Emotionen.

- Grundfrequenzhöhe, -Steigung und Verlaufstyp auf Phrasenebene:
  Paeschke & Sendlmeier (2000) unterscheiden zwischen zwei Subkategorien von Langeweile: gähnender und normale. Der Unterschied besteht im wesentlichen darin, dass gähnende Langeweile ein deutlich fallende Deklinationslinie aufweist, während sie ansonsten eher flach ist.

Stimmqualitative Merkmale

- Beschreibung des Anregungssignals:
  Nach Laver (1980, S. 126) ist knarrende Phonation für englische Sprecher ein Zeichen von gelangweilter Resignation. Die Intensität gelangweilter Sprechweise ist Bergmann et al. (1988) zufolge eher schwach.

- Lage der Formanten:
  Sendlmeier (1997) misst eine Vergrößerung des Abstands zwischen den ersten beiden Formanten.

- Energieverteilung der spektralen Bänder:
  Klasmeyer & Sendlmeier (2000) messen weniger Energie in den höheren Frequenzbändern.

Artikulatorische Merkmale

- Allgemein
  Laut Klasmeyer & Sendlmeier (1995) ist die Artikulation ungenau.
- De- bzw. Zentralisierung von Vokalen
  Kienast & Sendlmeier (2000) messen eine Zentralisierung der Formanten
- Artikulatorischer Aufwand bei Konsonanten
  Die spektrale Balance bei stimmlosen Frikativen ist geringer (Kienast & Sendlmeier (2000)). Nach Klasmeyer & Sendlmeier (1995) tendieren stimmlose Frikative in VCV-Verbindungen dazu, stimmhaft zu werden.
- Koartikulationseffekte und lautliche Elisionen und Epenthesen
  Kienast & Sendlmeier (2000) und Sendlmeier (1997) stellen anhand des Lautminderungsquotienten eine beträchtliche artikulatorische Reduktion fest.

4.8 Zusammenfassung und Thesen

Die gerade beschriebenen Ergebnisse aus der Literaturrecherche sind in Tab. 4.1 komprimiert dargestellt. Dabei sind die Ergebnisse so zusammengefasst worden, dass Widersprüche nicht berücksichtigt wurden. Gab es Widersprüche in der Literatur, z.B. in Bezug auf die Sprechgeschwindigkeit bei Ärger, so wurde das Merkmal so beschrieben, wie es die überwiegende Anzahl der Untersuchungen angibt. Unsichere Angaben sind durch ein 'eventuell' annotiert. Leere Tabellenzellen bedeuten, dass für dieses Merkmal bei der betreffenden Emotion keine Aussagen gefunden wurden. Die durch die Literaturrecherche gefundenen Ergebnisse bilden die Hypothesen für die akustischen Korrelate emotionaler Sprechweise und werden im weiteren Verlauf durch Simulation mit Sprachsynthese in Perzeptionsexperimenten überprüft.

> **These 1**: Die in der Literatur beschriebenen akustischen Korrelate emotionaler Sprechweise sind dazu geeignet, emotionalen Sprechausdruck zu simulieren.

Weitere Hypothesen betreffen die Simulierbarkeit emotionaler Sprecherzustände mit allgemeinen Sprachsynthesesystemen durch Anwendung von Modifikationsregeln. Diese Modifikationsregeln sollen die Steueregeln des Synthesesystems zur neutralen Sprachausgabe so modifizieren, dass perzeptiv ein erkennbarer emotionaler Sprechausdruck simuliert wird.

> **These 2**: Es ist möglich, die Korrelate zu operationalisieren und regelhaft zu formulieren, so dass ein emotionaler Sprechausdruck automatisch durch Anwendung dieser Regeln durch Sprachsynthesesysteme simuliert werden kann.

Weiterhin ergibt sich durch die Beschränkung zeitbereichsbasierter Synthesizer auf prosodische Modifikationen die Frage nach der Bedeutung stimmqualitativer und artikulatorischer Merkmale für emotionalen Sprechausdruck.

> **These 3**: Es ist möglich, emotionalen Sprechausdruck alleine durch Modifikation prosodischer Merkmale zu simulieren, dieser kann jedoch durch Verwendung parametrischer Synthesizer und Berücksichtigung stimmqualitativer und artikulatorischer Merkmale signifikant gesteigert werden.

Merkmal	Ärger	Angst	Freude	Langeweile	Trauer
Anzahl betonter Silben	größer	größer	größer	-	kleiner
Silbenlängen	kürzer	kürzer, betonte länger	kürzer	länger, vor allem bei betonten Silben	stark verlängert, Auftreten von Sprechpausen
F_0-Lage	angehoben	angehoben	angehoben	abgesenkt	abgesenkt
F_0-Range	breiter	breiter	breiter	schmaler	schmaler
F_0-Verlauf auf Silbenebene	angehoben, starke Bewegung, Abwärtstrend	betonte Silben eher steigend	angehoben, starke Bewegung, Aufwärtstrend	Häufung gerader Silbenverläufe	betonte Silben abgesenkt, Häufung fallender Konturen
F_0-Verlauf auf Phrasenebene	Absenkung am Ende	Anhebung am Ende	-	-	Anhebung am Ende
Stimmqualität	rauhe Stimme	Jitter, eventuell Falsett	leicht gespannt	eventuell Knarrphonation	Behauchung, eventuell Jitter
Lage der Formanten	F1 angehoben	größerer Abstand von F1 und F2, F2 angehoben	Formanten angehoben	größerer Abstand von F1 und F2	größerer Abstand von F1 und F2
Energie in bestimmten Frequenzbändern	stärkere Energie in höheren Frequenzen	stärkere Energie in höheren Frequenzen	stärkere Energie in höheren Frequenzen	weniger Energie in höheren Frequenzen	weniger Energie in höheren Frequenzen
Artikulationsgenauigkeit	eher elaboriert	eventuell Elaboration betonter Silben, sonst Reduktion	-	starke Reduktionen	reduziert

Tabelle 4.1: Zusammenfassung der akustischen Korrelate für vier Basisemotionen und Langeweile anhand der Literaturrecherche

Teil II

Experimenteller Teil

Kapitel 5

Perzeptionstest zur Variation von Merkmalen der Prosodie

5.1 Motivation

Dieses erste Hörexperiment stellt zunächst die direkte Erweiterung der Magisterarbeit (Burkhardt (1999)), bei der nur die Emotion 'Freude' betrachtet wurde, auf mehrere Emotionen dar. Es soll vor allem überprüft werden, inwieweit sich der dort entwickelte 'Emotionsfilter' zur Simulation von vier Basisemotionen eignet. Der Emotionsfilter, genannt 'EmoFilt', ermöglicht regelhafte Beschreibungen zur Modifikation prosodischer Parameter und wurde als Zusatzkomponente der gegebenen DSP-Komponente[1] MBROLA (Dutoit et al. (1996)) und der NLP-Komponente TXT2PHO (Portele (1996a)) konzipiert.

Die konkatenierende Synthese im Zeitbereich (vgl. Abschnitt 3.4.2) hat in den letzten Jahren gerade im kommerziellen Bereich eine starke Verbreitung gefunden, da sie eine hohe Synthesequalität mit geringem algorithmischen Aufwand verbindet. Es stellt sich nun die Frage, inwieweit eine Vergrößerung der Variationsbreite stimmlichen Ausdrucks durch Manipulation lediglich solcher Merkmale möglich ist, die sich zur Synthesezeit bei Zeitbereichsverfahren variieren lassen. Dabei handelt es sich zunächst ausschließlich um intonatorische Merkmale und solche, die mit den Lautdauern in Zusammenhang stehen. Merkmale der Lautintensität sind, wie in Abschnitt 3.3 beschrieben, normalerweise nicht zur Laufzeit manipulierbar, sondern werden bei der Erstellung der Segmentdatenbank festgelegt.

Eine weitere Möglichkeit in diesem Rahmen wäre die Simulation von Laut-Assimilationen oder -Elisionen durch Vertauschen oder Weglassen einzelner Phonbeschreibungen an der Schnittstelle zwischen NLP- und DSP-Komponente (wie sie z.B. Cahn (1990) implementiert hat). Dieses wurde aber hier aus Aufwandsgründen nicht berücksichtigt.

Eine zusätzliche Fragestellung für dieses Perzeptionsexperiment ergibt sich daraus, dass es grundsätzlich zwei unterschiedliche Ansätze gibt, natürlich wirkende Prosodieparameter für einen zu synthetisierenden Text zu gewinnen (vgl. Abschnitt 3.3):

- **Regelbasierte Algorithmen** wenden auf das Ergebnis einer syntaktischen Analyse des

[1] NLP steht für *Natural Language Processing*, DSP für *Digital Speech Processing* (vgl. Abschnitt 3.2).

Eingabetextes bestimmte Betonungsregeln an (z.B. Liberman & Church (1992)). Eine Erweiterung der Variation des sprecherischen Ausdrucks würde hier die Anwendung gewisser 'Emotionsregeln' aufgrund einer semantischen Analyse erfordern.

- **Datenbasierte Algorithmen** wenden in der Regel Entscheidungsbaumverfahren auf eine Sprachdatenbank an. Das Ziel ist es dabei, Prosodieparameter zu finden, die in einer syntaktisch möglichst ähnlichen Situation angetroffen wurden (z.B. Malfrère et al. (1999)). Um hier eine größere Variationsvielfalt zu erreichen, müssten die Lernalgorithmen mit unterschiedlichen Datenbanken trainiert werden, da Regeln nicht explizit zur Verfügung stehen.

Von daher stellt sich die Frage, inwieweit sich daten- und regelbasierte Emotionssimulation hinsichtlich ihrer Güte unterscheiden. Die Güte der Simulation soll anhand der Wiedererkennungsrate für eine intendierte Emotion gemessen werden.

5.2 Konzeption des Hörversuchs

Aus dem oben skizzierten Problemfeld wird folgenden zwei Fragen explizit nachgegangen:

1. Wie überzeugend lässt sich emotionale Sprechweise mit Zeitbereichs-Syntheseverfahren simulieren?

2. Inwieweit unterscheidet sich dabei daten- von regelbasierter Simulation?

Um diese Fragen zu beantworten, wurde der Versuch unternommen, stimmlichen emotionalen Ausdruck mit einem Zeitbereichs-Syntheseverfahren zu simulieren. Als Synthesizer wurde das Diphon-Konkatenationssystem MBROLA verwendet (Dutoit et al. (1996)). Zur Gewinnung der Steuerparameter wurden sowohl daten- als auch regelbasierte Verfahren angewendet. Dem regelbasierten Verfahren lag die NLP-Komponente TXT2PHO (Portele (1996a)) zugrunde. Die Schnittstelle zwischen den beiden Komponenten ist eine Signalbeschreibungsdatei, die eine phonetische Transkription des Eingabetextes und Prosodiesteueranweisungen umfasst. Der Kürze halber wird sie im folgenden als 'PHO-File' bezeichnet.

Berücksichtigte Parameter Wie oben ausgeführt, erlauben zeitbereichsbasierte Synthesizer eine Manipulation suprasegmenteller Merkmale wie Intonationskontur und Lautdauerstruktur. In eingeschränkter Form können segmentelle Phänomene wie Silben- oder Lautelision und Assimilation berücksichtigt werden. Zum Teil ist auch die Umsetzung nicht-prosodischer Merkmale möglich; so kann z.B. Jitter durch mikroprosodische Manipulation erzeugt werden.

Nicht möglich sind Variationen auf artikulatorischer oder stimmqualitativer Ebene wie etwa eine Beeinflussung der Phonationsart, der Artikulationsgenauigkeit oder eine Simulation von Nasalierung.

Berücksichtigte Emotionen Um einen Vergleich mit früheren Untersuchungen zu erleichtern, wurden als zu simulierende Emotionen die in der Literatur oft betrachteten vier Basisemotionen *Freude, Angst, Wut* und *Trauer* verwendet.

5.3 Generierung der Teststimuli

Zur Generierung der Stimuli wurden zwei Verfahren angewendet:

- **Datenbasierte Simulation**
  Um eine datenbasierte Simulation zu realisieren, wurden natürlichsprachliche emotionale Äußerungen hinsichtlich der Parameter Lautdauer, Intonationskontur und Laut-Elision bzw. -Assimilation kopiert. Diese Äußerungen wurden von Schauspielern simuliert und stammen aus einer im Rahmen eines DFG-Projekts erstellten Sprachdatenbank (siehe Abschnitt 4.6.1). Dabei wurden von vier Sprechern (zwei männliche und zwei weibliche) je zwei Sätze[2] als Grundlage für die Erstellung der von MBROLA benötigten PHO-Files verwendet. Die intendierten Emotionen der Äußerungen waren zuvor in einem Hörtest von mindestens 80 % der Hörer erkannt worden.

  Zur Berechnung der PHO-Files aus den Originalsignalen wurde ein Programm entwickelt, dass die Lautdauern und F_0-Werte automatisch aus den gelabelten Signaldateien durch die ESPS-Tools (Entropic Research Inc.) extrahiert und das PHO-File generiert. Durch Fehler bei der automatischen F_0-Detektion war allerdings eine manuelle Nachbearbeitung notwendig.

- **Regelbasierte Simulation**
  Um emotionale Sprechweise regelbasiert zu simulieren, wurde der von Burkhardt (1999) beschriebene Prosodiemanipulationsfilter 'EmoFilt' verwendet. Dieses Programm ermöglicht die regelhaft beschriebene Manipulation prosodischer Parameter für das MBROLA-Eingabeformat. Beispielsweise lassen sich dadurch Regeln wie: "*Vergrößere den F_0-Range um 20 %*" automatisch in eine geänderte Intonationsbeschreibung umsetzen.

  Als Anhaltspunkte für die Manipulationsparameter wurden Ergebnisse aus der Literatur bzgl. prosodischer Korrelate emotionaler Sprechweise verwendet (vgl. Tab. 4.1). Die Feineinstellungen wurden für die betrachteten Emotionen unter Verwendung des graphischen Frontends von EmoFilt von Hand optimiert. Für beide Sätze wurden dieselben Einstellungen verwendet.

  Die Manipulationen für die einzelnen Emotionen sind in Tabelle 5.1 zusammengefasst. Da der in Kap. 6 beschriebene Synthesizer eine Weiterentwicklung von EmoFilt darstellt, sind die Beschreibungen der Algorithmen für prosodische Modifikationen in den Abschnitten 6.3.1 und 6.3.2 gegeben. Der Emotionsfilter wurde auf neutrale Versionen der beiden Sätze angewendet, welche von der Phonemisierungskomponente TXT2PHO (Portele (1996a)) generiert wurden.

Die Stimuli, bei denen weibliche Sprecher kopiert wurden und die regelbasierten Versionen wurden mit einem weiblichen Diphoninventar synthetisiert, die anderen mit einem männlichen.

[2] Die Sätze lauteten: "*An den Wochenenden bin ich jetzt immer nach Hause gefahren und habe Agnes besucht.* " und "*Das schwarze Blatt Papier befindet sich dort oben neben dem Holzstück.* ".

Merkmal	Wut	Freude	Angst	Trauer
F_0-Lage	-	um 40 % angehoben	um 150 % angehoben	um 20 % abgesenkt
F_0-Range	um 50 % verbreitert	um 40 % verbreitert	um 40 % verbreitert	um 80 % verringert
F_0-Variabilität	um 10 % angehoben	um 100 % angehoben	-	-
F_0-Kontur über Phrasen	leicht fallend	-	am Ende steigend	-
starkbetonte Silben	leicht angehoben, Dauer um 10 % verlängert, fallende F_0-Kontur	leicht angehoben, Dauer um 20 % verlängert, steigende F_0-Kontur	Dauer um 20 % verkürzt	um 20 % verlängert, fallende F_0-Kontur
Sprechrate	unbetonte Silben 25 % schneller	unbetonte Silben 20 % schneller	alle Silben 30 % schneller	alle Silben 20 % langsamer
Lautdauern	Nasale, Frikative und Plosive um 10 % verkürzt	stimmhafte Frikative um 30 % verlängert	-	-
Sprechpausen	-	-	-	um 100 % verlängert
Jitter	3 % Jitter	-	8 % Jitter	10 % Jitter

Tabelle 5.1: Manipulationen der Stimuli bezogen auf neutrale Sprechweise

5.4 Durchführung des Hörtests

Um die Kontextabhängigkeit der Hörerurteile zu minimieren, wurden die Stimuli jedem Hörer in einer anderen Zufallsreihenfolge dargeboten. Dafür wurde ein Computerprogramm entwickelt, so dass die Tests in Einzelsitzungen an einem Terminal stattfanden. Es wurde ein forced-choice Test gewählt. Nach dem einmaligen Hören des Stimulus musste die Versuchsperson auf die Frage *"Wie klingt der/die Sprecher/in?"* durch die Angabe von 'neutral', 'wütend', 'freudig', 'ängstlich' oder 'traurig' reagieren.

Insgesamt waren 50 Stimuli zu beurteilen (vier Emotionen plus Neutral mal zwei Sätze mal vier datenbasierte plus der regelbasierten Version). Um die Hörer an das Beurteilungsverfahren und den Syntheseklang zu gewöhnen, wurden vier Stimuli vorangestellt, die bei der Auswertung nicht berücksichtigt wurden. Teilnehmer waren 10 männliche und 10 weibliche deutsche Muttersprachler im Alter zwischen 21 und 34 Jahren. Die Stimuli wurden mit Kopfhörern dargeboten.

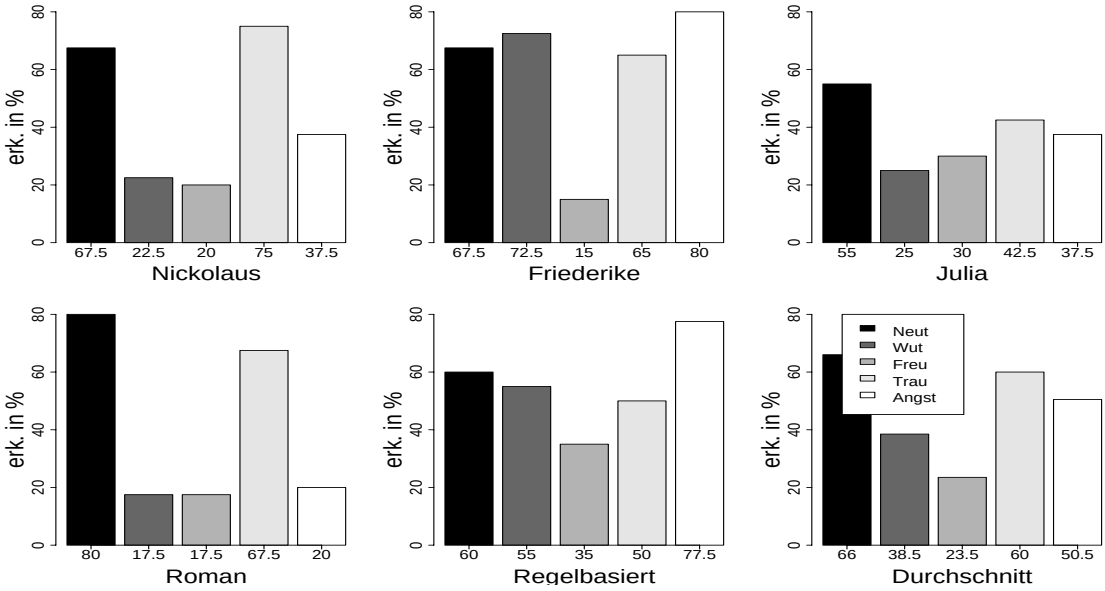

Abbildung 5.1: durchschnittliche Bewertungen der Stimuli nach Sprechern

5.5 Auswertung der Ergebnisse

Die Beurteilungen wurden mittels einer mehrfaktoriellen Varianzanalyse mit kompletter Messwiederholung ausgewertet (vgl. Diehl (1977, S. 272); Bortz (1993, S. 320); Werner (1997, S. 486)). Die drei Faktoren lauteten: *Emotion* (fünf Ausprägungen), *Sprecher* (fünf Ausprägungen) und *Satz* (zwei Ausprägungen). Es ergaben sich signifikante Effekte[3] für die Faktoren *Emotion* und *Sprecher* sowie die Interaktion zwischen *Sprecher* und *Emotion*. Der Faktor *Satz* und der Innersubjektfaktor *Geschlecht des Beurteilers* resultierten nicht in einer signifikanten Änderung des Beurteilerverhaltens.

In Abb. 5.1 sind die durchschnittlichen Bewertungen für die Sprecher in Abhängigkeit von der Emotion sowie die mittlere Bewertung für die einzelnen Emotionen dargestellt. Die Urteile für die beiden Sätze sind dabei zusammengefasst worden.

- **Unterschiede zwischen den Sprechern**
  Signifikante Unterschiede innerhalb der Sprecher ergeben sich zwischen zwei Gruppen: Sprecherin Friederike und die regelbasierte Version einerseits sowie die restlichen Sprecher andererseits. Somit sind die regelbasierten Versionen, wenn man von Sprecherin Friederike absieht, besser erkannt worden als die datenbasierten.

  Insgesamt gesehen haben die Schauspieler den emotionalen Ausdruck offensichtlich eher durch artikulatorische oder stimmqualitative Merkmale als durch prosodische Variation erzeugt. Da die regelbasierten Versionen sich gerade auf prosodische Manipulationen bezogen, war eine höhere Erkennungsrate zu erwarten.

[3] Für ein Signifikanzniveau $\alpha = .05$.

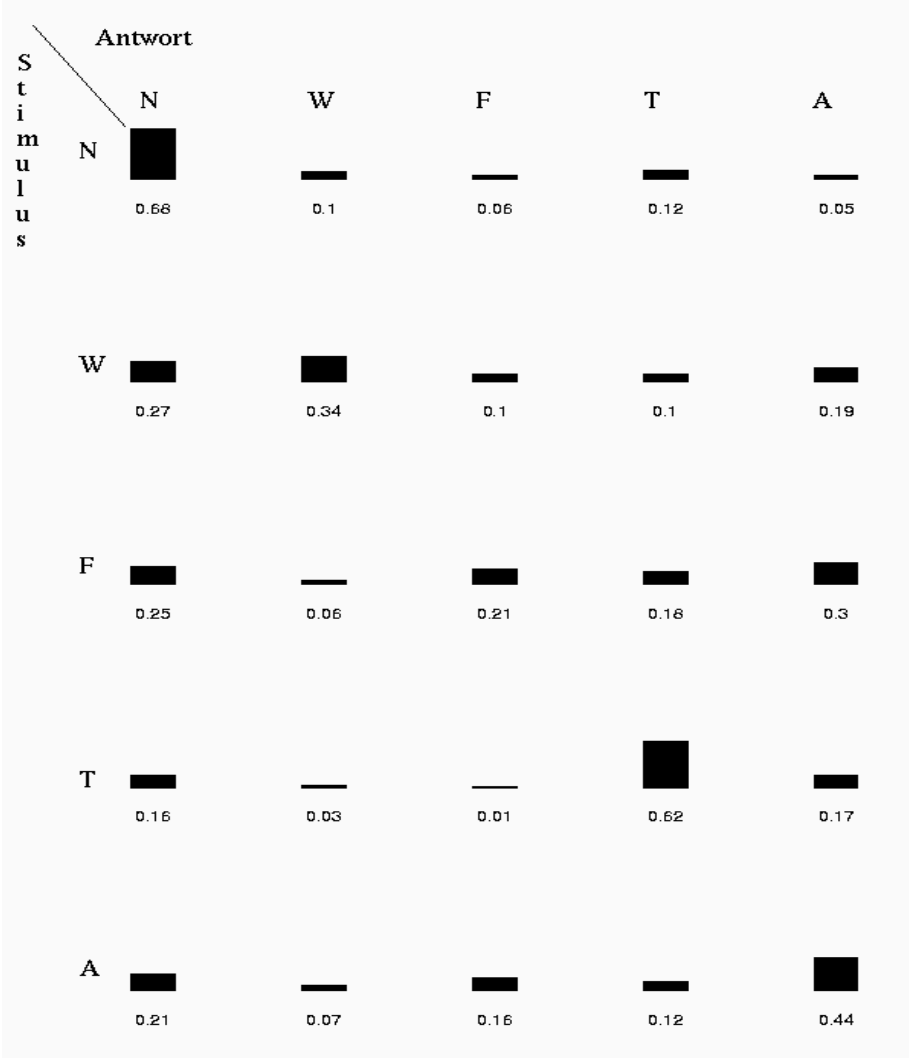

Abbildung 5.2: Konfusionsmatrix für die datenbasierten Stimuli. Die Zeilen entsprechen den intendierten Emotionen, die Spalten den erkannten (N=Neutral, W=Wut, F=Freude, T=Trauer, A=Angst).

- **Unterschiede zwischen den Emotionen**
  Die Erkennungsrate der Emotionen hat die Rangfolge neutral, traurig, ängstlich, wütend und freudig. Abgesehen von den Unterschieden zwischen neutral / traurig und traurig / ängstlich sind alle Unterschiede signifikant.

  Die freudigen Versionen liegen bei fast allen datenbasierten Versionen unter dem Zufallsniveau von 20 %. Auch bei den regelbasierten Versionen erreicht Freude eine Erkennungsrate von lediglich 35 %.

  Die einzigen Emotionen, bei denen die regelbasierten Versionen nicht signifikant besser erkannt wurden als die datenbasierten, sind Trauer und Neutral.

Bei den Angst simulierenden regelbasierten Stimuli ist die Erkennungsrate auffällig hoch. Anscheinend waren die zugrundeliegenden Regeln besonders adäquat.

- **Konfusionsmatrizen**
  In Abb. 5.2 und 5.3 sind Konfusionsmatrizen für die daten- und die regelbasierten Simulationen dargestellt. Bei der Konfusionsmatrix für die datenbasierten Stimuli fällt auf, dass allgemein die Emotionen häufig als Neutral klingend beurteilt wurden. Vermutlich zeigte sich hier der Effekt, dass bei Unsicherheit des Beurteilers für Neutral entschieden wurde. Ansonsten gab es die meisten Verwechslungen für Freude. Sie wurde am häufigsten mit Angst und am seltensten mit Wut verwechselt. Die Verwechslung mit Angst mag darauf zurückzuführen zu sein, dass beide Emotionen mit einer Anhebung der Grundfrequenz einhergehen.

  Bei den regelbasierten Versionen wurde Wut am häufigsten mit Neutral verwechselt und Freude mit Wut. Diese Verwechslung läßt sich durch eine ähnliche Erregungstufe der beiden Emotionen erklären. Banse & Scherer (1996, S. 632) weisen darauf hin, dass Verwechslungen in der Regel aufgrund einer Ähnlichkeit in einem der drei Bereiche: Qualität, Erregung und Valenz erfolgen. Alle anderen Verwechslungen lagen unter dem Zufallsniveau.

5.6 Bezug der Ergebnisse zu früheren Untersuchungen

Ähnliche Untersuchungen liegen von Schröder (1999); Heuft et al. (1998); Rank & Pirker (1998); Murray & Arnott (1995); Vroomen et al. (1993); Carlson et al. (1992) und Cahn (1990) vor. Gemeinsam ist fast allen Untersuchungen, dass verschiedene Emotionen sehr unterschiedlich gut erkannt wurden. Weiterhin liegen bei fast allen Untersuchungen die Erkennungsraten ähnlich hoch (im Schnitt zwischen 30 und 50 %). Die einzige Ausnahme ist hier Vroomen et al. (1993), bei der von überdurchschnittlich hohen Erkennungsraten berichtet wird.

Auch bei anderen Untersuchungen wurde Freude eher schlechter (Carlson et al. (1992); Murray & Arnott (1995); Heuft et al. (1998); Schröder (1999)) und Trauer eher besser (Cahn (1990); Murray & Arnott (1995); Rank & Pirker (1998); Schröder (1999)) erkannt. Scherer (1981) weist darauf hin, dass negative Emotionen anscheinend generell besser erkannt werden als positive.

Es findet sich allerdings relativ wenig Übereinstimmung zwischen den Ergebnissen dieser Untersuchung und früheren Ergebnissen in Bezug auf Verwechslungen zwischen den Emotionen. Offensichtlich sind die sprecherspezifischen Unterschiede im Bereich emotionaler stimmlicher Ausdrucksstrategien sehr groß.

Ein Vergleich zwischen verschiedenen Untersuchungen wird selbstverständlich dadurch erschwert, dass verschiedene Syntheseverfahren angewendet wurden, unterschiedlich viele Basisemotionen untersucht wurden und die Stimuli sowohl in Bezug auf den Inhalt als auf die Sprache her unterschiedlich waren.

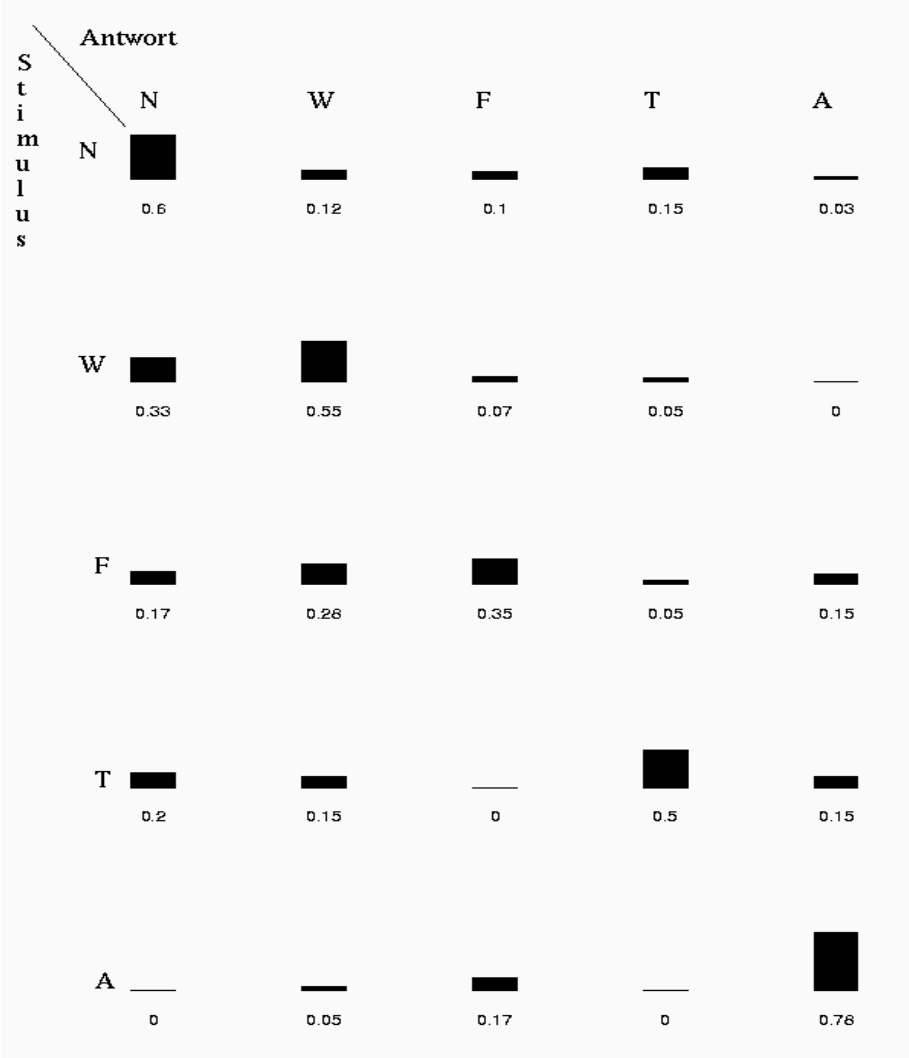

Abbildung 5.3: Konfusionsmatrix für die regelbasierten Stimuli. Die Zeilen entsprechen den intendierten Emotionen, die Spalten den erkannten (N=Neutral, W=Wut, F=Freude, T=Trauer, A=Angst).

5.7 Zusammenfassung und Diskussion

Die Ergebnisse zeigen, dass emotionale Sprechweise auch bei den beschränkten Möglichkeiten einer Zeitbereichssynthese erkennbar zu simulieren ist. Dabei gibt es allerdings beträchtliche Unterschiede zwischen einzelnen Emotionen.

Regelbasierte Modellierung ergibt dabei generell bessere Ergebnisse, da die Regeln speziell auf die Möglichkeiten des Synthesizers abgestimmt werden können.

Die großen Unterschiede zwischen den Sprechern deuten auf unterschiedliche Sprecherstrategien hin, Emotionen auszudrücken (vgl. auch Schröder (1999)).

Die erhebliche Diskrepanz der Erkennungsrate zwischen resynthetisierten und natürlichen Stimuli legt nahe, dass zur Simulation nicht nur erkennbarer, sondern auch natürlich wirkender emotionaler Sprechweise eine Modellierung stimmqualitativer und artikulatorischer Merkmale unumgänglich ist.

Insofern scheint die Verwendung von Synthesizern, die eine Variation segmenteller Merkmale ermöglichen, attraktiv. Bei Carlson et al. (1992) z.B. stieg die Erkennungsrate deutlich für Stimuli, bei denen auch das Anregungssignal eines Formantsynthesizers dem natürlichen Vorbild nachempfunden war.

Weitere Informationen zu EmoFilt (inklusive Quelltext) sowie die für dieses Experiment generierten Audiobeispiele können im Internet heruntergeladen werden. Die Adresse lautet:
http://www.kgw.tu-berlin.de/∼felixbur/emoFilt.html
bzw.
http://www.kgw.tu-berlin.de/∼felixbur/speechEmotions.html

Kapitel 6

EmoSyn: Ein Sprachsyntheseverfahren zur Simulation emotionaler Sprechweise

6.1 Einleitung

In diesem Kapitel wird ein Sprachsynthesizer beschrieben, der speziell für die Simulation emotionaler Sprechweise optimiert wurde. Vier Gründe führten dazu, der nicht unbeträchtlichen Anzahl an vorhandenen Sprachsynthesizern einen weiteren hinzuzufügen:

- Die gezielte Manipulation ausgewählter akustischer Parameter lässt sich sowohl durch die Anwendung von Resyntheseverfahren als auch durch Sprachsynthesizer erreichen (zur Definition vgl. Kap. 3). Unterschiedliche Resyntheseverfahren eignen sich jedoch unterschiedlich gut für verschiedene Manipulationen, was bei systematischer Variation von Merkmalen eine Kombination mehrerer Verfahren erfordert (wie etwa in Burkhardt & Sendlmeier (1999a) eine Kombination aus LPC-Resynthese und PSOLA-Manipulation). Durch die sehr unterschiedlichen Ansätze dieser Verfahren können dabei Störungsursachen sehr komplex werden. Daher ist die Verwendung eines einheitlichen Systems günstiger.

- Da aus der Literatur sehr viele Hinweise auf die Relevanz stimmqualitativer und artikulatorischer Merkmale bei emotionalem Sprechausdruck gefunden wurden, sollten diese Merkmale unbedingt modifizierbar sein. Das Syntheseverfahren, das die größte Flexibilität bzgl. dieser Merkmale mit der besten Synthesequalität verbindet, ist die Formantsynthese.

- Die Validität der gefundenen Korrelate sollte durch die Entwicklung von nachweisbar gültigen 'Emotionsregeln' für allgemeine Sprachsynthesizer überprüft werden. Vergleichbare Untersuchungen wie Murray & Arnott (1995) oder Cahn (1990) verwendeten ein kommerzielles System (DecTalk). Dabei wurde von Schwierigkeiten berichtet, die von Limitationen der Steuerbarkeit herrühren.

- Da dem Autor kein Formantsynthesesystem für das Deutsche zugänglich war, das ausreichende Modifizierungsmöglichkeiten, etwa durch Zugänglichkeit des Quellcodes,

ermöglicht hätte, wurde ein eigenes System mit der Bezeichnung *emoSyn* (Akronym für Emotions-Synthesizer) entwickelt.

Dabei handelt es sich um ein Programm, das aus einer Signalbeschreibungsdatei Parameterwerte für den Klatt-Formantsynthesizer KLSYN88 (Klatt & Klatt (1990); Klatt (1990, 1980)) erzeugt, also um ein Untermodul der DSP-Komponente. Die Simulation emotionaler Sprechweise mit Formantsynthesizern im allgemeinen (Carlson et al. (1992)) und mit dem Klattsynthesizer im speziellen (Murray & Arnott (1995); Rutledge et al. (1995) sowie Cahn (1990)) wurde bereits erfolgreich durchgeführt.

6.2 Beschreibung der Software

Das Programm wurde in der objektorientierten Sprache C++ (Stroustrup (1998)) implementiert, die einerseits durch die Unterstützung von Objektorientierung eine sinnvolle Datenrepräsentation ermöglicht und andererseits Optimierungsmöglichkeiten bietet, falls die Berechnungszeit zum Problem werden sollte. Ein weiterer Vorteil des objektorientierten Ansatzes ist die leichtere Erweiterbarkeit. Zudem ist C++ sehr weit verbreitet, was die zukünftige Nutzung einzelner Module des Programms in anderer Umgebung erleichtert.

Das für die Formantsynthese benötigte Wissen um die akustischen Eigenschaften des Sprachsignals wird für stimmhafte Signalanteile konzeptionell einer Diphondatenbank entnommen und für stimmlose regelhaft beschrieben. Allerdings wäre die Anfertigung einer Diphondatenbank für das Deutsche über den Rahmen dieser Arbeit hinaus gegangen. Zudem ist es fraglich, ob eine reine Diphonkonkatenation bei einer Formantsynthese für das Deutsche überhaupt einen günstigen Ansatz darstellt (vgl. Abschnitt 3.4.2). Das Problem der idealen Segmentgröße für einen Formantsynthesizer zur Synthese von deutscher Sprache stellt jedoch ein äußerst anspruchsvolles Thema für sich dar. Deshalb wurde für die Experimente lediglich das für die Testsätze benötigte Inventar erzeugt.

Die Segmente wurden dabei aus demselben Satz extrahiert, der synthetisiert werden sollte. Dies steigert die Synthesequalität natürlich außerordentlich (und mag zur Überprüfung eines TTS-Systems ungeeignet erscheinen). Es wurde der Standpunkt eingenommen, dass zum Zweck der Simulation emotionaler Sprechweise vom Problem der idealen Segmentdatenbank abstrahiert werden kann.

6.2.1 Datenstruktur

Die Datenstruktur von emoSyn ist in Abb. 6.1 dargestellt. Ein Satz besteht aus einer Reihe von Silben, die wiederum aus einer Reihe von Lauten bestehen. Jeder Silbe ist eine der drei Betonungsstufen ’unbetont’, ’wortbetont’ und ’phrasenbetont’ zugeordnet. Jeder Laut wird durch die Dauer und eine Reihe von Vektoren beschrieben, in denen die Steuerparameter für den Formantsynthesizer gespeichert sind. Zusätzlich sind jedem Laut ein Reihe von Merkmalen zugeordnet, die phonatorische oder artikulatorische Eigenheiten beschreiben (im folgenden *phone-features* genannt). Zur Synthesezeit werden aufgrund dieser Eigenheiten die Merkmalsvektoren modifiziert und dem Formantsynthesizer übergeben. Dieser Aufbau ermöglicht eine hierarchische

Beschreibung der Änderungsmethoden. Soll z.B. die Intonationskontur der Äußerung modifiziert werden, so wird dafür die Methode *change_intonationskontur* der Klasse Satz aufgerufen. Diese ruft die Methode *change_intonationskontur* bei den enthaltenen Silben auf, was wiederum einen Aufruf der Methode *change_intonationskontur* der in der Silbe enthaltenen Laute zur Folge hat. Damit können sich Manipulationen konzeptuell auf die gesamte Äußerung, einzelne

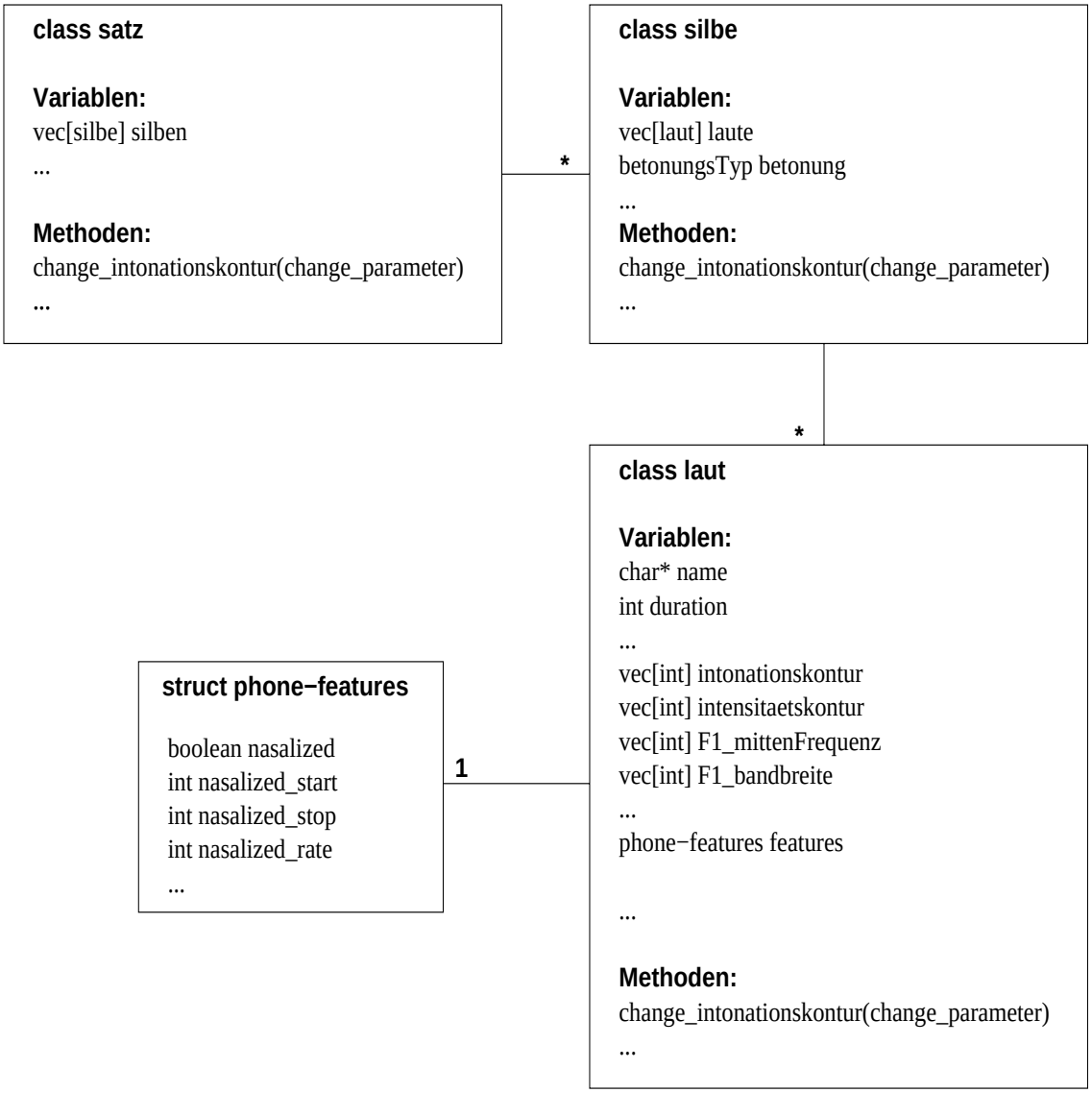

Abbildung 6.1: Objektmodell von emoSyn (unvollständig), siehe Text

Silben oder Laute beziehen. Eine Erweiterung dieser Datenstruktur um abstraktere Konzepte (wie z.B. die Zusammenfassung mehrerer Silben zu *Akzentstrukturen*, vgl. Abschnitt 2.2.1) wäre bei zukünftigen Versionen eventuell vorzunehmen.

In den phone-features sind sowohl phonatorische (wie z.B. creaky, breathy oder voi-

ced/unvoiced) als auch artikulatorische (wie z.B. nasalized, spread lips oder formant-target undershoot) Eigenschaften zusammengefasst, die jeweils durch einen Wahrheitswert für das Auftreten, eine Start- und Endzeit des Auftretens und einen Stärkegrad (*rate*) beschrieben sind.

6.2.2 Eingabeformat

Das Eingabeformat stellt eine um Silbenkennzeichnung erweiterte Form der PHO-Files von MBROLA (vgl. Abschnitt 5.2) dar. Das folgende Beispiel für das Wort "*Wochenenden*" soll das Eingabeformat demonstrieren:

```
v 35 0 95 14 95 28 95 42 95 56 95 70 95 84 99
O 55 0 99 9 99 18 99 27 103 36 107 45 111 54 111 ...
+overshoot 30 40
- 2
x 50
@ 30 0 178 16 175 32 174 48 168 64 165 80 160
n 45 0 153 11 156 22 153 33 151 44 149 55 143  ...
- 0
E 65 0 115 7 116 14 109 21 109 28 112 35 112 42 112 ...
-lar 30 50
n 40 0 117 12 123 24 121 36 116 48 115 60 113 72 108 84 113
- 1
? 35
n 125 0 129 4 126 8 129 12 129 16 126 20 126 24 128 ...
- 0
```

Dabei ist jedem Laut zunächst sein Name (in Sampa), eine Dauerangabe (in ms) und eine Intonationsbeschreibung in der Form von [Zeitpunkt (in %)/F_0-Wert]-Paaren zugeordnet. Weiterhin können beliebig viele Einträge für artikulatorische oder phonatorische Eigenschaften (vgl. Abschnitt 6.3) mit Angabe des Bereichs im Laut zugeordnet werden. So bedeuten etwa die Zeilen:

```
E 65 0 115 7 116 14 109 21 109 28 112 35 112 42 112 ...
-lar 30 50
```

dass das [ɛ] eine Dauer von 65 ms hat, bei Beginn mit 115 Hz und bei 42 % der Dauer mit 112 Hz gesprochen werden soll. Weiterhin soll es bis zur 30. ms zu 50 % laryngalisiert werden. Silbengrenzen sind durch einen Bindestrich gekennzeichnet, gefolgt von einer Kennzahl für die Betonung (0=unbetont, 1=wortbetont und 2=satzbetont). Diese bezieht sich auf die vorangegangene Silbe.

Dieses Format entspricht abgesehen von den Silbenmarkierungen und den zusätzlichen Merkmalen dem MBROLA-Format (Dutoit et al. (1996)). Um es automatisch durch die Phonemisierungskomponente TXT2PHO (Portele (1996a)) erzeugen zu können, wurde ein Programm entwickelt, das für das MBROLA-Format regelbasiert Silbenmarkierungen und Betonungsstufen generiert.

6.2.3 Ausgabeformat

Als Formantsynthesizer werden zwei Implementationen des Klatt-Formantsynthesizers unterstützt: die Version von Iles & Simmons (1995) und die kommerzielle Version der Firma Sensymetrics Inc. (SENSYMETRICS (1999)). Da bei der Version von Iles & Simmons allerdings nicht alle in Klatt & Klatt (1990) beschriebenen Parameter (vor allem solche zur Modellierung des Glottissignals) implementiert sind, wurde bei den in Kap. 7 und 8 beschriebenen Perzeptionsexperimenten die Version von Sensimetrics verwendet. Diese Version erwartet als Eingabe 12 konstante und 48 variable Parameter, die jeweils alle 5 ms aktualisiert werden.

6.2.4 Steuerparameter für periodische Signalanteile: Die Segmentdatenbank

Um eine Äußerung zu synthetisieren, erwartet das Programm für jede vorkommende Kombination zweier Laute einen Eintrag in der Diphondatenbank. Eine Diphonbeschreibung umfasst für jeweils fünf ms Werte für die originale F_0 (zu Resynthesezwecken), die relative Amplitude in dB (normalisiert auf max. 60 dB) und Frequenzwerte für die ersten fünf Formanten.

Zwar ist für die Vokalverständlichkeit im Prinzip lediglich eine Information über die ersten 2-3 Formantfrequenzen notwendig (vgl. Fant (1960, 48)), es ergibt sich jedoch eine höhere Natürlichkeit der Ausgabe bei dynamischen Verläufen auch der oberen Formanten. Da der Klatt'sche Formantsynthesizer für eine Bandbreite bis 10 kHz ausgelegt ist, reicht eine Spezifizierung bis zum fünften Formanten.

Auf eine Angabe der Bandbreiten wurde aus zwei Gründen verzichtet. Erstens ist die Abschätzung von Formantbandbreiten überaus aufwendig und fehleranfällig (vgl. Hawks & Miller (1995) oder Fujimura & Lindqvist (1971)), und zweitens sind Schwankungen der Bandbreiten perzeptiv erst ab Änderungen von etwa 40 % bemerkbar (Carlson et al. (1979)). Statt dessen wurde die in Fujimura & Lindqvist (1971, S. 547) dargestellte Punktwolke, bei der die Bandbreite des ersten Formanten gegen seine Mittenfrequenz dargestellt ist, durch ein Parabelstück approximiert. Dabei ergab sich folgende Formel:

$$BW_1 = 4 * 10^{-4} F_1^2 - 0,45 F_1 + 153$$

die die Bandbreite (BW_1) des ersten Formanten als Funktion seiner Frequenz (F_1) berechnet. Diese Funktion wurde bei emoSyn getestet. Nach informellem Höreindruck ergibt sich daraus jedoch keine perzeptiv wahrnehmbare Verbesserung der Ausgabe (obwohl die Bandbreiten gemäß der Formel immerhin zwischen 40 und 80 Hz liegen, also Schwankungen von bis zu 50 % aufweisen).

Das folgende Beispiel des Eintrags für das Diphon v-O (der Übergang von [v] zu [ɔ]) soll als Beispiel für einen Eintrag in der Segmentdatenbank dienen:

```
name v-O
frameNumber 9
t1 3
border 4
```

```
t2 6
1017 42 301 1232 2356 3337 4602
...
1075 59 463 1076 2356 3397 4665
```

Dabei gibt der Wert neben **t1** die Zeit (in Frames) an, ab der der Steady-State-Bereich des ersten Phon endet, der Wert neben **border** gibt die Grenze zwischen den Lauten an und der neben **t2** den Zeitpunkt, an dem der Steady-State des hinteren Lautes beginnt. Die Transition liegt demnach zwischen **t1** und **t2**. Anschließend folgen die Parameterwerte. Dabei entspricht jede Zeile einem Frame. Die Werte sind in der Reihenfolge: F_0, Intensität, F1, ..., F5 angeordnet. Die F_0 ist in Zehntel-Hertz angegeben.

Zur Extraktion von Steuerparametern für die Formantsynthese aus natürlicher Sprache sind verschiedene Ansätze entwickelt worden (z.B. Simpson (1998) oder Mannell (1998), vgl Abschnitt 3.5.3). Auf eine Erstellung einer kompletten Datenbank für den Synthesizer wurde, wie oben erwähnt, aus Aufwandsgründen verzichtet. Die Frage nach den günstigsten Segmenten für eine Formantsynthese für das Deutsche ist bereits ein äußerst anspruchsvolles Problem für sich. Statt dessen wurde für ausgewählte Beispielsätze je eine eigene begrenzte Datenbank von Hand generiert. Formantverläufe aus einer Originaläußerung eben des zu synthetisierenden Satzes wurden dafür durch das Signalverarbeitungsprogramm *Xwaves* stilisiert (vgl. Abb. 6.2) und der Datenbank hinzugefügt. Hierfür wurden eine Reihe von Hilfsprogrammen in der Programmiersprache *Perl* (Wall et al. (1998)) entwickelt, die automatisch aus den Xwaves-Beschreibungsdateien Diphon-Dateien erzeugen.

6.2.5 Steuerparameter für rauschhafte Signalanteile

Die Steuerparameter für aperiodische Signalanteile sind fest im Synthesizer kodiert. Jeder Laut ist gemäß Artikulationsart, -Ort und Stimmhaftigkeit klassifiziert. Das Merkmal Stimmhaftigkeit wurde dabei mit der Artikulationsart zusammengefasst, so dass sich die Einteilung in: Vokal, Liquid, Nasal, stimmhafter Frikativ, stimmloser Frikativ, stimmhafter Plosiv und stimmloser Plosiv ergibt. Als mögliche Artikulationsorte sind bilabial, labial, labiodental, alveolar, postalveolar, palatal, velar, uvular und glottal vorgesehen. Die Kombination dieser Merkmale ermöglicht eine eindeutige Zuordnung aller 46 möglichen Lautsymbole. Gemäß dieser Einteilung werden die Merkmalsvektoren für den parallelen Zweig des Formantsynthesizers (der vor allem für Konsonanten benutzt wird) mit adäquaten Werten belegt. Diese wurden wiederum durch optischen Vergleich von natürlichen Spektrogrammen und solchen des Synthesizers sowie durch auditive Überprüfung ermittelt.

Zur Überprüfung der Merkmalsvektoren wurde ein Visualisierungsprogramm namens **viewTracks** entwickelt, mit dem sich die Parameterverläufe über den Lautsymbolen darstellen lassen. Ein Screenshot davon ist in Abb. 6.3 dargestellt.

Alle Steuerparameter werden zunächst mit Defaultwerten initialisiert, die in einer entsprechenden Datei festgehalten sind. Das Testen bestimmter Parameterbelegungen wird durch Modifikation dieser Default-Datei und Ausgabe eines statischen Testsignals von 2 sek Länge erleichtert.

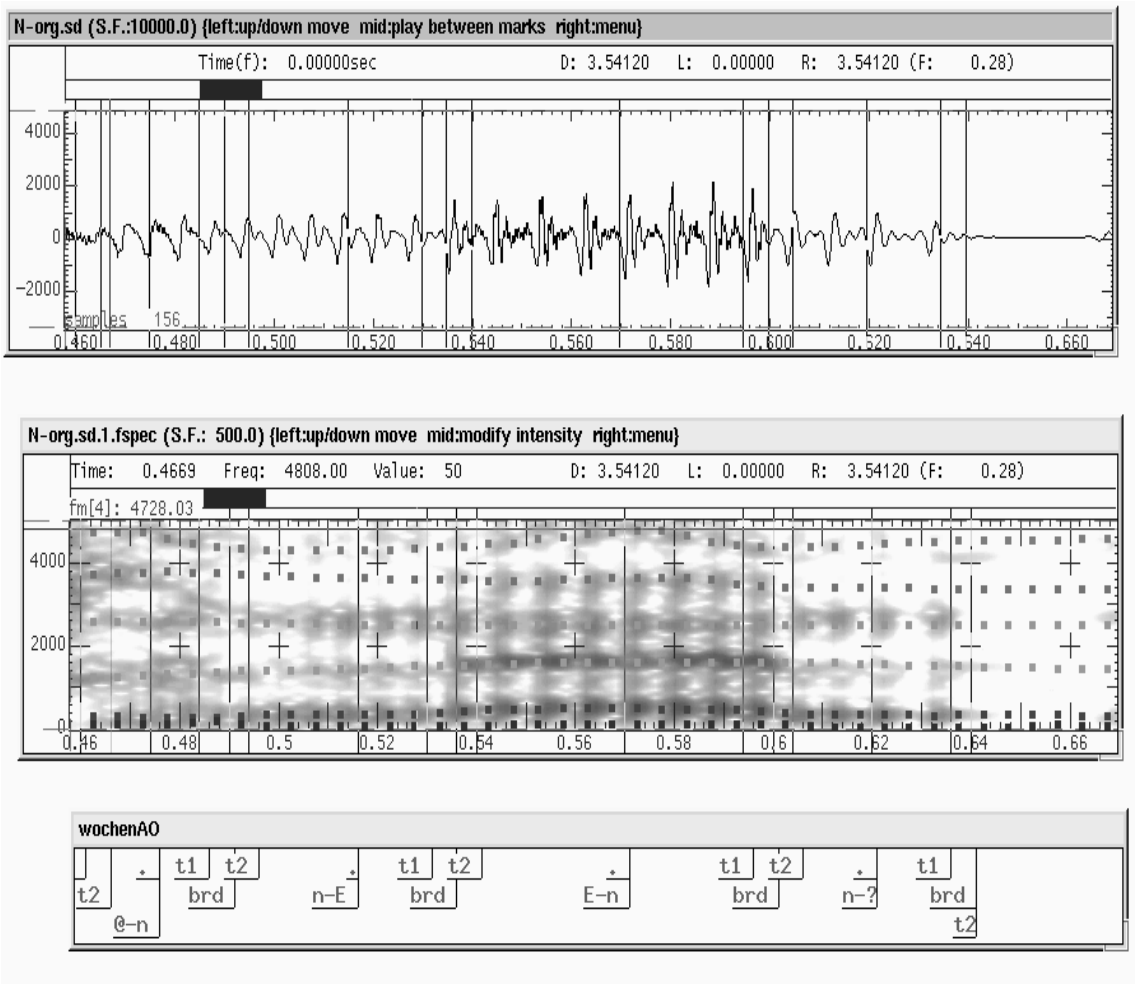

Abbildung 6.2: Stilisierung der Formantverläufe zur Segmentdatenbankgewinnung anhand des Spektrogramms

6.2.6 Emotionssimulation: Die MOD-Files

Um mit dem Synthesizer emotionale Sprechweise zu simulieren, ist die regelhafte Beschreibung der Merkmale in Modifikations-Dateien (MOD-Files) vorgesehen. Das Programm arbeitet die im MOD-File beschriebenen Änderungen sequentiell ab. Die Reihenfolge der Manipulationen ist also durchaus von Bedeutung. Wird z.B. zuerst die F_0-Lage verändert und dann die Intonationskontur manipuliert, so ist der Effekt ein anderer als wenn erst die Intonationskontur und dann die Lage geändert wird, da die Änderungsregeln sich auf prozentuale Anteile der F_0-Werte beziehen. Beispiele für MOD-Files sind im Anhang, Abschnitt A, dargestellt.

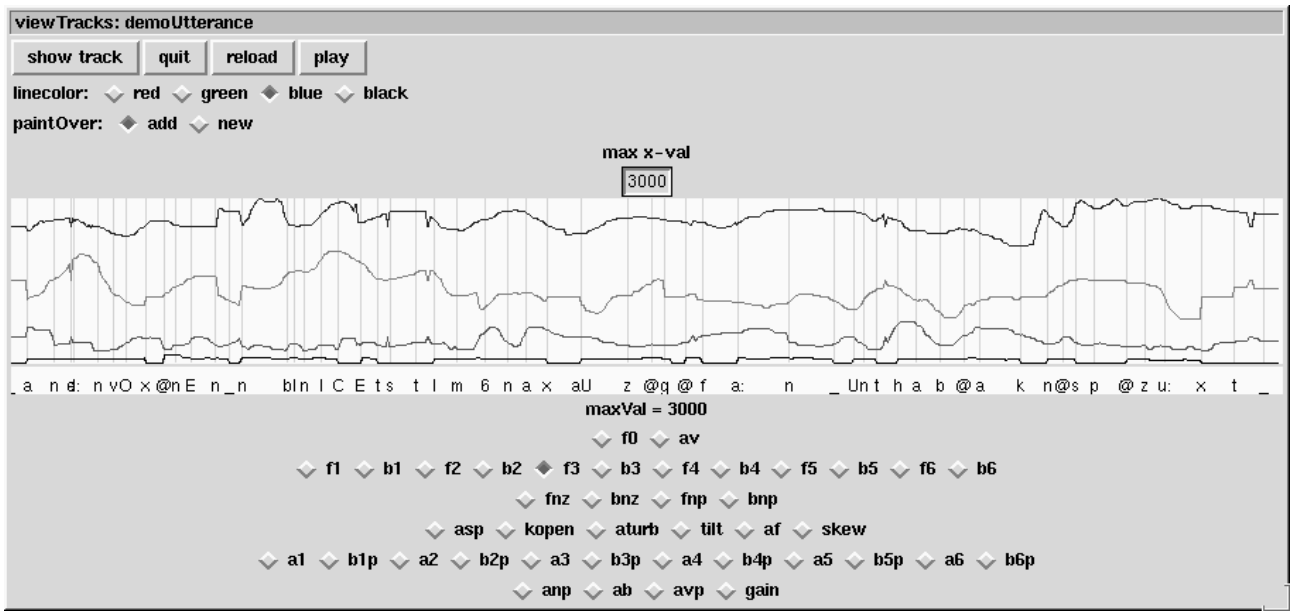

Abbildung 6.3: Screenshot des Hilfsprogramms viewTracks mit Darstellung der Grundfrequenz und der ersten drei Formantkonturen

6.3 Manipulierbare Parameter

Die zur Erzeugung emotionalen Sprecheindrucks manipulierbaren Parameter wurden in fünf Hauptgruppen eingeteilt: Dauer, Intonation, Intensität, Stimmqualität und Artikulation. Die Parameter wirken im Allgemeinen nicht unabhängig voneinander. Ein banales Beispiel ist die Verkürzung stimmhafter Frikative, die die Sprechrate insgesamt verändert. Weiterhin korrelieren einige Phonationsarten und durchschnittliche Grundfrequenz miteinander: Knarrphonation geht mit abgesenkter und Falsettphonation mit angehobener F_0 einher.

6.3.1 Dauermerkmale

Sprechrate

Eine Änderung der Sprechrate kann zum einen einfach durch eine prozentuale Änderung aller Lautdauern erreicht werden. Sprechpausen werden dabei nicht modifiziert. Zum anderen kann die Sprechrate phonetisch adäquat modifiziert werden (vgl. Bergmann et al. (1988, S. 197)). Die prozentuale Veränderung wirkt sich dann am stärksten auf betonte Vokale gefolgt von Vokalen allgemein und am geringsten auf Konsonanten aus. Plosive werden durch Einfügen von Frames innerhalb der Okklusionsphase verlängert. Bei kontinuanten Lauten kann die Änderung durch Weglassen bzw. Einfügen von Frames erreicht werden oder alternativ dazu durch eine Expandierung bzw. Komprimierung der Formantverläufe.

Lautdauern

Die durchschnittliche Dauer unterschiedlicher Lautklassen lässt sich durch Angabe der Änderung in Prozent manipulieren. Zudem lassen sich Dauerveränderungen für betonte und unbetonte Vokale getrennt angeben.

Sprechpausen

Sprechpausen werden in Bezug auf ihre Dauer wie Laute gehandhabt. Ein Einfügen von Pausen ist allerdings nicht implementiert, da dies zumindest eine syntaktische Analyse der Äußerung erfordern würde, idealerweise sogar eine semantische.

6.3.2 Intonationsmerkmale

Die folgenden Manipulationen beziehen sich auf die F_0-Targetwerte. Für alle Manipulationen gelten die Grenzwerte von 50 bzw. 500 Hz. Würden einzelne F_0-Werte aufgrund einer Manipulation außerhalb dieses Bereichs verlegt, werden sie auf die Extremwerte gesetzt.

Durchschnittliche F_0-Lage

Eine Änderung der durchschnittlichen Grundfrequenz lässt sich direkt durch ein Anheben bzw. Absenken aller F_0-Werte erreichen (in Prozent anzugeben).

F_0-Range

Eine Änderung des Range hat ebenfalls eine Verschiebung aller F_0-Werte zur Folge. Allerdings werden die Werte je nachdem, ob sie über- oder unterhalb des Durchschnittswerts der finalen Silbe liegen, prozentual abgesenkt oder angehoben. Dieses Vorgehen entspricht in seiner Auswirkung dem in Bergmann et al. (1988, S. 171) beschriebenen Verfahren, das auf dem *Peak-Feature* Modell basiert. Der Betrag der Veränderung berechnet sich dabei aus dem Produkt eines anzugebenden Faktors mit der Distanz des jeweiligen Wertes zum Mittelwert. Ein Faktor mit dem Wert 0 setzt alle Werte auf den Bezugswert[1].

F_0-Variabilität

Eine Manipulation der Variabilität wurde als Rangeänderung auf Silbenebene realisiert, d.h. die F_0-Werte werden silbenweise mit dem Silbendurchschnittswert als Bezugspunkt verschoben.

Intonationskontur auf Silbenebene

Für die satzbetonten Silben oder alle betonten Silben ist einer der drei Konturtypen 'gerade', 'steigend' und 'fallend' zu wählen. Zu fallenden oder steigenden Typen ist zusätzlich die Steigung in Halbtönen pro Sekunde (ST/s) anzugeben. Soll eine steigende Kontur modelliert wer-

[1] Der Algorithmus entspricht im übrigen einer Kontrastveränderung bei der digitalen Bildverarbeitung.

den, wird zwischen dem ersten F_0-Wert des stimmhaften Anteils der Silbe und dem gemäß der Steigung berechneten Endwert linear interpoliert. Für fallende Konturen wird als Anfangswert der Mittelwert der Silbe verwendet, ansonsten wird wie bei steigenden verfahren. Gerade Verläufe werden einfach dadurch erzeugt, dass alle F_0-Werte auf den Silbendurchschnittswert gesetzt werden. Zusätzlich laßt sich der perzeptiv sehr relevante Konturverlauf der letzten Silbe verändern. Eine weitere Möglichkeit besteht darin, die F_0-Kontur betonter Silben generell anzuheben bzw. abzusenken.

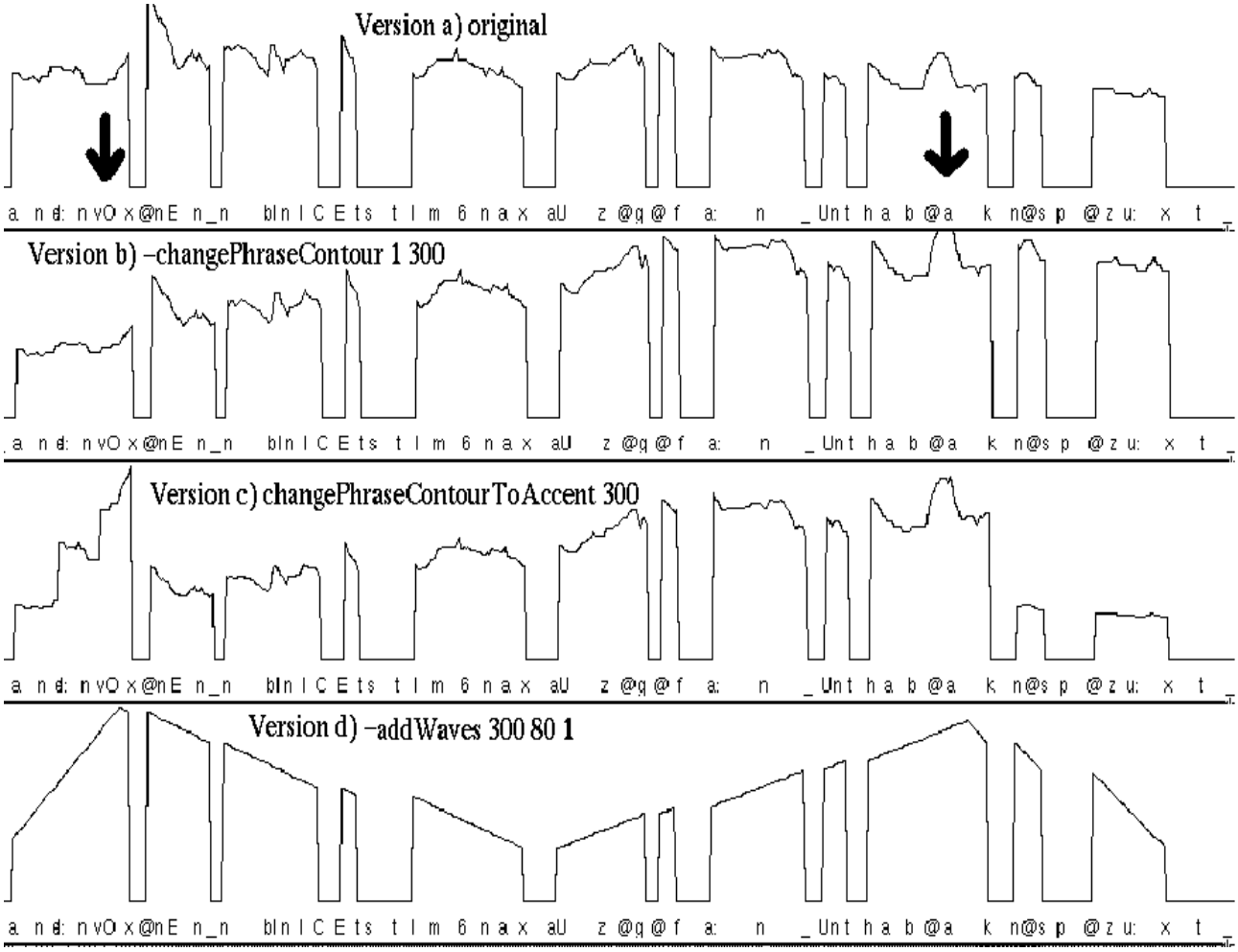

Abbildung 6.4: Vier Intonationsverläufe als Ergebnis der Phrasenkontur-Manipulationen, die Satzakzente sind mit Pfeilen markiert.

Intonationskontur auf Phrasenebene

Der Intonationsverlauf der gesamten Äußerung bzw. Unterabschnitte davon kann den Konturtypen ’steigend’, ’fallend’ und ’gerade’ zugeordnet werden. In der Abb. 6.4 sind die im folgenden

besprochenen Konturmanipulationen für den Beispielsatz *"An den Wochenenden bin ich jetzt immer nach Hause gefahren und habe Agnes besucht."* dargestellt. Als satzbetonte Silben wurden dabei [vɔ] aus *Wochenenden* und [a] aus *Agnes* festgelegt. Die verschiedenen Versionen sind jeweils mit der Beschreibungszeile aus dem MOD-File annotiert.

Zur Änderung der gesamten Intonationskontur gibt es drei Möglichkeiten. Einmal kann die gesamte Kontur einem Steigungstyp zugeordnet werden (vgl. Abb. 6.4, Version b). Alternativ dazu werden die satzbetonten Silben jeweils als Zielwerte angenommen, so dass quasi (falls mehrere Satzbetonungen vorhanden sind) eine stufige Kontur entsteht (vgl. Abb. 6.4, Version c). Sie werden dazu um einen anzugebenden Faktor angehoben oder abgesenkt. Alle Silben, die zwischen dem Zielwert und der vorigen Betonung bzw. dem Anfang der Äußerung liegen, werden prozentual zur Entfernung zur Zielsilbe in ihrer Lage verändert. Für einen geraden Konturverlauf werden alle Silbendurchschnittswerte auf den Gesamtdurchschnittswert durch Anhebung oder Absenkung aller Werte gebracht.

Als dritte Möglichkeit, den gesamten Intonationsverlauf einer Äußerung zu verändern, wurde das sogenannte 'Wellenmodell' implementiert (vgl. Abb. 6.4, Version d). Dabei werden die hauptbetonten Silben um einen bestimmten Prozentsatz angehoben und die mittig dazwischen liegenden Silben abgesenkt. Alle anderen Silben werden gemäß einer gedachten Linie zwischen diesen Extrempunkten entweder in ihrer Lage abgesenkt oder angehoben. Dabei kann die Silbenkontur erhalten werden (indem die Silben insgesamt angehoben oder abgesenkt werden) oder sämtliche F_0-Werte werden gemäß einer Interpolation zwischen den angehobenen und den abgesenkten Silben geglättet, wie in der Abb. dargestellt. Die Anzahl der Wellenhügel hängt somit von der Anzahl der hauptbetonten Silben in der Äußerung ab.

F_0-Perturbation (Jitter)

Unregelmäßigkeiten in der Grundfrequenz werden durch Schwankungen der F_0-Werte erzeugt. Es stehen dafür zwei unterschiedliche Methoden zur Verfügung. Bei der ersten Methode werden alle Werte eines stimmhaften Lautes um einen anzugebenden Prozentsatz abwechselnd angehoben und abgesenkt. Bei der zweiten Methode wird der Parameter FL des Klatt-Formantsynthesizers auf einen vorgegebenen Wert gesetzt. Dieser erzeugt Schwankungen in der F_0 durch Verschiebung um Werte, die aus zwei überlagerten Sinusschwingungen berechnet werden (Klatt & Klatt (1990)). Die F_0-Schwankungen können auf Silben bestimmter Betonungsstufen begrenzt werden.

6.3.3 Intensitätsmerkmale

Als einziger Parameter, der ausschließlich die Intensität (subjektiv die wahrgenommene Lautheit) beeinflusst, ist die Manipulation der betonen Silben vorgesehen. Eine Änderung ist in dB anzugeben. Ansonsten ist eine generelle Anhebung der stimmhaften Amplitude bei gespannter Stimme und eine generelle Absenkung bei behauchter Stimme gegeben (siehe unten).

6.3.4 Stimmqualitative (phonatorische) Merkmale

Stimmqualitätsparameter im engeren Sinne zeichnen sich dadurch aus, dass sie das Anregungssignal modifizieren. Akustische Eigenschaften von Phonationsarten wurden unter anderem in Laver (1980); Klatt & Klatt (1990); Ni Chasaide & Gobl (1997); Gobl & Ni Chasaide (1992); Karlsson (1992) sowie Hermes (1991) untersucht. Die dort gefundenen Hinweise wurden in Änderungen der Quellsignalparameter, der Lautheitswerte und der Bandbreiten umgesetzt. Die genaue Bedeutung der einzelnen Parameter ist unter Abschnitt 3.5.3 beschrieben.

Behauchte Phonation (*breathy voice*)

Behauchung kann mit einer Stärke zwischen 0 und 100 erzeugt werden. Dabei werden die Parameter OQ (Öffnungsquotient), TL (spektrale Dämpfung), B1 (Bandbreite des ersten Formanten), AV (Amplitude des stimmhaften Anteils) und AH (Amplitude des rauschhaften Anteils) nach folgenden Formeln für jeden Frame modifiziert:

$$\begin{aligned} OQ_{add} &= (OQ_{max} - OQ_{glob}) * rate/100 \\ TL_{add} &= (TL_{max} - TL_{glob}) * rate/100 \\ B1_{add} &= (B1_{max} - B1_{glob}) * rate/100 \\ AV_{sub} &= 6 * rate/100 \\ AH_{add} &= (AMP - 3) * rate/100 \end{aligned}$$

$OQ_{add}, TL_{add}, B1_{add}$ und AH_{add} geben dabei die Summanden zu den jeweiligen Parametern und AV_{sub} den Subtrahenden von AV an.
OQ_{glob} und TL_{glob} stehen für die voreingestellten Werte für OQ und TL. AMP ist der Lautheitswert des Frames und $rate$ die Stärke der Behauchung in Prozent. OQ_{max} (100 %), TL_{max} (24 dB) und $B1_{max}$ (250 Hz) geben die maximalen Werte nach der Modifikation an.
Weiterhin werden durch tracheale Kopplung auftretende Polstellen (nach Klatt (1990)) simuliert. Dafür werden die Frequenzen der nasalen Pol- und Nullstelle (FNP und FNZ) auf 550 Hz und die vorgesehene tracheale Pol- und Nullstelle (FTP und FTZ) auf 2100 Hz gesetzt (vgl. Klatt (1990, S. 68)). Die Bandbreiten der Polstellen werden dafür nach folgenden Formeln reduziert:

$$\begin{aligned} BTP_{sub} &= (BTP_{glob} - BTP_{min}) * rate/100 \\ BNP_{sub} &= (BNP_{glob} - BNP_{min}) * rate/100 \end{aligned}$$

Hierbei stehen BTP_{sub} und BNP_{sub} für die Subtrahenden, BTP_{glob} und BNP_{glob} (180 und 100 Hz) für die voreingestellten Werte, und BTP_{min} und BNP_{min} (30 Hz) für die minimalen Werte der Bandbreiten.

Der Amplitudenverlauf der stimmhaften Komponente für behauchten Anregung ist in Abb. 6.5 c) dargestellt. Es ist deutlich die eher sinusförmige Anregungsform wahrzunehmen, die aus der Dämpfung der höheren Frequenzbereiche resultiert.

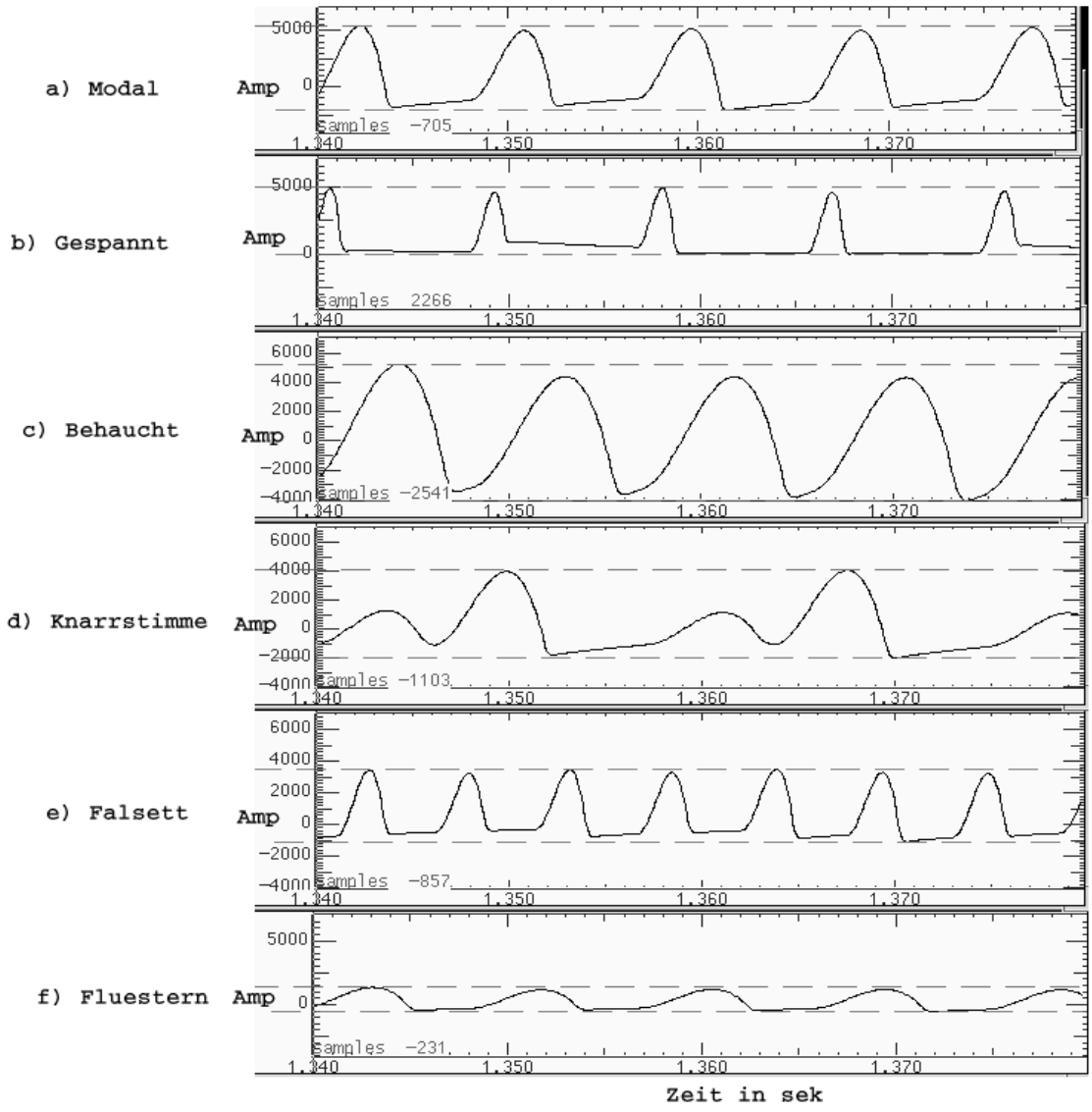

Abbildung 6.5: Periodische Komponente des Anregungssignals für unterschiedliche Phonationsarten

Gespannte Stimme (*tense voice*)

Das Merkmal *gespannte Stimme* ist streng genommen keine Phonationsart, da hiervon auch supralaryngale Komponenten des Sprechapparats betroffen sind. Laver (1980) führt als Phonationsart mit ähnlichen Merkmalen *harsh voice* an, wohingegen *tense voice* als globales Setting bezeichnet wird. Es wurde hier hinzugenommen (wie auch in anderen Arbeiten, z.B. Gobl & Ni Chasaide (1992)), da es stark mit der Erregungsdimension korreliert und sich von daher Hypothesen über die emotionale Eindruckswirkung von gespannter Sprechweise bilden lassen.

Die Parameter AV, OQ, TL und B1 bis B5 werden dabei wie folgt modifiziert:

$$\begin{aligned} AV_{add} &= 6 * rate/100 \\ OQ_{sub} &= (OQ_{glob} - OQ_{min}) * rate/100 \\ TL_{sub} &= TL_{glob} * rate/100 \\ B1_{sub} &= (B1_{glob} - B1_{min}) * rate/100 \\ &... \\ B5_{sub} &= (B5_{glob} - B5_{min}) * rate/100 \end{aligned}$$

Die Amplitude des stimmhaften Anteils wird um maximal 6 dB angehoben. Der Öffnungsquotient wird bis zum minimal sinnvollen Wert verringert ($OQ_{min} = 10$) und die spektrale Dämpfung im Extremfall von $rate = 100$ völlig abgeschaltet.

Um weiterhin die durch Anspannung des Gewebes schwächere Dämpfung zu berücksichtigen, werden die Bandbreiten aller Formanten bis zu einem minimalen Wert (BX_{min} =30-60 Hz) reduziert.

Eine Darstellung der periodischen Komponente für gespannte Stimme bei einer Rate von 70 % ist in Abb. 6.5 b) gegeben. Man sieht im Signal die im Vergleich mit der modalen Anregung (Abb. a)) deutlich kürzere Öffnungsphase und steilere Flanken.

Flüstern (*whispery voice***)**

Flüsternde Stimme simuliert das Programm vor allem durch eine Ersetzung des stimmhaften Anteils AV durch glottales Rauschen AH:

$$\begin{aligned} AV_{sub} &= AV * rate/100 \\ AH_{add} &= AV * rate/100 \end{aligned}$$

Weiterhin wird der Öffnungsquotient vergrößert, die spektrale Dämpfung verstärkt und die Bandbreiten aller Formanten verbreitert:

$$\begin{aligned} OQ_{add} &= (OQ_{max} - OQ_{glob}) * rate/100 \\ TL_{add} &= (TL_{max} - TL_{glob}) * rate/100 \\ B1_{add} &= (B1_{max} - B1_{glob}) * rate/100 \\ &... \\ B5_{add} &= (B5_{max} - B5_{glob}) * rate/100 \end{aligned}$$

Die verbleibenden periodischen Anteile (ohne Rauschanteile) des Anregungssignals bei 30 % Flüstern sind in Abb. 6.5 f) dargestellt. Die Amplitude ist deutlich geringer und der Verlauf eher sinusförmig.

Knarrstimme (*creaky voice***)**

Knarrphonation zeichnet sich vor allem durch eine unregelmäßige und sehr tiefe Grundfrequenz aus. Es kann dabei zum Effekt des Doppelpulsierens (*diplophonic double pulsing*) kommen.

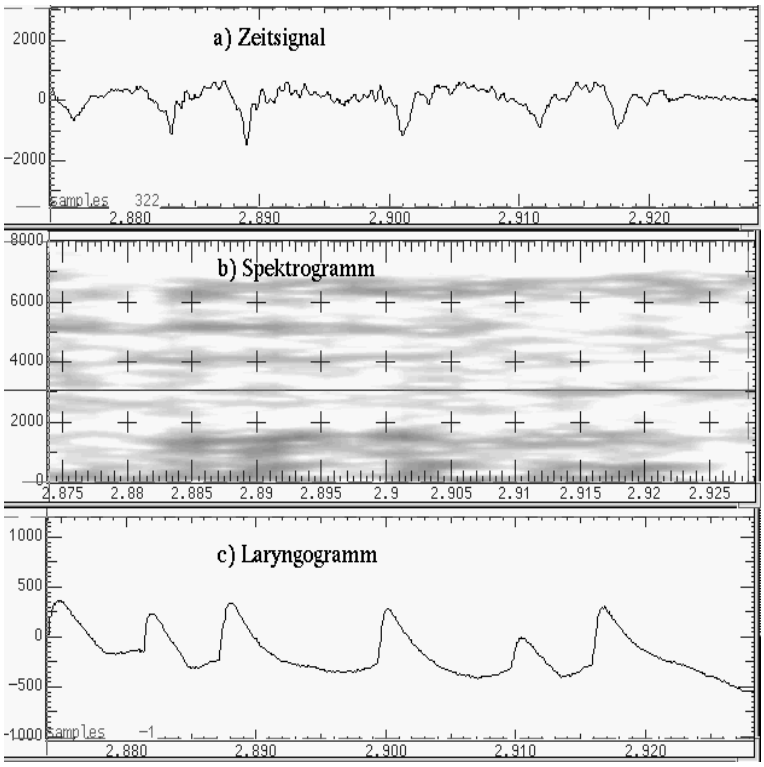

Abbildung 6.6: Oszillogramm, Spektrogramm und Laryngogramm für weiche Knarrstimme bei einem menschlichen Sprecher

Dieser ist bereits bei der zweiten Version (Klatt & Klatt (1990)) des Klatt-Synthesizers implementiert (vgl. Abschnitt 3.5.3, Parameter DI). Dabei wird paarweise jede erste Periode näher zur zweiten hin verschoben und in ihrer Amplitude abgesenkt. Dadurch werden die beiden Perioden als eine wahrgenommen und die Grundfrequenz effektiv halbiert.

In der Regel zeichnet sich die Knarrstimme durch eine kurze glottale Öffnungsphase mit steilflankigen Impulsen aus (vgl. Laver (1980, S. 125) oder Klatt (1990, S. 7)). Dies resultiert aus der relativ starken medialen Kompression und adduktiven Spannung. Allerdings ist die Knarrphonation äußerst variabel (Ni Chasaide & Gobl (1997, S. 450)). Ist der subglottale Druck eher gering und die muskuläre Spannung nicht so stark, können die Flanken auch weniger steil sein und die höheren Frequenzbereiche energielos (Klasmeyer (1999, S. 44))[2]. Die große Variabilität der einzelnen Impulse bei Knarrphonation wird auch durch Messungen von Gobl & Ni Chasaide (1992) belegt.

Um nicht in Konflikt mit Lavers Terminologie zu geraten, wird diese Art der Phonation im Folgenden als '*weiche Knarrstimme*' bezeichnet. Ein Beispiel dafür, dass diese Anregungsart bei emotionaler Sprechweise auftritt, ist in Abb. 6.6 dargestellt. Die Abbildung zeigt das Zeitsi-

[2]So berichten z.B. Ni Chasaide & Gobl (1997, S. 453) von einer *breathy-creaky voice* , die in !Xóõ, einer südafrikanischen Sprache, Phonemstatus hat. Nach Lavers Terminologie müsste sie *whispery-creaky voice* heißen.

gnal, ein Breitbandspektrogramm und das Laryngogramm[3] für den Vokal [a] aus einem traurig intendierten Satz einer Schauspielerin aus dem emotionalen Sprachkorpus. Man sieht im Laryngogramm, dass die Öffnungsphase sehr lang ist und sich die Stimmlippen langsam öffnen, was auf geringen subglottalen Druck hinweist. Bei der Grundfrequenz fallen vor allem die sehr starken Unregelmäßigkeiten auf. Im Spektrogramm sind deutlich Rauschanteile in den höheren Frequenzbereichen zu erkennen, während die Energie im Bereich der unteren Formanten eher gedämpft ist.

Da die Hypothesen für knarrende Anregung dahin gehen, dass diese Phonation eher für Trauer oder Langeweile adäquat ist und diese Emotionen mit einer geringen Erregung verbunden sind, wurde die Knarrstimme mit verlängerter Öffnungsphase (bis zur Hälfte des maximalen Wertes) implementiert:

$$\begin{aligned} DI &= rate \\ OQ_{add} &= (OQ_{max} - OQ_{glob}) * rate/200 \\ AV_{sub} &= 6 * rate/100 \\ B1_{add} &= (B1_{max} - B1_{glob}) * rate/100 \end{aligned}$$

Durch die verlängerten Öffnungsphasen vergrößert sich die Bandbreite des ersten Formanten. Dies unterstützt weiterhin die relativ schwache Energie im Bereich der fundamentalen Komponente (Klatt (1990)), weswegen auch die Amplitude der stimmhaften Anregung abgesenkt wird. Eine Darstellung der periodischen Anregung bei einer Rate von 70 % ist in Abb. 6.5 d) gegeben. Die jeweils erste Periode ist abgeschwächt und etwas verschoben. Die Öffnungsphase ist verlängert und die Flanken weniger steil.

Die Knarrstimme tritt wie bemerkt auch bei emotional neutraler Sprechweise recht häufig auf, vor allem am Äußerungsende. Bei bestimmten emotionalen Zuständen verstärkt sich dann die Wahrscheinlichkeit, mit Knarrphonation zu sprechen (Laver (1980, S. 126) prädiziert dies z.B. für Langeweile). Da der durchgehende Gebrauch der Knarrstimme eher unnatürlich wirkt, ist bei emoSyn eine zweite Möglichkeit vorgesehen, bei der die Knarrstimme nur bei der ersten Hälfte von Vokalen, die einer stimmlosen Periode folgen, simuliert wird (also beim Stimmeinsatz).

Falsett (*falsetto*)

Falsett ist akustisch vor allem durch eine hohe Grundfrequenz und starke Dämpfung der höheren Frequenzbereiche gekennzeichnet. Dies wird dadurch modelliert, dass die Grundfrequenz um maximal 100 % angehoben und die spektrale Dämpfung auf maximal 24 dB erhöht wird.

Da es teilweise gar nicht mehr zu einem vollständigen Verschluss an der Glottis kommt (Laver (1980, S. 118)), wird die Öffnungsphase verlängert. Falsett ist den Hypothesen nach für ängstliche Sprechweise charakteristisch, zudem ist eine Irregularität der Perioden aus dem teilweise unvollständigen Verschluss abzuleiten. Deshalb wird der Parameter FL, der eine leichte

[3] Da das Laryngogramm den durch den Larynx fließenden elektrischen Strom L_x darstellt, ist (im Gegensatz zur Darstellung des Anregungssignals durch den Luftdruck) die Glottis im Maximum geschlossen.

Verschiebung der einzelnen Perioden bewirkt, auf den vorgegebenen Wert der Falsettrate gesetzt.

$$\begin{aligned} F0_{add} &= F0 * rate/100 \\ OQ_{add} &= (OQ_{max} - OQ_{glob}) * rate/100 \\ TL_{add} &= (TL_{max} - TL_{glob}) * rate/100 \\ FL &= rate \end{aligned}$$

Da die Rauschanteile, die durch die ständige tendenzielle Öffnung der Glottis entstehen, bei geringerem subglottalem Druck eher schwach sind (Laver (1980, S. 119)), wurden sie nicht modelliert.

Die periodische Anregung (bei 70 % Falsett) ist in Abb. 6.5 e) dargestellt. Die angehobene Grundfrequenz ist deutlich zu erkennen.

6.3.5 Artikulatorische Merkmale

Target Undershoot

Ein Target Undershoot der Formanten bei stimmhaften Lauten tritt auf, wenn die charakteristische Artikulationstraktkonfiguration im dynamischen Verlauf nicht ganz erreicht wird. Dies wird dadurch simuliert, dass die ersten beiden Formanten zu den Zielwerten des neutralen Ansatzrohrs hin verschoben werden.

Die Verschiebung berechnet sich aus der Differenz des jeweiligen Formantwertes zu den neutralen Werten, gewichtet durch die Verschiebungsrate:

$$\begin{aligned} F1_{diff} &= (F1 - F1_{center}) * rate/100 \\ F2_{diff} &= (F2 - F2_{center}) * rate/100 \end{aligned}$$

Diese Differenz wird dann vom Wert subtrahiert, so dass dieser je nach Vorzeichen der Differenz angehoben oder abgesenkt wird.

Der Parameter $rate$ gibt dabei den Grad der Verschiebung in Prozent an. $F1_{center} = 500$ Hz und $F2_{center} = 1500$ Hz wurden auf die theoretischen Werte eines geraden Ansatzrohrs von etwa 17.5 cm Länge gesetzt. Die Simulation des Formant-Target Undershoot kann wiederum den einzelnen Silbenbetonungstypen zugewiesen werden. Soll die gesamte Äußerung mit Formant-Target Undershoot gesprochen werden, so ist der Eintrag im MOD-File dreimal (für alle Betonungsstufen) vorzunehmen, wobei selbstverständlich jeder Betonungsstufe eine eigene Rate zugewiesen werden kann.

Target Overshoot

Als Target Overshoot wird praktisch das Gegenteil von Undershoot bezeichnet, wenn also die Artikulatoren sozusagen über ihr Ziel hinausschießen. Analog zum Target Undershoot wird dies durch eine Verschiebung der ersten beiden Formanten simuliert. Die mit einer Rate gewichtete Differenz wird dabei allerdings zu dem vorigen Wert addiert.

Lippenspreizung

Eine Spreizung der Lippen, wie sie z.B. bei lächelnder Sprechweise aufritt, verursacht zunächst vor allem eine Verkürzung des Ansatzrohrs. Dadurch werden generell alle Formanten angehoben. Allerdings gilt dies nicht für Frequenzen, deren Wellenlänge kürzer als ein Viertel der tiefsten resonierenden Frequenz ist, da bei diesen die Voraussetzung der eindimensionalen Ausbreitung nicht mehr gegeben ist (vgl. Fant (1960, S. 64)). Rechnerisch ergibt sich ein Grenzfrequenzwert von

$$f = \frac{c}{\lambda} = \frac{330\ \mathrm{m/s}}{0,17\ \mathrm{m}} = 1941\ \mathrm{Hz}.$$

Damit wirkt sich die Änderung im Ansatzrohr vor allem auf die ersten beiden Formanten aus. Burkhardt & Sendlmeier (1999a) und Kienast (1998) stellten auch bei Messungen lächelnder Sprechweise eine Anhebung lediglich von F1 und F2 fest. Von daher werden zur Simulation von Lippenspreizung nur die ersten beiden Formanten gemäß einer anzugebenen Rate angehoben:

$$\begin{aligned} F1_{add} &= F1 * rate/100 \\ F2_{add} &= F2 * rate/100 \end{aligned}$$

Eine Simulation gerundeter Lippen wurde nicht implementiert, da es dafür keine Hinweise in der Literatur bzgl. eines verstärkten Vorkommens bei emotionaler Sprechweise gab.

6.4 Zusammenfassung

emoSyn ist also ein Synthesizer, der die regelhaft beschriebene Generierung von sprachlichen Äußerungen mit emotionalem Gehalt ermöglicht. Es handelt sich um eine Zwischenkomponente eines TTS-Systems, zu der bereits mehrere NLP- wie DSP-Komponenten existieren. Zum Ausbau als Synthesesystem für unbegrenzte Sprachausgabe wäre dazu lediglich eine geeignete Segmentdatenbank zu entwickeln, wobei nicht verschwiegen werden soll, dass es sich dabei um eine äußerst umfangreiche Aufgabe handelt. In den folgenden beiden Kapiteln werden jeweils Perzeptionsexperimente beschrieben, bei denen die Stimuli durch emoSyn generiert wurden. Dabei hat sich gezeigt, dass die oben beschriebenen Manipulationsmöglichkeiten tatsächlich ausreichen, um emotionale Sprechweise in sehr differenzierter Form wiedererkennbar zu simulieren.

Audiobeispiele sowie weitere Informationen zu emoSyn (inklusive des dokumentierten Quelltextes) können im Internet heruntergeladen werden. Die Adresse lautet:
http://www.kgw.tu-berlin.de/∼felixbur/emoSyn.html

Kapitel 7

Systematische Variation von Merkmalen

7.1 Auswahl der manipulierten Parameter und ihrer Ausprägungen

Der in Kap. 5 geschilderte Hörtest hatte zum Ziel, eine generelle Einschätzung der Relevanz unterschiedlicher Merkmale auf die Perzeption emotionaler Sprechweise zu liefern.

Aufgrund der Ergebnisse des Hörtest und der Literatur (vgl. Abschnitt 4.7) wurden fünf Merkmale aus verschiedenen Bereichen ausgewählt, die miteinander kombiniert werden sollten. Das Ziel der Merkmalsauswahl lag darin, einerseits möglichst viele Bereiche abzudecken und andererseits die Merkmalsmenge möglichst gering und allgemein zu halten.

Folgende Merkmale wurden berücksichtigt: Grundfrequenzhöhe (drei Ausprägungen), Grundfrequenzrange (drei Ausprägungen), Phonationsart (fünf Ausprägungen), Lautdauern (vier Ausprägungen) und Vokalqualität (drei Ausprägungen). Das komplette Versuchsdesign (alle Merkmale mit allen Ausprägungen mit allen kombinieren) hätte allerdings über 2000 Stimuli ergeben, ein solch umfangreiches Experiment ist aus forschungsökonomischen Gesichtspunkten kaum durchführbar. Um geringere Stimulusmengen zu erreichen, wurden die Merkmale in drei Obergruppen eingeteilt und diese miteinander kombiniert: In der Obergruppe *Intonation* wurden Grundfrequenzhöhe und -range zusammengefasst, das Merkmal *Phonationsart* bildet eine eigene Obergruppe und die Kategorie *Segmentelle Eigenschaften* umfasst Lautdauern und Vokalqualität. Diese drei Hauptgruppen wurden dann miteinander kombiniert, so dass daraus drei Experimente mit jeweils 45, 108 und 60 Stimuli resultierten.

7.2 Beschreibung der durchgeführten Manipulationen

Die Manipulationen wurden mit dem unter Kap. 6 beschriebenen Sprachsynthesizer durchgeführt. Da zu diesem Zeitpunkt noch kein Diphoninventar für den Synthesizer existierte, wurde als Ausgangspunkt für die Manipulationen ein neutraler Satz eines männlichen Sprechers aus der emotionalen Sprachdatenbank (vgl. Abschnitt 4.6.1) durch emoSyn kopiesynthetisiert. Der Testsatz lautete:
"*An den Wochenenden bin ich jetzt immer nach Hause gefahren und habe Agnes besucht*".

Er wurde ausgewählt, weil er semantisch als emotional unbestimmt gelten kann (vgl. Abschnitt 4.5.2) und zweiphrasig ist, so dass prosodische Manipulationen deutlich bemerkbar sind.

- **Intonation**
  Als einer der wichtigsten Parameter der Intonationskontur wurde die **durchschnittliche Grundfrequenz** des gesamten Satzes einmal um 50 % angehoben und einmal um 30 % abgesenkt. Die Werte für Anhebung und Absenkung ergaben sich aus den geforderten Randbedingungen, dass die Stimuli einerseits noch natürlich klingen und andererseits deutlich wahrnehmbare Variationen aufweisen sollen. Das Einhalten dieser Forderung wurde für alle Manipulationen angestrebt. Als weiteres wichtiges Merkmal für Intonation wurde der **Grundfrequenzrange** der Äußerung einmal um 50 % verringert und einmal um 50 % verbreitert.

- **Phonation**
  Das Merkmal Phonationsart wurde in fünf Ausprägungen variiert. Details zur Simulation von Phonationsarten sind in Abschnitt 6.3 beschrieben. Die Ausprägungen lauteten: Modal, Behaucht (70 %), Knarrstimme (30 %), Falsettstimme (70 %) und gespannte Stimme (70 %). Die Knarrstimme wird nur mit einem Faktor von 30 % erzeugt, da ein höherer Faktor (vor allem in Kombination mit abgesenkter F_0-Lage) unnatürlich wirkt. Ein Faktor von 70 % stellt für die übrigen Phonationsarten einen guten Kompromiss aus deutlicher Wahrnehmbarkeit und natürlich wirkendem Ausgabesignal dar. Obwohl das Merkmal *gespannte (tense) Stimme* eigentlich keine Phonationsart ist, wird es hinzugenommen[1], da es in der Literatur einige Hinweise auf die Relevanz gespannter Stimme für emotionale Sprechweise gibt (vgl. Abschnitt 4.7).

- **Segmentelle Eigenschaften**
  Als segmentelle Merkmale wurden Lautdauern und Merkmale der Artikulationsgenauigkeit zusammengefasst. Die **Lautdauern** werden auf drei Arten variiert. Einmal werden sie um 20 % verkürzt und einmal um 20 % verlängert. Zusätzlich wird eine Variante erstellt, in der alle betonten Silben um 20 % verlängert und alle anderen um 20 % verkürzt werden. Diese Manipulation ergibt sich aus Hinweisen aus der Literatur, nach denen diese Kombination für freudige und ärgerliche Eindruckswirkung günstig ist (siehe Abschnitt 4.7).

  Die **Artikulationsgenauigkeit** wird für Vokale auf zwei Arten variiert. Einmal werden die ersten beiden Formanten der stimmhaften Laute der betonten Silben ausgehend von der Lage des neutralen Lautes [ə] von diesen Werten weggeschoben (Formant-Target Overshoot). Desweiteren werden die ersten beiden Formanten aller stimmhaften Laute zur Lage des [ə] hin verschoben (Formant-Target Undershoot). Eine genauere Beschreibung der Verfahrensweise ist in Abschnitt 6.3 gegeben. Der Overshoot-Faktor beträgt dabei für die satzbetonten Silben 80 % und für die wortbetonten Silben 30 %. Der Undershoot-Faktor beträgt für unbetonte Silben 50 % und für betonte 20 %. Auch diese

[1] Eine ausführlichere Begründung für die Einordnung von gespannter Stimme zu den Phonationsarten wird in Abschnitt 6.3.4 gegeben.

Merkmalsausprägungen ergaben sich durch Optimierung der Natürlichkeit und Variationsbreite für den Testsatz.

7.3 Beschreibung der Durchführung des Hörtests

Für die Durchführung der Hörtests wurde ein forced-choice Test gewählt, bei der die Beurteiler für jeden angebotenen Stimulus genau eine von sechs möglichen Einschätzungen angeben mussten. Die Alternativen lauteten: *neutral, ängstlich, freudig, gelangweilt, traurig* und *wütend/ärgerlich*.

Zur Durchführung des Tests wurde ein Computerprogramm entwickelt, das für jeden Beurteiler eine neue Zufallsreihenfolge der Stimuli generierte (vgl. Abschnitt 5.4). Jeder Hörer bekam jeden Stimulus dabei genau einmal angeboten, eine Wiederholung war nicht möglich. Zur Eingewöhnung wurden zwei Sätze vorangestellt, deren Beurteilung nicht in der Auswertung berücksichtigt wurde. Die Tests wurden in Einzelsitzungen mit Kopfhörern in ruhiger Umgebung durchgeführt. Da alle drei Hörtests zusammen über 200 Stimuli umfassen, wurden sie in zwei Sitzungen aufgeteilt.

Bei der ersten Sitzung gab es 30 Teilnehmer, 14 Frauen und 16 Männer, im Alter zwischen 12 und 48 Jahren. Das Durchschnittsalter betrug 30,2 Jahre. Der zweite Test wurde von 31 Hörern absolviert, davon waren 15 Männer und 16 Frauen. Das Alter lag zwischen 21 und 38 Jahren, im Schnitt betrug es 28,8 Jahre. Alle waren deutsche Muttersprachler, es gab sowohl Experten als auch naive Hörer.

7.4 Auswertung und Interpretation der Ergebnisse

Die Ergebnisse wurden durch eine univariate mehrfaktorielle Varianzanalyse mit kompletter Messwiederholung ausgewertet (vgl. Werner (1997, S. 486), Bortz (1993, S. 320) und Diehl (1977, S. 272)). Dies geschah automatisiert durch das Statistik-Programm SPSS. Ausgewertet wurden die Ergebnisse der multivariaten Tests nach Pillai-Spur (vgl. Bühl & Zöfel (1999, S. 374)). Das Ergebnis dieser Tests ist ein Niveau α, auf dem der Einfluss der Faktoren auf die abhängige Variable (in diesem Fall der Beurteilung der Stimuli als Emotion x) signifikant ist. Das Signifikanzniveau für die einzelnen Effekte ist jeweils in der Überschrift angegeben.

Da SPSS für Interaktionen zweiter Ordnung keine Signifikanzen ausgibt, wurden für solche Fälle beidseitige T-Tests für abhängige Stichproben (Bortz (1993, S. 135)) angewendet. Dafür wurde das Statistik-Paket R (ein freier Clone von S-Plus) verwendet. Das Ergebnis des T-Tests ist als p-Wert angegeben, dieser entspricht dem Signifikanzniveau α.

Für jeden Versuch und jede Emotion wurde eine eigene Varianzanalyse durchgeführt, für die die Nullhypothesen jeweils lauteten: die Ausprägungen der unabhängigen Variablen haben keinen Einfluss auf die Bewertung der Stimuli als Emotion x.

Die durchschnittlichen Antworthäufigkeiten für alle drei Hörtests sind in Tab. 7.1 dargestellt. Ungefähr ein Drittel aller Stimuli wurden als neutral klingend bewertet, ein Indiz dafür, das hierunter alle nicht-entscheidbaren Antworten fallen.

	Neutral	Angst	Freude	Langeweile	Trauer	Wut
Versuch 1: Intonation und Phonation	33,9 %	17 %	7,4 %	14 %	20,7 %	7 %
Versuch 2: Intonation und segmentelle Eigenschaften	31,6 %	12,2 %	9,9 %	15,9 %	20,2 %	10,2 %
Versuch 3: segmentelle Eigenschaften und Phonation	32.4 %	14.3 %	7.4 %	16.7 %	15.4 %	13.6 %

Tabelle 7.1: Antworthäufigkeiten in Prozent

Im Folgenden werden die Ergebnisse aller Versuche für jede Emotion getrennt betrachtet. Dabei sind Abbildungen für Interaktionen zweiter Ordnung jeweils für die entsprechende Emotion einzeln gegeben. Die Darstellungen der Ergebnisse von Haupteffekten sind für alle relevanten[2] Emotionen zusammengefasst.

Da drei Merkmalsgruppen in verschiedenen Kombinationen untersucht wurden, gab es natürlich jeweils zwei mögliche Ergebnisse für Haupteffekte. Es wurden jeweils die berücksichtigt, die ein höheres Signifikanzniveau erreichten[3]. Erreichte ein Haupteffekt (oder eine Interaktion) nur bei einem der Versuche eine Signifikanz, so ist das im Text vermerkt.

Die Werte geben jeweils die durchschnittliche Antworthäufigkeit in Prozent an, d.h. ein Wert von 29,6 bei Angst und angehobener Grundfrequenz besagt z.B., dass 29,6 % der Stimuli mit angehobener Grundfrequenz als ängstlich beurteilt wurden. Um die Ergebnisse vergleichbarer zu machen, sind alle Abbildungen mit einer Ordinate bis zu 60 % dargestellt.

7.4.1 Angst

Für den Eindruck der Stimuli als ängstlich wurden die Haupteffekte F_0-Lage, F_0-Range, Lautdauern und Phonation signifikant. Als Interaktionen zweiter Ordnung erreichten weiterhin F_0-Lage/F_0-Range, F_0-Lage/Phonation, F_0-Lage/Lautdauern und Lautdauern/Phonationsart ebenfalls Signifikanz.

Effekte der F_0-Lage ($\alpha < .001$)

Ein Anheben der F_0-Lage (siehe Abb. 7.21) wurde von den Hörern als günstiger für ängstliche Sprechweise beurteilt als eine abgesenkte oder neutrale F_0-Lage ($\alpha < .001$). Weiterhin führte eine Absenkung der F_0-Lage dazu, die Stimuli fast nie als ängstlich wahrzunehmen ($\alpha < .001$). Dieses Ergebnis geht mit Analysen von Produktionsdaten (z.B. Paeschke et al. (1999) oder

[2]Es wurden nur signifikante Ergebnisse mit $\alpha < .05$ betrachtet. Alle Ergebnisse mit einem Signifikanzniveau von $\alpha < .05$ werden als signifikant und Ergebnisse mit einem Signifikanzniveau von $\alpha < .01$ werden als hoch signifikant bezeichnet.

[3]Daraus erklärt sich auch, dass sich die Mittelwerte für ein Merkmal nicht immer zu 100 % addieren, da sie eventuell aus unterschiedlichen Versuchen stammen

Banse & Scherer (1996)), nach denen Angst zu einem starken Anstieg der Grundfrequenz führt, absolut konform.

Effekte des F_0-Range ($\alpha < .001$)

Stimuli mit breiterem Range werden von Hörern öfter mit dem emotionalen Sprecherzustand Angst in Verbindung gebracht als solche mit neutralem ($\alpha = .021$) und solche mit schmalerem Range ($\alpha < .001$) (siehe Abb. 7.22). War der Range schmaler, wurden die Stimuli seltener als ängstlich bewertet als bei neutralem Range ($\alpha = .003$). Auch dieses Resultat entspricht früheren Ergebnissen aus der Literatur. Verglichen mit dem Effekt der F_0-Lage scheint die Auswirkung aber nicht so groß zu sein. Während 29 % der Stimuli mit angehobener F_0 als ängstlich beurteilt wurden, trifft dies nur auf 16 % derer mit breiterem Range zu.

Effekte der Phonationsarten ($\alpha < .001$)

Für die Manipulation der Phonationsart ergab sich eine klare Präferenz der Hörer für eine ängstliche Sprechweise bei Falsettphonation ($\alpha < .001$) (siehe Abb. 7.25). Der Mittelwert ist sogar so hoch (fast 50 % aller Stimuli mit Falsettanregung wurden als ängstlich beurteilt), dass dem Faktor Phonationsart eine besondere Relevanz bei der Beurteilung von gesprochener Sprache als ängstlich zuzuschreiben ist. Dieses Ergebnis bestätigt Messungen von Klasmeyer & Sendlmeier (2000). Da Falsett allerdings auch zu einem Anstieg der Irregularität in der Anregung führt (vgl. Abschnitt 6.3), werden ebenso die Resultate von Williams & Stevens (1972); Murray & Arnott (1995) und Murray & Arnott (1993) bestätigt.

Das Resultat erklärt sich weiterhin aus der Tatsache, dass eine Änderung der Phonationsart zu Falsett mit einer Rate von 70 % zu einem Anstieg der Grundfrequenz um ebenfalls 70 % führt (vgl. Abschnitt 6.3). Dass eine Anhebung der F_0-Lage für Angst günstig ist, hatte sich bereits bestätigt. Die Anhebung war aber dabei nur um 30 % erfolgt. Offensichtlich führt auch eine deutlich stärkere Anhebung der Grundfrequenz zu einer Steigerung ängstlicher Eindruckswirkung.

Effekte der Lautdauern ($\alpha < .01$)

Für die Dauerveränderungen ergab sich, dass das gemischte Modell, bei dem unbetonte Silben schneller und betonte langsamer waren, eher als ängstlich attribuiert wird als neutrale Lautdauern ($\alpha < .001$) oder langsamere Sprechweise ($\alpha < .01$) (siehe Abb. 7.23). Entgegen der Erwartung wurde also generell schnellere Sprechweise nicht signifikant als ängstlicher beurteilt als neutrale. Allerdings führt auch das gemischte Modell zu einer Anhebung der Sprechrate, da ja die Anzahl der unbetonten Silben überwiegt.

Dieses Modell war für Emotionen konzipiert, bei denen der Sprecher zwar starker Erregung ausgesetzt ist (die generell mit höherer Sprechgeschwindigkeit korreliert), ihm aber die Verständlichkeit seiner Aussagen wichtig genug erscheint, um Inhaltswörter deutlicher und damit langsamer zu artikulieren. Ein solcher Zustand mag für Angst durchaus adäquat erscheinen. Auch bei Bergmann et al. (1988, S. 187) führte eine Dehnung der phrasenbetonenden Silben zu einer Steigerung ängstlicher Eindruckswirkung. Zudem lässt sich das Resultat auch negativ

werten (vor allem im Hinblick auf die durchweg eher niedrigen Mittelwerte). Demnach führt neutrale und langsamere Sprechweise sehr viel seltener zu ängstlicher Attribuierung als erhöhte Sprechgeschwindigkeit.

Effekte des F_0-Range in Abhängigkeit von der F_0-Lage ($\alpha < .001$)

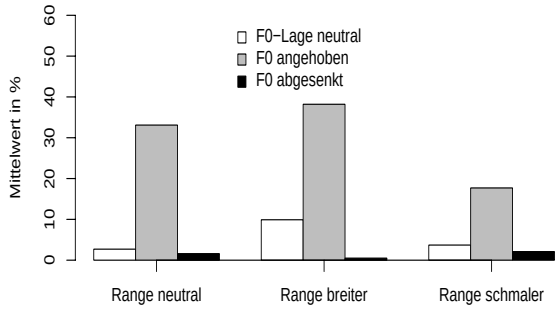

Abbildung 7.1: Mittelwerte der F0-Range-Varianten in Abhängigkeit von der F0-Lage für Angst

Die Mittelwerte der F_0-Range-Varianten in Abhängigkeit von der F_0-Lage (siehe Abb. 7.1) deuten auf eine Prominenz der F_0-Lage hin. Eine angehobene Grundfrequenz wirkt ängstlicher bei neutralem oder breitem Range als bei schmalem ($p < .001$). Ein Unterschied zwischen neutralem und breiterem Range wird dabei allerdings nicht mehr signifikant. Dies lässt sich vielleicht dadurch erklären, dass bei Falsettanregung und angehobener Grundfrequenz ein breiterer Range perzeptiv kaum noch bemerkbar ist.

Effekte der Phonationsarten in Abhängigkeit von der F_0-Lage ($\alpha < .01$)

Die Falsettphonation wirkt erwartungsgemäß bei angehobener Grundfrequenz deutlich ängstlicher als bei neutraler ($p = .012$) (siehe Abb. 7.2). 66 % dieser Stimuli wurden der Emotion Angst zugeschrieben. Dies ist ein weiteres Indiz dafür, dass eine extrem erhöhte Stimmlage ängstlich wirkt. Die F_0-Lage dieser Stimuli wurden zunächst um 30 % erhöht und dann noch einmal um 70 %, was einer Gesamtanhebung um 121 % entspricht. Dieses Ergebnis unterstreicht das Resultat aus dem Hörexperiment zu Merkmalen der Prosodie (Kap. 5), wo unter anderem eine Anhebung der Grundfrequenz um 150 % zu einer Erkennungsrate der ängstlich intendierten Stimuli von rund 80 % geführt hat.

Zunächst scheint das Merkmal Phonationsart wichtiger als die durchschnittliche Grundfrequenz. Stimuli mit Falsettphonation und neutraler F_0-Lage wurden sehr viel häufiger als ängstlich attribuiert als solche mit behauchter Phonation und angehobener Grundfrequenz ($p < .001$).

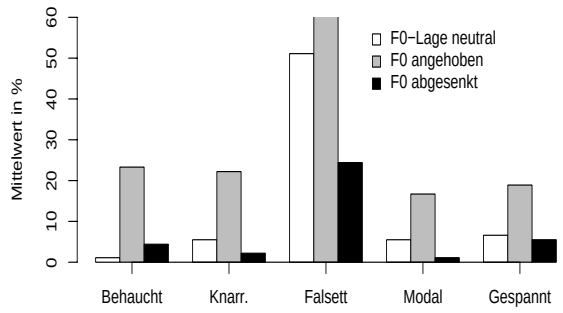

Abbildung 7.2: Mittelwerte der Phonations-Varianten in Abhängigkeit von der F0-Lage für Angst

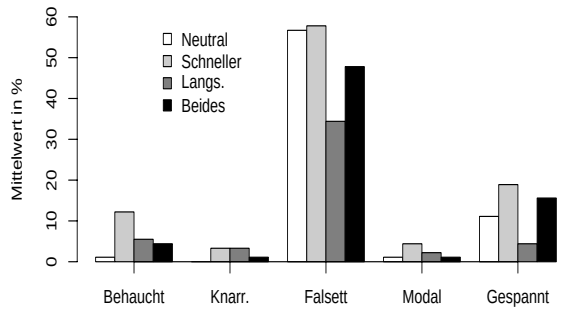

Abbildung 7.3: Mittelwerte der Phonations-Varianten in Abhängigkeit von den Dauer-Varianten für Angst

Andererseits ist es möglich, dass die für Angst günstige Wirkung der Merkmalsausprägung Falsett vor allem auf die angehobene F_0-Lage zurückzuführen ist, und diese war ausgeprägter als beim Merkmal F_0-Lage selber.

Effekte der Phonationsarten in Abhängigkeit von den Dauervarianten ($\alpha = .037$)

Auch in Abhängigkeit von den Dauermodellen zeigt sich die Prominenz von Falsettphonation für ängstliche Eindruckswirkung (siehe Abb. 7.3). Die Eindruckswirkung verstärkt sich allerdings bei schnellerer Sprechweise gegenüber langsamerer ($p < .01$) und auch bei dem gemischten Modell ($p < .05$). Ein Unterschied zwischen neutral, schnellerer und gemischter Sprech-

weise ist bei Falsettanregung nicht signifikant.

Effekte der F_0-Lage in Abhängigkeit von den Dauervarianten ($\alpha = .048$)

Es zeigt sich, dass die F_0-Lage wichtiger für ängstliche Eindruckswirkung ist als die Lautdauern (siehe Abb. 7.4). Immerhin wirkt die angehobene Grundfrequenz ängstlicher bei dem gemischten Dauermodell als bei neutraler Sprechrate ($p = .001$). Der Unterschied zwischen der gemischten Variante und allgemein kürzeren Lautdauern bei angehobener F_0 ergab lediglich eine Tendenz ($p = .09$) dafür, dass eine Verlängerung betonter Silben bei gleichzeitiger Verkürzung aller anderen Silben ängstlicher als eine allgemeine Verkürzung wirkt.

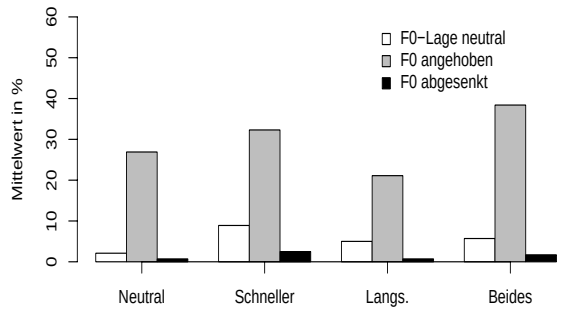

Abbildung 7.4: Mittelwerte der F0-Lage in Abhängigkeit von den Dauer-Varianten für Angst

Zusammenfassung der Ergebnisse für Angst

Insgesamt wurden für alle Merkmale die in der Literatur gefundenen Resultate bestätigt. Damit zeigt sich ein deutliches Profil einer Emotion, deren Eindruck vor allem durch stark erhöhte Grundfrequenz und Irregularitäten im Anregungssignal geprägt ist. Weiterhin führte ein breiterer Range sowie schnellere Sprechweise, bei der allerdings die betonten Silben verlängert sind, zu einer deutlichen Steigerung ängstlichen Sprecherausdrucks.

Von speziellem Interesse ist die günstige Wirkung des gemischten Dauermodells für ängstliche Attribuierung, da diese Merkmalsausprägung in der Literatur bisher wenig berücksichtigt wurde.

Das Merkmal Vokalqualität wurde hier zwar nicht signifikant, aber die Hinweise aus der Literatur sind ohnehin widersprüchlich: während Murray & Arnott (1995) eine präzisere Artikulation zugrunde legen, geht Kienast (1998) von starker segmenteller Reduktion aus. Eventuell ist ein Merkmal wie Artikulationsgenauigkeit entweder für ängstliche Eindruckswirkung nicht entscheidend oder es ist nicht genügend differenziert worden.

7.4.2 Freude

Der emotionale Sprecherzustand Freude wurde bei dieser Testreihe am seltensten attribuiert (vgl. Tab. 7.1). Dies schlägt sich in relativ geringen Antworthäufigkeiten nieder. Trotzdem erreichten alle Haupteffekte ein Signifikanzniveau von unter 5 % und es lassen sich zumindest Aussagen über Parameterausprägungen ableiten, die deutlich nicht freudig wirken.

Effekte der F_0-Lage ($\alpha = .035$)

Das Merkmal F_0-Lage ergab für Freude nur beim ersten Versuch (in Kombination mit Phonationsart) eine signifikante Beurteilung (siehe Abb. 7.21). Demnach wirkt eine abgesenkte Grundfrequenz weniger freudig als eine neutrale ($\alpha = .048$) und als eine angehobene ($\alpha = .017$).

Es gilt als allgemein gesicherte Erkenntnis, dass Freude zu einem Anstieg der F_0 führt (vgl. Abschnitt 4.7). Dass dies in diesem Test nicht signifikant wurde, mag daran liegen, dass dieser Anstieg der Grundfrequenz im Prinzip für alle Emotionen mit hohem Erregungsgrad gilt, bei Freude nun aber gerade auch andere Merkmale eine wichtige Rolle spielen. Die generell niedrigen Antworthäufigkeiten für Freude lassen darauf schließen, dass für freudigen Sprechausdruck wichtige Parameter bzw. Ausprägungen (wie z.B. bestimmte Intonationskonturen oder artikulatorische Eigenschaften) bei dieser Testreihe nicht berücksichtigt wurden.

Effekte des F_0-Range ($\alpha < .001$)

Die Unterschiede der F_0-Range-Varianten erreichten für Freude alle ein Signifikanzniveau von unter oder gleich .001 (siehe Abb. 7.22). Das Ergebnis geht auch mit sämtlichen Aussagen aus der Literatur konform: ein breiterer Range wirkt freudiger und ein schmaler unfreudiger. Da diese Ausprägungen vor allem Kennzeichen starker Erregung sind, gelten sie allerdings ebenso für Angst und Wut.

Zur Manipulation der F_0-Lage und des Range ist zu bemerken, dass sowohl die Anhebung als auch die Verbreiterung eher moderat waren, während diese Merkmale laut Literatur deutlich stärker ausgeprägt sein müssten. Aus forschungsökonomischen Gesichtspunkten heraus wurde aber auf eine mehrstufige Anhebung der F_0-Lage verzichtet.

Effekte der Phonationsarten ($\alpha = .001$)

Die Auswirkungen der Phonationsarten auf freudigen Sprechausdruck erreichten nur beim ersten Versuch (in Kombination mit Merkmalen der Intonation) einen signifikanten Effekt (siehe Abb. 7.25). Demnach wirkt modale Phonation eher freudig als behauchte oder knarrende ($\alpha <= .001$) und Falsett ($\alpha = .016$). Weiterhin wirkt auch eine gespannte Phonation eher freudig als behauchte oder knarrende Anregung ($\alpha = .002$). Der stärkere Eindruck von Freude bei modaler und gespannter Phonation bestätigt die Vorhersagen von Klasmeyer & Sendlmeier (2000). Einen Widerspruch zur Literatur stellt die ausgesprochen unfreudige Bewertung behauchter Stimmqualität dar. Allerdings ist behauchte Anregung ein Kennzeichen entspannter Muskulatur und mag damit ein Korrelat der Valenzdimension sein, aber nur bei fehlender Erregung (z.B. Wohlbehagen).

Effekte der Lautdauern ($\alpha < .001$)

Der Einfluss der Sprechrate auf die Attribuierung der Stimuli als freudig ist in Abb. 7.23) dargestellt. Demzufolge wirkt eine erhöhte Sprechrate freudiger als eine geringere ($\alpha < .001$), neutrale ($\alpha = .002$) und auch als die Version, bei der die betonten Silben verlängert und alle anderen verkürzt wurden ($\alpha = .009$).

Langsamere Sprechweise wirkt weniger freudig als alle anderen Varianten ($\alpha < .001$). Dies wiederspricht Ergebnissen aus der Literatur (Öster & Riesberg (1986) und Carlson et al. (1992)). Andere Quellen gehen jedoch von einer Erhöhung der Sprechrate aus (z.B. Banse & Scherer (1996); Scherer (1995) oder Fonagy (1981)). Da aber Sprechgeschwindigkeit stark mit der Erregungsdimension korreliert, mag hier dasselbe wie bei der behauchten Phonation gelten: langsamere Sprechweise kann eventuell mit Wohlbehagen einhergehen, aber sicherlich nicht mit starker Freude.

Das gemischte Modell, dass bereits in Burkhardt & Sendlmeier (1999a) für Freude bewertet wurde, erreichte hier keine Steigerung freudiger Eindruckswirkung gegenüber neutraler Sprechweise. Eventuell ist eine generell erhöhte Sprechweise (die dort nicht bewertet wurde), eher geeignet, freudige Eindruckswirkung zu erzeugen.

Effekte der Vokalqualität ($\alpha = .004$)

Manipulationen der Vokalqualität brachten nur beim zweiten Versuch, im Zusammenhang mit Intonationsvariationen, eine signifikante Änderung der Hörerurteile bzgl. freudiger Sprechweise (siehe Abb. 7.24). Beim dritten Hörtest, bei dem Vokalqualität unter anderem mit unterschiedlichen Phonationsarten kombiniert wurde, ergab sich dafür lediglich eine Tendenz.

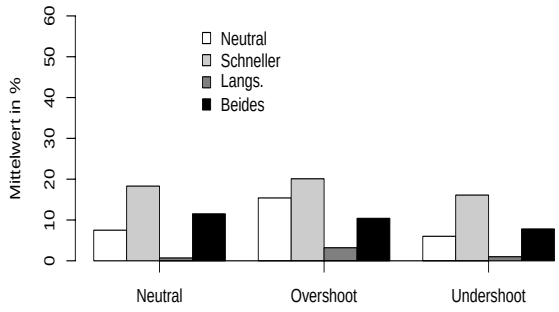

Abbildung 7.5: Mittelwerte für die Vokalqualität in Abhängigkeit von den Dauervarianten für Freude ($\alpha = .098$)

Die Unterschiede besagen, dass Sprechweise eher freudig wirkt, wenn die Vokale der betonten Silben elaboriert artikuliert werden als bei neutraler Sprechweise ($\alpha = .025$). Folgerichtig

wirkt eine auf die Qualität des [ə] hin zentralisierte Sprechweise der Vokalqualitäten (Formant-Target Undershoot) ebenso deutlich unfreudiger wie eine explizitere Sprechweise ($\alpha = .001$). Die Erwartung ging dahin, dass ein Formant-Target Overshoot auf betonten Silben zusammen mit dem gemischten Dauermodell einen signifikanten Effekt auf emotionale Eindruckswirkung haben würde, da Overshoot bei Emotionen mit hohem Erregungsgrad eher auftritt und sich das gemischte Dauermodell bei Burkhardt & Sendlmeier (1999a) als günstig für freudige Eindruckswirkung gezeigt hat. Zwar ergab der kombinierte Einfluss von Lautdauern und Vokalqualität tendenziell einen Unterschied für die Beurteilung der Stimuli als freudig ($\alpha = .098$) (siehe Abb. 7.5), aber nicht hinsichtlich einer Steigerung freudiger Eindruckswirkung bei der Kombination von Formant-Target Overshoot und gemischter Dauervariante.

Effekte der Phonationsarten in Abhängigkeit von den Dauervarianten ($\alpha = .033$)

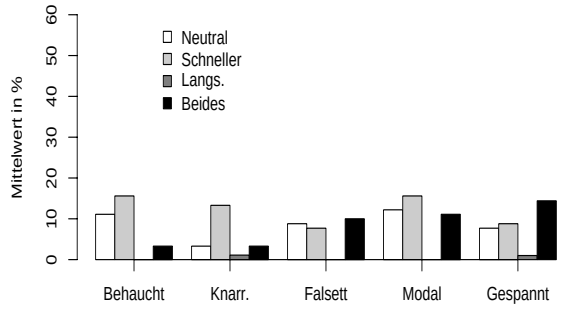

Abbildung 7.6: Mittelwerte der Phonationsarten in Abhängigkeit von den Dauervarianten für Freude

Der Zusammenhang zwischen Phonationsart und Lautdauern ist in Abb. 7.6 dargestellt. Der Faktor Sprechrate scheint wichtiger für freudigen Sprechausdruck zu sein als die Phonationsart. Schnellere Sprechweise wirkt z.B. bei behauchter Phonation genauso freudig wie bei gespannter Stimme. Die Daten zeigen auch, dass das gemischte Dauermodell bei behauchter Phonation weniger freudig wirkt als bei unveränderten Lautdauern ($p = .051$). Zudem wirkt schnellere Sprechweise außer bei behauchter und knarrender Phonation nicht freudiger als neutrale oder das gemischte Modell.

Effekte des F_0-Range in Abhängigkeit von den Dauervarianten ($\alpha = .001$)

Der Einfluss des F_0-Range scheint eine ähnliche starke Bedeutung auf freudige Eindruckswirkung zu haben wie der Einfluss der Dauermodelle (siehe Abb. 7.7). Einerseits führt der Range bei langsamerer Sprechweise nicht mehr zu einer Beeinflussung der Hörerurteile, andererseits

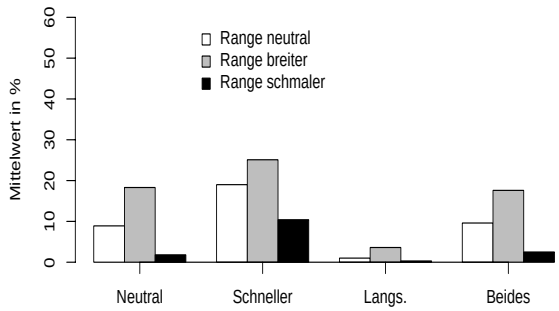

Abbildung 7.7: Mittelwerte für den F0-Range in Abhängigkeit von den Dauervarianten für Freude

bleiben die Mittelwertsrelationen für den Range bei anderen Dauermodellen annähernd erhalten.

Effekte des F_0-Range in Abhängigkeit von der Phonationsart ($\alpha = .005$)

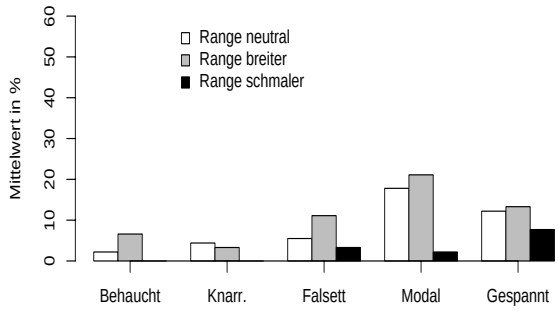

Abbildung 7.8: Mittelwerte für den F0-Range in Abhängigkeit von der Phonationsart für Freude

In Interaktion mit der Phonationsart wird der Unterschied zwischen breiterem und neutralem Range nicht mehr signifikant (siehe Abb. 7.8). Ist die Phonationsart gespannt, führt eine Änderung des F_0-Range nicht mehr zu einer Änderung des Beurteilerverhaltens für freudige Eindruckswirkung. Allerdings sind die Mittelwerte insgesamt so gering, dass die Resultate schwer zu interpretieren sind. Sogar bei Falsettanregung, bei der die Mittelwertsdifferenz am

größten ist, gibt es lediglich eine Tendenz dafür, dass ein breiterer Range freudiger wirkt als ein neutraler ($p = .071$).

Zusammenfassung der Ergebnisse für Freude

Schon die Tatsache, dass Freude generell die niedrigsten Antworthäufigkeiten hatte, lässt erahnen, dass bei dieser Testreihe Merkmale bzw. Merkmalsausprägungen nicht berücksichtigt wurden, die für freudigen Sprecherausdruck wichtig sind. So wurde z.B. die Intonationskontur an sich nicht verändert, sie wurde lediglich als Ganzes verschoben und in der Variationsbreite gestaucht oder gestreckt. Weiterhin wurden wichtige Merkmale wie die höhere Lage der unteren Formanten, verursacht durch lächelnde Sprechweise, nicht modelliert.

Dennoch haben sich ein breiterer Range und eine schnellere Sprechweise als für freudige Eindruckswirkung günstig bewährt. Eine neue Erkenntnis stellt die bessere Bewertung von Stimuli mit Formant-Target Overshoot dar. Dieser Effekt war allerdings vom Mittelwert her nicht sehr ausgeprägt und bedarf weiterer Überprüfung.

7.4.3 Langeweile

Für Langeweile ergaben sich ebenfalls für alle variierten Merkmale signifikante Unterschiede bzgl. der Hörerurteile. Weiterhin wurde die Interaktion der Dauerveränderungen mit den Merkmalen Phonationsart, Vokalqualität und F_0-Lage signifikant. Auch die Interaktion zwischen F_0-Range und Phonationsart sowie F_0-Lage führte zu einem deutlich veränderten Urteilsverhalten.

Effekte der F_0-Lage ($\alpha < .001$)

Eine abgesenkte F_0-Lage wirkt eher gelangweilt als eine neutrale ($\alpha < .024$) (siehe Abb. 7.21). Ist die Grundfrequenz angehoben, so wirkt der Stimulus seltener gelangweilt als bei neutraler ($\alpha < .001$) oder abgesenkter F_0-Lage ($\alpha < .001$). Dieses Ergebnis entspricht der dimensionalen Einordnung von Langeweile als eine Emotion, die mit geringer Erregung verbunden ist. Auch frühere Studien kamen zu ähnlichen Resultaten (vgl. Abschnitt 4.7).

Effekte des F_0-Range ($\alpha < .001$)

Die Mittelwerte der Hörerurteile für die F_0-Range-Veränderungen stimmen mit den Ergebnissen von Bergmann et al. (1988) überein (siehe Abb. 7.22). Sämtliche Unterschiede zwischen den drei Modellen erreichten eine Signifikanz von $\alpha < .001$. Stimuli mit schmalem Range werden öfter als gelangweilt empfunden als solche mit neutralem Range und diese wirken eher gelangweilt als solche mit größerer Tonhöhenvariabilität.

Effekte der Phonationsarten ($\alpha < .001$)

Behauchte Anregung ist für gelangweilte Eindruckswirkung günstiger als Falsett ($\alpha < .001$), modale Phonation ($\alpha = .034$) und gespannte Stimme ($\alpha = .001$) (siehe Abb. 7.25). Knarrstimme führt ebenfalls eher zu gelangweilter Wirkung als Falsett ($\alpha < .001$), modale Phonation

($\alpha = .002$) und gespannte Stimme ($\alpha < .001$). Die Falsettphonation führt weiterhin weitaus seltener zu einer Beurteilung der Stimuli als gelangweilt als modale Anregung ($\alpha < .001$).

Den Ergebnissen folgend sind also Knarrstimme und behauchte Stimme günstiger für gelangweilte Eindruckswirkung als andere Phonationsarten, während Stimuli mit Falsettanregung oder gespannter Stimme äußerst selten als gelangweilt beurteilt werden. Hierzu sind dem Autor zwar keine früheren Untersuchungen bekannt, aber da behauchte und die hierbei implementierte (vgl. Abschnitt 6.3.4) Knarrstimme die Phonationsarten mit dem geringsten *vocal effort* sind, passt dies zu der Tatsache, dass Langeweile eine Emotion mit sehr geringer Erregung ist. Klasmeyer & Sendlmeier (2000) bezeichnen zwar die Phonationsart als modal, messen aber geringere Energie in den höheren Frequenzbereichen, was sowohl für behauchte Anregung als auch für Knarrstimme gegeben ist.

Effekte der Lautdauern ($\alpha < .001$)

Der Einfluss der Lautdauervariation auf die Beurteilung der Stimuli als gelangweilt besagt, dass langsamere Sprechweise gelangweilter wirkt als alle anderen Modelle (jeweils $\alpha < .001$). Folgerichtig wirkt schnellere Sprechweise am wenigsten gelangweilt (ebenfalls jeweils $\alpha < .001$). Auch das gemischte Modell wirkt noch weniger gelangweilt als neutrale Lautdauern ($\alpha = .013$). Zwar sind von einer Verlängerung der Silben laut Literatur bei gelangweilter Sprechweise vor allem die betonten Silben betroffen, doch scheint eine Verkürzung der restlichen Silben einem Eindruck von Langeweile auf jeden Fall abträglich zu sein.

Effekte der Vokalqualität ($\alpha < .001$)

Langeweile ist neben Freude die einzige Emotion, bei der die Variation der Vokalqualität zu einer signifikanten Beeinflussung der Hörerurteile geführt hat (siehe Abb. 7.24). Danach führt Formant-Target Undershoot eher zu einer gelangweilten Eindruckswirkung als normale Vokalartikulation ($\alpha < .001$) und als Overshoot ($\alpha < .001$). Dieses Ergebnis unterstützt die Ergebnisse von Klasmeyer & Sendlmeier (2000) und Kienast & Sendlmeier (2000), laut der die Artikulation hier eher ungenau ist und unterstreicht die geringe Erregung, die mit Langeweile einhergeht.

Effekte der Phonationsarten in Abhängigkeit vom F_0-Range ($\alpha = .012$)

In Abb. 7.9) sind die Mittelwerte der Phonationsarten in Abhängigkeit vom F_0-Range dargestellt. Es zeigt sich deutlich die Präferenz für schmaleren Range mit behauchter Anregung oder Knarrstimme. Dass die Verschiebung der F_0-Lage in diesem Zusammenhang nicht mehr signifikant geworden ist, mag daran liegen, dass die Knarrstimme ohnehin bereits zu einer erheblichen Absenkung der F_0 führt.

Effekte der Phonationsarten in Abhängigkeit von den Dauermodellen ($\alpha < .001$)

Im Zusammenhang mit den Phonationsart-Variationen zeigt sich nochmals die Bedeutung langsamerer Sprechweise auf gelangweilte Eindruckswirkung (siehe Abb. 7.10). Bei verringertem

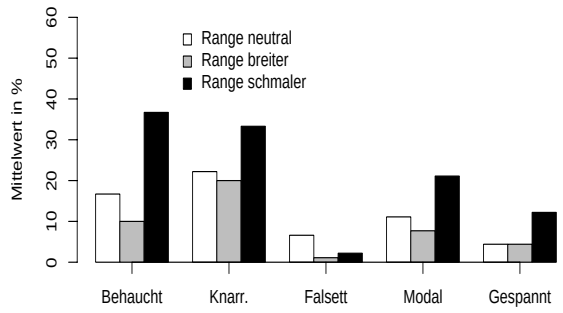

Abbildung 7.9: Mittelwerte der Phonationsarten in Abhängigkeit vom F0-Range für Langeweile

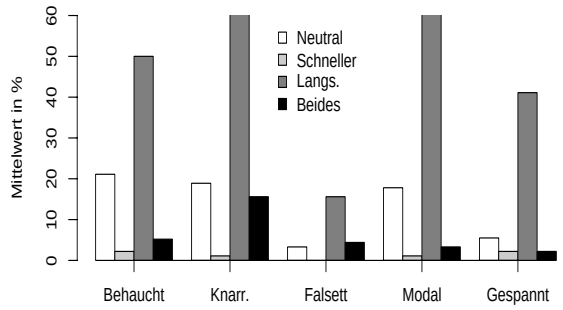

Abbildung 7.10: Mittelwerte der Dauermodelle in Abhängigkeit von der Phonationsart für Langeweile

Sprechtempo spielt es für die Beurteilung der Stimuli als gelangweilt keine Rolle, ob die Phonationsart behaucht, knarrend oder modal ist.

Effekte der Dauermodelle in Abhängigkeit von der Vokalqualität ($\alpha < .001$)

Eine Signifikanz für den Einfluss der Lautdauern in Abhängigkeit von der Vokalqualität ergab sich nur beim zweiten Versuch, in Kombination mit den Intonationsvariationen (siehe Abb. 7.11). Es zeigt sich, dass langsamere Sprechweise bei allen Varianten der Vokalqualität zu einem stärkeren Höreindruck von Langeweile führt, wobei dies bei Formant-Target Undershoot aber deutlicher ist als bei neutraler Artikulationsgenauigkeit oder bei Overshoot.

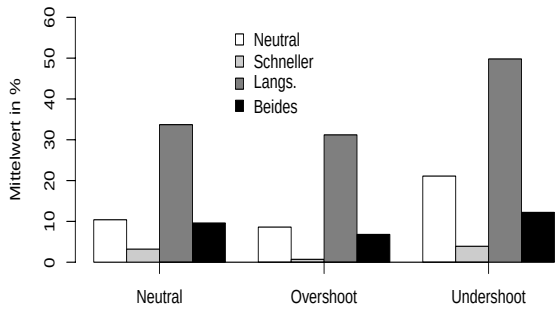

Abbildung 7.11: Mittelwerte der Dauermodelle in Abhängigkeit von der Vokalqualität für Langeweile

Effekte der F_0-Lage in Abhängigkeit von den Dauermodellen ($\alpha < .001$)

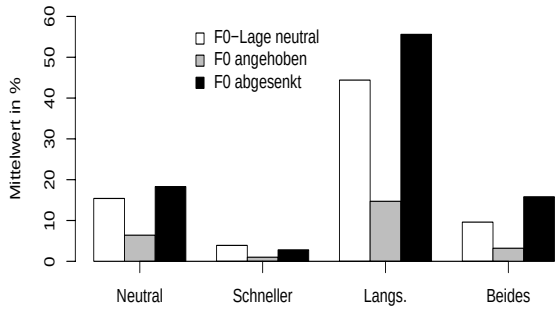

Abbildung 7.12: Mittelwerte der F0-Lage in Abhängigkeit von den Dauermodellen für Langeweile

Bei langsamerer Sprechweise ($p = .005$) und in Kombination mit dem gemischten Modell ($p = .024$) führt eine abgesenkte F_0-Lage zu einer Erhöhung gelangweilter Eindruckswirkung gegenüber neutraler F_0-Lage ($p = .005$), während dies bei den anderen Dauermodellen nicht der Fall ist (siehe Abb. 7.12). Dieses Ergebnis unterstützt die Aussagen der Untersuchung der Haupteffekte.

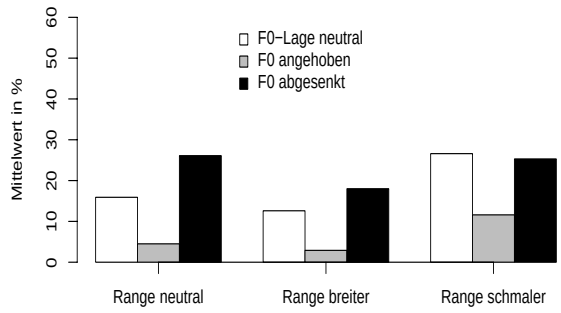

Abbildung 7.13: Mittelwerte der F0-Lage in Abhängigkeit vom F0-Range für Langeweile

Effekte der F_0-Lage in Abhängigkeit vom F_0-Range ($\alpha = .017$)

Die Mittelwerte für Effekte der F_0-Lage in Abhängigkeit vom Range sind in Abb. 7.13) dargestellt. Es zeigt sich, das eine abgesenkte F_0-Lage zwar bei neutralem ($p < .001$) und breiterem ($p = .015$) Range eher gelangweilt wirkt, nicht aber bei schmalerem. Andererseits wirkt ein schmalerer Range nicht bei abgesenkter Grundfrequenz gelangweilter als ein neutraler. Die beiden Faktoren scheinen sich in ihrer Wirkung auf die Attribuierung der Stimuli als gelangweilt nicht zu addieren, sondern es reicht, wenn eines der Merkmale eine für Langeweile günstige Ausprägung aufweist.

Zusammenfassung der Ergebnisse für Langeweile

Langeweile wurde bei früheren Untersuchungen oft nicht berücksichtigt, da sie nicht zu den klassischen vier Basisemotionen zählt. In dieser Versuchsreihe hat sich ein deutliches akustisches Profil gezeigt, das vor allem durch Merkmale geringer Erregung geprägt ist. Dazu zählen eine behauchte Anregung, ein schmaler F_0-Range und abgesenkte F_0-Lage, weiterhin eine verlangsamte Sprechweise und eine eher undeutliche Vokalartikulation.

Eventuell könnte dieses Profil auch auf *Müdigkeit* zutreffen, wobei dieser Begriff in keiner dem Autor bekannten Studie abgefragt wurde. Oft hat die Kategorie Langeweile auch eine ärgerliche Komponente (*genervt sein*), was zu einer teilweisen Verwischung dieses Profils geführt haben mag (z.B. dass eine abgesenkte F_0-Lage in Kombination mit schmalerem Range nicht gelangweilt wirkte).

7.4.4 Trauer

Für Trauer waren die Haupteffekte F_0-Lage und Range, Phonationsart und Lautdauer eine signifikant. Die Interaktionen der Veränderungen der Sprechrate mit allen anderen Merkmalen sowie

die Interaktion zwischen F_0-Lage und Range erbrachten ebenfalls eine signifikante Änderung im Beurteilerverhalten.

Effekte der F_0-Lage ($\alpha < .001$)

Die Signifikanz des Haupteffekts F_0-Lage ergab sich nur beim zweiten Test, in Kombination mit segmentellen Eigenschaften (siehe Abb. 7.21). Demzufolge ist eine angehobene F_0-Lage günstiger für traurige Eindruckswirkung als eine neutrale ($\alpha < .001$) oder eine abgesenkte ($\alpha < .001$). Die abgesenkte Variante wirkt weniger traurig als die neutrale ($\alpha = .009$). Dies steht völlig im Gegensatz zu allen Ergebnissen aus der Literatur, offensichtlich wurde dort eine andere Art der Trauer beschrieben als die Beurteiler bei diesem Experiment erkannt haben. Da auch die Ergebnisse bei den Phonationsarten den Erwartungen widersprachen, wird das Thema dort näher diskutiert.

Effekte des F_0-Range ($\alpha = .001$)

Der Effekt des Merkmals F_0-Range wurde ebenfalls lediglich beim zweiten Experiment signifikant (siehe Abb. 7.22). Demzufolge wirkt ein schmaler Range trauriger als ein breiter ($\alpha < .001$) und tendenziell auch trauriger als ein neutraler ($\alpha = .075$). Darüber hinaus wirkt ein breiter Range seltener traurig als ein neutraler ($\alpha < .001$).

Obwohl die sonstigen Ergebnisse der Merkmalsvariation für Trauer den Hypothesen eher widersprechen, ist dies beim F_0-Range nicht der Fall. Dies mag als Indiz dafür gelten, dass der F_0-Range nicht in erster Linie mit der Aktivitätsdimension (Erregungsgrad) korreliert, sondern eher mit der Valenz- oder Potenzdimension.

Effekte der Phonationsarten ($\alpha = .002$)

Die Mittelwerte für die Attribuierung der Stimuli als traurig in Abhängigkeit von der Phonationsart sind in Abb. 7.25 dargestellt. Die Resultate der Signifikanztests besagen, dass modale Anregung weniger traurig wirkt als behauchte ($\alpha = .001$), knarrende ($\alpha = .023$) oder Falsett-Anregung ($\alpha < .001$). Damit wirkten im Prinzip alle Phonationsarten eher traurig als die modale, so unterschiedlich sie auch sein mögen. Es scheint klar, dass die Kategorie Trauer für dieses Experiment eine zu grobe Klassifizierung darstellt.

In der Literatur wird für traurige Sprechweise vor allem behauchte Anregung zugrunde gelegt, und in der Tat ergab sich die höchste Erkennung bei Behauchung für Trauer. Erstaunlich ist zunächst die häufige (an die 30 %) Nennung von Trauer bei Falsettanregung. Dieses Ergebnis, wie auch die Beurteilung einer angehobenen F_0-Lage als günstig für traurigen Sprecherausdruck, legt den Verdacht nahe, dass Falsettanregung eine in der Literatur selten beachtete Art der Trauer begünstigt.

Der Autor schlägt deshalb eine Unterscheidung zwischen *stiller* und *weinerlicher* Trauer vor (analog zur weitverbreiteten Unterscheidung zwischen *cold* und *hot anger*), mit jeweils unterschiedlichen akustischen Profilen. Dabei würden sich die beiden Arten ebenfalls vor allem durch ihre Erregung unterscheiden, wobei weinerliche Trauer oft mit Angst vermischt auftritt.

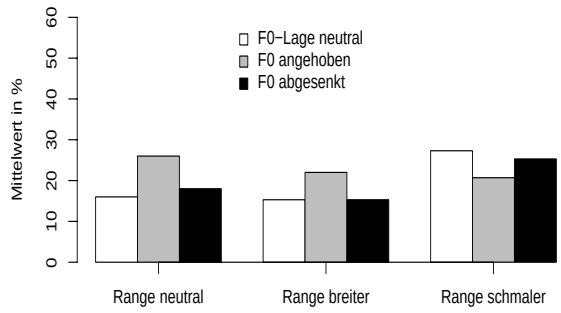

Abbildung 7.14: Mittelwerte des F0-Range in Abhängigkeit von der F0-Lage für Trauer

Damit ist sie im Gegensatz zu stiller Trauer durch eine angehobene F_0-Lage charakterisiert. Eventuell ist dabei auch Falsettanregung möglich.

Effekte der Lautdauern ($\alpha < .001$)

Die Ergebnisse für den Effekt der Dauervariationen auf die Erkennung als Trauer sind in Abb. 7.23 dargestellt. Sämtliche Unterschiede bis auf den zwischen neutralen Dauern und dem gemischten Modell, bei dem betonte Silben verlängert und alle anderen verkürzt wurden ($\alpha = .012$), erreichten dabei ein Signifikanzniveau von $\alpha < .001$.

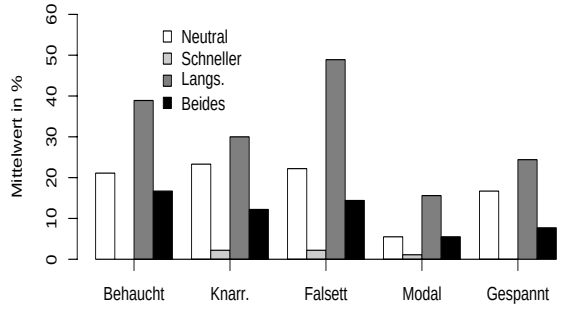

Abbildung 7.15: Mittelwerte der Phonationsarten in Abhängigkeit von den Dauervarianten für Trauer

Daraus ergibt sich ein direkter Zusammenhang zwischen abnehmender Sprechrate und zu-

nehmendem Eindruck von Traurigkeit. Die Ergebnisse bestärken erwartungsgemäß alle früheren Untersuchungen dahingehend, dass langsamere Sprechweise sehr viel eher traurig wirkt als schnellere.

Effekte des F_0-Range in Abhängigkeit von der F_0-Lage ($\alpha = .004$)

Die Auswirkungen von F_0-Lage und Range wurden zwar im ersten Experiment als Haupteffekte nicht signifikant, doch ihre Kombination erreichte durchaus eine hoch signifikante Änderung im Beurteilerverhalten bezüglich trauriger Sprechweise (siehe Abb. 7.14). Laut Hypothesen sollte ein schmalerer Range und eine abgesenkte F_0-Lage eher traurig wirken. und tatsächlich wirkt eine abgesenkte Grundfrequenz bei schmalem Range trauriger als bei breitem ($p = .015$) und tendenziell auch trauriger als bei neutralem F_0-Range ($p = .09$). Bei breiterem oder schmalerem Range wird der Einfluss der F_0-Lage allerdings nicht signifikant, lediglich bei neutralem Range gibt es Signifikanzen dafür, dass eine angehobene (!) F_0-Lage als trauriger (im Sinne einer weinerlichen Trauer) empfunden wird als neutrale ($p = .022$) und tendenziell auch als eine abgesenkte ($p = .083$). Dieses Resultat hatte sich auch im zweiten Hörversuch bei der Auswertung der F_0-Lage als Haupteffekt ergeben.

Wiederum scheinen mehrere Arten von Trauer mit unterschiedlichen akustischen Korrelaten ein signifikant eindeutiges Merkmalsprofil erreicht zu haben. Einige Testhörer berichteten von der Schwierigkeit, beim Hören der Stimuli zwischen Angst und Trauer zu unterscheiden.

Effekte der Phonationsarten in Abhängigkeit von den Dauervarianten ($\alpha = .004$)

Da die Ergebnisse der Wirkung der Dauervariation und der Phonationsart auf die Bewertung der Stimuli als traurig nur den Schluss zulassen, dass mindestens zwei Arten von Trauer (stille und weinerliche) identifiziert wurden, werden Hypothesen nun dazu einzeln verfolgt.

Der Eindruck von Trauer ist stärker bei behauchter Anregung, wenn die Sprechgeschwindigkeit langsamer ist als bei neutraler ($p = .004$) und schnellerer Sprechrate ($p < .001$). Weiterhin wirkt bei behauchter Phonation langsame Sprechweise trauriger als die gemischte Variante ($p < .001$) (siehe Abb. 7.15). Dies entspricht den Hypothesen für stille Trauer.

Der Unterschied zwischen langsamerer Sprechweise bei behauchter Phonation und Falsett ist zwar nicht signifikant, aber immerhin wurden 48 % der Stimuli mit niedriger Sprechrate und Falsettanregung im dritten Experiment als traurig klingend beurteilt. Da bei den Ergebnissen aus früheren Studien Falsettanregung für Trauer nicht charakteristisch war, ist zu vermuten, dass der Untersuchungsgegenstand dabei jeweils stille Trauer war.

Effekte der Dauermodelle in Abhängigkeit von der Vokalqualität ($\alpha = .023$)

Die Interaktion zwischen den Lautdauern und der Vokalqualität wurde nur im zweiten Experiment, in Kombination mit Merkmalen der Intonation, signifikant (siehe Abb. 7.16). Demnach wirkt Formant-Target Undershoot bei langsamer Sprechweise weniger traurig als Overshoot ($p = .017$) und als neutrale Artikulation ($p = .014$).

Dieses zunächst verwirrende Ergebnis lässt sich dadurch erklären, dass bei langsamer Sprechweise 75 % der Stimuli als traurig oder gelangweilt beurteilt wurden. Kam zu der lang-

samen Sprechweise nun auch noch eine Reduktion der Vokalqualität (Formant-Target Undershoot) hinzu, so wurde zu 49.8 % auf Langeweile entschieden, so dass sich bei dieser Kombination signifikant weniger Beurteilungen für die Kategorie ”traurig” zeigten. Eine Variation der Vokalqualität, zumindest in der hier implementierten Form, hat somit keinen Einfluss auf traurige Eindruckswirkung[4].

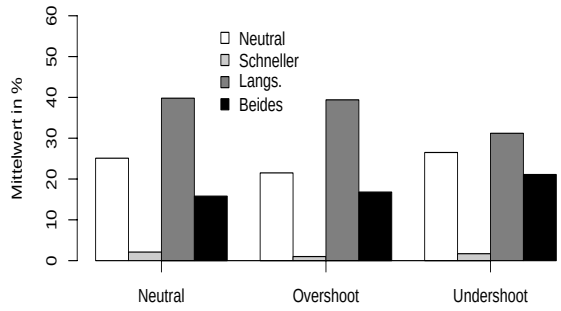

Abbildung 7.16: Mittelwerte der Dauermodelle in Abhängigkeit von der Vokalqualität für Trauer

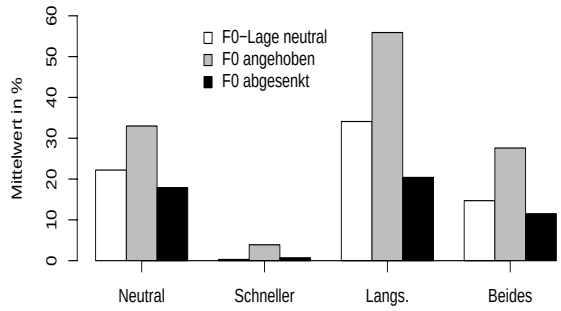

Abbildung 7.17: Mittelwerte der F0-Lage in Abhängigkeit von den Dauermodellen für Trauer

[4]Dies folgert auch aus der Tatsache, dass der Effekt der Vokalqualität an sich in keinem der beiden möglichen Experimente zu signifikanten Ergebnissen führte.

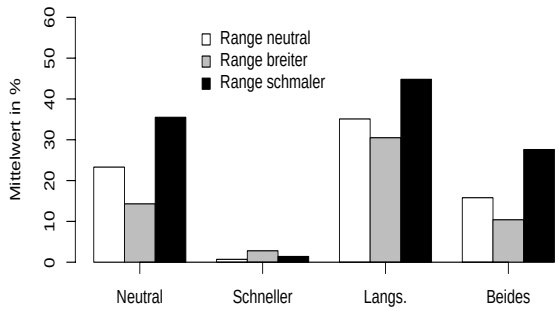

Abbildung 7.18: Mittelwerte des F0-Range in Abhängigkeit von den Dauermodellen für Trauer

Effekte der F_0-Lage in Abhängigkeit von den Dauermodellen ($\alpha < .001$)

Die Untersuchung der gemeinsamen Wirkung von F_0-Lage und den verschiedenen Dauermodellen ergibt, dass eine angehobene F_0-Lage bei langsamerer Sprechweise trauriger wirkt als eine neutrale ($p < .001$) und als eine abgesenkte Grundfrequenz ($p < .001$) (siehe Abb. 7.17). Als weiteres Resultat führte eine abgesenkte F_0-Lage bei niedrigem Sprechtempo entgegen der Hypothesen signifikant seltener zu einer Beurteilung des Stimulus als traurig als der neutrale Grundfrequenzverlauf ($p < .001$).

Das erste Ergebnis ist konsistent mit dem bisherigen Profil einer weinerlichen Trauer, die sich durch angehobene Grundfrequenz aber geringer Sprechrate auszeichnet. Das zweite Ergebnis lässt sich wiederum aus der Konkurrenzsituation erklären, die Trauer gegenüber Langeweile bei den Stimuli mit verlängerten Lautdauern hatte. Die Stimuli mit abgesenkter Grundfrequenz und langsamer Sprechweise wurden so häufig als gelangweilt bewertet, dass gegenüber denen mit neutraler F_0-Lage signifikant weniger als traurig klingend beurteilt wurden.

Effekte des F_0-Range in Abhängigkeit von den Dauermodellen ($\alpha = .001$)

Ein schmalerer Range wirkt bei langsamerer Sprechweise trauriger als ein neutraler ($p = .009$) und als ein breiterer F_0-Range ($p < .001$) (siehe Abb. 7.18). Dieses Resultat entspricht den Hypothesen und ist auch konsistent mit den Ergebnissen der Haupteffekte.

Zusammenfassung der Ergebnisse für Trauer

Die Analyse der Hörerurteile für Trauer stellt zunächst ein verwirrendes Bild dar, welches sich aber auflöst, wenn man Trauer in zwei Subkategorien aufteilt.

Gemeinsam ist diesen beiden eine verlangsamte Sprechweise und ein schmalerer Range. Weinerliche Trauer wird durch eine angehobene F_0-Lage und Falsettphonation charakterisiert und kann damit eher als Emotion mit relativ hohem Erregungsgrad gelten, während stille Trauer

eher von behauchter Phonation und dementsprechend von geringer Aktivierung gekennzeichnet ist.

7.4.5 Wut

Für Wut wurden lediglich zwei der Faktoren hoch signifikant: Phonationsart und Dauer. Die Interaktion zwischen diesen Faktoren wie auch die zwischen den Merkmalen Lautdauer und F_0-Lage erreichte ebenfalls Signifikanz.

Effekte der Phonationsarten ($\alpha < .001$)

In Abb. 7.25 sind die Mittelwerte für den Einfluss der Phonationsarten auf den Eindruck von Wut bzw. Ärger dargestellt. Die Signifikanztests zeigen, dass eine gespannte Stimme eher wütend wirkt als alle anderen Phonationsarten (jeweils $\alpha < .001$). Dieses begründet sich daraus, dass die gespannte Stimme bei der Aktivitätsdimension auf Erregung und bei der Potenzdimension auf Dominanz hinweist.

Es steht auch im Einklang mit anderen Ergebnissen: Merkmale wie schwacher erster Formant (Williams & Stevens (1972)) und stärkere Energie in höheren Frequenzbändern (etwa Banse & Scherer (1996)) sind bei gespannter Phonation gegeben. Eine gespannte Artikulation (Murray & Arnott (1995)) erhöht weiterhin die Wahrscheinlichkeit des Auftretens gespannter Phonation.

Effekte der Lautdauern ($\alpha < .001$)

Der Einfluss des Faktors Lautdauern zeigt eine ähnlich deutliche Wirkung wie die Wirkung der Phonationsarten (siehe Abb. 7.23). Es zeigt sich eine starke Präferenz für schnellere Sprechweise, wobei sämtliche Unterschiede ein Signifikanzniveau von $\alpha < .001$ erreichen. Dieses Ergebnis ist mit fast allen früheren Untersuchungen (vgl. Abschnitt 4.7) konsistent. Lediglich Williams & Stevens (1972) berichten von verlangsamter Sprechweise bei Ärger, jedoch erstens nicht für alle Sprecher und zweitens wurde nicht zwischen *hot* und *cold anger* unterschieden.

Effekte der Phonationsarten in Abhängigkeit von den Dauervarianten ($\alpha = .021$)

Der Effekt der Phonationsarten in Abhängigkeit von den Dauervarianten ist in Abb. 7.19 dargestellt. Dabei zeigt sich (in Konsistenz mit den Ergebnissen der Haupteffekte), dass schnellere Sprechweise bei gespannter Phonation wütender wirkt als bei allen anderen Phonationsarten (alle Unterschiede hatten ein Signifikanzniveau von $\alpha < .001$).

64 % der Stimuli mit gespannter Stimme und schnellerer Sprechweise wurden als wütend klassifiziert. Damit scheint die Kombination aus Anregung und Lautdauern bereits auszureichen, eine starke Eindruckswirkung von Wut zu erzeugen.

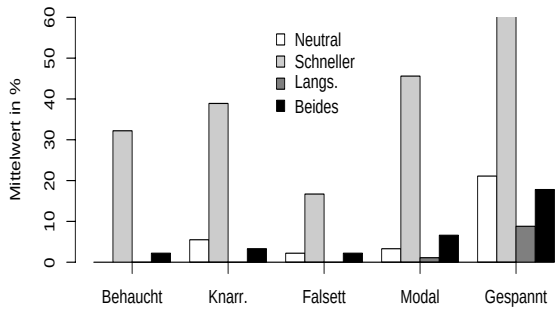

Abbildung 7.19: Mittelwerte der Phonationsarten in Abhängigkeit von den Dauervarianten für Wut

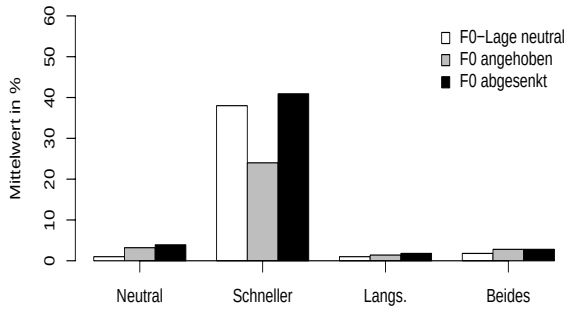

Abbildung 7.20: Mittelwerte der F0-Lage in Abhängigkeit von den Dauervarianten für Wut

Effekte der F_0-Lage in Abhängigkeit von den Dauervarianten ($\alpha = .002$)

Bei höherer Sprechrate erzielen Stimuli mit angehobener F_0-Lage bei höherer Sprechrate seltener eine Beurteilung als wütend als bei neutraler ($p < .001$) oder abgesenkter Grundfrequenz ($p < .001$) (siehe Abb. 7.20).

Diese Ergebnis steht zunächst im Widerspruch zu den meisten Resultaten aus der Literatur (von den berücksichtigten Arbeiten geht nur Cahn (1990) von einer tieferen Grundfrequenz bei Ärger aus). Dieser Widerspruch ist allerdings in der Literatur bekannt (vgl. z.B. Banse & Scherer (1996)) und lässt sich dadurch auflösen, dass man zwischen einer mehr nach außen gerichteten (*hot anger*) und einer eher unterdrückten Form von Ärger (*cold anger*) unterscheidet. Bei diesem Experiment ist die Ursache aber wohl eher darin zu suchen, dass eine höhe-

re Sprechgeschwindigkeit mit angehobener F_0-Lage bei modaler Phonation eher ängstlich als wütend wirkt, zumal Stimuli mit schnellerer Sprechweise und angehobener F_0-Lage nicht mit unterschiedlichen Phonationsarten kombiniert wurden.

Zusammenfassung der Ergebnisse für Wut

Für Wut waren die Ergebnisse zwar nicht so zahlreich wie für die anderen Emotionen, dafür waren diese Resultate aber sehr deutlich ausgeprägt. Der Eindruck von Wut bzw. Ärger wird damit durch eine gespannte Phonation und ein erhöhtes Sprechtempo begünstigt. Die Tatsache, dass ein breiterer Range entgegen den Hypothesen nicht zu einer Steigerung wütender Eindruckswirkung geführt hat, mag daran liegen, dass eine Änderung des F_0-Range dafür nicht ausgeprägt genug war. Dass weiterhin auch die Anhebung der Grundfrequenz nicht wütend wirkte, ist als Indiz dafür zu werten, dass eine Anhebung der F_0-Lage nicht alleine, sondern nur in Kombination mit anderen akustischen Merkmalen zu einer Eindruckswirkung von Sprachsignalen als wütend führt.

7.5 Zusammenfassung

Die Ergebnisse dieser Untersuchung waren sehr vielschichtig, sowohl in Bezug auf akustische Korrelate einzelner Emotionskategorien als auch auf die Bedeutung von bestimmten Sprachsignalmerkmalen für emotionale Hörerperzeption an sich. Die gefundenen akustischen Profile für die untersuchten Emotionen lassen sich wie folgt zusammenfassen:

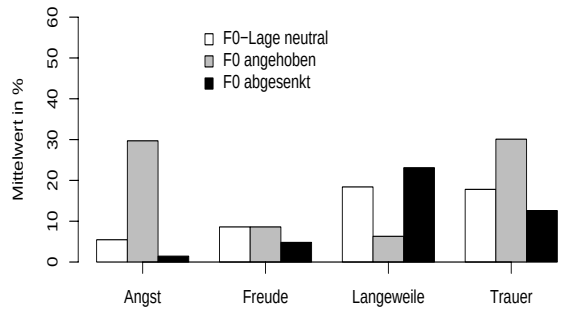

Abbildung 7.21: Mittelwerte der F0-Lage-Varianten für signifikante Emotionen

Der Eindruck von Angst im Sprachsignal wird durch eine stark angehobene F_0-Lage, einen verbreiterten Range, Irregularitäten im Anregungssignal sowie einer beschleunigten Sprechweise begünstigt, bei der allerdings die betonten Silben eher verlängert sind. Wenn auch die Resultate für Freude nicht sehr ausgeprägt waren, so zeigten sich doch deutliche Hinweise darauf,

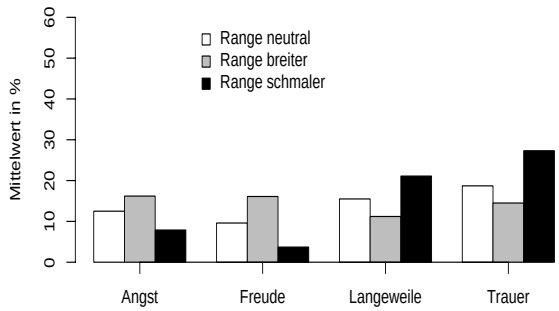

Abbildung 7.22: Mittelwerte der Range-Varianten für signifikante Emotionen

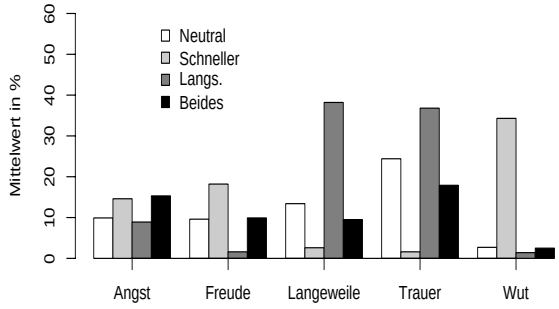

Abbildung 7.23: Mittelwerte der Dauer-Varianten für signifikante Emotionen

dass freudige Sprechweise durch eine angehobene Sprechrate, einen verbreiterten Range und überdeutlicher Artikulation der betonten Silben charakterisiert ist. Deutliche Antworthäufigkeiten gab es für die Kategorie Langeweile. Diese ist vor allem durch eine behauchte oder knarrende Anregung, eine abgesenkte Grundfrequenz, einen schmalen F_0-Range sowie langsamere Sprechweise und undeutliche Artikulation geprägt. All diese akustischen Ausprägungen könnten auch für Trauer gelten. Aus dieser Konkurrenzsituation heraus wurde von den meisten Hörern eine weinerliche Trauer mit relativ starkem Erregungsgrad erkannt. Diese ist ebenso wie stille Trauer durch eine verlangsamte Sprechweise und einen schmaleren F_0-Range gekennzeichnet, wird aber im Gegensatz zu dieser durch eine angehobene F_0-Lage und irreguläre Anregung begünstigt. Als charakteristisch für wütenden Sprecherausdruck präferierten die Testhörer deutlich eine gespannte Phonation und ein erhöhtes Sprechtempo. Andere für starke Wut typische Merkmalsausprägungen wie eine angehobene F_0-Lage und ein breiter F_0-Range wurden

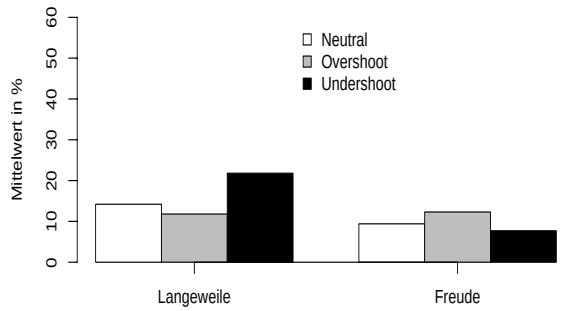

Abbildung 7.24: Mittelwerte der Vokalqualitäts-Varianten für Freude und Langeweile

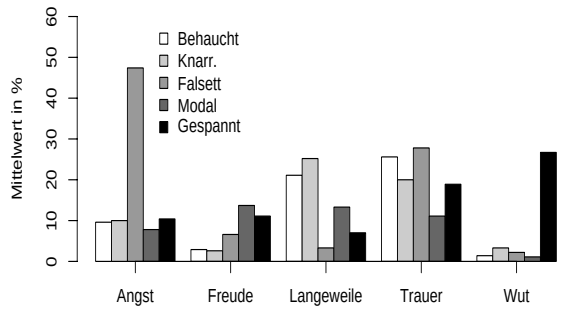

Abbildung 7.25: Mittelwerte der Phonations-Varianten für signifikante Emotionen

in diesem Zusammenhang nicht als wütend, sondern als ängstlich klingend beurteilt.

Als eher selten untersuchtes Merkmal hat sich die Relevanz der Vokalqualität erwiesen. Bei Kienast & Sendlmeier (2000) wird eine Zentralisierung der Formanten für die Emotionen Angst, Trauer und Langeweile prädiziert, während sich bei diesem Experiment nur für Langeweile ein signifikanter Effekt ergab. Bezüglich der Wichtigkeit der einzelnen akustischen Merkmale für die Erkennung emotionaler Sprechweise hat sich gezeigt, dass prosodische Merkmale wie Intonation und Lautdauern keine stärkere Wirkung auf die Einschätzung von Stimuli als emotional haben als stimmqualitative Merkmale. Dabei ist selbstverständlich zu bedenken, dass intonatorische und stimmqualitative Merkmale nicht unabhängig voneinander sind.

Kapitel 8

Überprüfung spezieller Hypothesen zur Emotionssimulation

8.1 Motivation und Konzeption

Der in Kap. 7 geschilderte Hörtest beinhaltete die systematische Variation ausgewählter Merkmale. Die Menge der untersuchten Merkmale und ihrer Ausprägungen musste gering gehalten werden, um die Stimulusmenge nicht zu groß werden zu lassen. Einige Emotionen, speziell Freude, wurden eher selten erkannt. Dies legt nahe, dass die zur Erkennung von Freude wichtigen Parameterausprägungen in der Stimulusmenge nicht vorhanden waren. Weiterhin war die Menge der Emotionskategorien offensichtlich nicht groß genug; so kam es z.B. zu widersprüchlichen Ergebnissen für Trauer, welche sich nur dadurch erklären lassen, dass hier Emotionen mit sehr unterschiedlichen akustischen Profilen einer begrifflichen Kategorie zugeordnet wurden.

Um feinere Differenzierungen vornehmen zu können und weitere Ergebnisse aus der Literatur auf perzeptive Relevanz hin zu überprüfen, werden daher weitere Hörexperimente durchgeführt, für die gezielt Prototypen für bestimmte Emotionen hergestellt wurden. Der Ansatz bei diesem Experiment gegenüber dem aus Kap. 7 unterscheidet sich also grundlegend. Wurden dort zunächst Stimuli generiert und die Rezipienten dann befragt, welcher Emotion sie die Stimuli zuordnen würden, so werden nun gezielt Stimuli auf den Ausdruck einer bestimmten Emotion hin optimiert. Es wird dann geprüft, inwieweit die einzelnen Varianten die Eindruckswirkung der intendierten Emotion verstärken.

Zudem wird die Menge der Emotionskategorien erweitert. Banse & Scherer (1996) weisen darauf hin, dass man bei einer zu geringen Auswahlmöglichkeit an abgefragten Emotionskategorien nicht davon ausgehen kann, dass die Hörer tatsächlich Emotionen erkannt haben, sondern dass sie lediglich zwischen ihnen diskriminieren konnten[1]. Ein weiterer Vorteil einer erweiterten Kategorienmenge liegt darin, dass zunächst widersprüchlich erscheinende Ergebnisse sinnvoll interpretiert werden können, wenn man die Basisemotionen in Subkategorien unterteilt. Paeschke & Sendlmeier (2000) z.B. kommen zu eindeutigeren Ergebnissen bei der Analyse emotio-

[1] Wobei hierin wohl generell die Problematik von *forced-choice* Tests liegt.

naler Sprechweise, indem sie die Menge der Emotionen vergrößern. Im vorherigen Experiment hatte sich gezeigt, dass es zu widersprüchlichen Hörerbeurteilungen aufgrund ungenügender Spezifizierung der gemeinten Emotion kommen kann. Aus diesen Gründen werden einige der ursprünglich fünf Basisemotionen um Varianten mit unterschiedlichem Erregungsgrad ergänzt. Die Emotionsfamilie Ärger wird durch Zorn und Ärger repräsentiert. Bei Freude wird zwischen Freude und Zufriedenheit unterschieden und Trauer wird durch weinerliche und stille Trauer subkategorisiert. Angst und Langeweile werden nicht weiter unterschieden, da bei ihnen ein emotionaler Zustand mit niedriger bzw. hoher Erregung schwer vorstellbar ist. Es wird sich allerdings zeigen, dass die Unterscheidung hinsichtlich der Erregungsdimension bereits für Zorn vs. Ärger nur graduell ist, während bei Langeweile eine Unterscheidung zwischen gähnender und normaler Langeweile durchaus sinnvoll gewesen wäre.

Letztendlich dient dieses Experiment auch dazu, die Güte des verwendeten Sprachsynthesizers EmoSyn hinsichtlich der Simulation emotionaler Sprechweise zu überprüfen. Aus den oben genannten Beschränkungen des vorherigen Hörexperiments ergibt sich, dass keine optimalen emotionalen Prototypen in der Stimulusmenge vorhanden waren. Die Bewertung der speziell für die Eindruckswirkung spezifischer Emotionen generierten Stimuli soll auch dazu dienen, generell den Ansatz der 'Emotionsregeln' und das verwendete System zu evaluieren und Schwachpunkte aufzudecken.

8.2 Herstellung der Teststimuli

Untersuchungsgegenstand sind somit die Emotionen Zorn, Ärger, Freude, Zufriedenheit, weinerliche Trauer, stille Trauer, Angst und Langeweile. Als Vorlage zur Generierung der Stimuli diente wiederum die bereits im vorigen Experiment (Kap. 7) verwendete emotional neutrale Äußerung *"An den Wochenenden bin ich jetzt immer nach Hause gefahren und habe Agnes besucht."*. Sie wurde zunächst durch EmoSyn copysynthetisiert. Diese Kopie diente einerseits als neutraler Referenzstimulus und andererseits als Vorlage für die durchgeführten Modifikationen. Als satzbetonte Silben wurden dabei [vɔ] aus *Wochenenden* und [a] aus *Agnes* festgelegt.

Der Suche nach den optimalen Stimulusvarianten liegt folgende Heuristik zugrunde. Die genauen Ausprägungen der Merkmale wurden durch manuelle Überprüfung mit direkter akustischer Rückmeldung ermittelt. Für jede Variante wurden die Merkmalsausprägungen so gewählt, dass sie einen möglichst guten Kompromiss aus 1. Deutlichkeit bzgl. der Wahrnehmbarkeit, 2. Natürlichkeit des resultierenden sprecherischen Ausdrucks und 3. Deutlichkeit bzgl. der intendierten Emotion darstellen. Zum Beispiel hatte sich gezeigt, dass eine Anhebung des gesamten Grundfrequenzverlaufs von über 200 % mehrheitlich als ängstlich empfunden wird. Da jedoch unter einer derart drastischen Manipulation die Natürlichkeit der Ausgabe leidet, wird bei den ängstlich intendierten Varianten die F_0 lediglich um 150 % angehoben. Im Fall von weinerlicher Trauer werden die Varianten mit fallenden Intonationskonturen mit 20 ST/s modelliert, während bei stiller Trauer die Neigung 30 ST/s beträgt. Dies liegt ebenfalls darin begründet dass bei weinerlicher Trauer, deren Varianten durch eine angehobene F_0-Kontur gekennzeichnet sind, ein zu starkes Intonationsgefälle auf den betonten Silben unnatürlich wirkt. Die genaue Bedeutung der durchgeführten Modifikationen ist der Beschreibung des verwendeten Synthesizers (Ab-

schnitt. 6.3) zu entnehmen. Im folgenden werden die durchgeführten Modifikationen für jede der berücksichtigten Emotionen beschrieben. Die genauen Ausprägungen der Merkmale sind jeweils in Klammern vermerkt. Die dabei verwendeten MOD-Files (siehe Abschnitt 6.2.6) sind im Anhang, Abschnitt A dargestellt.

- **Zorn**
  Zorn (Wut, starker Ärger) ist durch eine starke Aktivierung des sympathischen Nervensystems gekennzeichnet. Dies liegt daran, dass er eine spontane Reaktion auf ein bedrohliches, aber scheinbar handhabbares Ereignis darstellt und der Körper daher Angriffspotential bereitstellt. Somit ist die erregte Wut ebenso wie Angst durch eine erhöhte Sprechgeschwindigkeit und eine angehobene F_0-Lage gekennzeichnet. Weiterhin zeichnet sich Zorn durch intonatorische Merkmale wie ein breiter F_0-Range und stark bewegte Intonationskonturen über vermehrt auftretenden stark betonten Silben aus. Das Auftreten von gespannter Phonation ist eine weiteres Zeichen der starken Erregung. Allerdings hatte ein breiterer F_0-Range und eine angehobene Grundfrequenz im vorangegangenen Hörexperiment nicht zu einer Steigerung wütender Eindruckswirkung geführt. Es ist also zu überprüfen, ob dies in Kombination mit den anderen Merkmalen wie Stimmgebung (Phonationsart), Lautdauern und adäquaten Silbenintonationsverläufen gelingt.

  Der Basistyp für Zorn zeichnet sich durch eine gespannte Stimme (50 %), eine generell angehobene F_0-Lage (50 %) und eine schnellere Sprechweise (30 %) aus. Hiervon ausgehend wurden fünf Varianten erzeugt. Der F_0-Range wurde einmal generell verbreitert (200 %) und einmal wurden die betonten Silben angehoben (50 %). Weiterhin wurden alle betonten Silben mit fallenden Intonationskonturen versehen (30 ST/s). In einer weiteren Variante wurde der Effekt von Formant-Target Overshoot der betonten Silben (50 % für satz- und 30 % für wortbetonte) und Undershoot der unbetonten (20 %) getestet. Letztendlich wurden die Intensität der betonten Silben um 9 dB angehoben.

- **Ärger**
  Die Unterscheidung zwischen *hot* und *cold anger* hat sich in der Literatur durchgesetzt, um kontrastierende Beobachtungen bzgl. der F_0-Lage zu erklären (vgl. etwa Banse & Scherer (1996)). Daher unterscheidet sich der Prototyp für Ärger von dem für Zorn lediglich dadurch, dass die Grundfrequenz nicht angehoben, sondern abgesenkt wird. Da jedoch auch diese Emotion mit einem (in Bezug auf neutrale Sprechweise) höheren Erregungsgrad einhergeht, ist die Sprechgeschwindigkeit ebenfalls erhöht. Auch hierbei wurde der Effekt unterschiedlicher Silbenintonationsverläufe bislang nicht untersucht, dieses soll im folgenden Experiment nachgeholt werden. Weiterhin wird durch die Literatur für Ärger eine erhöhte Artikulationsgenauigkeit prädiziert (Kienast (1998); Murray & Arnott (1995, 1993); Cahn (1990) und Williams & Stevens (1972)), die sich etwa durch Auftreten von Target Overshoot der Formanten bemerkbar machen würde. Dies ließ sich im vorigen Experiment nicht nachweisen. Der Grund dafür ist eventuell im Fehlen einer generell günstigen Parameterkonstellation für Ärger zu suchen, hier besteht also ebenfalls weiterer Überprüfungsbedarf.

  Alle Varianten weisen eine erhöhte Sprechgeschwindigkeit (30 %), eine abgesenkte

Grundfrequenz (20 %) und gespannte Stimmcharakteristik (50 %) auf. Die Varianten enstehen (analog zu Zorn) durch eine Verbreiterung des F_0-Range (100 %), die Modellierung fallender Intonationskonturen (30 ST/s), einer Manipulation der Artikulationsgenauigkeit und einer Intensitätssteigerung der betonten Silben (9 dB).

- **Freude**
  Auch Freude ist eine Emotion, die von relativ starker Erregung geprägt ist. Die F_0-Lage ist angehoben und die Sprechrate schneller. Weiterhin ist freudige Sprechweise durch einen breiten Range und vermehrt auftretende steigende Silbenintonationsverläufe gekennzeichnet (Murray & Arnott (1993)). Dieses so beschriebene akustische Profil wurde allerdings im vorigen Versuch nicht überprüft. Eine günstige Wirkung von gesteigerter Artikulationsgenauigkeit für freudige Eindruckswirkung hatte sich zwar ergeben, dieser Effekt war jedoch nicht allzu deutlich und sollte ebenfalls überprüft werden. Weiterhin blieben die akustischen Charakteristika gespreizter Lippen (vgl. Schröder et al. (1998); Laver (1991) und Tartter (1980)) in dieser Untersuchung bislang unberücksichtigt.

  Alle Varianten sind durch eine erhöhte Sprechgeschwindigkeit (30 %), eine angehobene F_0 (50 %) und einen verbreiterten F_0-Range (100 %) gekennzeichnet. Die Variationen betreffen die Simulation gespreizter Lippen (10 %) und die Modellierung steigender Intonationsverläufe auf den betonten Silben (50 ST/s). Weiterhin wurde die Wirkung von Formant-Target Overshoot (50 % für haupt- und 30 % für wortbetonte Silben) überprüft. Als zusätzliche Variation für die gesamte Intonationskontur wurde das Wellenmodell angewendet (100 % Anhebung, 20 % Absenkung, linear interpoliert).

- **Zufriedenheit**
  Der Unterschied zwischen Zufriedenheit und Freude besteht vor allem in einem geringeren Erregungsgrad. Damit ist die F_0-Lage niedriger als bei Freude und die Sprechgeschwindigkeit langsamer. Bei starker Entspannung können behauchte Phonation und vermehrte Nasalierung (Sendlmeier & Heile (1998)) auftreten. Ansonsten gelten die intonatorischen Kennzeichen hoher Valenz wie bei starker Freude. Wie bei starker Freude sind eine Anhebung der Formanten (Wirkung des Lächelns), steigende Silbenintonationsverläufe und glatte F_0-Konturen auf günstige Eindruckswirkung für Zufriedenheit hin zu überprüfen.

  Beim Basistyp für Zufriedenheit wird lediglich die Sprechweise verlangsamt (20 %). Die Varianten unterscheiden sich dadurch, dass der F_0-Range verbreitert wird (100 %), Lippenspreizung simuliert wird (10 %), steigende Intonationskonturen auf den betonten Silben (50 ST/s) modelliert werden und eine behauchte Phonation (50 %) zugrundegelegt wird. Als zusätzliche Variation kommt wie bei Freude das Intonationskontur-Wellenmodell zur Anwendung.

- **Weinerliche Trauer**
  Weinerliche Trauer unterscheidet sich von stiller Trauer vor allem durch den höheren Erregungsgrad. Das Hauptdiskriminationsmerkmal ist von daher eine angehobene F_0-Lage.

Da diese Emotion allerdings sehr leicht mit Angst zu verwechseln ist[2], wird versucht, eine Abgrenzung gegenüber ängstlicher Sprechweise durch eine Verlangsamung der Sprechweise anstatt der für einen hohen Erregungsgrad eigentlich typischen Beschleunigung der Sprechrate zu erreichen. Weiterhin sollten die akustischen Korrelate von stiller Trauer wie ein schmaler F_0-Range, ein vermehrtes Auftreten fallender Intonationskonturen sowie behauchte Phonation oder F_0-Irregularitäten bei weinerlicher Trauer ebenfalls auftreten.

Für die Basisversion für weinerliche Trauer wird eine langsamere Sprechweise (40 %), eine angehobene F_0-Lage (100 %), ein schmalerer F_0-Range (20 %) und eine geringere F_0-Variabilität (20 %) zugrundegelegt. Die Varianten sind durch fallende Intonationskonturen auf betonten Silben (20 ST/s), F_0-Irregularität (FL=200) und behauchte Phonation sowie Falsett-Anregung gekennzeichnet.

- **Stille Trauer**
  Stille Trauer ist vor allem von Kraftlosigkeit geprägt und somit als Emotion mit geringer Erregung einzustufen. Die F_0-Lage ist also eher abgesenkt, der F_0-Range schmal und die Sprechweise langsam, eventuell auch stockend. Im übrigen sind die Merkmalsausprägungen wie bei weinerlicher Trauer einzuschätzen. Wie dort bleibt die Frage nach der approbaten Form der Silbenintonationsverläufe. Darüberhinaus besteht die Schwierigkeit, diese Emotion deutlich von Langeweile abzugrenzen. Beide Emotionen beinhalten das Element der *Müdigkeit*, aber sie unterscheiden sich bezüglich der Subdominanz. Eventuell drückt sich dies gerade durch ein vermehrtes Auftreten von fallenden Konturen aus.

  Die Grundversion zeichnet sich durch eine langsamere Sprechrate (40 %), eine abgesenkte Grundfrequenz (20 %) sowie einen schmaleren F_0-Range und geringere Variabilität (beide 20 %) aus. Die Varianten ergeben sich durch die Modellierung fallender Intonationskonturen (30 ST/s), F_0-Irregularitäten (FL=300) sowie behauchte Phonation (50 %). Zusätzlich wurden bei einer Variante behauchte Phonation und F_0-Irregularitäten kombiniert.

- **Angst**
  Die Emotionskategorie Angst wurde nicht weiter differenziert. Eine Variante mit geringer Erregung ist nicht denkbar, da Angst stark mit Fluchtmechanismen einhergeht und die Muskulatur deswegen stark angespannt ist. Daher ist die F_0-Lage angehoben und die Sprechgeschwindigkeit schneller. Der F_0-Range ist eher breiter. Es bleibt zu überprüfen, ob eine Häufung gerader Silbenintonationskonturen (Paeschke et al. (1999)) sowie das Wegfallen der finalen F_0-Absenkung zu einem stärkeren Eindruck von Angst im Sprachsignal führt. Weiterhin soll die Frage geklärt werden, ob F_0-Irregularitäten bei stark angehobener Grundfrequenz tatsächlich häufiger zu Erkennung von Angst führen. Es besteht die Gefahr, dass Jitter eher den Eindruck weinerlicher Trauer begünstigt.

  Der Basistyp für Angst ist durch eine erhöhte Sprechgeschwindigkeit (30 %) und einen breiteren F_0-Range (20 %) charakterisiert. Die Variationen werden durch ein angehobene F_0 (150 %), F_0-Irregularitäten und gerade Intonationsverläufe auf den betonten Silben in

[2] Beide Emotionen stimmen ja hinsichtlich der Dimensionen Valenz, Erregung und Potenz weitgehend überein.

Kombination mit einer steigenden Kontur auf der letzten Silbe erzeugt. Zusätzlich wurde eine Äußerung mit Falsettphonation generiert. Bei dieser wurde die durchschnittliche Grundfrequenz nur um 80 % angehoben, um bei allen Stimuli die gleiche F_0-Anhebung zu erreichen. Es hatte sich zwar gezeigt, dass eine F_0-Anhebung von über 200 % den Eindruck von Angst deutlich verstärkt, aber da die resultierenden Stimuli bei derart drastischer Manipulation an Natürlichkeit einbüßen, wird die Anhebung hier auf 150 % beschränkt. Außerdem zeigt die Analyse ängstlicher natürlichsprachiger Aufnahmen, dass die Grundfrequenzanhebung bei diesen zwar deutlich, aber nicht in extrem hohem Maße erfolgt. Paeschke & Sendlmeier (2000) etwa messen bei Freude und Ärger eine signifikant stärkere Anhebung der F_0-Lage als bei Angst.

- **Langeweile**
  Auch die Kategorie Langeweile wurde nicht weiter unterteilt. Langeweile ist nicht mit hoher Erregung vereinbar, da sie einen Zustand darstellt, der vor allem von Desinteresse geprägt ist. Kennzeichen sind hiermit vor allem eine abgesenkte F_0-Lage, ein schmalerer F_0-Range sowie eine verlangsamte Sprechweise. Die Aktivierung kann so schwach sein, dass geknarrte Phonation oder behauchte Stimmgebung auftritt. Die Artikulation ist extrem unpräzise und ein Targetundershoot der Formanten tritt gehäuft auf. Es soll vor allem überprüft werden, wie sich Langeweile von stiller Trauer abgrenzen lässt. Dies sollte sich vor allem durch Merkmale fehlender Subdominanz ausdrücken lassen. Daher sollten die Silbenintonationsverläufe nicht primär fallende Kontur aufweisen und sich die gesamte F_0-Kontur am Phrasenende deutlich absenken.

  Alle Varianten sind durch eine langsamere Sprechweise (20 %), eine Verlängerung der hauptbetonten Silben (40 %), ein Absenkung der Grundfrequenz sowie einen schmaleren F_0-Range (50 %) und geringere F_0-Variabilität (20 %) gekennzeichnet. Variiert wurden behauchte und geknarrte (nur beim Stimmeinsatz) Phonation (beide 50 %) sowie Formant-Target Undershoot. Eine letzte Variante wurde durch die Kombination von Knarrstimme, behauchter Phonation und Target-Undershoot generiert.

8.3 Durchführung des Hörexperiments

Das Perzeptionsexperiment wurde an einem Computerterminal mit Kopfhörern in ruhiger Umgebung in Einzelsitzungen durchgeführt. Das Testprogramm erzeugt für jeden Beurteiler eine eigene Zufallsreihenfolge (vgl. Kap. 7). Teilgenommen haben 23 Frauen und 19 Männer im Alter zwischen 12 und 60 Jahren. Das Durchschnittsalter betrug 30 Jahre. Einige waren Experten (4) aber die meisten Laien (38). Die im Vergleich zu den vorigen Experimenten höhere Anzahl an Versuchspersonen begründet sich aus der sehr starken Streuung der Urteile. Um den Vergleich mit neutraler Sprechweise zu ermöglichen, wurde die neutrale Äußerung viermal in die zu beurteilende Stimulusmenge aufgenommen. Zur Eingewöhnung der Hörer wurden vier zufällig ausgewählte Stimuli vorangestellt, deren Beurteilung bei der Auswertung nicht berücksichtigt wurde. Die Aufgabe der Hörer bestand zunächst darin, den dargebotenen Stimulus einer der acht Emotionen oder 'Neutral' zuzuordnen. Eine wiederholtes Hören war nicht

möglich. Weiterhin wurden die Hörer befragt, wie überzeugend die Emotion ihrer Meinung nach dargestellt war (binäre Antwortmöglichkeit). Abgesehen von dem Bildschirmtext wurden die Hörer mündlich darauf hingewiesen, dass sie die Stimuli nach Möglichkeit der Emotion zuweisen sollten, die der wahrgenommenen am nächsten kommt. Sie sollten explizit nicht die Kategorie 'Neutral' für schwer zu klassifizierende Stimuli verwenden. Als Strategie wurde vorgeschlagen, zunächst den beim Hören spontan entstehenden Eindruck zu benennen und diese Benennung dann in einem zweiten Schritt einer der neun Emotionen zuzuordnen.

8.4 Auswertung der Ergebnisse

Die Ergebnisse wurden durch Varianzanalysen mit kompletter Messwiederholung (vgl. Werner (1997, S. 486), Bortz (1993, S. 320) und Diehl (1977, S. 272)) auf Signifikanz hin überprüft. Für jede Emotion wurde eine Varianzanalyse durch das Statistikprogramm SPSS berechnet. Dabei fungierte die Erkennung der Emotion als abhängige und die Variation des Stimulus als unabhängige Variable. Die Ergebnisse der Varianzanalyse sind in den Tabellen 8.1 - 8.8 zusammengefasst. Die Tabellen stellen jeweils in der ersten Spalte die durchschnittliche Erkennungsrate für die einzelnen Varianten sowie die Verwechslung mit Neutral dar. In den weiteren Spalten ist die Signifikanz des Unterschiedes zwischen den Variationen und 'Neutral' dargestellt. Dabei stehen drei Sterne für einen höchst signifikanten Unterschied ($\alpha <= .001$), zwei Sterne für eine hoch signifikanten ($\alpha <= .01$) und ein Stern für einen signifikanten Unterschied ($\alpha <= .05$). Bei Signifikanzniveaus, die eine Tendenz aufzeigen ($\alpha <= .1$), ist der entsprechende Wert angegeben.

Zusätzlich wurden Verwechslungen zwischen den Emotionen ausgewertet. Das Zufallsniveau beträgt bei diesem Test 11 %. Die Verwechslungen sind in den Abb. 8.1 - 8.8 dargestellt. Bewertungen unterhalb des Zufallsniveaus wurden bei den Abbildungen auf Null gesetzt.

Da die Unterschiede zwischen den Varianten sehr oft nicht signifikant waren, wurden zwei weitere Reihen von Varianzanalysen berechnet. Bei der einen wurden die Ergebnisse für Erkennung und Überzeugung zusammengefasst, d.h. ein Stimulus bekam zwei Punkte wenn er erkannt und als überzeugend eingestuft wurde. Weiterhin wurden die qualitativ paarweisen Emotionen zusammengefasst, ein Stimulus gilt somit als erkannt wenn er einer der Emotionen derselben Qualität zugeordnet wurde. Die Ergebnisse werden im folgenden für jede Emotion einzeln diskutiert.

8.4.1 Zorn

Abgesehen von Variante 4 (vgl. Tab. 8.1 und Abb. 8.1), die sich (neben den Grundmanipulationen) lediglich durch fallende Intonationskonturen auszeichnet, wurden alle Varianten überzufällig erkannt. Die Mittelwerte liegen jedoch nicht sehr hoch. Abb. 8.1 ist zu entnehmen, dass alle Varianten oft mit Ärger verwechselt wurden. Der Unterschied zwischen Zorn und Ärger ist sicherlich gradueller als der zwischen Freude und Zufriedenheit oder weinerlicher und stiller Trauer. Der Grund für die seltene Zuweisung als zornig mag darin liegen, dass die begrenzte Bandbreite des Synthesizers (5 kHz max. Frequenz) eine für zornigen Stimmausdruck adäquate

Stimulus	⌀[%]	Signifikanz des Unterschiedes				
Neutral:	0,6	V1: ***	V2: *	V3: **	V4: ,06	V5: ***
V1: FR(300%)	23,8	N: ***	V2:-	V3:-	V4: ,057	V5: -
V2: PCS(fall, 30 St/sek), LSC, IS(9 dB)	14,3	N: *	V1: -	V3: -	V4: -	V5: -
V3: PCS(fall, 30 St/sek), LSC, OV	21,4	N: **	V1: -	V2: -	V4: -	V5: -
V4: PCS(fall, 30 St/sek), LSC	9,5	N: ,06	V1: ,057	V2: -	V3: -	V5: *
V5: SF0(50% angehoben)	28,6	N: ***	V1: -	V2: -	V3: -	V4: *

Tabelle 8.1: Ergebnisse der Varianzanalyse für Zorn, alle Varianten: SR(30% schneller), F0(50% angehoben), P(tense). Bedeutung der Abk.: SR=speech rate, F0=⌀F0, FR=F0-range, PCS=pitch contour of stressed syllables, LSC=last syllable contour, SF0=⌀F0 of stressed syllables, IS-intensity of stressed syllables, OV=vowel target overshoot

Modellierung nicht gestattet[3]. Eine weitere Ursache für die schlechte Erkennung von Zorn ist die Tatsache, dass starker Ärger mit einer Häufung starkbetonter Silben einhergeht (Paeschke et al. (1999) und Williams & Stevens (1972)). Einer der Nachteile eines Synthesekonzepts ohne semantisches Wissen ist die Unfähigkeit, neue mögliche Betonungen zu finden[4]. Die Versionen mit Target Overshoot (V3) oder lauteren betonten Silben (V2) sowie die mit verbreitertem F_0-Range (V1) wurden auch mit Angst verwechselt. Dies lässt sich durch eine Ähnlichkeit hinsichtlich starker Erregung und negativer Valenz erklären. Die auftretenden Verwechslungen mit Zufriedenheit (V1 und V4) liegen knapp über dem Zufallsniveau (11,9 %) und können wohl ignoriert werden.

8.4.2 Ärger

Stimulus	⌀[%]	Signifikanz des Unterschiedes				
Neutral:	0	V1: ***	V2: ***	V3: ***	V4: ***	V5: ***
V1: FR(100%), LSC(fall, 80 ST/sek)	42,9	N: ***	V2: -	V3: -	V4: -	V5: -
V2:	45,2	N: ***	V1: -	V3: -	V4: -	V5: -
V3: PCS(fall, 30 St/sek), LSC, IS(9 dB)	45,2	N: ***	V1: -	V2: -	V4: -	V5: -
V4: PCS(fall, 30 St/sek), LSC, OV	59,5	N: ***	V1: -	V2: -	V3: -	V5: -
V5: PCS(fall, 30 St/sek), LSC	52,4	N: ***	V1: -	V2: -	V3: -	V4: -

Tabelle 8.2: Ergebnisse der Varianzanalyse für Ärger, alle Varianten: SR(30% langsamer), F0(20% abgesenkt), P(tense). Bedeutung der Abk.: SR=speech rate, F0=⌀F0, FR=F0-range, PCS=pitch contour of stressed syllables, LSC=last syllable contour, IS-intensity of stressed syllables, OV=vowel target overshoot, P=phonation type

[3]Eine ähnliche Beobachtung wurde bei Rutledge et al. (1995) gemacht.

[4]Cahn (1990) hat dieses Problem umgangen, indem sie für ihren Synthesizer ein Concept-To-Speech System zugrunde gelegt hat.

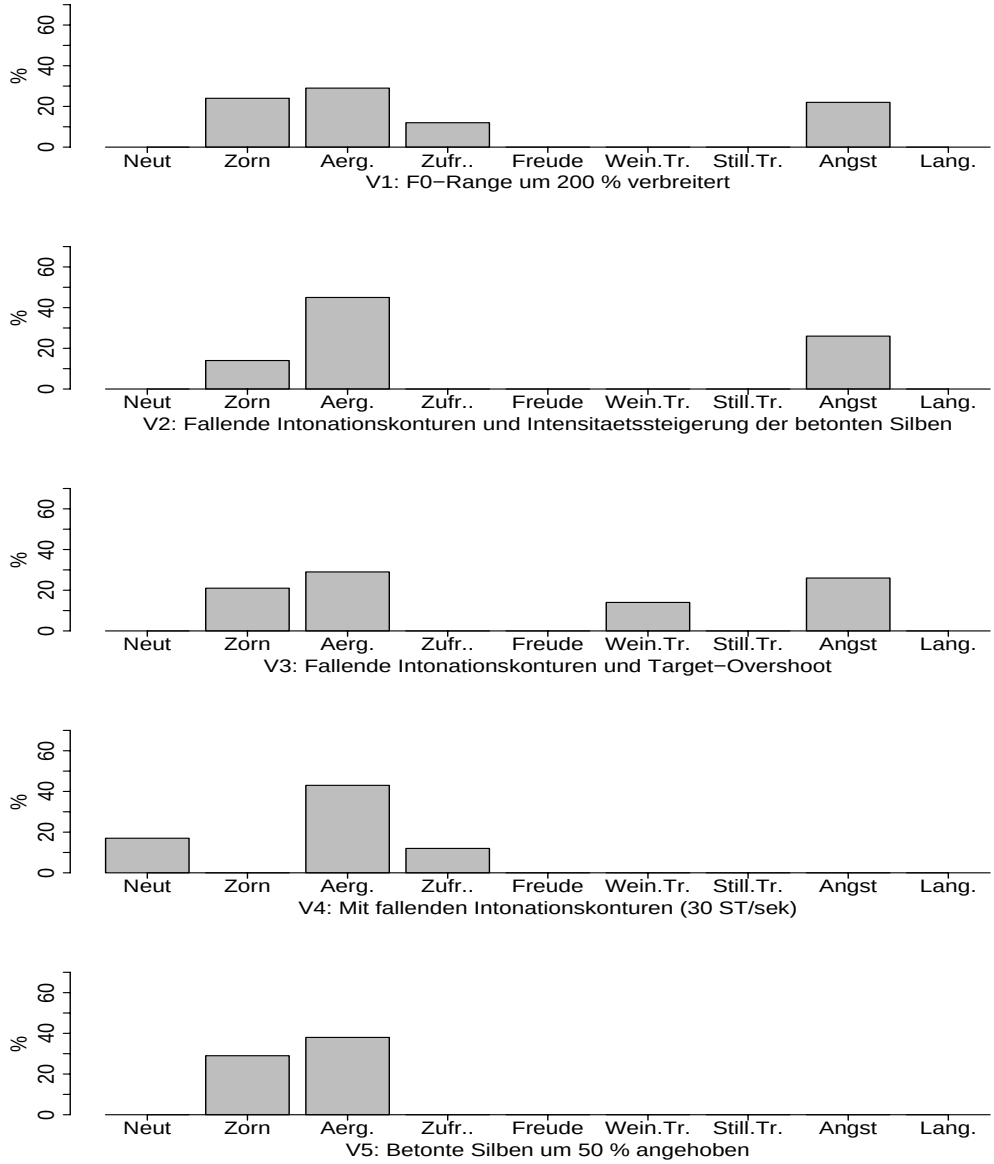

Abbildung 8.1: Verwechslungen der Varianten für Zorn (alle Varianten: schnellere Sprechweise (30 %), angehobene F_0 (50 %) und gespannte Phonation)

Die ärgerlich intendierten Stimuli wurden alle weit über dem Zufallsniveau erkannt. Sie unterscheiden sich jedoch nicht signifikant voneinander. Alle Varianten wurden mit Neutral verwechselt (durchschnittlich 19,5 %), was bei den zornig intendierten nicht der Fall war. Offensichtlich führte das Nichtanheben der Grundfrequenz zu diesem Urteil. Betrachtet man die Ergebnisse für den Fall, dass entweder Ärger oder Zorn erkannt wurde, ergibt sich immerhin eine Signifikanz dafür, dass die Variante mit Target-Overshoot (V4) besser erkannt wurde als

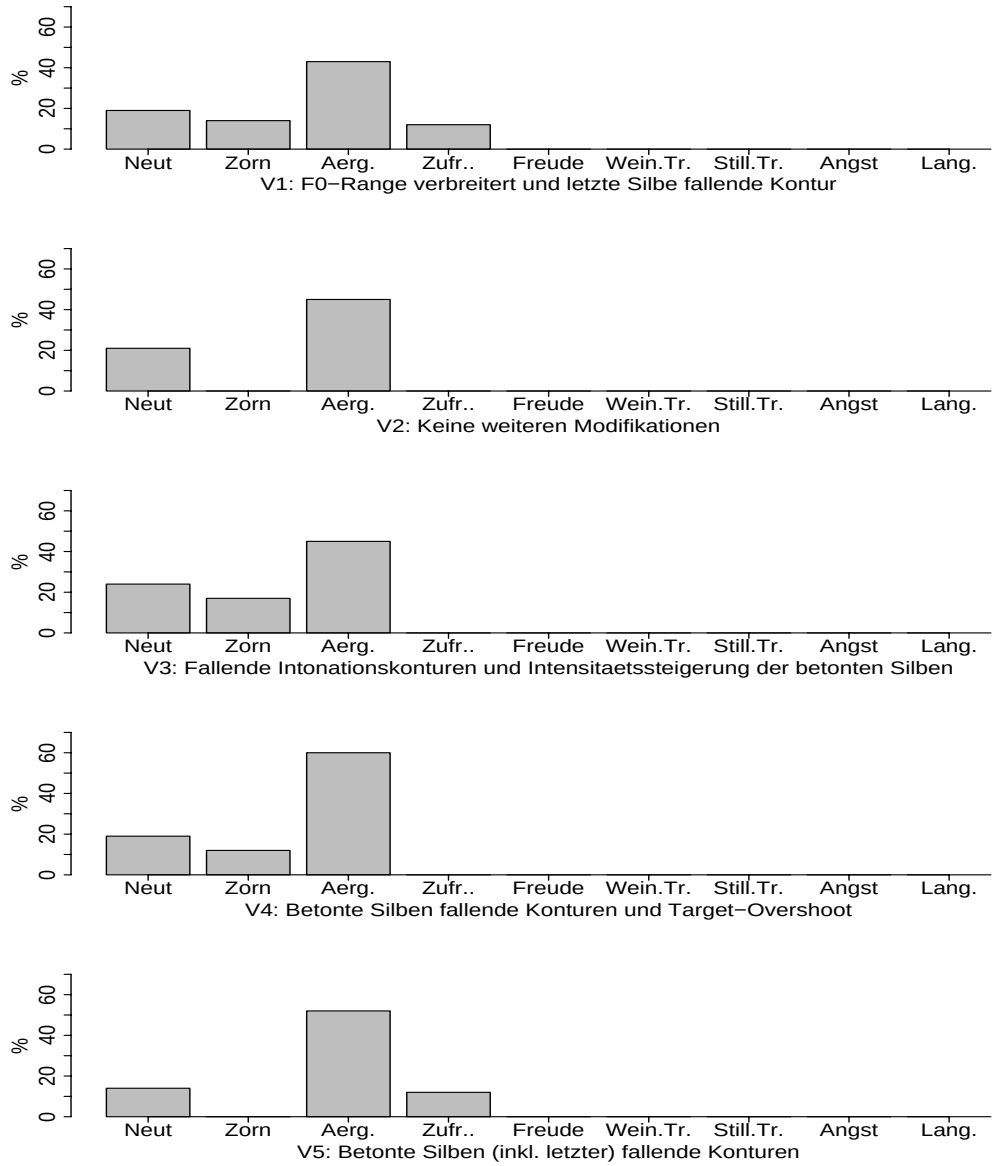

Abbildung 8.2: Verwechslungen der Varianten für Ärger (alle Varianten: schnellere Sprechweise (30 %), abgesenkte F_0 (20 %) und gespannte Phonation)

die mit Minimaländerung (V2). Eine Tendenz dafür ergibt sich ebenfalls, wenn sowohl Erkennung als auch Überzeugung bei der Analyse berücksichtigt werden. Die Kombination aus fallenden Intonationskonturen und Target-Overshoot ergibt also durchaus eine Steigerung ärgerlich/zorniger Eindruckswirkung. Die Verwechslungen (V1 und V5) mit Zufriedenheit liegen wie bei Zorn so knapp über dem Zufallsniveau, dass sie ignoriert werden können. Die Tatsache, dass Zorn weitaus häufiger mit Ärger verwechselt wurde als umgekehrt, lässt darauf schließen,

dass durch das Synthesesystem ein eher gemäßigter emotionaler Ausdruck generiert wurde. Wie oben erwähnt lag den Manipulationen als Kriterium zugrunde, nicht durch übermäßige Ausprägung der Merkmale die Natürlichkeit der Ausgabe zu mindern.

8.4.3 Freude

Stimulus	⌀[%]	Signifikanz des Unterschiedes				
Neutral:	0	V1: ,08	V2: ***	V3: ***	V4: ***	V5: ***
V1:	7,1	N: ,08	V2: **	V3: *	V4: *	V5: ***
V2: PCS(stei, 50 ST/sek), SL(10%), OV	31	N: ***	V1: **	V3: -	V4: -	V5: ***
V3: PCS(stei, 50 ST/sek), SL(10%)	28,6	N: ***	V1: *	V2: -	V4: -	V5: ***
V4: SL(10%)	23,8	N: ***	V1: *	V2: -	V3: -	V5: ***
V5: W, SL(10%)	81	N: ***	V1: ***	V2: ***	V3: ***	V4: ***

Tabelle 8.3: Ergebnisse der Varianzanalyse für Freude, alle Varianten: SR(30% schneller), F0(50% angehoben), FR(100% breiter). Bedeutung der Abk.: SR=speech rate, F0=⌀F0, FR=F0-range, PCS=pitch contour of stressed syllables, SL=spread lips, W=intonation wave-model, OV=vowel target overshoot

Für die freudig gemeinten Varianten zeigen sich deutlichere Unterschiede als bei Zorn und Ärger. Die Basisversion (V1) wurde nicht erkannt, sondern mit Neutral, Ärger, weinerlicher Trauer oder Angst verwechselt (vgl. Abb. 8.3 und Tab. 8.3). Da lediglich Korrelate körperlicher Erregung modelliert wurden, war eine Verwechslung mit Emotionen, die sich ebenfalls durch eine stärkere Erregung auszeichnen, zu erwarten. Offensichtlich perzipieren die Hörer freudigen Sprechausdruck nur dann, wenn zusätzlich akustische Korrelate positiver Valenz gegeben sind. Die Varianten, bei denen weitere Merkmale als lediglich solche starker Erregung modelliert wurden, wurden deutlich über dem Zufallsniveau erkannt.

Allein die Modellierung gespreizter Lippen (V4) führt bereits zu einem freudigen Hörereindruck, wie auch in Burkhardt & Sendlmeier (1999a) und Kienast & Sendlmeier (2000) prädiziert. Diese Variante wird weiterhin mit Zufriedenheit und Ärger verwechselt. Die zusätzlichen Modifikationen von steigenden Silbenkonturen (V3) oder Target-Overshoot (V2) erhöhen zwar leicht die Erkennung, die Änderung wird aber nicht signifikant. Interessanterweise wird die Variante mit Target-Overshoot als einzige mit Zorn verwechselt. Eine deutliche Erkennung für Freude ergibt sich erst durch die Modellierung der Intonationskontur nach dem Wellenmodell (V5). Da die Anwendung des Modells mit einer Glättung der Kontur verbunden ist, bleibt allerdings unklar, ob sich der günstige Effekt primär aus der Glättung der Kontur oder dem gleichmäßigen Fallen und Steigen ergibt. Es liegt die Vermutung nahe, dass es gerade die Kombination aus beidem ist: nach informellem Höreindruck wirkt weder ein Glätten der Kontur alleine noch ein Verschieben der Silben ohne Änderung der Silbenkontur deutlich freudiger. Lediglich eine Variante (V4) wird mit Zufriedenheit verwechselt. Offensichtlich bewerten die Hörer den Unterschied zwischen Freude und Zufriedenheit nicht so graduell wie den zwischen Zorn und Ärger.

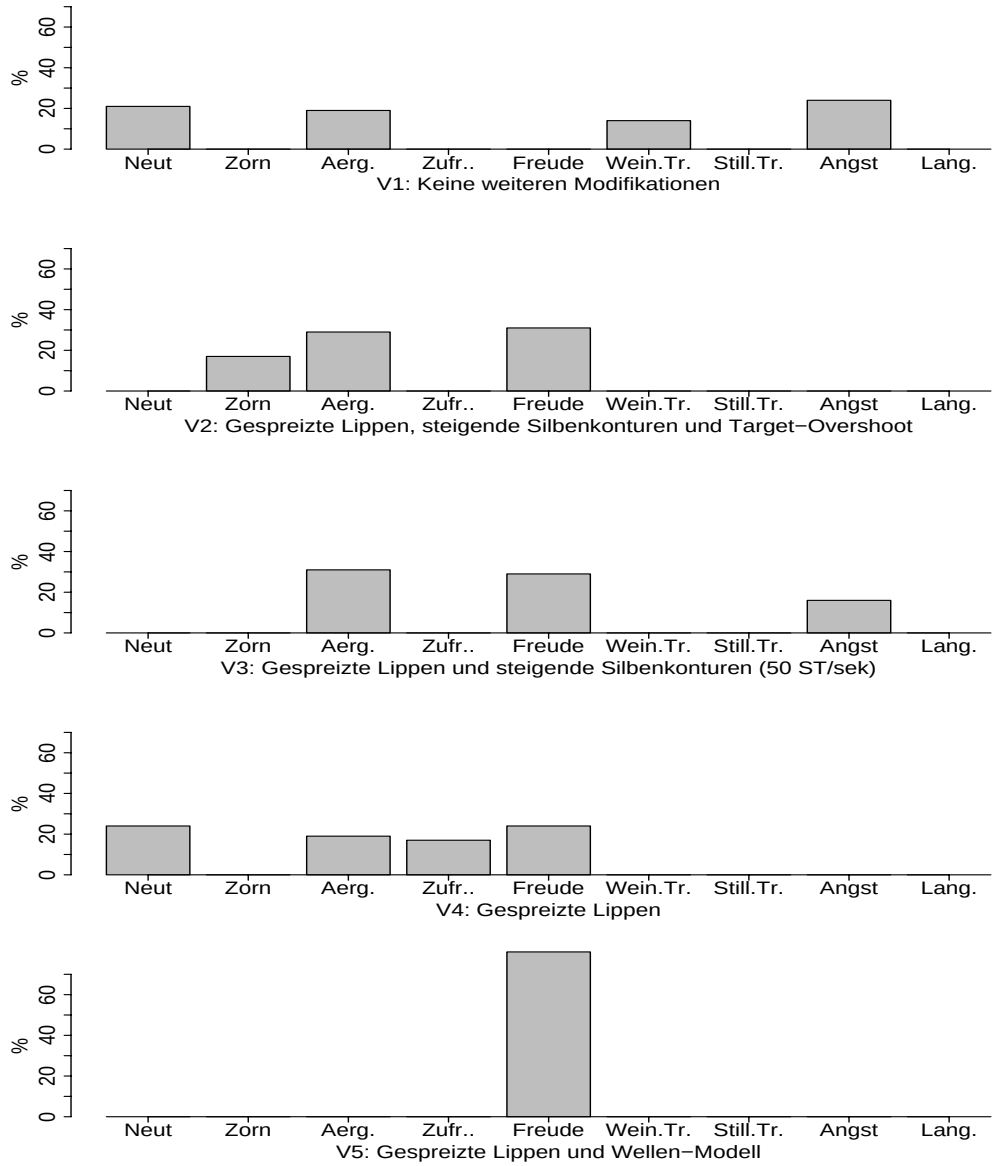

Abbildung 8.3: Verwechslungen der Varianten für Freude (alle Varianten: schnellere Sprechweise (30 %), angehobene F_0 (50 %) und breiterer F_0-Range (100 %))

8.4.4 Zufriedenheit

Für die Varianten mit intendierter Zufriedenheit ergibt sich ein ähnliches Bild wie bei Freude. Die Basisversion (V1) wurde nicht über dem Zufallsniveau erkannt, sondern mit Neutral, Trauer oder Langeweile verwechselt (vgl. Abb. 8.4 und Tab. 8.4). Sie ist ja auch lediglich durch eine Verlangsamung der Sprechrate sowie eine leichte Verbreiterung des F_0-Range gekennzeichnet, so dass eine Zuweisung für Emotionen mit niedrigerem Erregungsgrad als plausibel erscheint.

Stimulus	⌀[%]	Signifikanz des Unterschiedes				
Neutral:	21,4	V1: ,01	V2: -	V3: -	V4: -	V5: ***
V1: FR(100% breiter)	7,1	N: ,01	V2: *	V3: -	V4: ***	V5: ***
V2: FR(100%), SL(10%)	21,4	N: -	V1: *	V3: -	V4: -	V5: ***
V3: FR(100%), SL(10%), P(breathy)	14,3	N: -	V1: -	V2: -	V4: *	V5: ***
V4: PCS(stei, 50 ST/sek), SL(10%)	33,3	N: -	V1: ***	V2: -	V3: *	V5: **
V5: W, SL(10%)	61,9	N: ***	V1: ***	V2: ***	V3: ***	V4: **

Tabelle 8.4: Ergebnisse der Varianzanalyse für Zufriedenheit, alle Varianten: SR(20% langsamer). Bedeutung der Abk.: SR=speech rate, F0=⌀F0, FR=F0-range, PCS=pitch contour of stressed syllables, SL=spread lips, W=intonation wave-model, OV=vowel target overshoot

Dass Zufriedenheit nicht darunter ist[5], bestätigt einmal mehr die These, dass positive Emotionen nur im Zusammenhang mit deutlichen Korrelaten positiver Valenz perzipiert werden.

Eine Erkennung von Zufriedenheit ergibt sich auch hier erst durch die Modellierung der gespreizten Lippencharakteristik (V2). Allerdings führt weder die Verbreiterung des F_0-Range (V2) noch die Modellierung behauchter Phonation (V3) dazu, dass die Varianten öfter mit Zufriedenheit als mit Neutral, Trauer oder Langeweile attribuiert werden. Werden Zufriedenheit und Freude zusammengefasst, so wird die Version mit steigenden Silbenkonturen (V4) immerhin signifikant öfter erkannt als mit Neutral oder Ärger verwechselt. Die perzeptive Relevanz steigender Intonationskonturen für freudig/zufriedenen Sprecherausdruck kann also bestätigt werden. Interessanterweise ist dies auch die einzige Variante, die fast nie (4,8 %) mit Langeweile verwechselt wird. Die Variante mit behauchter Phonation wurde als einzige mit Angst verwechselt (von 14,3 % der Hörer). Behauchung ist zwar nicht mit dem hohen Erregungsgrad ängstlicher Sprechweise vereinbar, der akustische Eindruck ist aber flüsternder Phonation recht ähnlich. Klasmeyer & Sendlmeier (2000) beobachten, dass ängstliche Sprechweise mit geflüsterter Anregung einhergehen kann. Wie bei Freude ergibt sich eine wirklich deutliche Erkennung für Zufriedenheit nur bei der Variante mit dem Intonationskontur-Wellenmodell (V5).

8.4.5 Weinerliche Trauer

Alle Varianten für weinerliche Trauer wurden weit über dem Zufallsniveau erkannt (vgl. Abb. 8.5 und Tab. 8.5). Es gibt eine Tendenz dafür, dass die Version mit behauchter Stimme (V4) seltener erkannt wurde als die Variante ohne Behauchung (V2). Sie wurde auffällig oft mit stiller Trauer verwechselt. Da eine behauchte Phonation physiologisch nicht mit einer hohen Erregung vereinbar ist, ist die Verwechslung mit stiller Trauer durchaus konsequent. Die Modellierung fallender Konturen führte ebenfalls nicht zu einer Steigerung des Eindrucks von weinerlicher Trauer. Diese Variante (V2) wurde sogar tendenziell als weniger überzeugend bewertet.

Lediglich Falsett-Phonation (V5) und F_0-Schwankungen (V3) erhöhten den Eindruck weinerlicher Trauer gegenüber der Grundkonfiguration (V1). Da beide von F_0-Irregularitäten ge-

[5]Wie auch die Stimuli, die lediglich akustische Korrelate von Erregung aufweisen, nicht als freudig bewertet wurden.

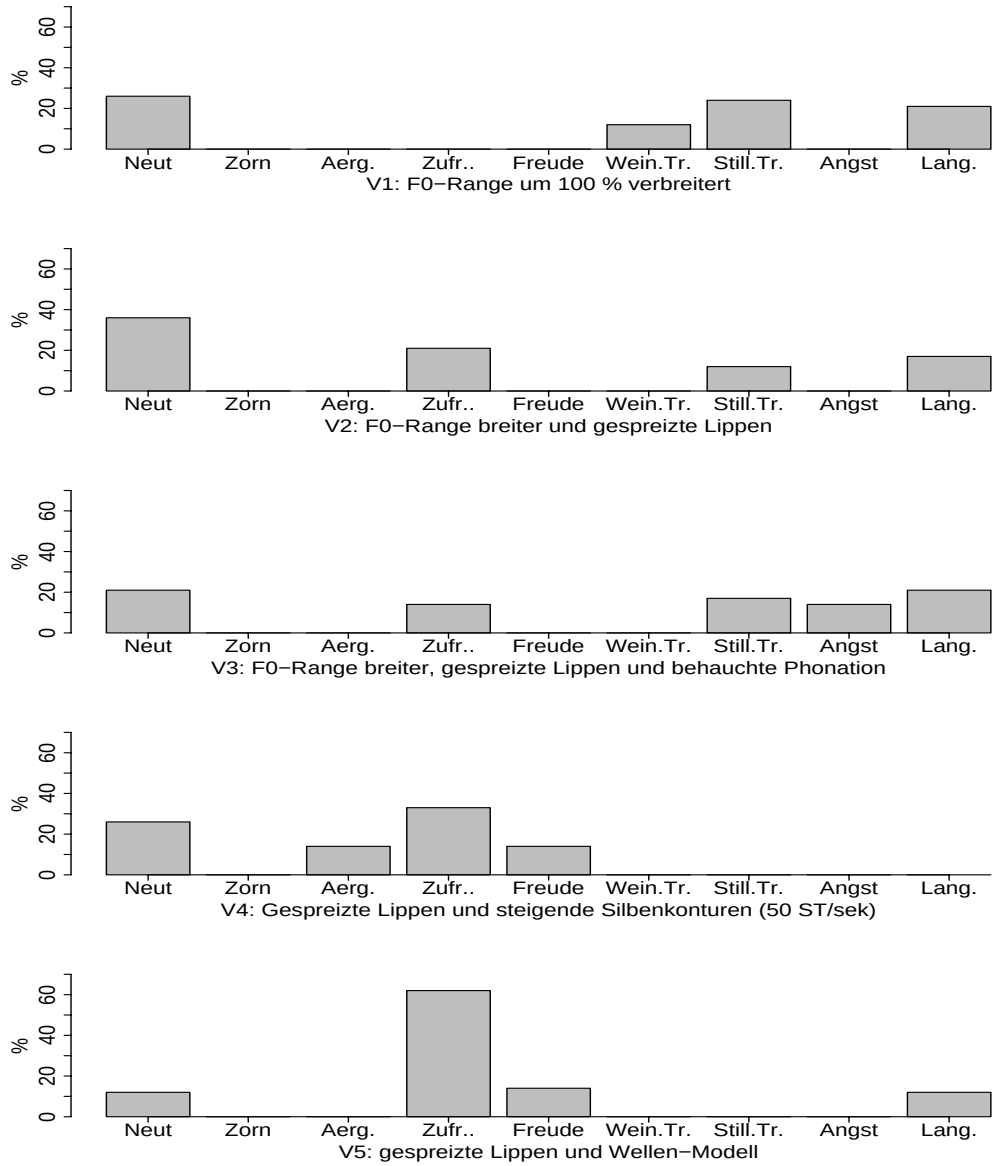

Abbildung 8.4: Verwechslungen der Varianten für Zufriedenheit (alle Varianten: langsamere Sprechweise (20 %))

kennzeichnet sind, kann dies wohl als ein perzeptiv relevantes Merkmal weinerlicher Trauer gelten. Abgesehen von diesen beiden Stimuli wurden alle auch mit stiller Trauer verwechselt. Weiterhin wurden alle Varianten auch mit Angst verwechselt (durchschnittlich von 19,5 % der Hörer), was sich leicht durch eine Ähnlichkeit in Bezug auf negative Valenz, höhere Erregung und Subdominanz erklären lässt. Da keine weiteren Verwechslungen vorkamen, ist davon auszugehen, dass akustische Korrelate negativer Valenz und Subdominanz in allen Varianten über-

Stimulus	⌀[%]	Signifikanz des Unterschiedes				
Neutral:	1,2	V1: ***	V2: ***	V3: ***	V4: ***	V5: ***
V1:	54,8	N: ***	V2: -	V3: ,08	V4: ,08	V5: -
V2: PCS(fall, 20 ST/sek)	40,5	N: ***	V1: -	V3: **	V4: -	V5: *
V3: PCS(fall, 20 ST/sek), J(200)	69	N: ***	V1: ,08	V2: **	V4: **	V5: -
V4: PCS(fall, 20 ST/sek), P(breathy)	35,7	N: ***	V1: ,08	V2: -	V3: **	V5: *
V5: PCS(fall, 20 ST/sek), P(falsett)	61,9	N: ***	V1: -	V2: *	V3: -	V4: *

Tabelle 8.5: Ergebnisse der Varianzanalyse für weinerliche Trauer, alle Varianten: SR(40% langsamer), F0(100% angehoben), FR(20% schmaler), FV(20% geringer). Bedeutung der Abk.: SR=speech rate, F0=⌀F0, FR=F0-range, PCS=pitch contour of stressed syllables, J-jitter, P=phonation type.

zeugend enkodiert sind.

8.4.6 Stille Trauer

Stimulus	⌀[%]	Signifikanz des Unterschiedes				
Neutral:	7,1	V1: ***	V2: ,07	V3: ***	V4: ***	V5: *
V1:	40,5	N: ***	V2: *	V3: -	V4: -	V5: -
V2: PCS(fall, 30ST/sek)	19	N: ,07	V1: *	V3: ,08	V4: *	V5: -
V3: PCS(fall, 30 ST/sek), J(300)	38,1	N: ***	V1: -	V2: ,08	V4: -	V5: -
V4: PCS(f., 30 ST/sek), J(300), P(br.)	38,1	N: ***	V1:	V2: *	V3: -	V5: -
V5: PCS(fall, 30 ST/sek), P(breathy)	26,2	N: *	V1: -	V2: -	V3: -	V4: -

Tabelle 8.6: Ergebnisse der Varianzanalyse für stille Trauer, alle Varianten: SR(40% schneller), F0(20% abgesenkt), FR(20% schmaler), FV(20% geringer). Bedeutung der Abk.: SR=speech rate, F0=⌀F0, FR=F0-range, FV=F0-variability, PCS=pitch contour of stressed syllables, J-jitter, P=phonation type.

Bei stiller Trauer wurde ebenfalls die Variante mit fallenden Intonationskonturen (V2) deutlich schlechter erkannt als die anderen (vgl. Abb. 8.6 und Tab. 8.6). Dafür wurde sie auffällig oft mit Langeweile verwechselt (von 62 % der Hörer). Damit kann diese Version schon als guter Prototyp für Langeweile gelten. Generell ist die Konkurrenzsituation zwischen stiller Trauer und Langeweile, die bereits in Kap. 7 auffiel, deutlich geworden. Es gab keine Variante, die verglichen mit Langeweile signifikant häufiger als stille Trauer erkannt wurde. Lediglich die Versionen mit F_0-Irregularität (V3 und V4) werden, wenn man beide Trauerversionen zusammenfasst, öfter als traurig empfunden. Somit kann das Merkmal F_0-Schwankung, wie auch schon bei weinerlicher Trauer bemerkt, als relativ deutliches Korrelat trauriger Sprechweise gelten. Es scheint damit durchaus geeignet, zwischen Langeweile und Trauer zu diskriminieren. Für das Merkmal Behauchung (V5) hat sich weiterhin keine signifikante Steigerung des

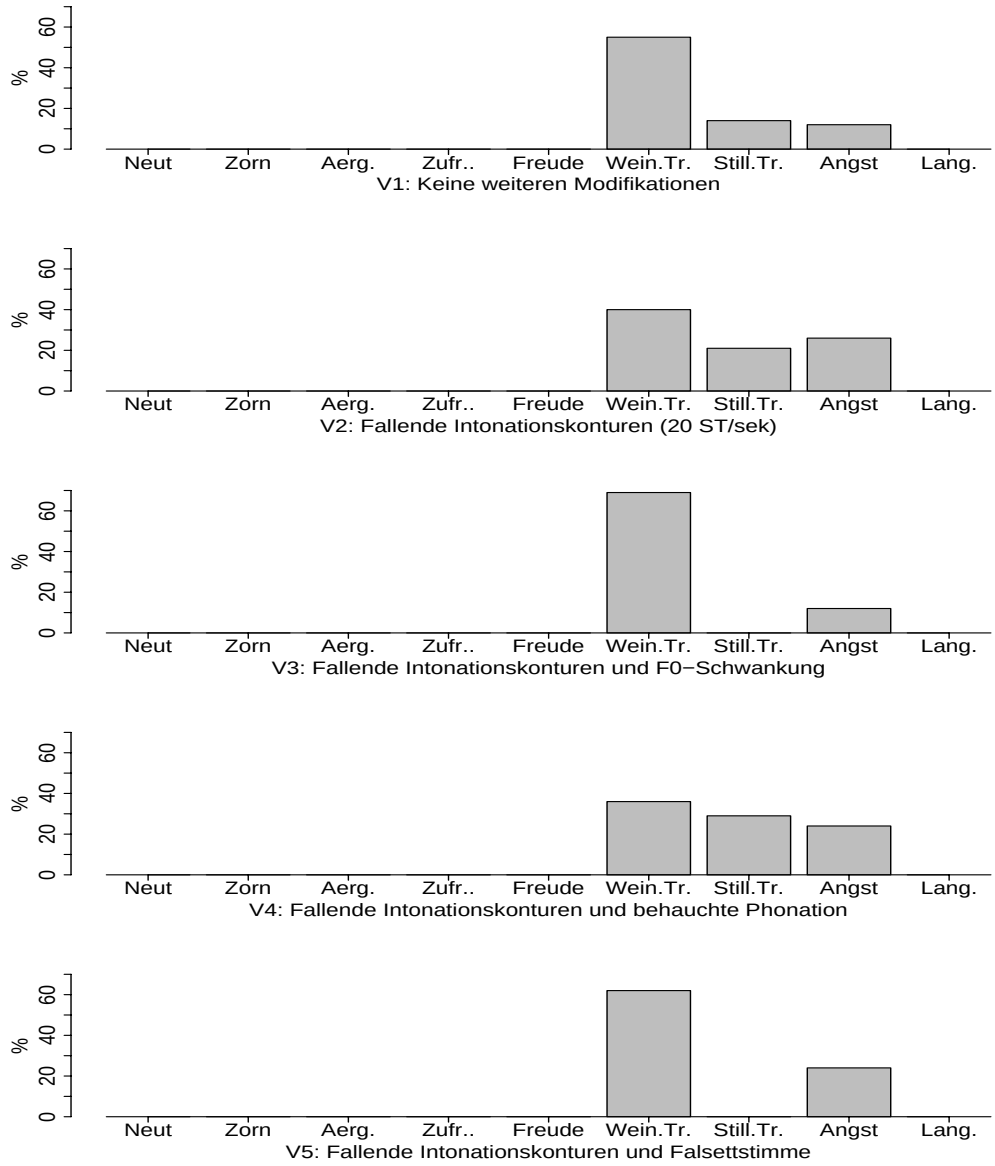

Abbildung 8.5: Verwechslungen der Varianten für weinerliche Trauer (alle Varianten: langsamere Sprechweise (40 %), angehobene F_0 (100 %), schmalerer F_0-Range und Variabilität (20 %))

Eindrucks von Trauer ergeben, obwohl dieses Merkmal im Zusammenhang mit weinerlicher Trauer eine Verwechslung mit stiller Trauer begünstigt hat. Immerhin scheint sich bei behauchter Phonation der Eindruck von Langeweile etwas abzuschwächen, wenn man Version 5 mit Version 2 vergleicht.

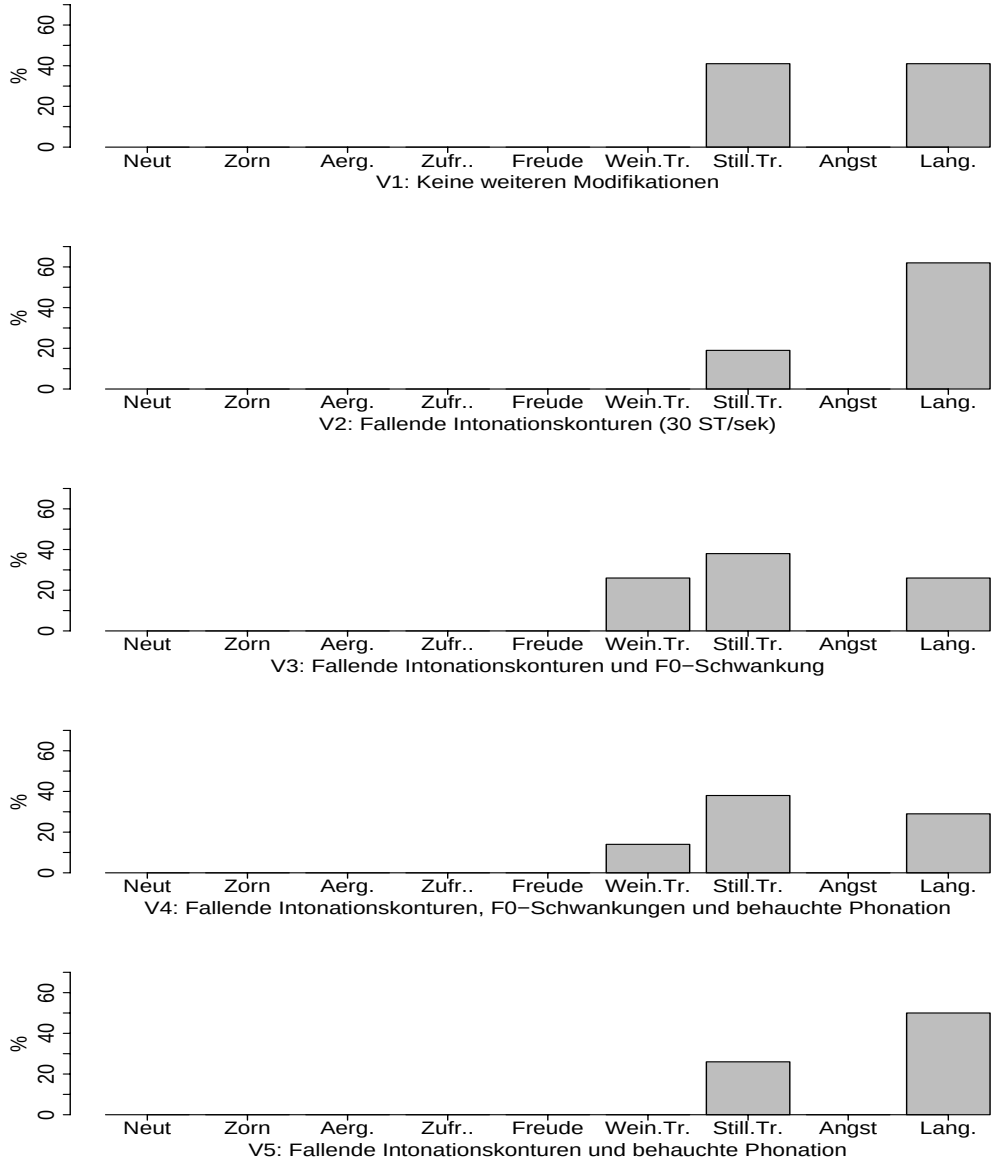

Abbildung 8.6: Verwechslungen der Varianten für stille Trauer (alle Varianten: langsamere Sprechweise (40 %), abgesenkte F_0 (20 %), schmalerer F_0-Range und geringere Variabilität (20 %))

8.4.7 Angst

Alle als ängstlich intendierten Varianten wurden deutlich erkannt (vgl. Abb. 8.7 und Tab. 8.7). Somit sind die Ergebnisse aus Kap. 7 noch einmal bestätigt worden; alleine eine schnellere Sprechweise, eine angehobene Grundfrequenz und ein breiterer F_0-Range führen zu einem Eindruck von Angst. Die Modellierung gerader Intonationskonturen (V3) hat nicht zu einer

Stimulus	⌀[%]	Signifikanz des Unterschiedes			
N:	1,7	V1: ***	V2: ***	V3: ***	V4: ***
V1: F0(150%)	42,9	N: ***	V2: -	V3: -	V4: -
V2: F0(150%), J(300)	42,9	N: ***	V1: -	V3: -	V4: -
V3: F0(150%), PCS(ger.), LSC(stei.,30 ST/sek)	35,7	N: ***	V1: -	V2: -	V4: 0,7
V4: F0(50%), P(Falsett, 100 %)	52,4	N: ***	V1: -	V2: -	V3: 0,7

Tabelle 8.7: Ergebnisse der Varianzanalyse für Angst, alle Varianten: SR(30% schneller), FR(20% breiter). Bedeutung der Abk.: SR=speech rate, F0=⌀F0, FR=F0-range, J=jitter, PCS=pitch contour stressed, LSC=last syllable contour, P=phonation type

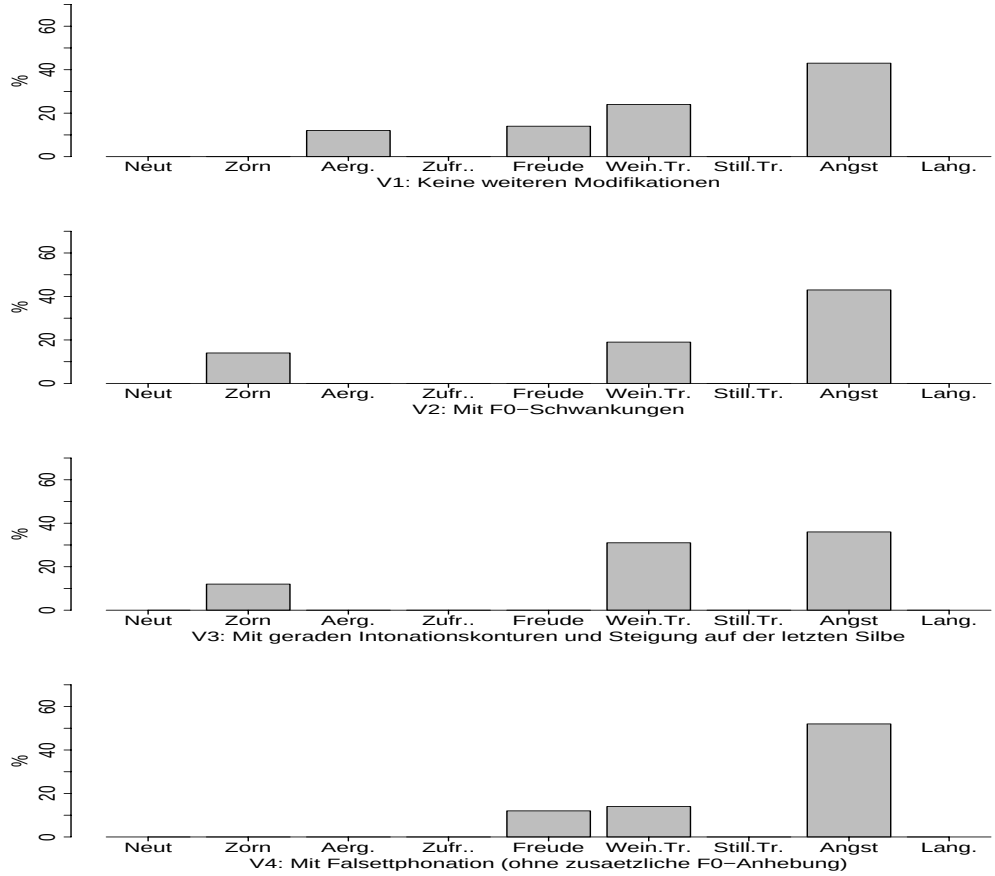

Abbildung 8.7: Verwechslungen der Varianten für Angst (alle Varianten: schnellere Sprechrate (30 %), breiterer F_0-Range (20 %), angehobene F_0 (150 %))

Steigerung der Erkennung geführt, sondern eher zu einer Abschwächung (vgl. mit V4). Interessanterweise ist diese Variante besonders häufig mit weinerlicher Trauer verwechselt worden. Eventuell führen gerade Intonationskonturen eher zu einer Steigerung trauriger Eindruckswir-

kung als fallende. In jedem Fall scheint dieses Merkmal ungeeignet zu sein, um zwischen Angst und weinerlicher Trauer zu diskriminieren.

Die maximale Erkennungsrate liegt allerdings mit 52 % (für V4) unter dem Schnitt der Erkennungsraten für die anderen Emotionen. Dies liegt unter anderem sicher daran, dass die Grundfrequenz aus den oben erwähnten Gründen bei allen Varianten nicht übermäßig angehoben wurde. Daher scheint es lohnenswert, nach weiteren Korrelaten ängstlicher Sprechweise zu suchen (und zwar solchen, die Angst vom Eindruck weinerlicher Trauer abgrenzen). Am deutlichsten grenzen sich interessanterweise die Varianten mit Grundfrequenzirregularitäten (V2 und V4) von weinerlicher Trauer ab. Hieraus ergibt sich ein weiterer Hinweis darauf, dass die Merkmale nicht unabhängig voneinander, sondern in Kombination mit anderen wirken. F_0-Schwankungen bei schmalem Range hatten sich als geeignetes Merkmal gezeigt, um Trauer von Langeweile zu diskriminieren. In Kombination mit einem breiteren Range zeigt sich nun, dass es ebenfalls geeignet ist, um Angst von weinerlicher Trauer abzugrenzen.

8.4.8 Langeweile

Stimulus	⌀[%]	Signifikanz des Unterschiedes				
Neutral:	12,5	V1: ***	V2: ***	V3: ***	V4: ***	V5: ***
V1:	61,9	N: ***	V2: -	V3: -	V4: -	V5: -
V2: P(breathy)	45,2	N: ***	V1: -	V3: -	V4: *	V5: -
V3: P(creaky)	47,6	N: ***	V1: -	V2: -	V4: *	V5: -
V4: UV(30%)	71,4	N: ***	V1: -	V2: *	V3: *	V5: *
V5: UV(30%), P(creaky, breathy)	45,2	N: ***	V1: -	V2: -	V3: -	V4: *

Tabelle 8.8: Ergebnisse der Varianzanalyse für Langeweile, alle Varianten: SR(20% langsamer), SRS(40% langsamer), F0(20% abgesenkt), FR(50% schmaler), FV(20% geringer). Bedeutung der Abk.: SR=speech rate, SRS=stressed syllables speech rate, F0=⌀F0, FR=F0-range, FV=F0-variability, UV=vowel target undershoot, P=phonation type.

Alle Varianten mit intendierter Langeweile wurden weit über dem Zufallsniveau erkannt (vgl. Abb. 8.8 und Tab. 8.8). Die Versionen mit manipulierter Phonationsart (V2, V3 und V5) wurden allerdings ebenso häufig mit stiller Trauer verwechselt wie erkannt. Dies weist darauf hin, dass eine entspannte Phonationsart wie Knarrstimme[6] oder Behauchung eher ein Korrelat von stiller Trauer als von Langeweile ist. Bei den Varianten für stille Trauer hat sich gezeigt, dass diejenigen mit fallenden Silbenkonturen auffällig oft mit Langeweile verwechselt wurden. Der Verdacht liegt nahe, das die Merkmale hier gerade für die jeweils andere Emotion gelten: Langeweile sollte ohne eine Änderung der Stimmqualität aber mit fallenden Silbenkonturen und stille Trauer mit behauchter/geknarrter Stimmgebung aber ohne fallende Konturen modelliert werden. Es ist damit eine Variante von Langeweile denkbar, für die das Merkmal monotone

[6] Für eine Diskussion bzgl. Knarrstimme mit entspannter Larynxmuskulatur siehe Kap. 6.3.

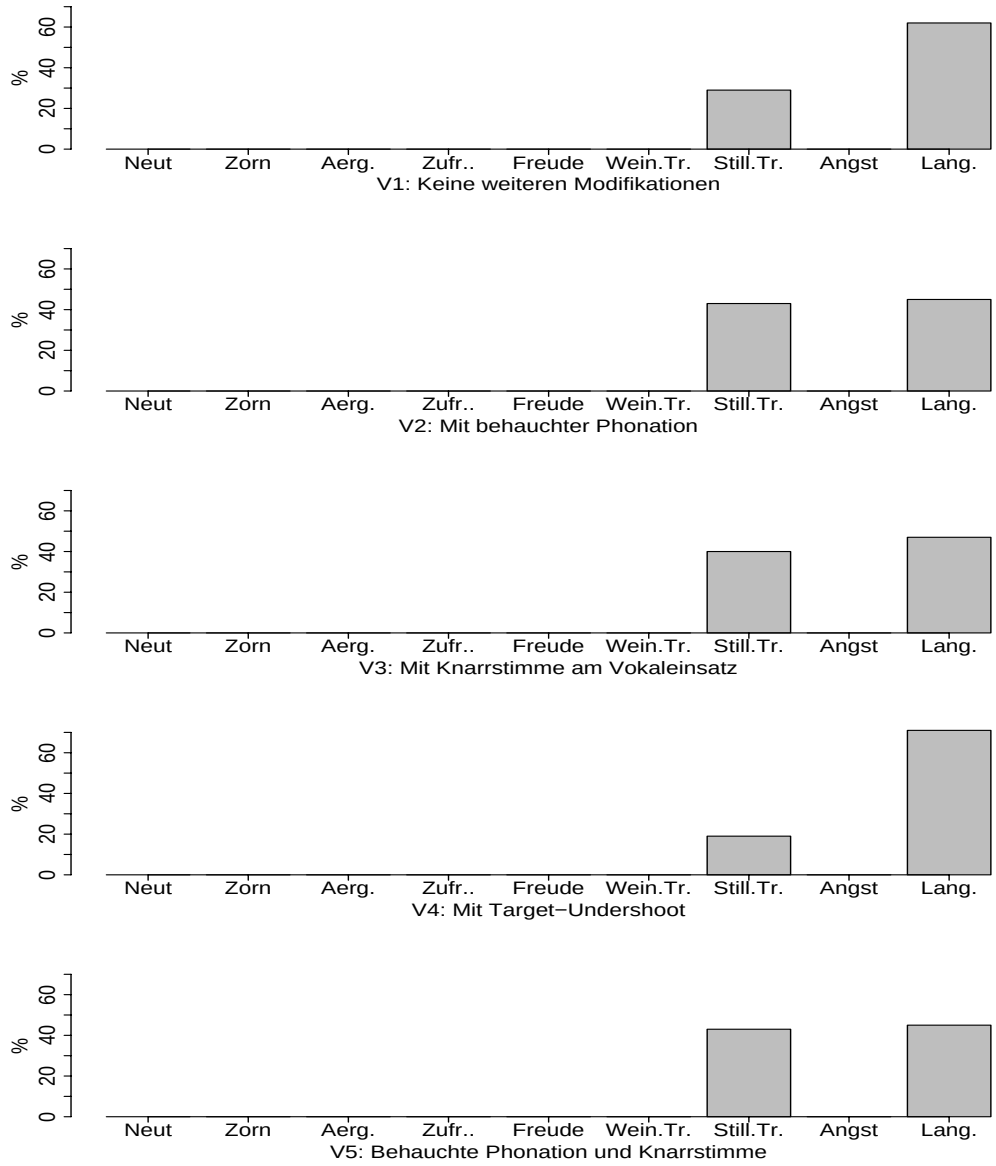

Abbildung 8.8: Verwechslungen der Varianten für Langeweile (alle Varianten: langsamere Sprechweise (20 %) und betonte Silben (40%), abgesenkte F_0 (20 %), schmalerer F_0-Range (20 %) und geringere Variabilität (50 %))

Intonationskontur nicht gilt. Paeschke & Sendlmeier (2000) kommen bei der Analyse natürlichsprachiger Äußerungen zu dem selben Schluss. Die Modellierung von Formant-Target Undershoot führte in diesem Zusammenhang nicht zu einer signifikanten Steigerung gelangweilter Eindruckswirkung (V4 gegenüber V1). Dies ist allerdings nicht als ein Zeichen mangelnder perzeptiver Relevanz des Merkmals Formant-Target Undershoot zu bewerten. Vielmehr erklärt

sich diese Tatsache dadurch, dass bereits die Grundversion so oft erkannt wurde, dass sich eine Steigerung des gelangweilten Ausdrucks nicht mehr signifikant im Beurteilerverhalten ausgedrückt hat.

8.5 Zusammenfassende Diskussion der Ergebnisse

Die Resultate dieses Experimentes lassen sich hinsichtlich zweier Fragestellungen interpretieren. Einerseits ging es darum, die Relevanz spezieller akustischer Korrelate für ausgewählte Emotionen zu überprüfen. Andererseits konnte durch die gezielte Manipulation der Stimuli die Güte des Synthesesystems EmoSyn hinsichtlich der Simulation emotionaler Sprechweise evaluiert werden. Diese beiden Aspekte werden im Folgenden getrennt diskutiert.

8.5.1 Akustische Profile der einzelnen Emotionskategorien

Bei den Emotionen Zorn und Ärger hat sich die bereits im vorherigen Experiment nachgewiesene günstige Wirkung schnellerer Sprechweise und gespannter Phonation deutlich bestätigt. Eine Trennung zwischen Zorn und Ärger konnte allerdings nicht erreicht werden. Nach Meinung des Autors liegt dies vor allem daran, dass das System dahingehend verbessert werden müsste, dass extremere Veränderungen des neutralen Ausgangsstimulus möglich werden, ohne dabei die Natürlichkeit der Ausgabe übermäßig zu beeinträchtigen. Solche Manipulationen führen leicht zu Artefakten und es bedarf eines beträchtlichen Forschungsaufwandes, diese abzuschwächen. Weiterhin ergaben sich bei beiden Emotionen keine signifikanten Unterschiede bzgl. der Erkennung für die einzelnen Varianten. Daraus ist allerdings nicht zu schließen, dass die variierten Merkmale keine perzeptive Relevanz haben. Vielmehr sind Hörexperimente, bei denen vor allem die Auswahl an intendierten Emotionen groß ist, ungeeignet, feinere Differenzierungen im Hörerverhalten aufzudecken. Wären die Beurteiler lediglich mit der Frage: ”Welcher Stimulus klingt zorniger/ärgerlicher?” konfrontiert worden, so hätten sich sicherlich signifikante Unterschiede ergeben. Bei diesem Testdesign hat sich lediglich eine Tendenz für die perzeptive Relevanz von Formant-Target Overshoot ergeben (wie sie ja von Kienast & Sendlmeier (2000) prädiziert wurde).

Bei dem Emotionspaar Freude/Zufriedenheit hat sich - wie oben ausgeführt - das Intonations-Wellenmodell als äußerst günstig für die Erkennung erwiesen. Der Autor geht wie gesagt davon aus, dass es gerade die Kombination aus gleichmäßiger F_0-Bewegung und glatten F_0-Verläufen ist, die diese Wirkung hervorruft. So beschreiben Abadjieva et al. (1993) eine gleichmäßige Betonungsstruktur als typisch für freudige Sprechweise und in Murray & Arnott (1995, 1993) werden ’*smooth, upward inflections*’ als charakteristisch für Freude erwähnt. Sehr deutlich hat sich bei beiden Emotionen die Relevanz der akustischen Charakteristika von Lippenspreizung bestätigt. Weiterhin ergaben sich Tendenzen für die Relevanz steigender Intonationskonturen auf den betonten Silben. Dabei haben sich die Profile für die beiden Emotionen als durchaus unterschiedlich erwiesen, da beide Emotionen sehr selten miteinander verwechselt wurden.

Für weinerliche bzw. stille Trauer konnte die günstige Wirkung von F_0-Irregularitäten nachgewiesen werden. Zudem gab es einige Hinweise auf die Relevanz behauchter Phonation. Auch dieses Emotionspaar scheint für die Hörer deutlich zwei Kategorien zu bilden. Es gab zwar Verwechslungen, weinerliche Trauer wurde jedoch häufiger mit Angst und stille Trauer eher mit Langeweile verwechselt[7]. Weinerliche Trauer wurde insgesamt zuverlässiger erkannt. Allein die Kombination aus angehobener F_0-Lage, langsamem Sprechtempo und schmalerem F_0-Range genügte bereits für eine deutliche Erkennung. Ein weiteres Resultat ist dadurch gegeben, dass fallende Intonationskonturen die Erkennung eher vermindern, da die Stimuli dann eher mit Langeweile verwechselt wurden. Die Tatsache, dass die ängstlich intendierten Stimuli mit geraden Intonationskonturen auffallend häufig mit weinerlicher Trauer verwechselt wurden, lässt darauf schließen, dass gerade Intonationskonturen eher einen Eindruck von Trauer fördern.

Die ängstlich gemeinten Stimuli wurden ebenfalls bereits durch eine Verbreiterung des F_0-Range, Anhebung der F_0-Lage und Verkürzung der Silbendauern zuverlässig erkannt. Dabei wirkten gerade Intonationsverläufe einer Erkennung entgegen. Da die durchschnittliche Erkennung allerdings verglichen mit den anderen Emotionen geringer ist, scheinen wichtige akustische Korrelate unberücksichtigt geblieben zu sein. Dies sind vor allem solche, die Angst von weinerlicher Trauer diskriminieren.

Die Ergebnisse für Langeweile weisen darauf hin, dass eine behauchte Phonation eher bei trauriger Sprechweise auftritt. Der günstige Effekt von Formant-Target Undershoot konnte repliziert werden. Da einige der Prototypen anderer Emotionen (z.B. Zufriedenheit oder stille Trauer) mit bewegteren Intonationskonturen mit Langeweile verwechselt wurden, scheint ein relativ monotoner F_0-Verlauf kein notwendiges Merkmal gelangweilten Sprechausdrucks zu sein. Offensichtlich ist die Kategorie 'Langeweile' nicht genügend spezifiziert, um zuverlässige Aussagen über die Intonationskontur zu machen. Es scheint mehrere Kategorien zu geben, für die jeweils ein bewegterer oder ein monotoner F_0-Verlauf adäquat ist (vgl. auch Paeschke & Sendlmeier (2000)).

8.5.2 Gütebewertung des verwendeten Synthesesystems

Eine Verwechslungsmatrix für die Stimuli, die jeweils am häufigsten erkannt wurden, ist in Abb. 8.9 dargestellt. Es zeigt sich, dass es durchaus gelungen ist, durch das verwendete Synthesesystem fast alle Emotionen deutlich voneinander abzugrenzen. Fasst man die Erkennungsraten für die paarweisen Emotionen zusammen (siehe Abb. 8.10), so wurden alle Basisemotionen außer Angst von über 70 % der Hörer erkannt. Die Erkennungsraten liegen damit in einem Bereich, der auch bei Stimuli von menschlichen Enkodierern erreicht wird (vgl. etwa Banse & Scherer (1996)).

Allerdings trugen viele der Manipulationen nicht zu einer Steigerung und manche sogar zu einer Verringerung der Erkennung bei. Dies liegt wie gesagt zum einen daran, dass die Komplexität der Fragestellung sehr hoch war. Einzeltests sind eher dazu geeignet, die Relevanz relativ subtiler Merkmale zu überprüfen. Zum anderen haben sich einige Merkmale als weniger geeignet erwiesen. Durch die hohe Komplexität der Wechselwirkung der Merkmale ließen sich

[7]Die aufgetretenen Verwechslungen bestätigen übrigens recht deutlich die Validität des Dimensionsmodells. Verwechslungen traten fast immer bei eng benachbarten Emotionen auf (vgl. Abb. 1.1) .

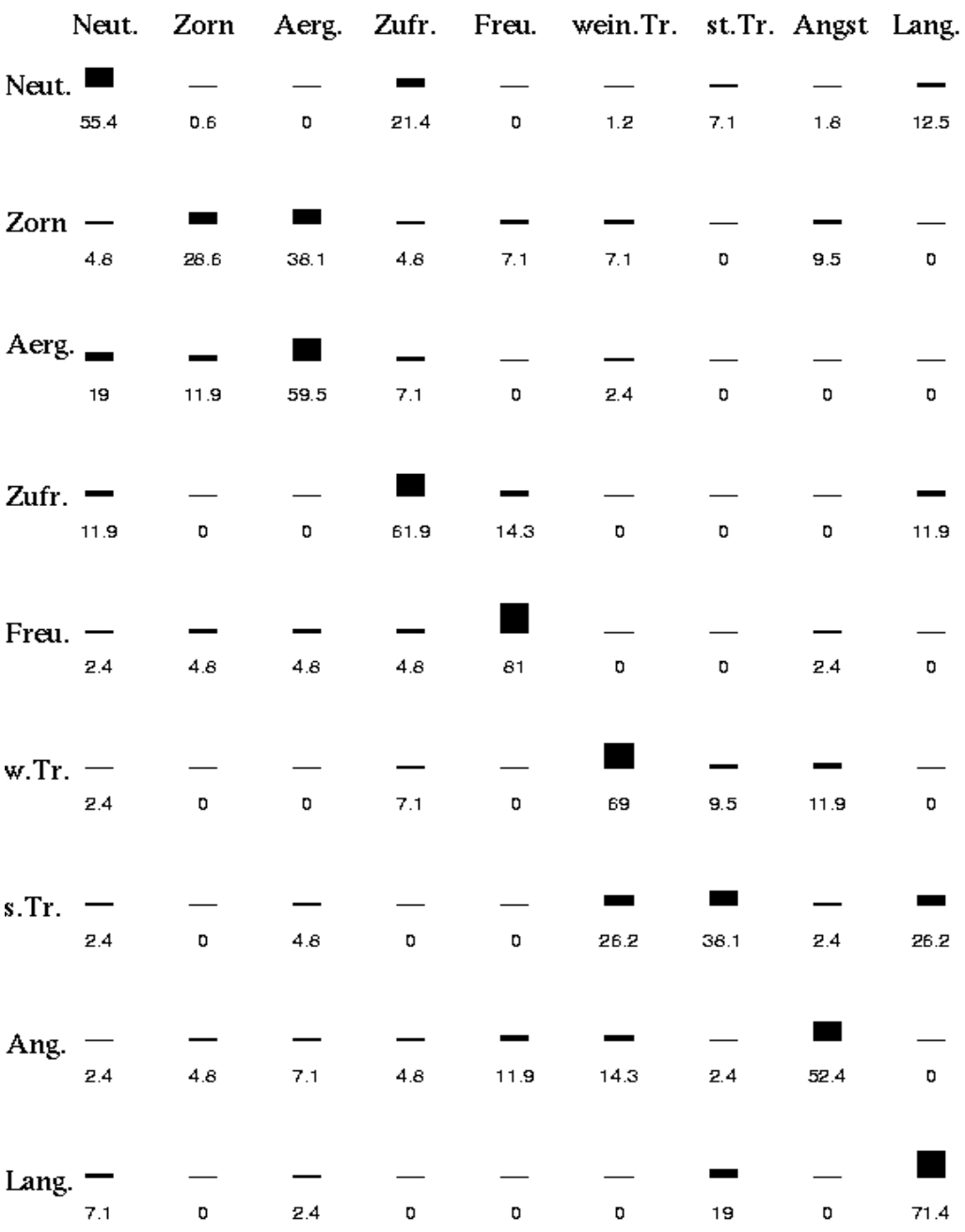

Abbildung 8.9: Verwechslungsmatrix für die Stimuli mit den höchsten Erkennungsraten. Die Reihen stehen für die wahrgenommene Emotion, die Spalten für die intendierte

allerdings beliebig viele Versuche durchführen, die jeweils eine Erkennung bestimmter Emotionen förderlich wären.

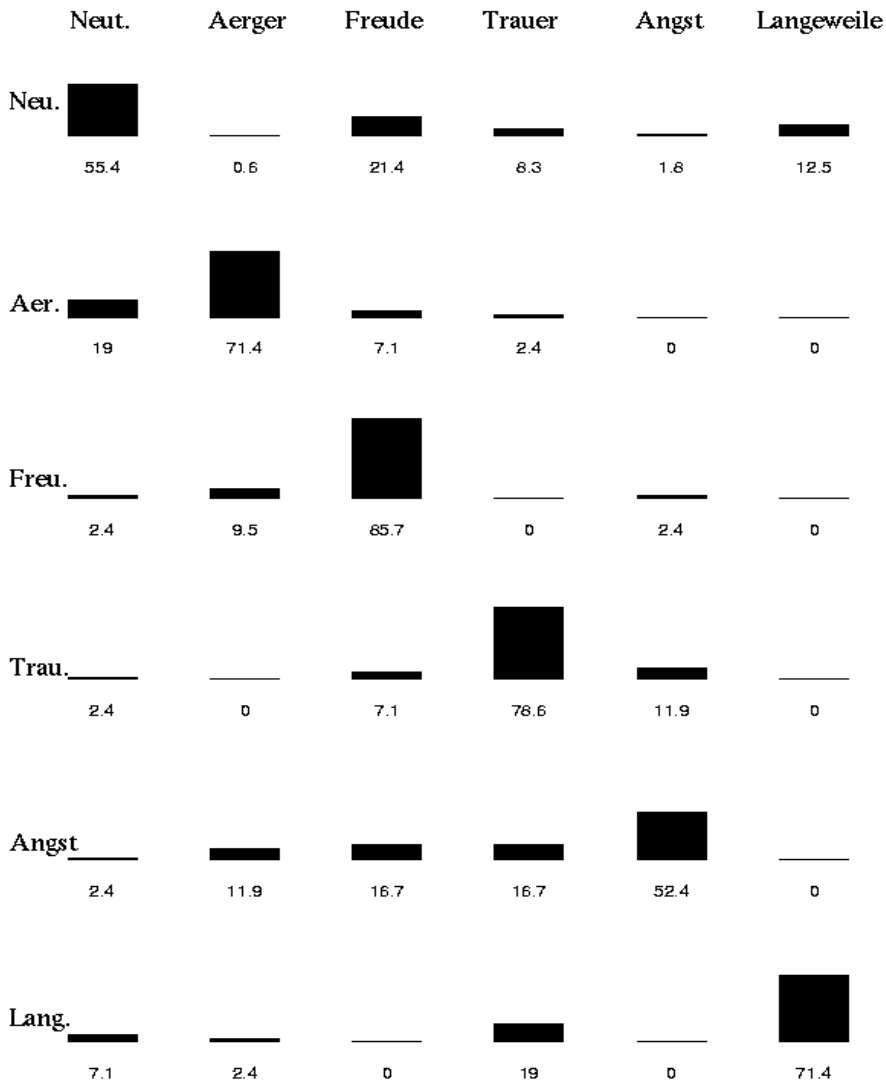

Abbildung 8.10: Verwechslungsmatrix für die Stimuli mit den höchsten Erkennungsraten, wobei qualitativ ähnliche Emotionen zusammengefasst sind. Die Reihen stehen für die wahrgenommene Emotion, die Spalten für die intendierte.

Einige Schwächen des Systems haben sich abgezeichnet. Um Emotionen mit starkem Erregungsgrad simulieren zu können, ist eine bessere Berücksichtigung der entstehenden Artefakte notwendig. Im Hinblick auf die Modellierung der Intonationskontur erscheint die Einteilung der Intonationseinheiten in Phrase, Silbe und Laut als unzureichend. Bei Paeschke & Sendlmeier (2000) wird etwa das Konstrukt Akzent (vgl. Abschnitt. 2.2.1) verwendet, um Betonungsgruppen zu definieren. Ähnliche Phrasenabschnitte sollten in der Datenstruktur des Synthesesystems aufgenommen werden.

Kapitel 9

Zusammenfassung und Ausblick

Das Anliegen der vorliegenden Arbeit hatte im wesentlichen zwei Aspekte. Erstens sollte die perzeptive Relevanz akustischer Korrelate von emotionaler Sprechweise mit Sprachsyntheseexperimenten überprüft werden (Analysis-by-Synthesis). Die zu überprüfenden Merkmale stammen dabei vorrangig aus empirischen Untersuchungen, bei denen emotionales Sprachmaterial analysiert wurde. Der zweite Aspekt dieser Untersuchung betraf die Umsetzung der überprüften Eigenschaften in Regeln, die dazu geeignet sind, die Ausgabe bestehender Sprachsynthesizer zu emotionalisieren.

Um sich dem Gegenstandsbereich anzunähern, wurde zunächst im 1. Kapitel der Begriff 'Emotion' diskutiert. Dabei ergab sich, dass die in Kleinginna & Kleinginna (1981) gegebene Arbeitsdefinition den Anforderungen der vorliegenden Arbeit genügt. Eine weitere Folgerung aus diesem Kapitel besteht darin, dass sowohl der kategoriale als auch der dimensionale Ansatz zur Emotionsbeschreibung bei der vorliegenden Fragestellung günstig sind. Mit der Component Process Theorie von Scherer (1984) wurde ein Modell vorgestellt, welches emotionale Zustände mit physiologischen Auswirkungen verbindet. Der Stand der Forschung wurde in Kap. 4 reflektiert, welches damit sowohl die zu überprüfenden Korrelate liefert als auch Ergebnisse früherer Arbeiten mit ähnlicher Fragestellung diskutiert.

Synthesesysteme, die auf Segmentverkettung im Zeitbereich basieren, sind derzeit vor allem im kommerziellen Bereich stark verbreitet. Um ihre Eignung bzgl. der Simulation emotionaler Sprechweise zu evaluieren, wurde in Kap. 5 ein Perzeptionsexperiment vorgestellt, bei dem die Ausgabe eines solchen Systems unter zwei verschiedenen Voraussetzungen (regel- und datenbasiert) überprüft wurde. Dabei ergab sich, dass es zwar grundsätzlich möglich ist, wiedererkennbar emotionalen Sprechausdruck für vier Basisemotionen zu simulieren. Jedoch lagen die Erkennungsraten deutlich unter solchen, die bei menschlichen Enkodierern gefunden wurden. Darüberhinaus ergaben sich beträchtliche Unterschiede zwischen den Emotionen. So wurden Freude und Wut deutlich schlechter als Trauer und Angst erkannt, was auf die Relevanz unberücksichtigter Merkmale hinweist. Zudem ergab sich die Erkennung vor allem bei der regelbasierten Simulation. Die reine Copy-Synthese der prosodischen Merkmale natürlicher Sprachaufnahmen wurde deutlich schlechter erkannt, obwohl die Originalaufnahmen Erkennungsraten von mindestens 80 % erreichten. Speziell hier zeigt sich, dass der emotionale Ausruck stark in nichtprosodischen Spracheigenschaften enkodiert sein muss.

Ein speziell für die vorliegende Arbeit entwickelter Sprachsynthesizer ist in Kap. 6 beschrieben. Er basiert auf der Formantsynthese, die zur Zeit den besten Kompromiss aus Flexibilität in der Generierung des Sprachsignals und Qualität bzgl. der Ausgabe bei unbegrenzter Synthese darstellt. Mittels dieses Synthesesystems wurde ein Perzeptionsexperiment durchgeführt, das in Kap. 7 beschrieben ist. Dabei wurden fünf Merkmale aus den Bereichen Prosodie, Stimmqualität und Artikulation systematisch variiert. Die Testhörer sollten die resultierenden Stimuli einer von fünf Emotionen oder 'Neutral' zuweisen. Für alle Merkmale konnten signifikante Haupteffekte bei den meisten Emotionen nachgewiesen werden. Daraus ergaben sich bereits zwei Ergebnisse bzgl. der grundlegenden Fragestellung der vorliegenden Arbeit. Erstens sind auch eher subtile und bislang seltener untersuchte Merkmale wie artikulatorische Reduktion bzw. Elaboration perzeptiv relevant. Zweitens ermöglicht der verwendete Synthesizer eine erkennbare Simulation emotionalen Sprechausdrucks, was in hohem Maße durch die Modellierung von unterschiedlichen Phonationsarten ermöglicht wird, da ja die Variation dieses Merkmals bei allen Emotionen Haupteffekte ergab.

Allerdings ergibt sich bei systematischer Variation relativ grundlegender Parameter das Problem, dass die Anzahl der variierten Merkmale und ihrer Ausprägungen aufgrund von forschungsökonomischen Beschränkungen gering gehalten werden muss. Dies hat zur Folge, dass für einige der Emotionen deutlich günstige Ausprägungskombinationen in der Stimulusmenge nicht vorkommen und weiterhin zahlreiche Eigenschaften nicht überprüft werden können. Zudem zeigte sich bei dem Perzeptionsexperiment, dass einige Emotionsbegriffe nicht ausreichend spezifiziert worden waren. Daher wurde ein weiteres Perzeptionsexperiment durchgeführt, welches im 8. Kapitel beschrieben ist. Dabei wurde einmal die Menge der Emotionen von fünf auf acht erweitert, indem Freude, Wut und Trauer durch jeweils zwei Subkategorien repräsentiert wurden, die sich hauptsächlich durch den Erregungsgrad unterscheiden. Weiterhin wurden diesmal nicht nur einzelne Merkmale variiert, sondern Prototypen miteinander verglichen, die in ausgewählten Merkmalen variierten. Auf diese Art lassen sich spezifischere Fragen beantworten (z.B.: "*Führt das Merkmal gespreizter Lippen eher zu einer Attribuierung als Freudig?*"), als dies bei systematischer Variation einfacher Merkmale möglich wäre. Zudem lässt sich die Fähigkeit des Synthesizers, emotionalen Sprechausdruck zu simulieren, so besser evaluieren. Tatsächlich gab es für alle Emotionen außer Zorn, der oft mit Ärger verwechselt wurde, mindestens einen Prototypen, dessen Erkennungsrate deutlich über dem Zufallsniveau lag und der am häufigsten der intendierten Emotion zugewiesen wurde. Die Erkennungsraten lagen insgesamt durchaus in einem Bereich, der auch bei menschlichen Enkodierern erzielt wird, vor allem, wenn man von der Differenzierung bzgl. des Erregungsgrads abstrahiert. Interessante Hinweise für die Wirkung der Merkmalsausprägungen auf das Beurteilerverhalten haben sich auch aus den Verwechslungen ergeben, so wurden einige der Emotionen bei bestimmten Varianten zwar nicht häufiger erkannt, aber seltener verwechselt, was auf eine ungünstige Wirkung dieser Merkmalsausprägung für die zuvor verwechselte Emotion hinweist.

Der Nachteil dieser Prototypenbildung besteht natürlich darin, dass die variierten Merkmale äußerst komplex sind und daher der Bezug von Ursache und Wirkung unklar wird. So müssen Perzeptionsexperimente immer einen *trade-off* zwischen der Deutlichkeit der Erkennungsrate und geringer Komplexität der Variation eingehen. Speziellere Fragen lassen sich weiterhin leichter beantworten, wenn man die Menge der Emotionen stark einschränkt. Letzt-

endlich ist wegen des unscharfen Emotionsbegiffs und der unendlichen Menge der Merkmale die Durchführung des 'ultimativen' Perzeptionsexperiments ohnehin nicht möglich, sondern es können jeweils nur klar umrissene Fragestellungen beantwortet werden.

9.1 Zusammenfassung der Ergebnisse

Im folgenden werden die Ergebnisse der drei Perzeptionstests (vgl. Kap. 5, 7 und 8) für die einzelnen Emotionen zusammengefasst.

- **Ärger**
  Für Ärger zeigt sich bei der Simulation mit prosodischen Merkmalen eine starke Diskrepanz bzgl. der Erkennung bei daten- (34 %) und regelbasierter (55 %) Simulation. Da die regelbasierten Stimuli deutlich besser erkannt wurden, müssen die Schauspieler den emotionalen Ausdruck vor allem durch stimmqualitative und artikulatorische Merkmale erzeugt haben. Immerhin zeigt das Ergebnis, dass Ärger auch durch rein prosodische Variation recht gut simuliert werden kann. Diese Aussagen wurden bei dem Experiment mit systematischer Variation der Merkmale bestätigt. Die Variation der Merkmale Phonationsart und Sprechgeschwindigkeit ergab sehr deutlich, dass eine gespannte Stimme sowie schnellere Sprechweise wütend wirken, die Kombination aus beidem führte zu einer Erkennung von 64 %. Die Ergebnisse für das dritte Perzeptionsexperiment unterstreichen die günstige Wirkung gespannter Phonation für wütenden Sprechausdruck. Alle Versionen wurden als ärgerlich erkannt, unabhängig davon, ob die F_0-Lage angehoben oder abgesenkt und ob die Sprechgeschwindigkeit schneller oder langsamer ist. Es gelang dabei nicht, den Unterschied zwischen Zorn (*hot anger*) und Ärger (*cold anger*) zu simulieren, fast alle Varianten wurden mehrheitlich als ärgerlich klassifiziert. Als weiteres Resultat ergab sich die Tendenz, dass die Modellierung von Formant-Target Overshoot bei den Vokalen eher ärgerlich wirkt.

- **Freude**
  Für Freude ergab sich bei der Simulation mit Zeitbereichssynthese sowohl bei den daten- (21 %) als auch bei den regelbasierten (35 %) Stimuli eine im Vergleich mit den anderen Emotionen eher geringe Erkennungsrate, was auf die Prominenz nichtprosodischer Parameter bzw. komplexerer Intonationsmodelle hinweist. Dies bestätigte sich bei der systematischen Variation, bei der ebenfalls Freude am schlechtesten erkannt wurde. Immerhin ergab sich dabei eine signifikante Steigerung der Antworthäufigkeiten bei breiterem F_0-Range, schnellerer Sprechweise und Formant-Target Overshoot. Die Relevanz des letzten Merkmals konnte allerdings bei dem Perzeptionsexperiment mit gezielter Variation nicht bestätigt werden, vielmehr wurde der Prototyp mit Formant-Target Overshoot als einziger mit Zorn verwechselt. Die Anhebung der ersten zwei Formanten sowie das Intonationskontur-Wellenmodell führten deutlich zu einer Steigerung freudig/zufriedener Eindruckswirkung. Darüberhinaus konnte die günstige Wirkung von steigenden Intonationskonturen auf den betonten Silben nachgewiesen werden. Die Trennung des Konzepts 'Freude' in eine aktivere (*Freude*) und eine passivere (*Zufriedenheit*) Komponente auf-

grund gegensätzlicher Modellierung von Merkmalen der Erregung gelang dabei sehr gut, da die Prototypen fast nie miteinander verwechselt wurden.

- **Trauer**
  Die Simulation von Trauer zeigte bei der zeitbereichsbasierten Simulation sowohl für die kopierten als auch für die anhand von Regeln generierten Stimuli eine relativ hohe Erkennungsrate (62 bzw. 50 %). Dabei wurde stille Trauer simuliert. Da die Kategorie 'Langeweile' bei diesem Versuch nicht zur Auswahl angeboten wurde, kam es nicht zu Verwechslungen. Bei dem Hörversuch mit systematischer Merkmalsvariation führte die Verwechslung mit Langeweile jedoch dazu, dass bei den in der Literatur prädizierten Merkmalskombinationen nicht für Trauer entschieden wurde. Grundsätzlich führten ein schmalerer F_0-Range und eine Verlangsamung der Sprechgeschwindigkeit häufiger zu trauriger Eindruckswirkung. Bezüglich der Phonationsarten ergab sich, dass sowohl Falsettanregung als auch behauchte oder Knarrphonation eher traurig wirken. Da diese Merkmalsausprägungen widersprüchlich sind, wurde als Konsequenz daraus die Kategorie Trauer bei dem Experiment mit variierten Prototypen in *stille* und *weinerliche* Trauer unterteilt. Dabei wurde weinerliche Trauer oft mit Angst verwechselt und stille Trauer mit Langeweile. Für beide Emotionen konnte die günstige Wirkung von F_0-Irregularitäten nachgewiesen werden. Die Auswertung der Verwechslung mit Langeweile oder Angst weist darauf hin, dass gerade Intonationsverläufe auf Silbenebene eher traurig wirken als fallende. Auch bei diesem Emotionspaar konnten die Testhörer deutlich zwischen der erregteren und der ruhigeren Form diskriminieren.

- **Angst**
  Auch bei Angst ergab sich, ähnlich wie bei Ärger, dass die Simulation mit dem Zeitbereichsverfahren zwar auch für die datenbasierten Stimuli erkannt wurde (44 %), die regelbasierten Stimuli jedoch eine deutlich höhere Erkennungsrate aufwiesen (78 %). Das Hauptmerkmal bei den regelbasierten Stimuli bestand in der sehr starken Anhebung der Grundfrequenz und einer schnelleren Sprechgeschwindigkeit. Auch für das Perzeptionsexperiment mit systematischer Variation ergab sich, dass die Stimuli mit stark erhöhter Grundfrequenz (bei Kombination aus angehobener F_0-Lage und Falsettanregung) zu 66 % mit Angst attribuiert wurden. Weiterhin führte ein breiterer F_0-Range und eine schnellere Sprechweise, auch bei langsameren betonten Silben, zu einer häufigen Erkennung der Stimuli als ängstlich. Diese Resultate wurden bei dem Hörversuch mit Prototypenvariation bestätigt. Zusätzlich ergab sich, dass Intonationskonturen mit geradem Verlauf auf den betonten Silben einer ängstlichen Eindruckswirkung entgegen wirken und eher eine Verwechslung mit weinerlicher Trauer begünstigen.

- **Langeweile**
  Bei der systematischen Merkmalsvariation ergab sich eine Steigerung von gelangweilter Eindruckswirkung bei behauchter Phonation, schmalerem F_0-Range, abgesenkter F_0-Lage, verlangsamter Sprechweise und Formant-Target Undershoot der Vokale. Damit waren alle Merkmale muskulärer Entspannung einem Ausdruck von Langeweile förderlich. Auch bei der Variation von Prototypen wurden alle Stimuli aufgrund der abgesenkten

F_0-Lage, des schmaleren F_0-Range und stark verlangsamter Sprechweise erkannt. Die Versionen mit manipulierter Phonationsart wurden allerdings häufig mit stiller Trauer verwechselt. Da andererseits einige der Prototypen für andere Emotionen wie Zufriedenheit oder stille Trauer mit bewegteren Intonationskonturen mit Langeweile verwechselt wurden, scheint der Ausdruck von Langeweile auch durch nicht-monotone Intonationskonturen möglich zu sein. Diese Beobachtung wurde auch bei Paeschke & Sendlmeier (2000) gemacht. Die Autoren lösen den scheinbaren Widerspruch durch Subkategorisierung auf (’gähnende’ vs. ’normale’ Langeweile).

9.2 Ausblick

Die vorliegende Arbeit ließe sich in viele Richtungen ausweiten:

- Besonders das Perzeptionsexperiment mit Merkmalsvariation hat gezeigt, dass für spezielle Fragestellungen die Untersuchung einer einzelnen Emotion günstiger ist.
- Zukünftige Forschung könnte sich auf die Untersuchung bislang unberücksichtigter Merkmale konzentrieren. So wurden etwa die von Sendlmeier (1997) beschriebenen Merkmale hinsichtlich der Asynchronizität der Amplituden- und F_0-Gipfel nicht auf perzeptive Relevanz hin überprüft. Weiterhin wurden Merkmale der Assimilation oder Elision auf Laut- und Silbenebene nicht berücksichtigt, die z.B. in Kienast & Sendlmeier (2000) untersucht wurden.
- Im Hinblick auf die Erweiterung der Flexibilität von Sprachsynthesizern wäre eine ’Renaissance’ der Formantsynthese wünschenswert. So ist etwa eine Beschränkung auf eine maximale Frequenz von 5 kHz nicht mehr zeitgemäß. In den letzten Jahren stand die Entwicklung von Sprachsynthesizern, die auf der Konkatenation im Zeitbereich basieren, stark im Vordergrund, da sie eine relativ hohe Sprachgüte mit relativ simplen Berechnungsalgorithmen verbinden. Sie sind jedoch bzgl. der Modellierung extralinguistischer Merkmale eher unflexibel.
- Auch die artikulatorische Synthese könnte große Dienste bei der Simulation von Merkmalen leisten, die in direktem Zusammenhang mit motorischer Kontrolle stehen. Dies setzt allerdings voraus, dass es gelingt, zumindest für einzelne Testsätze eine natürlich klingende Sprachausgabe mit artikulatorischer Synthese zu erzielen.
- Der in der vorliegenden Arbeit entwickelte Sprachsynthesizer könnte bzgl. der Umsetzung komplexer Merkmale, wie etwa der Phonationsarten oder bestimmter Intonationskonturen, deutlich verbessert werden. Speziell extreme Manipulationen der Grundfrequenz führen leicht zu Artefakten, was aus dem resultierenden Missverhältnis aus F_0 und Spektrum resultiert.
- Weiterhin könnte der Synthesizer um linguistische Abstraktionseinheiten erweitert werden, wie z.B. die unter Abschnitt 2.2.1 beschriebene Begrifflichkeit ’Akzent’. Dadurch würde die Modellierung intonatorischer Phänomene erheblich vereinfacht.

- Bislang wurde der Einfluss emotionaler Erregung auf das Sprechverhalten fast immer anhand einzelner Sätze überprüft. Gerade im Hinblick auf mögliche Anwendungen wäre es aber wichtig, den Einfluss auch über längere Zeiträume hin zu untersuchen. Wenn z.B. das Ziel einer solchen Forschung darin besteht, die Monotonie von Sprachsynthesizern zu verringern, so ist die Entwicklung von Intonationsmodellen für größere Sprechabschnitte ein wichtiger Schritt.

- Es wäre von Interesse, den Einfluss der unterschiedlichen Emotionsdimensionen auf das Sprachsignal genauer zu überprüfen. Bislang wurde vor allem der Einfluss der Erregungsdimension auf das Sprechverhalten betrachtet, die am deutlichsten mit Muskelspannungszuständen korreliert. Die akustischen Korrelate von Subdominanz oder Valenz sind jedoch noch weitgehend unerforscht.

- Bei der Simulation emotionaler Sprechweise mit TTS-Systemen wäre zu beachten, bei welchem Grad der Modfifikation eine Stimme so verändert klingt, dass sie nicht mehr als Ausdruck einer Person, sondern als der einer anderen Person wahrgenommen wird.

- Als ”Blick über den Tellerrand” (für den Phonetiker) wäre es äußerst interessant , computerlinguistische Modelle zur semantischen Analyse auf die Interpretation emotionalen Ausdrucks hin zu erweitern. Derartige Modelle wären für ”intelligente” Vorlesemaschinen unabdingbar.

Literaturverzeichnis

E. Abadjieva, I. R. Murray & J. L. Arnott. Applying Analysis of Human Emotional Speech to Enhance Synthetic Speech. In *Proc. 1993 Eurospeech, Berlin*, Seiten 909–912, 1993.

T. V. Ananthapadmamnabha. Acoustic Analysis of Voice Source Dynamics. *Speech Transmiss. Lab. - Q.Prog.Stat.Rep (STL-QPSR), Royal Institute of Technology, Stockholm*, 2/3:1–24, 1984.

R. Banse & K. R. Scherer. Acoustic Profiles in Vocal Emotion Expression. *Journal of Personality and Social Psychology*, 70(3):614–636, 1996.

R. Benzmüller & W. J. Barry. Mikrosegmentsynthese - Ökonomische Prinzipien bei der Konkatenation Subphonemischer Spracheinheiten. In D. Mehnert, Hrsg., *Elektronische Sprachsignalverarbeitung*, 13, Seiten 85–93. Humbold-Universität Berlin, 1996. Studientexte zur Sprachkommunikation, Tagungsband der 7. Konferenz.

G. Bergmann, T. Goldbeck & K. R. Scherer. Emotionale Eindruckswirkung von Prosodischen Sprachmerkmalen. *Zeitschrift für experimentelle und angewandte Psychologie*, 35:167–200, 1988.

C. A. Bickley, K. N. Stevens & D. R. Williams. A Framework for Synthesis of Segments Based on Pseudoarticulatory Parameters. In J. P. H. Van Santen, R. W. Sproat, O. J. P. & J. Hirschberg, Hrsg., *Progress in Speech Synthesis*, Seiten 211–220. Springer Verlag, Berlin Heidelberg New York, 1997.

J. Bortz. *Statistik für Sozialwissenschaftler*. Springer Verlag, Berlin Heidelberg New York, 4. Auflage, 1993.

A. Bühl & P. Zöfel. *SPSS Version 8, Einführung in die Moderne Datenanalyse unter Windows*. Addison-Wesley, 1999.

F. Burkhardt & W. F. Sendlmeier. Simulation des Emotionalen Ausdrucks ‚Freude' mit Sprachsyntheseverfahren. In *Forum Acusticum Berlin '99*, 1999.

F. Burkhardt & W. F. Sendlmeier. Simulation Emotionaler Sprechweise mit Konkatenierender Sprachsynthese. Vortrag auf der 30. Jahrestagung der GAL in Frankfurt, 1999.

F. Burkhardt. Simulation des Emotionalen Sprecherzustands ‚Freude' mit einem Sprachsyntheseverfahren. Magisterarbeit, TU-Berlin, Institut für Kommunikationswissenschaft, 1999.

J. E. Cahn. The Generation of Affect in Synthesized Speech. *Journal of the American Voice I/O Society*, 8:1–19, 1990.

W. N. Campbell. CHATR: A High-Definition Speech Re-Sequencing System. In *Proc. Acoustic Society of America and Acoustic Society of Japan, Third Joint Meeting, Honolulu, Hawaii*, 1996.

R. Carlson, G. Granström & D. H. Klatt. Vowel Perception: The Relative Perceptual Salience of Selected Acoustic Manipulations. *Speech Transmiss. Lab. - Q.Prog.Stat.Rep (STL-QPSR), Royal Institute of Technology, Stockholm*, 2-3:73–83, 1979.

R. Carlson, G. Granström & L. Nord. Experiments With Emotive Speech, Acted Utterances And Synthesized Replicas. *Speech Communication*, 2:347–355, 1992.

M. J. Charpentier & M. G. Stella. Diphone Synthesis using an Overlap-Add Technique for Speech Waveform Concatenation. In *Proc. 1984 IEEE Internat. Conf. on Acoust., Speech and Signal Processing (ICASSP), San Diego*, Seiten 2015–2018, 1984.

C. d'Alessandro & P. Mertens. Automatic Pitch Contour Stylization Using a Model of Tonal Perception. *Computer Speech and Language*, 9:257–288, 1995.

J. R. Davitz. Personality, Perceptual and Cognitive Correlates of Emotional Sensitivity. In *The Communication of Emotional Meaning*. Mc-Graw-Hill, New York, 1964.

F. Dellaert, T. Polzin & A. Waibel. Recognizing Emotion in Speech. In *Proc. 1996 Internat. Conf. on Spoken Language Processing (ICSLP), Philadelphia*, 1996.

J. M. Diehl. *Varianzanalyse*. Band 3 der Reihe "Methoden in der Psychologie". Fachbuchhandlung für Psychologie, 1977.

R. E. Donovan. *Trainable Speech Synthesis*. Diss., Cambridge University, 1996.

T. Dutoit & H. Leich. MBR-PSOLA : Text-To-Speech Synthesis based on an MBE Re-Synthesis of the Segments Database. *Speech Communication*, 13, 1993 a.

T. Dutoit, V. Pagel, N. Pierret, F. Bataille & O. Van der Vreken. The MBROLA Project: Towards a Set of High-Quality Speech Synthesizers Free of Use for Non-Commercial Purposes. In *Proc. 1996 Internat. Conf. on Spoken Language Processing (ICSLP), Philadelphia*, Band 3, Seiten 1393–1396, 1996.

T. Dutoit. *An Introduction to Text-To-Seech Synthesis*. Kluwe Academic Publishers, 1997.

G. Fant, J. Liljencrants & Q. Lin. A Four-Parameter Model of Glottal Flow. *Speech Transmiss. Lab. - Q.Prog.Stat.Rep (STL-QPSR), Royal Institute of Technology, Stockholm*, 4:1–13, 1985.

G. Fant. Modern Instruments and Methods for Acoustic Studies of Speech. In *Proc. of the 8th International Congress of Linguists*, Seiten 282–388, 1957.

G. Fant. *Acoustic Theory of Speech Production*. Mouton, 's-Gravenhage, 1960.

G. Fant. The Voice Source Theory and Acoustic Modeling. In I. R. Titze & R. C. Scherer, Hrsg., *Vocal Fold Physiology: Biomechanics, Acoustics and Phonatory Control*, Seiten 453–464. The Denver Center for the Performing Arts, Denver, CO, 1985.

K. Fellbaum. *Sprachverarbeitung und Sprachübertragung*. Springer Verlag, Berlin Heidelberg New York, 1984.

J. Flanagan & K. Ishizaka. Computer Model to Characterize the Air Volume Displaced by the Vibrating Vocal Cords. *Journal of the Acoustical Society of America (JASA)*, Seiten 1559–1565, 1978.

I. Fonagy & K. Magdics. Emotional Patterns in Intonation and Music. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung*, 16:293–326, 1963.

I. Fonagy. Emotions, Voice and Music. In *Research Aspects on Singing*, Band 33, Seiten 51–79. Royal Swedish Academy of Music, 1981.

C. S. Fordyce & M. Ostendorf. Prosody Prediction for Speech Synthesis using Transformational Rule-Based Learning. In *Proc. 1998 Internat. Conf. on Spoken Language Processing (ICSLP), Sidney*, 1998.

R. W. Frick. Communicating Emotion: The Role of Prosodic Features. *Psychological Bulletin*, 97(3):412–429, 1985.

O. Fujimura & J. Lindqvist. Sweep-Tone Measurements of Vocal-Tract Characteristics. *Journal of the Acoustical Society of America (JASA)*, 49(2):541–558, 1971.

H. Fujisaki. Modeling the Process of Fundamental Frequency Contour Generation. In Y. Tohkura, E. Vatikiotis-Bateson & Y. Sagisaka, Hrsg., *Speech, Perception, Production and Linguistic Structure*, Seiten 314–326. Ohmsha (u. a.), 1992.

C. Gobl & A. Ni Chasaide. Acoustic Characteristics of Voice Quality. *Speech Communication*, 11:481–490, 1992.

B. Gold & L. R. Rabiner. Analysis of Digital and Analog Formant Synthesizers. In *IEEE Trans. Audio Electroacoust.*, Band AU-16, Seiten 81–94, 1968.

B. Granström. The Use Of Speech Synthesis In Exploring Different Speaking Styles. In *Proc. 1992 Internat. Conf. on Spoken Language Processing (ICSLP), Banff, Kanada*, Band 1, Seiten 671–674, 1992.

M. Grice, M. Reyelt, R. Benzmüller, J. R. Mayer & A. Batliner. Consistency in Transcription and Labeling of German Intonation with GToBI. In *Proc. 1996 Internat. Conf. on Spoken Language Processing (ICSLP), Philadelphia*, 1996.

J. W. Hawks & J. D. Miller. A Formant Bandwidth Estimation Procedure for Vowel Synthesis. *Journal of the Acoustical Society of America (JASA)*, 97:1343–1344, 1995.

M. L. Hecker, K. N. Stevens, G. von Bismark & C. E. Williams. Manifestations of Task-Induced Stress in the Acoustic Speech Signal. *Journal of the Acoustical Society of America (JASA)*, 44:993–1001, 1968.

D. J. Hermes. Synthesis of Breathy Vowels: Some Research Methods. *Speech Communication*, Seiten 497–502, 1991.

B. Heuft, T. Portele & M. Rath. Emotions in Time Domain Synthesis. In *Proc. 1998 Internat. Conf. on Spoken Language Processing (ICSLP), Sidney*, 1998.

B. Hildebrandt. Die Arithmetische Bestimmung der Durativen Funktion: Eine Neue Methode der Lautdauerbewertung. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung*, 14:328–336, 1963.

J. N. Holmes. The Influence of the Glottal Waveform on the Naturalness of Speech from a Parallel Formant Synthesizer. In *IEEE Trans. Audio Electroacoust. Au 21*, Seiten 298–305, 1973.

A. Iida, N. Campbell & M. Yasumura. Designing and Testing a Corpus of Emotional Speech. In *Oriental COCOSDA*, 2000.

J. Iles & N. Simmons. Klatt: A Klatt-Style Speech Synthesizer Implemented in C, 1995. http://www. cs. cmu. edu/afs/cs. cmu. edu/project/ai-repository/ai/areas/speech/systems/-klatt/0. html.

T. Johnstone & K. R. Scherer. The Effect of Emotions on Voice Quality. In *Proc. 1999 Internat. Congress of Phonetic Sciences (ICPhS), San Francisco*, 1999.

I. Karlsson. Modelling Voice Variations in Female Speech Synthesis. *Speech Communication*, 11:491–495, 1992.

KAY. Instruction Manual von ASL, 1992.

M. Kienast & W. F. Sendlmeier. Acoustical Analysis of Spectral and Temporal Changes in Emotional Speech. In *Proc. 2000 Internat. Speech Communication Association (ISCA) Workshop on Speech and Emotion, Belfast*, 2000.

M. Kienast, A. Paeschke & W. F. Sendlmeier. Articulatory Reduction in Emotional Speech. In *Proc. 1999 Eurospeech, Budapest*, 1999.

M. Kienast. Segmentelle Reduktion bei Emotionaler Sprechweise. Magisterarbeit, Institut für Kommunikationswissenschaft der TU-Berlin, 1998.

M. Kienast. *Segmentelle Reduktion und Elaboration bei Emotionaler Sprechweise*. Diss., TU-Berlin, Institut für Kommunikationswissenschaft, noch nicht erschienen.

G. Klasmeyer & W. F. Sendlmeier. Objective Voice Parameters to Characterize the Emotional Content in Speech. In *Proc. 1995 Internat. Congress of Phonetic Sciences (ICPhS), Stockholm*, Seiten 182–185, 1995.

G. Klasmeyer & W. F. Sendlmeier. The Classification of Different Phonation Types in Emotional and Neutral Speech. *Forensic Linguistics*, 4:104–124, 1997.

G. Klasmeyer & W. F. Sendlmeier. Voice and Emotional States. In R. D. Kent & M. J. Ball, Hrsg., *Voice Quality Measurement*, Seiten 339–357. Singular Publ. Group, 2000.

G. Klasmeyer. The Perceptual Importance of Selected Voice Quality Parameters. In *Proc. 1997 IEEE Internat. Conf. on Acoust., Speech and Signal Processing (ICASSP), München*, 1997.

G. Klasmeyer. *Akustische Korrelate des Stimmlichen Emotionalen Ausdrucks in der Lautsprache*. Hector, FFM, 1999.

D. H. Klatt & L. C. Klatt. Analysis, Synthesis and Perception of Voice Qualtity Variations among Female and Male Talkers. *Journal of the Acoustical Society of America (JASA)*, 87(2):820–856, 1990.

D. H. Klatt. Software for a Cascade/Parallel Formant Synthesizer. *Journal of the Acoustical Society of America (JASA)*, 67(3):971–959, 1980.

D. H. Klatt. Review of Text-To-Speech Conversion for English. *Journal of the Acoustical Society of America (JASA)*, 82(3):737–793, 1987.

D. H. Klatt. *KLATTALK, The Conversion of English Text to Speech*, Kapitel 3: Description of the Cascade/Parallel Formant Synthesizer. unpublished, 1990.

P. R. Kleinginna & A. M. Kleinginna. A Categorized List of Emotion Definitionns, with Suggestions for a Consensual Definition. *Motivation & Emotion*, Seiten 345–379, 1981.

B. J. Kröger, G. Schröder & C. Opgen-Rhein. A Gesture-Based Dynamic Model Describing Articulatory Movement Data. *Journal of the Acoustical Society of America (JASA)*, 98:1878–1889, 1995.

D. R. Ladd. Phonological Features of Intonational Peaks. *Language*, 95(3):721–759, 1983.

J. Laver. *The Phonetic Description of Voice Quality*. Cambridge University Press, 1980.

J. Laver. *The Gift Of Speech*. Edinburgh University Press, 1991.

J. Laver. *Principles of Phonetics*. Cambridge University Press, 1994.

M. J. Liberman & K. W. Church. Text Analysis and Word Pronounciation in TTS-Synthesis. In S. Furuy & M. M. Sondhi, Hrsg., *Advances in Speech Signal Processing*, Seiten 791–831. Dekker, New York, 1992.

F. Malfrère & T. Dutoit. Speech Synthesis for TTS-Alignment and Prosodic Feature Extraction. In *Proc. 1998 Internat. Conf. on Spoken Language Processing (ICSLP), Sidney*, 1998.

F. Malfrère, T. Dutoit & P. Mertens. Automatic Prosody Generation Using Suprasegmental Unit Selection. In *Proc. of the 3. ESCA/IEEE Workshop in Speech Synthesis*, 1999.

R. H. Mannell. Formant Diphone Parameter Extraction Utilising a Labelled Single-Speaker Database. In *Proc. 1998 Internat. Conf. on Spoken Language Processing (ICSLP), Sidney*, 1998.

J. D. Markel & A. H. Gray. *Linear Prediction of Speech*. Springer Verlag, Berlin Heidelberg New York, 1976.

J. M. Montero, J. Guitérrez-Arriola, S. Palazuelos, E. Enríquez, S. Aguilera & J. M. Pardo. Emotional Speech Synthesis: From Speech Database to TTS. In *Proc. 1998 Internat. Conf. on Spoken Language Processing (ICSLP), Sidney*, 1998.

J. M. Montero, J. Guitèrrez-Arriola, J. Colás, J. Maciás, E. Enríquez & J. M. Pardo. Development of an Emotional Speech Synthesizer in Spanish. In *Proc. 1999 Eurospeech, Budapest*, 1999.

E. S. Morton. On the Occurance and Significance of Motivation-Structural Rules in some Birds and Mammal Sounds. *American Naturalist*, 111:855–869, 1977.

K. Morton. Adding Emotion To Synthetic Speech Dialogue Systems. In *Proc. 1992 Internat. Conf. on Spoken Language Processing (ICSLP), Banff, Kanada*, Seiten 675–678, 1992.

E. Moulines & F. Charpentier. Pitch-Synchronous Waveform Processing Techniques for Text-To-Speech Synthesis using Diphones. *Speech Communication*, 9:453–467, 1990.

S. J. L. Mozziconacci & D. J. Hermes. A Study of Intonation Patterns in Speech Expressing Emotion or Attitude: Production and Perception. *IPO Annual Progress Report*, 32:154–160, 1997.

I. R. Murray & J. L. Arnott. Toward the Simulation of Emotion in Synthetic Speech: A Review of the Literature on Human Vocal Emotion. *Journal of the Acoustical Society of America (JASA)*, 2:1097–1107, 1993.

I. R. Murray & J. L. Arnott. Implementation and Testing of a System for Producing Emotion-by-rule in Synthetic Speech. *Speech Communication*, 16:369–390, 1995.

I. R. Murray & J. L. Arnott. Synthesizing Emotions in Speech: Is It Time to get Excited? In *Proc. 1996 Internat. Conf. on Spoken Language Processing (ICSLP), Philadelphia*, 1996.

I. R. Murray, J. L. Arnott & E. A. Rohwer. Emotional Stress in Synthetic Speech: Progress and Future Directions. *Speech Communication*, 20:85–91, 1996.

A. Ni Chasaide & C. Gobl. Voice Source Variation. In W. J. Hardcastle & J. Laver, Hrsg., *The Handbook of Phonetic Sciences*. Blackwell, Oxford, 1997.

J. J. Ohala. The Voice of Dominance. *Journal of the Acoustical Society of America (JASA)*, 72:S66, 1982.

D. O'Shaughnessy, L. Barbeau, D. Bernardi & D. Archambault. Diphone Speech Synthesis. *Speech Communication*, 7(1):56–65, 1987.

D. O'Shaughnessy. *Speech Communication: Human and Machine*. Addison-Wesley, 1987.

A. Öster & A. Riesberg. The Identification of the Mood of a Speaker by Hearing Impaired Listeners. *Speech Transmiss. Lab. - Q.Prog.Stat.Rep (STL-QPSR), Royal Institute of Technology, Stockholm*, Seiten 79–90, 1986.

A. Paeschke & W. F. Sendlmeier. Prosodic Characteristics of Emotional Speech: Measurements of Fundamental Frequency Movements. In *Proc. 2000 Internat. Speech Communication Association (ISCA) Workshop on Speech and Emotion, Belfast*, 2000.

A. Paeschke, M. Kienast & W. F. Sendlmeier. F0-Konturs in Emotional Speech. In *Proc. 1999 Internat. Congress of Phonetic Sciences (ICPhS), San Francisco*, 1999.

A. Paeschke. *Prosodische Indikatoren Emotionaler Sprechweise*. Diss., TU-Berlin, Institut für Kommunikationswissenschaft, noch nicht erschienen.

B. Peters. Prototypische Intonationsmuster in Deutscher Lese- und Spontansprache. In K. J. Kohler, Hrsg., *Arbeitsberichte des Institut für Phonetik und digitale Sprachverarbeitung der Universität Kiel*, 34. Universität Kiel, 1999.

T. Portele, W. F. Sendlmeier & W. Hess. HADIFIX: A System for German Speech Synthesis Based on Demisyllables, Diphones, and Suffixes. In *Proc. 1990 European Speech Communication Association (ESCA)-Workshop on Speech Synthesis, Autrans, Frankreich*, Seiten 161–164, 1990.

T. Portele. Das Sprachsynthesesystem Hadifix. *Sprache und Datenverabeitung*, 21:5–22, 1996.

T. Portele. *Ein Phonetisch-Akustisch Motiviertes Inventar zur Sprachsynthese Deutscher Äußerungen*. Max Niemeyer Verlag, Tübingen, 1996.

E. Rank & H. Pirker. Generating Emotional Speech with a Concatenative Synthesizer. In *Proc. 1998 Internat. Conf. on Spoken Language Processing (ICSLP), Sidney*, 1998.

E. Rank. Erzeugung Emotional Gefärbter Sprache mit dem VieCtoS-Synthesizer. *Austrian Research Institute for Artificial Intelligence, Vienna, TR-99-01*, 1999.

W. J. Revers. *Die Psychologie der Langeweile*. Meisenheim, Hain, 1949.

J. C. Rutledge, K. E. Cummings, D. A. Lambert & M. A. Clements. Synthesizing Styled Speech Using the Klatt Synthesizer. In *Proc. 1995 IEEE Internat. Conf. on Acoust., Speech and Signal Processing (ICASSP), Detroit*, 1995.

M. T. M. Scheffers & A. P. Simpson. LACS: Label Assisted Copy Synthesis. In *Proc. 1995 Internat. Congress of Phonetic Sciences (ICPhS), Stockholm*, Band 2, Seiten 346–349, 1995.

K. R. Scherer & J. S. Oshinsky. Cue Utilization in Emotion Attribution from Auditory Stimuli. *Motivation and Emotion*, 1:331–346, 1977.

K. R. Scherer & U. Scherer. Speech Behaviour and Personality. In J. K. Darby, Hrsg., *Speech Evaluation in Psychiatry*, Seiten 117–135. Grune & Stratton, 1981.

K. R. Scherer, D. R. Ladd & K. E. A. Silverman. Vocal Cues to Speaker Affect: Testing Two Models. *Journal of the Acoustical Society of America (JASA)*, 76:1346–1356, 1984.

K. R. Scherer, R. Banse & H. G. Wallbott. Emotion Inferences from Vocal Expression Correlate across Languages and Cultures. *Journal of Cross-Cultural Psychology*, 2000.

K. R. Scherer. Speech and Emotional States. In J. K. Darby, Hrsg., *Speech Evaluation in Psychiatry*, Seiten 189–220. Grune & Stratton, 1981.

K. R. Scherer. Akustische Parameter der Vokalen Kommunikation. In K. R. Scherer, Hrsg., *Vokale Kommunikation*, Seiten 122–137. Beltzverlag, 1982.

K. R. Scherer. On the Nature and Function of Emotion: a Component Process Approach. In K. R. Scherer & P. Ekman, Hrsg., *Approaches to Emotion*, Seiten 293–318. Hillsdale, New York, 1984.

K. R. Scherer. Vocal Affect Expression: A Review and a Model for Future Research. *Psychological Bulletin*, 99(2):143–165, 1986.

K. R. Scherer. Vocal Correlates of Emotional Arousal and Affective Disturbance. In H. Wagner & A. Manstead, Hrsg., *Handbook of Social Psychophysiology*, Seiten 165–197. John Wiley & Sons Ltd., 1989.

K. R. Scherer. How Emotion is Expressed in Speech and Singing. In *Proc. 1995 Internat. Congress of Phonetic Sciences (ICPhS), Stockholm*, Band 3, Seiten 90–96, 1995.

H. Schlosberg. Three Dimensions of Emotions. *Psychological Review*, 61(2):81–88, 1954.

M. Schröder, V. Aubergé & M. Cathiard. Can We Hear Smile? In *Proc. 1998 Internat. Conf. on Spoken Language Processing (ICSLP), Sidney*, 1998.

M. Schröder. Can Emotions be Synthesized without Controlling Voice Quality? In *Phonus 4, Forschungsbericht Institut für Phonetik, Universität des Saarlandes*, 1999.

V. K. Sedlácek & A. Sychra. Die Melodie als Faktor des Emotionellen Ausdrucks. *Folia Phoniatrica*, 15:89–98, 1963.

T. Sejnowski & C. R. Rosenberg. Parallel Networks that Learn to Pronounce English Text. *Complex Systems*, 1:145–168, 1987.

W. F. Sendlmeier & A. Heile. Nasalität und Behauchung als Indikatoren für den stimmlichen Ausdruck körperlichen Wohlbehagens. In H. W. Wodarz, Hrsg., *Festschrift Georg Heike*, Seiten 1–14. Forum Phoneticum 66, FFM, 1998.

W. F. Sendlmeier. *Von Sprechkunst und Normphonetik*, Kapitel Phonetische Reduktion und Elaboration bei Emotionaler Sprechweise, Seiten 169–178. Werner Dausin, Hanau und Halle, 1997. Festschrift zum 65. Geburtstag von Eva-Maria Krech.

SENSYMETRICS, 1999. http://www. sens. com/Sensyn. htm.

A. Simpson. Wissensbasierte Gewinnung von Steuerparametern für die Formantsynthese. In K. J. Kohler, Hrsg., *Phonetische Datenbanken des Deutschen in der Empirischen Sprachforschung und der Phonologischen Theoriebildung*, Arbeitsberichte 33. Institut für Phonetik und digitale Sprachverarbeitung der Universität Kiel, 1998.

M. Slaney & G. McRoberts. Baby Ears: a Recogintion System for Affective Vocalizations. In *Proc. 1998 IEEE Internat. Conf. on Acoust., Speech and Signal Processing (ICASSP), Seattle*, 1998.

H. Strik & L. Boves. On the Relation between Voice Source Characteristics and Prosody. In *Proc. 1991 Eurospeech, Genova*, Band 3, Seiten 1145–1148, 1991.

B. Stroustrup. *Die C++ Programmiersprache*. Addison-Wesley, 1998.

V. C. Tartter. Happy Talk: Perceptual and Acoustic Effects of Smiling on Speech. *Percept. Psychophys.*, 27:24–27, 1980.

J. t'Hart, R. Collier & A. Cohen. *A Perceptual Study of Intonation*. Cambridge University Press, 1990.

B. Tischer. *Die Vokale Kommunikation Von Gefühlen*. Psychologie-Verlag-Union, 1993.

F. Trojan. *Biophonetik*. Wissenschaftsverlag, 1975.

G. Ungeheuer. *Elemente einer Akustischen Theorie der Vokalartikulation*. Springer Verlag, Berlin Heidelberg New York, 1962.

H. Valbret, E. Moulines & J. E. Tubach. Voice Transformation using PSOLA Technique. *Speech Communication*, 11:175–187, 1992.

R. J. J. H. Van Son & L. C. W. Pols. An Acoustic Description of Consonant Reduction. *Speech Communication*, 28:125–140, 1999.

J. Vroomen, R. Collier & S. Mozziconacci. Duration and Intonation in Emotional Speech. In *Proc. 1993 Eurospeech, Berlin*, Band 1, Seiten 577–580, 1993.

L. Wall, T. Christiansen & R. L. Schwartz. *Programming Perl*. O'Reilly, 2. Auflage, 1998.

J. Werner. *Lineare Statistik*. Psychologie Verlags Union, 1997.

C. E. Williams & K. N. Stevens. Emotions and Speech: Some Acoustical Correlates. *Journal of the Acoustical Society of America (JASA)*, 52:1238–1250, 1972.

C. E. Williams & K. N. Stevens. Vocal Correlates of Emotional Speech. In J. K. Darby, Hrsg., *Speech Evaluation in Psychiatry*. Grune & Stratton, 1981.

E. Zwicker & H. Fastl. *Psychoacoustics - Facts and Models*. Springer Verlag, Berlin Heidelberg New York, 1990.

Anhang A

Die in Kap. 8 verwendeten MOD-Files

A.1 Zorn

1. Version

```
-changeSpeechRate 70 1 1
-changeMeanF0 150
-changeF0Range 300
-tense 50
```

2. Version

```
-changeSpeechRate 70 1 1
-stressedIntensity 9 1
-changeLastSylContour 2 20
-changePitchContourStressed 2 30 1
-changeMeanF0 150
-tense 50
```

3. Version

```
-changeSpeechRate 70 1 1
-changeLastSylContour 2 20
-changePitchContourStressed 2 30 1
-changeMeanF0 150
-undershootVowels 0 30
-overshootVowels 1 30
-overshootVowels 2 50
-tense 50
```

4. Version

```
-changeSpeechRate 70 1 1
-changeLastSylContour 2 20
-changePitchContourStressed 2 30 1
-changeMeanF0 150
-tense 50
```

5. Version

```
-changeSpeechRate 70 1 1
-changeMeanF0 150
-changeStressedMeanF0 150 1
-tense 50
```

A.2 Ärger

1. Version

```
-changeSpeechRate 70 1 1
-changeLastSylContour 2 80
-changeF0Range 200
-changeMeanF0 80
-tense 50
```

2. Version

```
-changeSpeechRate 70 1 1
-changeMeanF0 80
-tense 50
```

3. Version

```
-changeSpeechRate 70 1 1
-changeLastSylContour 2 80
-changePitchContourStressed 2 30 1
-changeMeanF0 80
-stressedIntensity 9 1
-tense 50
```

4. Version

```
-changeSpeechRate 70 1 1
-changeLastSylContour 2 80
-changePitchContourStressed 2 30 1
-changeMeanF0 80
-undershootVowels 0 30
-overshootVowels 1 30
-overshootVowels 2 50
-tense 50
```

5. Version

```
-changeSpeechRate 70 1 1
-changeLastSylContour 2 80
-changePitchContourStressed 2 30 1
-changeMeanF0 80
-tense 50
```

A.3 Freude

1. Version

```
-changeSpeechRate 70 1 1
-changeDurFricVL 150 1
-changeF0Range 200
-changeMeanF0 150
```

2. Version

```
-changeSpeechRate 70 1 1
-changeDurFricVL 150 1
-changePitchContourStressed 1 50 1
-changeF0Range 200
-changeMeanF0 150
-undershootVowels 0 30
-overshootVowels 1 30
-overshootVowels 2 50
-spreadLips 10
```

3. Version

```
-changeSpeechRate 70 1 1
-changeDurFricVL 150 1
```

```
-changePitchContourStressed 1 50 1
-changeF0Range 200
-changeMeanF0 150
-spreadLips 10
```

4. Version

```
-changeSpeechRate 70 1 1
-changeDurFricVL 150 1
-changeF0Range 200
-changeMeanF0 150
-spreadLips 10
```

5. Version

```
-changeSpeechRate 70 1 1
-changeDurFricVL 150 1
-addWaves 200 80 1
-changeF0Range 200
-changeMeanF0 150
-spreadLips 10
```

A.4 Zufriedenheit

1. Version

```
-changeSpeechRate 120 1 1
-changeDurFricVL 150 1
-changeF0Range 200
```

2. Version

```
-changeSpeechRate 120 1 1
-changeDurFricVL 150 1
-changeF0Range 200
-spreadLips 10
```

3. Version

```
-changeSpeechRate 120 1 1
-changeDurFricVL 150 1
-changeF0Range 200
-spreadLips 10
-breathy 50
```

4. Version

```
-changeSpeechRate 120 1 1
-changeDurFricVL 150 1
-changePitchContourStressed 1 50 1
-spreadLips 10
```

5. Version

```
-changeSpeechRate 120 1 1
-changeDurFricVL 150 1
-addWaves 200 80 1
-spreadLips 10
```

A.5 Weinerliche Trauer

1. Version

```
-changeSpeechRate 120 1 1
-changeMeanF0 200
-changeF0Range 80
-changeF0Variability 80
```

2. Version

```
-changePitchContourStressed 2 20 1
-changeSpeechRate 120 1 1
-changeMeanF0 200
-changeF0Range 80
-changeF0Variability 80
```

3. Version

```
-changePitchContourStressed 2 20 1
-changeSpeechRate 120 1 1
-changeMeanF0 200
-changeF0Range 80
-changeF0Variability 80
-jitter 0 200 2
-jitter 1 200 2
-jitter 2 200 2
```

4. Version

```
-changePitchContourStressed 2 20 1
-changeSpeechRate 120 1 1
-changeMeanF0 200
-changeF0Range 80
-changeF0Variability 80
-breathy 50
```

5. Version

```
-changePitchContourStressed 2 20 1
-changeSpeechRate 120 1 1
-changeMeanF0 200
-changeF0Range 80
-changeF0Variability 80
-falsett 50
```

A.6 Stille Trauer

1. Version

```
-changeSpeechRate 140 1 1
-changeMeanF0 80
-changeF0Range 80
-changeF0Variability 80
```

2. Version

```
-changeSpeechRate 140 1 1
-changePitchContourStressed 2 30 1
-changeMeanF0 80
-changeF0Range 80
-changeF0Variability 80
```

3. Version

```
-changeSpeechRate 140 1 1
-changePitchContourStressed 2 30 1
-changeMeanF0 80
-changeF0Range 80
-changeF0Variability 80
-jitter 0 300 2
```

```
-jitter 1 300 2
-jitter 2 300 2
```

4. Version

```
-changeSpeechRate 140 1 1
-changePitchContourStressed 2 30 1
-changeMeanF0 80
-changeF0Range 80
-changeF0Variability 80
-breathy 50
```

5. Version

```
-changeSpeechRate 140 1 1
-changePitchContourStressed 2 30 1
-changeMeanF0 80
-changeF0Range 80
-changeF0Variability 80
-jitter 0 300 2
-jitter 1 300 2
-jitter 2 300 2
-breathy 50
```

A.7 Angst

1. Version

```
-changeSpeechRate 70 1 1
-changeMeanF0 250
-changeF0Range 120
```

2. Version

```
-changeSpeechRate 70 1 1
-changeMeanF0 300
-changeF0Range 120
-jitter 0 300 2
-jitter 1 300 2
-jitter 2 300 2
```

3. Version

```
-changeSpeechRate 70 1 1
-changePitchContourStressed 0 0 1
-changeLastSylContour 1 30
-changeMeanF0 300
-changeF0Range 120
```

4. Version

```
-changeSpeechRate 70 1 1
-changeMeanF0 150
-changeF0Range 120
-falsett 100
```

A.8 Langeweile

1. Version

```
-changeSpeechRate 120 1 1
-changeDurStressed 140 1 1
-changeMeanF0 80
-changeF0Range 50
-changeF0Variability 80
```

2. Version

```
-changeSpeechRate 120 1 1
-changeDurStressed 140 1 1
-changeMeanF0 80
-changeF0Range 50
-changeF0Variability 80
-breathy 50
```

3. Version

```
-changeSpeechRate 120 1 1
-changeDurStressed 140 1 1
-changeMeanF0 80
-changeF0Range 50
-changeF0Variability 80
-laryngealized 50
```

4. Version

```
-changeSpeechRate 120 1 1
-changeDurStressed 140 1 1
-changeMeanF0 80
-changeF0Range 50
-changeF0Variability 80
-undershootVowels 0 50
-undershootVowels 1 20
-undershootVowels 2 20
```

5. Version

```
-changeSpeechRate 120 1 1
-changeDurStressed 140 1 1
-changeMeanF0 80
-changeF0Range 50
-changeF0Variability 80
-undershootVowels 0 50
-undershootVowels 1 20
-undershootVowels 2 20
-breathy 50
-laryngealized 50
```