Simulation des emotionalen Sprecherzustands ‚Freude' mit einem Sprachsyntheseverfahren

Magisterarbeit

bei Professor Walter F. Sendlmeier
Technische Universität Berlin
FB 1 - Kommunikations- und Geschichtswissenschaften

Felix Burkhardt
Matrikel-Nr. 165824

17. Mai 1999

Inhaltsverzeichnis

Einleitung

Die künstliche Spracherzeugung hat seit Wolfgang von Kempelens sprechender Maschine (1791) durch die Verbreitung von Digital-Computern einen großen Sprung gemacht. Die Verständlichkeit synthetischer Sprache ist auch bei Text-to-Speech-Systemen[1] (im Folgenden kurz TTS-Systeme genannt) stark verbessert worden, und auch die Natürlichkeit hat durch die Fortschritte im Bereich der konkatenierenden Synthese sowie der Entwicklung komplexerer Modelle für das Glottissignal bei der regelbasierten Synthese eine bis dahin unbekannte Güte erreicht.

Klassische Anwendungsgebiete von TTS-Systemen sind die Vorlesemaschine für Blinde, die Verwendung als Frontend für automatische Übersetzung, der Einsatz in der Forschung bei „Analyse durch Synthese"-Experimenten sowie ganz allgemein eine Verbesserung 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, die selbstverständlich interagieren.

- 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 verschiedene Modelle, die Aspekte der Sprechsituation in die Synthese einzubeziehen. Menschen sprechen anders, wenn sie vorlesen, vor einem großen Auditorium sprechen müssen oder beispielsweise mit Menschen anderer Hierarchiestufen umgehen.

- Vermutlich gibt es keine emotional neutrale Sprache, der emotionale Gehalt von Sprachsignalen wird ständig durch generelle Einstellungen, momentane Stimmungen und weitere Einflüsse verändert.

Eine Verbesserung dieser Kriterien 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 z.B. lauter Umgebung oder bei langer Dauer zu gewährleisten.

Diese Arbeit soll sich nun mit dem letzten Punkt beschäftigen, nämlich der Simulation von emotionalem Ausdruck bei der Sprachsynthese.

[1] Im Gegensatz zu Sprachwiedergabesystemen

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 (1989)). 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 auf der Hörerseite an. Die zunächst für neutrale Sprechweise berechneten Merkmalsbeschreibungen für die Synthese werden den akustischen Korrelaten der zu simulierenden Emotion gemäß 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 (Morton (1992)).

Um den Rahmen der Arbeit einzugrenzen, wird beispielhaft für alle Emotionen die Simulation der Emotion Freude betrachtet. Die meisten Theorien über Emotionen gehen von einer Anzahl von Basisemotionen aus, zu denen, wenn auch Art und Anzahl umstritten sind, Freude gezählt werden kann. Freude zählt zu den schwerer simulierbaren Emotionen (Murray & Arnott (1995)). Dies hat zwei Gründe;

- Erstens wird Freude eher schlecht erkannt (vgl. etwa Heuft et al. (1998)). Scherer (1981) weist darauf hin, dass negative Emotionen anscheinend besser erkannt werden als positive. Bei allen Untersuchungen rangierte Freude bei der Erkennung im unteren Drittel.

- Zweitens gibt es weniger Untersuchungen darüber. Zu negativen Emotionen gibt es mehr natürliche Aufnahmen (aus Katastrophenberichten oder Pilotenaufnahmen), während die Gewinnung von Sprachäußerungen, die in echter freudiger Stimmung getätigt wurden, schwierig ist. (vgl. Scherer (1981, S. 201)).

Diese Tatsache war die Hauptmotivation dafür, Freude als Untersuchungsgegenstand dieser Arbeit zu nehmen.

Die Arbeit gliedert sich in einen theoretischen und einen experimentellen Teil. Im ersten Kapitel werden zunächst verschiedene Begriffe eingeführt. Es ist entsprechend der Vielschichtigkeit des Themas in mehrere Unterkapitel unterteilt. Zunächst wird der Versuch unternommen, eine dem Rahmen dieser Arbeit genügende Klärung der Begriffe „Emotion" und „Freude" zu finden. Daraufhin werden Zusammenhänge zwischen emotionalen und physiologischen Zuständen betrachtet. Wenn hier auch physiologische Vorgänge nicht direkt modelliert werden sollen, sind sie doch als Hinweis auf akustische Korrelate und zu ihrer Begründung äußerst wichtig.

Ein weiteres Unterkapitel beschäftigt sich mit den verschiedenen Parametern und Messmethoden akustischer Sprachsignale. Weiterhin wird eine kurze Einführung in das Gebiet der Text-to-Speech-Synthese gegeben. Verschiedene Ansätze werden im Hinblick auf ihre Eignung zur Simulation emotionaler Sprache besprochen. Der Schwerpunkt liegt dabei gemäß dem Thema auf High-Quality-Verfahren. Abschließend werden mit LPC und PSOLA zwei grundlegende Verfahren zur Resynthese digitaler Sprachsignale erläutert.

Das zweiten Kapitel gibt einen Überblick über den bisherigen Forschungsstand zum Thema, indem einige ausgewählte Experimente genauer besprochen werden. Es schließt mit einem Überblick über die akustischen Korrelate freudiger Sprechweise ab.

Mit dem dritten Kapitel beginnt der experimentelle Teil. Es beschäftigt sich mit einem analytischen Vergleich zwischen je neun neutralen und freudigen Äußerungen, der Merkmale freudiger Sprechweise aufdecken und Ergebnisse aus der Literatur bestätigen soll.

Im vierten Kapitel werden einige der gefundenen Korrelate mittels eines Perzeptionstest überprüft. Gegenstand des fünften Kapitels ist die Implementation eines „Emotion-by-rule“-Systems in einem Computerprogramm, dass ein gegebenes TTS-System dazu befähigen soll, freudigen Sprecherzustand zu simulieren.

Das sechste Kapitel handelt von der Evaluierung dieses Systems durch einen weiteren Perzeptionstest. Die Arbeit schließt mit einer kurzen Zusammenfassung der Ergebnisse und einem Ausblick für zukünftige Forschung.

Teil I

Theoretischer Teil

Kapitel 1

Grundlagen

1.1 Klassifikation von Emotionen

Die Definition von Emotionen ist in der Literatur sehr umstritten und es wurden viele Klassifizierungs- oder Kategorisierungssysteme vorgeschlagen (siehe Tischer (1993)). Abzugrenzen ist der Begriff der Stimmung, der sich auf emotionale Zustände bezieht, die für längere Zeiträume gelten. Zur Klassifizierung von Emotionen bietet sich einmal die Einteilung in diskrete Begriffe, eventuell unterteilt in ein System von Basis- und Nebenemotionen, an und andererseits die Beschreibung als Punkt in einem mehrdimensionalem Raum.

1.1.1 Dimensionaler Ansatz

Das Dimensionsmodell setzt in der Regel drei Dimensionen an (Tischer (1993)), von denen zwei relativ unumstritten sind; Valenz (angenehm/unangenehm) und Potenz (stark/schwach). Über die Art der dritten Dimension herrscht weitgehende Uneinigkeit. Schlosberg (1954) etwa schlägt als weitere Dimension die Aktivität (gespannt/gelöst) vor.

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

1.1.2 Kategorialer Ansatz

Im Gegensatz zum dimensionalen Ansatz werden hier die Emotionen nach nominalen Klassen unterteilt. Der Begriff 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 (Tischer (1993)). Wenn auch Art und Anzahl der Basisemotionen bei den verschiedenen Vertretern dieses Ansatzes stark umstritten sind, sind doch die Emotionen Ärger, Furcht, Ekel, Freude und Trauer in fast allen Modellen enthalten.[1]

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

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 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.
Gegenstand dieser Untersuchung soll eine Art von Freude sein, die wohl am besten mit „Fröhligkeit“ beschrieben werden kann, da es sich nicht um stille Freude handelt, sondern um eine kräftige, extrovertierte Emotion.

Insofern wird eine Mischform der obengenannten beiden Ansätze verwendet; Gegenstand der Untersuchung ist die Emotion „Freude“ mit relativ hoher Erregung.

1.2 Physiologische Grundlagen

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.

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. Trojan (1975); Williams & Stevens (1981)). Da sich zudem bei symphatischer 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 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 großer F_0-Range, größere Unterschiede zwischen betonten und unbetonten Silben und ein schnelleres Tempo zu erwarten. Dies gilt wie gesagt nicht nur für Freude, sondern allgemein für Emotionen mit hohem Erregungsgrad.
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 Ansatzrohres zur Folge hat (vgl. 2.1.3). 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 sozia-

len Norm sind oder sonstige körperunabhängige Faktoren. Sie sind meist von der Intention des Sprechers motiviert, beim Empfänger einen möglichst günstigen Eindruck zu erwecken[2].

1.3 Akustische Parameter

Die Spracherzeugung erfolgt durch ein Zusammenwirken von Atmung (Respiration), Stimmgebung (Phonation) und Lautbildung (Artikulation). Scherer (1989) schlägt eine Einteilung akustischer Parameter in die Kategorien Stimmintensität, Stimmfrequenz, Stimmqualität und Resonanzverteilung vor. Diese werden durch Respiration, Phonation und Artikulation beeinflusst. In Scherer (1982) wird eine Einteilung nach ihren physikalischen Dimensionen Zeit-, Energie- und Frequenzbereich vorgeschlagen.

Die Einteilung in dieser Arbeit ist durch die Wirkung auf emotionale Sprechweise motiviert, es werden Prosodieparameter, segmentale Parameter (wie z.B. durchschnittliche Dauern einzelner Lautklassen) und Stimmqualitätsparameter unterschieden. Bevor auf die Parameter im Einzelnen eingegangen wird, soll der wichtige Begriff der Prosodie geklärt werden.

1.3.1 Prosodie

Prosodische Merkmale sind solche, die vom spezifischen Lautinventar einer Sprache unabhängig sind, wie Sprechrhythmus und Sprechmelodie. Die Prosodie fasst Dauer, Lautheit und Tonhöhenverlauf (Intonation) auf suprasegmentaler Ebene[3] zusammen.

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 (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.

1.3.2 Intensität

Die Intensität von Sprachsignalen wird subjektiv als Lautheit wahrgenommen. Sie wird durch Respiration und Phonation beeinflusst. Gemessen wird sie als Schalldruck, entweder in dB oder Volt. 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.

Ein aus der Intensität abgeleiteter, für den emotionalen Ausdruck wichtiger Parameter ist der Shimmer, der das Vorhandensein kurzzeitiger Schwankungen der Amplitude angibt. Eine Erregung des sympathischen Nervensystems korreliert in der Regel mit einem Anstieg der Intensität. 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] Beispielsweise wird bei aggressiver Sprechweise die Grundfrequenz oft abgesenkt, um größer zu wirken.
[3] suprasegmental: über einzelnen Laute hinweg, Phrasen- oder Satzebene

1.3.3 Frequenz

Die Grundfrequenz (als äquivalenter Begriff wird hier auch F_0 verwendet) wird subjektiv als Tonhöhe wahrgenommen. Die wahrgenommene Tonhöhe hängt allerdings nicht nur von der F_0 ab, sondern wird auch durch die spektrale Zusammensetzung des Signals beeinflusst. Die Grundfrequenz hängt von Respiration und Phonation ab. Bei steigendem subglottalem Druck sowie bei einer Verstärkung der Spannung der laryngalen Muskulatur steigt die Grundfrequenz. Auch hier korreliert ein Anstieg mit der Erregung des Sympathikus.

Aus der Grundfrequenz lassen sich verschiedene Parameter ableiten wie etwa die durchschnittliche Tonhöhe, der Range[4] oder die Variabilität[5]. Um dem logarithmischen Hörempfinden des Menschen Rechnung zu tragen, sollten Grundfrequenzabstandswerte in Halbtönen oder Prozent angegeben werden. Ein für viele Emotionen wichtiger Parameter ist der Jitter (äquivalent dazu: F_0-Perturbation). Er beschreibt Schwankungen der Grundfrequenz von einer Periode zur nächsten. Weiterhin lässt sich aus der Grundfrequenzmessung der Konturverlauf (äquivalent dafür: die Intonation) von Äußerungen beschreiben. Dieser wird in der Regel getrennt auf silbischer Ebene und über ganze Äußerungen bzw. Phrasen hinweg betrachtet.

1.3.4 Dauern, Tempo

Eine weitere für emotionale Sprechweise wichtige Parameterklasse stellen die Dauern 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)).

1.3.5 Stimmqualität

Die Stimmqualität wird subjektiv als Timbre wahrgenommen. Sie wird durch Phonation und Artikulation (vgl. Laver (1991)) beeinflusst. Als akustische Parameter werden meist Relationen verschiedener spektraler Bänder verwendet, so hat z.B. eine schrille Stimme mehr Energie in den höheren Bereichen. Ein weiterer Parameter 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.

Man kann hier zwischen Effekten mit und ohne Sprecherkontrolle unterscheiden. Effekte ohne Sprecherkontrolle ergeben sich durch die anatomischen Unterschiede zwischen den Sprechern und können zur Sprecheridentifizierng genutzt werden (Valbret et al. (1992)). Effekte mit Sprecherkontrolle entstehen durch Spannungen bestimmter Muskeln am Kehlkopf oder am Artikulationstrakt und geben Stimmungen und Einstellungen des Sprechers wieder.

Stimmqualität alleine kann schon zu Erkennung von emotionalen Zuständen des Sprechers führen (Scherer (1995)). Die Stimmqualität ist ein sehr wichtiger Parameter für den emotionalen

[4] Der Range ist der Abstand zwischen geringstem und größtem Wert (äquivalent dazu: Variationsbreite).
[5] Die Variabilität ist ein weiterer Streuungsparameter, üblicherweise wird die Standardabweichung verwendet.

Gehalt, da sie einen geringen Einfluss auf die Verständlichkeit von Sprache hat, und nicht Träger linguistischer Information ist (vgl. Scherer et al. (1984, S. 1349)).

1.3.6 Resonanzverhalten

Das Resonanzverhalten des Ansatzrohres 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 und Diphthongen auf, wobei die Lage der untersten drei den Laut bestimmen und die Lage der höheren vor allem zum Klangeindruck beitragen. Als solche sind sie sprechercharakteristisch, da ihre Lage unter anderem durch die Länge des Artikulationstrakts festgelegt ist. Wenn auch ihre relative Lage weitgehend durch den artikulierten Vokal festgelegt ist, gibt es doch in der absoluten Lage durchaus Spielraum. Eine Verkürzung des Vokaltrakts, wie sie etwa durch lächelnde Sprechweise entsteht, führt z.B. zu einer generellen Erhöhung der Formanten (Fant (1957)).

Automatische Formantextraktion ist sehr fehleranfällig, speziell bei Sprache mit hoher Grundfrequenz, da der F_0 unter Umständen über dem ersten Formanten liegen kann. Formantlagen und Bandbreiten sind explizite Parameter von Formantsynthesizern (siehe 1.4.4), was diese zur Simulation emotionaler Sprechweise sehr geeignet macht (vgl. Karlsson (1992)).

1.4 Text-to-Speech-Verfahren

Text-to-Speech-Systeme (TTS-Systeme) zeichnen sich im Gegensatz zu Sprachwiedergabesystemen dadurch aus, dass die Eingabe grundsätzlich nur durch die jeweilige Sprache begrenzt ist. Sie kann in schriftlicher Form oder durch ein Textgenerierungssystem erfolgen. Der Prozess findet in zwei Stufen statt; Symbolverarbeitung und akustische Synthese. Diese beiden großen Teilbereiche werden auch als NLP (Natural Language Processing) und DSP (Digital Signal Processing) bezeichnet. Eine schematisch vereinfachte Darstellung eines TTS-Systems ist in Abbildung 1.1 dargestellt..

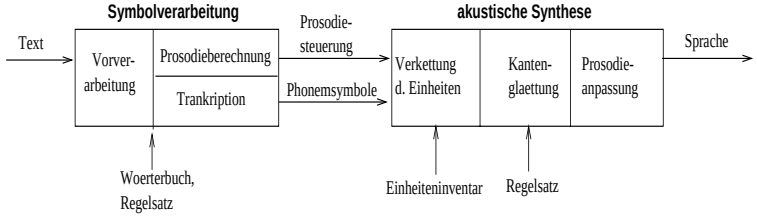

Abbildung 1.1: Allgemeines Schema für TTS-Systeme

1.4.1 Symbolverarbeitung

Der zu synthetisierende Text wird zunächst durch die Symbolverarbeitung in eine phonetische Darstellung gebracht. Zusätzlich wird eine Prosodiebeschreibung generiert. Dieser Prozess lässt sich wiederum in mehrere Schritte untergliedern.

1. In einer Vorverarbeitung werden Abkürzungen, Ziffern und Sonderzeichen in orthographische Form gebracht. Probleme gibt es hierbei vor allem mit der Vieldeutigkeit von Satzzeichen.

2. Danach wird in der Regel eine morphologische Analyse durchgeführt, bei der die Wörter in ihre Morpheme zerlegt und gemäß ihrer grammatischen Funktion klassifiziert werden. Dies geschieht mit Hilfe von Wörterbüchern.

3. Zusätzlich zur Analyse auf Wortebene wird eine kontextuelle Analyse durchgeführt, die die Wörter gemäß ihrem Kontext klassifiziert.

4. Die Wörter werden dann phonetisch transkribiert. Hier gibt es grundsätzlich den Unterschied zwischen regel- und datenbasierter Transkription. Regelbasierte Transkription wendet Phonemisierungsregeln an, während datenbasierte Einträge in Wörterbüchern sucht[6]. Beide Ansätze erfordern eine Mischform, so benötigt die datenbasierte Transkription Regeln für unbekannte Wörter und die Regelbasierte Wörterbücher für Ausnahmen. Die Entscheidung für einen der beiden Ansätze ist sprachabhängig. Oft gibt es eine Nachverarbeitung, die koartikulatorische Reduktionen berücksichtigt.

[6] Um das Lexikon nicht zu groß werden zu lassen, handelt es sich meist um Morphemwörterbücher.

5. Sind die Wörter nach grammatischer Funktion, Stellung im Satz und eventuell semantischem Gehalt klassifiziert, kann eine Berechnung der Prosodie (siehe 1.3.1) erfolgen. Die Beschreibung verschiedener Prosodieerzeugender Verfahren erfolgt unter 1.4.2 gesondert, da sie für die Generierung emotionaler Sprechweise von besonderer Bedeutung ist.

1.4.2 Prosodiesteuerung

Die Prosodie einer Äußerung hat viele Funktionen. 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. Zudem drückt sich in ihr auch der emotionale Zustand des Sprechers aus.

Um eine neutrale Prosodie mit geringem Aufwand generieren zu können, konzentrieren sich die meisten TTS-Systeme auf die Syntax. Eine einfache Methode ist der *Chinks'nChunks*-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[7].

Komplexere Systeme verwenden Ansätze des maschinellen Lernens wie z.B. statistische Entscheidungsbaumverfahren (CART[8]) (vgl. Malfrère et al. (1999); Fordyce & Ostendorf (1998)), oder neuronale Netze. Mittels dieser können aus großen Datenbanken jeweils approbate Prosodiebeschreibungen extrahiert werden. Ein großer Vorteil von Entscheidungsbaumverfahren gegenüber neuronalen Netzen liegt darin, dass aus dem Gelernten Regeln extrahiert werden können.

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[9].

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[10], 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 Akzent 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.

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

[8] Classification And Regression Tree

[9] 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. Von 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)

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

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 unter 1.3.2 beschriebenen Gründen Schwierigkeiten bereitet.

1.4.3 Akustische Synthese

Die akustischen Synthese generiert letztlich das Sprachsignal. Auf der Verkettungsstufe wird die diskrete Kette von phonetischen Symbolen in einen kontinuierlichen Datenstrom von Signalabtastwerten oder -parametern überführt. Je nach Synthesemethode werden 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. Grundsätzlich unterscheidet man zwischen regel- und datenbasierter Synthese.

1.4.4 Regelbasierte Systeme

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 tabellarisch erfasst sind. Regelbasierte Synthese eignet sich insofern zur Simulation emotionaler Sprechweise, als dass die Regeln zur Emotionserzeugung direkt auf die Generierungsregeln angewendet werden können. Es lassen sich akustisch-parametrische und artikulatorische Systeme unterscheiden.

Artikulatorische Synthese

Artikulatorische Synthese modelliert physiologische Vorgänge. Je nach Ansatzpunkt ist zwischen *neuromotor-command-*, *articulator-* und *vocal-tract-shape*-Synthese zu unterscheiden. Letztendlich werden Querschnittsverläufe für den Vokaltrakt[11] sowie ein Anregungssignal festgelegt, aus denen dann Parameter für Formantsynthese berechnet werden.

Physiologische Vorgänge, die für einzelne Emotionen charakteristisch sind, könnten hierbei auf höchster Ebene modelliert werden, was eine hohe Flexibilität ermöglichen würde. Langfristig ist dies ein vielversprechender Ansatz, natürlich klingende Sprache zu erzeugen. Allerdings ist das Wissen über die Zusammenhänge zwischen neuronalen Zuständen bzw. Artikulatorbewegungen und dem akustischen Signal zur Zeit noch zu gering, um auch nur bezogen auf die Verständlichkeit qualitativ hochwertige Sprachsynthese mit diesen Ansätzen zu erreichen.

Akustisch-parametrische Synthese

Bei den parametrischen Verfahren (Vocoder-Verfahren) sind vor allem LPC- und Formant-Synthese zu nennen. Die Spracherzeugung wird hierbei in ein Anregungssignal und einen Filter aufgeteilt (Quelle-Filter Modell). Dieses Modell ist beispielhaft für LPC-Synthese in Abbildung

[11] Mit Hilfe des Röhrenmodells, das das Ansatzrohr in 10-20 Röhrensegmente unterteilt

1.2 dargestellt. Das Anregungssignal besteht je nach Stimmhaftigkeit aus einer periodischen

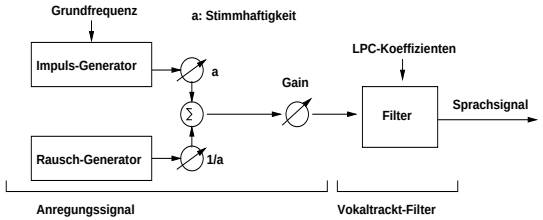

Abbildung 1.2: Einfaches LPC-Synthese Modell (nach O'Shaughnessy (1987))

Impulsfolge oder aus einer Zufallsfolge, die weißes Rauschen modelliert[12]. Formant-Synthese modelliert den Vokaltrakt durch eine Reihen- und/oder Parallelschaltung von digitalen Resonatoren, die die Frequenzen und Bandbreiten der ersten 3-4 Formanten repräsentieren. Höhere Formanten werden oft fest vorgegeben, da sie vor allem die Stimmqualität beeinflussen und nicht so sehr charakteristisch für einzelne Laute sind. Regelbasierte Synthese hat einen weitaus geringeren Speicheraufwand als datenbasierte, da das Einheiteninventar klein ist und stark komprimiert kodiert werden kann.

Ein für diese Arbeit wichtigerer Vorteil ist die Flexibilität bezüglich der akustischen Eigenschaften der Stimme, die sich aus der direkten Zugänglichkeit aller Parameter ergibt. Da die Stimmquelle durch mathematische Modelle erzeugt wird, kann die Stimmqualität direkt manipuliert werden (vgl. Karlsson (1992); Granström (1992)). So können akustische Phänomene wie z.B. Laryngalisierung erzeugt werden, die hauptsächlich von Unregelmäßigkeiten der Glottispulsfolge herrühren.

Allerdings ist regelbasierte Synthese oft von schlechterer Qualität als datenbasierte. Viele Koartikulationseffekte sind zwar durch einfache Regeln beschreibbar, wie z.B. bei gerundeten Lippen alle Formanten etwas abzusenken, für andere jedoch, wie etwa das Verkürzen zu Zielphonemen hin, gilt dies nicht. Zudem lassen Vereinfachungen wie das Modellieren der Luftstöße bei stimmhaften Lauten als Impulsfolge die resultierende Sprache unnatürlich klingen (O'Shaughnessy et al. (1987)). Bei Untersuchungen, die mit regelbasierten Systemen (wie z.B. DEC-Talk oder GLOVE) durchgeführt wurden (vgl. Murray & Arnott (1995); Carlson et al. (1992)), wurde bereits die neutrale Stimme von einem hohen Prozentsatz der Versuchshörern als traurig klingend bewertet, was sich vermutlich durch eine zu große Monotonie im Anregungssignal erklären lässt.

1.4.5 Datenbasierte Systeme

Datenbasierte Systeme generieren Sprache aus Grundeinheiten, die als Inventar vorliegen und miteinander konkateniert werden. Im Gegensatz zu den regelbasierten Systemen ist hier ein Großteil des phonetischen Wissens implizit in den Einheiten kodiert.

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

Wahl der Einheiten

Als Einheiten werden Silben, Halbsilben, Diphone, Phonemkluster oder Mischinventare verwendet (vgl. Portele (1996, S. 11); 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. Silben sind geeignet, weil sich Koartikulation vor allem innerhalb von Silben auswirkt. Ein Phonemkluster ist eine Kette von vokalischen oder konsonantischen Phonemen. Grundsätzlich wächst mit der Größe der Einheit das Inventar, aber die Glättungsregeln an den Übergängen werden einfacher, da es weniger Übergänge gibt.

Da Koartikulation sich oft über mehr als zwei Phone erstreckt, ist auch die Verwendung von Triphonen üblich. Zum Teil ist die Wahl des Inventars auch sprachabhängig. So sind Halbsilbensysteme attraktiv für Sprachen mit komplexen Konsonantenfolgen wie das Deutsche. Ein Halbsilbeninventar lässt sich reduzieren, indem z.B. finale Obstruenten abgetrennt werden oder Normschnittstellen zu homogenen Lauten verwendet werden.

Mischinventare verwenden etwa Anfangshalbsilben, gefolgt von Diphonen und Suffixen. Durch die enorme Leistungssteigerung der Hardware ist es auch möglich geworden, große gemischte Inventare zu verwenden, bei denen jeweils das am besten Passende herausgesucht wird. Eine Verfahren, bei dem subphonemische Einheiten verkettet werden, ist bei Benzmüller & Barry (1996) beschrieben.

Kodierung der Einheiten

Datenbasierte TTS-Systeme lassen sich zusätzlich zur Art der Einheiten danach klassifizieren, in welcher Form die zugrundeliegenden Einheiten gespeichert, verarbeitet und konkateniert werden (O'Shaughnessy et al. (1987)). Bei der Kodierung der Einheiten gibt es grundsätzlich zwei Möglichkeiten: parametrische und Waveform-Kodierung. Parametrische Kodierung spart Platz bei Speicherung und Übertragung und die Übergangsglättung und Anpassung an koartikulatorische Effekte und Intonation sind einfacher zu berechnen, da auf die spektralen Eigenschaften direkt zugegriffen werden kann. Dies stellt natürlich auch einen großen Vorteil bei der Simulation emotionaler Sprache dar. Allerdings macht sich der durch die Parametrisierung auftretende Informationsverlust bei der Resynthese störend bemerkbar.

Bei modernen Systemen wird in der Regel eine halbparametrische Kodierung verwendet, um die Vorteile beider Formen ausnutzen zu können (siehe Dutoit & Leich (1994)). Die Segmente werden etwa dadurch erzeugt, dass für untere Frequenzbereiche eine sinuidale Anregung und für höhere eine stochastische zugrunde gelegt wird (Harmonisch/Stochastisches Modell, siehe Dutoit & Gosselin (1996)). Hierbei wird die Voiced/Unvoiced-Entscheidung, die frühere Vocoder für ganze Sprachabschnitte getroffen haben, in einzelne Frequenzbereiche verlegt. Dieses Verfahren ist vom MBE-Modell[13] abgeleitet. Weitere halbparametrische Verfahren werden unter 1.5.1 beschrieben.

[13] Multi-Band-Excitation-Modell, siehe Dutoit & Leich (1993 b)

1.5 Resynthese- und Signalmanipulationsverfahren

Einige Verfahren zur Manipulation und Resynthese digitaler Sprachsignale werden hier näher erläutert, da sie nicht nur eine Bedeutung bei der automatischen Sprachgenerierung sondern auch bei der Durchführung mehrfaktorieller Versuchspläne mittels Resynthese haben (siehe 4.2).

1.5.1 PSOLA-Verfahren

Bei der akustischen Synthese in TTS-Systemen muss neben der Übergangsglättung an den Einheitengrenzen die Prosodie erzeugt werden. Bei parametrischer Darstellung ist dies relativ einfach, da alle Parameter der Prosodie (Dauer, Intensität, F_0) explizit in den Frames enthalten sind. Für die Waveform-Darstellung ist das erst durch den PSOLA-Algorithmus[14] 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 (siehe Abbildung 1.3). Durch einen zusätzlichen Faktor bei der Gewichtung wird die Energie der Signalabschnitte korrigiert.

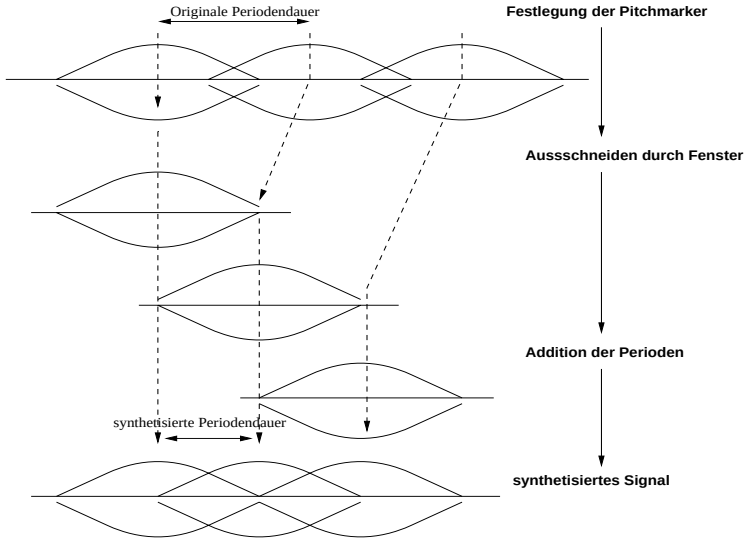

Abbildung 1.3: Erhöhung der Grundfrequenz durch das TD-PSOLA-Verfahren (nach Hirokawa & Hakoda, 1990)

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

[14] Pitch Synchronous Overlap Add

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[15] unterzogen werden. Die Syntheseergebnisse sind besser als beim TD-PSOLA, da die spektrale Glättung an den Übergängen der Fenster vereinfacht wird, jedoch ist das Verfahren äußerst aufwendig.

Beim LP-PSOLA (linear predictive PSOLA) wird das Signal durch LPC-Analyse (siehe 1.5.2) in ein Residual und ein Signalformungsfilter getrennt, der PSOLA-Algorithmus wird dann auf das Residual[16] 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[17], da es unter 6.1 verwendet wird. Es stellt eine Kombination dar aus TD-PSOLA und Verfahren, die das Signal in Frequenzbereiche zerlegen[18]. Einige Probleme des TD-PSOLA werden bei der Verwendung zur Sprachsynthese dabei dadurch vermieden, dass das Inventar mit konstanter Grundfrequenz vorliegt.

Zwei weitere auf PSOLA basierende Verfahren zur Veränderung der Charakteristik von Sprachsignalen werden von Valbret et al. (1992) vorgeschlagen. Das Spektrum des zu resynthetiserenden Signales wird dabei dem Spektrum eines Zielsignales 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.

1.5.2 LPC-Verfahren

Bei der linearen Prädiktion[19] wird die Signalformung als üblicherweise rein rekursives Filter aufgefasst, welches das Anregungssignal verändert. Das Residualsignal wird aus dem Originalsignal durch inverse Filterung mit den Filterkoeffizienten gewonnen. Dafür wird das Signal für die Dauer eines Analyse- bzw. Synthesefilters als stationär angenommen (siehe O'Shaughnessy (1987, S. 336)). Die Fensterlänge wird in der Regel zwischen 5 und 30 ms gewählt. Die Verschiebungszeit des Fensters kann ebenfalls festgelegt werden oder pitchsynchron erfolgen, was eine vorherige Festlegung von Pitch-Markern erfordert[20].

Bei der LPC-Resynthese wird als Anregungssignal nicht wie in Abbildung 1.2 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 Informationsverlust auftritt. Das liegt daran, dass Koeffizienten, die Energiehäufungen unter- oder oberhalb der Formantlagen beschreiben, ignoriert werden. Bei der Resynthese wird diese Verfälschung des Signals deutlich hörbar.

Eine weitere Fehlerquelle ist die Vereinfachung vieler LPC-Analyseverfahren, bei denen

[15] Fast-Fourier-Transformation

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

[17] Multi-Band- Resynthesis-TD-PSOLA (siehe Dutoit & Leich (1993 a)

[18] Es verwendet somit Einheiten in halbparametrische Darstellung.

[19] LPC: Linear Predictive Coding

[20] Die Lage der Pitch-Marker ist für eine gute Resynthesequalität kritisch und muss von Hand nachgebessert werden (vgl. KAY (1992)).

der Filter nur Pol- aber keine Nullstellen hat. Die Annahme, das Sprachspektrum hätte keine Nullstellen, stellt aber zumindest bei Nasalen eine grobe Vereinfachung dar (O'Shaughnessy (1987, S. 337)). Deshalb werden bei neueren LPC-Algorithmen die Nullstellen durch Polstellen modelliert.

Da die Frequenzenergie von Sprachsignalen mit etwa 6 dB pro Oktave abfällt, wird vor der LPC-Analyse Preemphase[21], auf das Signal angewendet, um auch die oberen Bereiche 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. Die LPC-Analyse wird charakterisiert durch:

- Die Art des Verfahrens zur Berechnung der Filterkoeffizienten, hier kommen vor allem Autokorrelations- und Kovarianzverfahren in Frage.
- Die Anzahl der Koeffizienten (Filterordnung). Bei der Umrechnung in Formantlagen und -breiten werden jeweils zwei Koeffizienten pro Formant benötigt, plus einigen zusätzlichen, um die Nullstellen zu modellieren. Üblicherweise wählt man die Filterordnung ungefähr entsprechend der Samplerate/1000.
- Die Form und Breite des Analysefensters, wobei üblicherweise für Sprachsignale ein Hamming- oder Hanningfenster verwendet wird. Die Breite muss natürlich mindestens dem Offset entsprechen und beträgt in der Regel das Anderthalbfache, damit sich die Fenster genügend überlappen.
- Die Entscheidung darüber, ob der Frameoffset pitchsynchron erfolgt oder, wenn er fest vorgegeben ist, den Abstand dafür. In der Regel beträgt der Frameoffset zwischen 5 und 25 ms.
- Die Wahl des Preemphasefaktors.

[21] Das Spektrum wird durch Hochpassfilterung geglättet.

Kapitel 2

Literaturbesprechung

2.1 Besprechung ausgewählter Experimente

Im Folgenden werden fünf Arbeiten aus der Literatur näher besprochen. Bei den drei ersten handelt es sich um Untersuchungen zu Korrelaten freudiger Sprechweise. Zwei davon (Scherer et al. und Granström et al.) verwendeten Resynthese, die dritte (Tartter) wird besprochen, da sie eine der wenigen Untersuchungen ist, die sich ausschließlich mit freudiger Sprechweise beschäftigt. Bei den letzten zwei (Murray & Arnott und Cahn) handelt es sich um Implementierungen von Computerprogrammen, die emotionale Sprache erzeugen sollen. Dies ist vor allem im Hinblick auf Kapitel 5 von Interesse.

2.1.1 Untersuchung allgemeiner prosodischer Merkmale

Bergmann et al. (1988) führten vier Experimente zu akustischen Korrelaten emotionaler Sprechweise 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. Die Versuchspersonen ordneten die Stimuli jeweils einer von sechs Emotionen zu[1]. Zusätzlich standen zur Beurteilung drei eher auf die Einstellung des Sprechers bezogene Skalen zur Verfügung[2]. Es wurden nur Ergebnisse angegeben, die unter einem Signifikanzniveau von 1 % lagen.

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. Als Ergebnis korrelierte „freudig“ am meisten mit „entgegenkommend“ (Korrelation von 0.38). Ansonsten war lediglich der Parameter F_0-Range für Freude signifikant, wobei ein schmaler Range als weniger freudig empfunden wurde als ein breiter. In einem zweiten Experiment wurde die

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

[2] nachdrrücklich, vorwurfsvoll und entgegenkommend

Akzenthöhe in fünf Varianten verändert, hierbei ergaben sich allerdings keinerlei signifikante Effekte für „freudig“.

Der dritte Versuch bestand darin, dass die Variablen Akzenthöhe, Akzentdauer sowie wiederum F_0-Perturbation variiert wurden. 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.

Sowohl das Merkmal Akzenthöhe als auch Dauer ergaben für „freudig“ eine signifikante Bewertung. Eine kurze Dauer und eine hohe F_0-Lage der betonten Silbe führte zu einer Attribution der Äußerung als freudig. Weder Interaktionen noch das Merkmal Perturbation ergaben einen signifikanten Unterschied für „Freude“.

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. Hierbei gab es keinerlei signifikante Effekte für „freudig“. 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.

2.1.2 Untersuchung zu F_0-Verlauf und Dauer

Carlson et al. (1992) führten Untersuchungen zum Einfluss der Prosodie auf emotionale Eindruckswirkung durch. Ausgangspunkt waren Stimuli aus einem früheren Experiment (Öster & Riesberg (1986)). Schauspieler hatten hierbei die Emotionen Angst, Traurigkeit, Freude und Neutral in mehreren Sätzen simuliert. Eine erste Analyse der in einem Hörversuch am besten wiedererkannten Sätze ergab eine höhere Energie und eine gegenüber der neutralen Versionen etwas langsamere Sprechweise der freudigen Äußerungen.

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 F_0-Verlaufs- und Dauerwerte, die Formanten und Parameter für das Glottissignal. 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. Die freudige Version erreichte eine Wiedererkennungsrate von gerade mal 42 % , was eine schlechte Synthesequalität vermuten lässt. Auch die mittels freudiger Werte manipulierten Versionen der anderen Emotionen wurden überwiegend nicht als freudig beurteilt. Freude wurde oft mit Ärger verwechselt, was sich durch einen ähnlichen Erregtheitsgrad der beiden Emotionen erklären lässt.

2.1.3 Untersuchung zur akustischen Wirkung von Lächeln

Tartter (1980) untersuchte die Auswirkungen von Lächeln auf das Sprachsignal. Lächeln verändert die Physiognomie auf drei Arten: die Mundöffnung wird breiter, die Länge des Ansatzrohres wird verkürzt und seine Breite vergrößert.

Als Stimuli wurden 29 Äußerungen verwendet (25 Nonsenseinsilber und vier Sätze), die jeweils mit und ohne Lächeln von sechs Sprechern realisiert wurden. Drei Gruppen von je 12 Versuchshörern wurden aufgefordert, die Äußerungen zu klassifizieren. Sie wurden jeweils paarweise angeboten. Die erste Gruppe sollte entscheiden, welche Äußerungen lächelnd realisiert wurden, die zweite, welche Äußerungen freudiger wirken und die dritte Gruppe, welche trauriger klingen.

Es stellte sich heraus, dass in der ersten Gruppe für alle Sprecher die lächelnd gesprochenen Äußerungen erkannt wurden. In der zweiten Gruppe wurde für fast alle Sprecher die Stimuli mit Lächeln als freudiger beurteilt. Auch die dritte Gruppe beurteilte zum großen Teil diejenigen Äußerungen als traurig klingend, welche mit starrer Miene gesprochen wurden.

Die Äußerungen wurden daraufhin auf Lagen der ersten drei Formanten, F_0-Verlauf, Dauern und Amplitude analysiert. Für alle Sprecher lagen alle drei Formanten höher, und zwar am stärksten bei denen, die auch am meisten als lächelnd bzw. freudig erkannt wurden. Auch die mittlere Grundfrequenz lag höher, was Tartter dadurch erklärt, dass sich die Spannung der Gesichtsmuskeln beim Lächeln auf die Stimmlippen auswirkt. Bei vier Sprechern gab es einen signifikanten Anstieg der Amplitude, der vermutlich vom größeren Öffnungsgrad des Mundes herrührt. Die Unterschiede hinsichtlich der Dauern ergaben zwar auch signifikante Werte, allerdings weder konsistent zwischen den Sprechern noch hinsichtlich ihrer Beurteilung. Der Autor kam zu dem Schluss, dass die Dauer kein ausschlaggebender Parameter für die Freudigkeit einer Äußerung ist.

2.1.4 Erzeugung von Freude in Text-to-Speech-Systemen

Cahn 1989

Ein System zur Generierung von Steuerinstruktionen für Sprachsynthesizer, die sie befähigen soll, Emotionen zu simulieren, wurde von Cahn (1989) in ihrer Magisterarbeit vorgestellt. Cahn 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. 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. Der Affect Editor analysiert diese Äußerung nach möglichen Betonungs- und Pausenorten, erzeugt einen Prosodieverlauf und belegt dann die oben genannten Parameter mit für die gewünschte Emotion approbaten Werten. Der Output besteht aus einem Gemisch von Text und Phonemen und 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 den Grad 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 mit Ärger klassifiziert wie mit 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 meisten verwechselten Urteilen angibt. Weiterhin fällt auf, dass unter den angebotenen Klassifizierungen Neutral nicht vorkam, wohingegen Überraschung[3] als Nebenemotion gegenüber den anderen Basisemotionen herausfällt.

Murray und Arnott 1995

Murray & Arnott (1995) stellten ein Softwaresystem namens HAMLET[4] vor. Dieses dient zur automatischen Anwendung von „Gefühlsregeln“, um den Output von TTS-Systemen zu emotionalisieren. Das System stellt zunächst 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“[5]. 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.

Als Ergänzung zur automatischen Anwendung wurde eine interaktive Umgebung entwickelt, in der Benutzer die Parameter manipulieren können, so dass HAMLET auch als Forschungsinstrument dienen kann. 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. Vorgesehen waren sechs Emotionen: Wut, Freude, Trauer, Angst, Ekel und starke Trauer[6]. Es wurden Testsätzen mit emotional unbestimmten Inhalt und solche mit emotionalem Inhalt verwendet. Daraus wurde eine Stimulimenge von 39 Sätzen erstellt, die jeweils neutral und emotional synthetisiert wurden. Insgesamt waren also 78 Sätze zu beurteilen. Von den 39 Sätzen waren 18

[3] original: *surprise*
[4] HAMLET: Helpful Automatic Machine for Language and Emotional Talk
[5] Es werden drei Arten von Betonung unterschieden: nebenbetont, hauptbetont und emphatisch.
[6] original: *grief*

emotional unbestimmten Inhalts und 21 vom Inhalt her einer Emotion zuzuordnen, so dass es vier Mengen von Testsätzen gab. Es wurden nacheinander zwei Tests durchgeführt, ein free-response- und ein forced-choice-Test (Begriffserklärung siehe 4.3). Beim forced-choice-Test wurden neben den sechs zu simulierenden Emotionen noch „neutral“, „andere“ und sieben weitere Distraktoremotionen zur Wahl gestellt.

Zu den Ergebnissen ist zunächst zu bemerken, dass bereits den emotional unbestimmten, neutral synthetisierten Sätzen von etwa zwei Dritteln der Versuchshörer Emotionen zugeordnet wurden, und zwar fast ausschließlich negative[7]. 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. Freude wurde am meisten mit Furcht verwechselt, allerdings streuten die Urteile sehr. Bei den Sätzen mit emotionalem Inhalt war das Ergebnis erwartungsgemäß deutlich besser. Mit Ausnahme von Ekel[8] lagen alle Maxima auf den entsprechenden Emotionen.

Die Autoren kommen zu dem Schluss, dass ihr System erkennbar emotionale Sprechweise generieren kann, was für Wut und Trauer auch sicherlich zutrifft. Die Evaluierung mittels Text, dem die gewünschte Emotion bereits durch seine Bedeutung inhärent ist, ist allerdings problematisch. Schließlich wissen die Hörer hierbei, welche Emotion intendiert ist, was ihr Urteil sicherlich erheblich beeinflusst. Dass dem so war zeigt ja die Tatsache, dass deutlich mehr Sätze mit emotionaler Semantik aber neutraler Sprechweise ihren korrespondierenden Emotionen zugeordnet wurden als Sätze mit neutralem Inhalt aber emotionaler Sprechweise. Die Autoren haben es sich andererseits auch durch die Hinzunahme von zahlreichen Distraktoremotionen als Antwortmöglichkeiten nicht gerade leichtgemacht.

2.2 Zusammenfassung von Ergebnissen

Es folgt eine Zusammenfassung der Ergebnisse aus verschiedenen weiteren Untersuchungen.

- **Grundfrequenz:**
  durchschnittliche F_0:
  Laut Öster & Riesberg (1986); Trojan (1975); Tartter (1980) führt Freude zu einem Anstieg der Grundfrequenz.
  Range:
  Laut Bergmann et al. (1988) werden Äußerungen mit breiterem Range eher als freudig empfunden.
  Variabilität:
  Laut Scherer (1981) führt Freude zu höherer Variabilität.
  segmentelle Konturen:
  Nach Mozziconacci & Hermes (1997) sind segmentelle Intonationsmuster für unterschiedliche Emotionen charakteristisch. Die Autoren haben z.B. für Freude mehr ansteigende Konturen auf den ersten Betonungen festgestellt als bei neutraler Sprechweise.

- **Betonungen:**
  Laut Trojan (1975) vergrößern sich die Unterschiede zwischen unbetonten und betonten

[7] Angst, Trauer und Ekel
[8] Ekel wurde oft mit Freude verwechselt

Silben. Nach Bergmann et al. (1988) sollten phrasenbetonende Silben von hoher Grundfrequenz und kurzer Dauer sein.

- **Sprechtempo:**
  Laut Öster & Riesberg (1986); Carlson et al. (1992) ist freudige Sprechweise durch ein langsames Tempo gekennzeichnet. Fonagy (1981); Davitz (1964) allerdings beschreiben das Tempo als lebendig bzw. weisen auf eine Erhöhung der Sprechrate hin.

- **Intensität, Lautheit:**
  Die Intensität nimmt laut Davitz (1964); Carlson et al. (1992) bei freudiger Sprechweise zu.

- **Stimmqualität**
  Formanten
  Laut Fant (1957, S. 19); Tartter (1980); Laver (1991, S. 219) führt lächelnde Sprechweise zu einer Anhebung der Formanten.
  Phonation
  Nach Fonagy & Magdics (1963) weist freudige Sprechweise eine behauchte Phonation auf.
  Energieverteilung in spektralen Bändern
  Klasmeyer & Sendlmeier (1999) messen bei freudigen Äußerungen einen erhöhtem Anteil der Frequenzen zwischen drei und fünf kHz.

Teil II

Experimenteller Teil

Kapitel 3

Akustische Analyse von neun Satzpaaren

3.1 Vorbemerkung

Um konkrete Anhaltspunkte zu bekommen, welche Parameter wie verändert werden müssen, um freudigen Sprechausdruck zu simulieren, wurde emotionales Sprachmaterial analysiert. Ziel war hierbei, Hinweise aus der Literatur zu bestätigen, zu erweitern und zu konkretisieren.

3.1.1 Gewinnung der Daten

Zur Untersuchung emotionaler Sprechweise wäre es natürlich am besten, emotionale Äußerungen aus echten Situationen zu verwenden. Dies ist aber in der Regel bei dem Vergleich mehrerer Emotionen nicht praktikabel. Erstens sind Aufnahmen in freier Umgebung für genauere Analysen nicht zu gebrauchen, da es zu viele Störgeräusche gibt und kein konstanter Mikrophonabstand garantiert werden kann. Weiterhin ist es unmöglich, auf diese Art Äußerungen mit unterschiedlichem Emotionsgehalt aber gleichlautendem Text zu erhalten.

Als Alternative werden deshalb in der Regel von professionellen Sprechern simulierte Sätze verwendet, deren Validität durch Hörerurteile absichert wird. Um das Hineinversetzen in emotionale Zustände zu erleichtern, werden die Äußerungen oft in kleine Szenen eingebettet (vgl. Williams & Stevens (1972); Scherer et al. (1984)).

Für diese Untersuchung wurde eine Datenbank verwendet, die im Rahmen eines DFG-Projekts am Institut erstellt wurde (Sendlmeier et al. (1998)). Es wurden zehn einphrasige und zehn zweiphrasige Sätze von sieben Schauspielern und Schauspielschülern in den Emotionen Freude, Trauer, Angst, Langeweile, Ekel und Ärger realisiert. Die Aufnahmen wurden mit einem U 87-Mikrofon und einem Sony-Datrecorder in einem reflexionsarmen Raum durchgeführt. Im Anschluss daran wurden die Äußerungen in einem Hörversuch evaluiert.

Für die folgende Analyse wurden alle Sätze verwendet, deren neutrale und freudige Version eine Erkennungsrate von über 98 % hatten und die zweiphrasig waren. Dies war dadurch motiviert, dass es bei mehrphrasigen Sätzen mehr Möglichkeiten der Phrasenbetonungen gibt, deren Veränderung Gegenstand der Untersuchung war. Diese Auswahl ergab 18 Sätze = 9 Paare. Insgesamt gab es fünf Sprecher, drei Frauen und zwei Männer. Die Sätze sind unter A.2 aufgelistet.

3.1.2 Überblick der vorgenommenen Messungen

An diesen Sätzen wurden folgende Messungen mit Xwaves und Praat (siehe A.1) vorgenommen:

- Dauern der Laute
- Dauern der Silben
- Für jede Silbe drei F_0-Werte (Anfang, Mitte und Ende)[1]
- Kennzeichnen von Wort- und Phrasenbetonungen
- Durchschnittliche Lage der ersten fünf Formanten
- Differenz der Energie in verschiedenen Bändern

Die Messwerte für Gesamtdauern und globale F_0-Parameter für alle Sätze sind unter A.7 dargestellt.

Alle Ergebnisse wurden zunächst für jeden Sprecher einzeln betrachtet. Gab es mehrere Sätze eines Sprechers, so wurden die Einzelwerte der Sätze gemittelt. Gegenstand der Analyse waren Dauerparameter, Grundfrequenzparameter und spektrale Eigenschaften. Eine genauere Analyse der Artikulationsgenauigkeit bzw. der Reduktion oder Elaboration von Silben wurde nicht durchgeführt, da sich beim Labeln der Sätze keine Hinweise auf derartige Unterschiede ergaben. Zudem hatte auch die Untersuchung von Kienast (1998), die zum Teil dasselbe Datenmaterial verwendete, keine signifikanten Unterschiede dieser Parameter für Freude ergeben.

Bei emotionaler Sprechweise treten oft extralinguistische Laute wie Schmatzen, Lachen und weitere auf (Granström (1992)). Diese wurden in dieser Untersuchung aber nicht berücksichtigt[2], da sie in der Regel mit TTS-Systemen nicht zu simulieren sind.

Weiterhin wurden eventuelle Unterschiede bezüglich der Lautheit nicht untersucht. Die objektive Messung (vgl. 1.3.2) ist problematisch und in der Regel ist bei TTS-Systemen eine Variation der Lautheit innerhalb von Äußerungen nicht vorgesehen.

Die Untersuchungen stellen jeweils einen Vergleich zwischen freudiger und neutraler Sprechweise dar, angegeben sind in der Regel die Veränderungen zwischen den korrespondierenden Sätzen in Prozent. Dies entspricht der Zielrichtung der Arbeit, Regeln für die Transformation neutraler Sprechweise in freudige zu finden.

Ergebnisse wurden mittels eines T-Test für abhängige Variablen (paired t-test) auf Signifikanz geprüft. Verwendet wurde die Statistiksoftware R (siehe A.1). Als Ergebnis wird der p-Wert des Tests angegeben. Dieser gibt die Wahrscheinlichkeit eines Irrtums bei Zurückweisung der Nullhypothese an. Bei p-Werten von unter 0,05 gilt das Ergebnis als signifikant, unterhalb von 0,01 als sehr bzw. hoch signifikant. Liegt der p-Wert unter 0,1, kann man von einer Tendenz sprechen. Dies entspricht aber keinesfalls einer gesicherten Erkenntnis. Wenn nicht anders angegeben, sind alle T-Tests zunächst zweiseitig angelegt, dass heißt, es wurde kein Vorwissen benutzt. Die Alternativhypothese besagt jeweils, dass die Werte für freudige versus neutrale Sprechweise sich signifikant unterscheiden, aber nicht, ob sie größer oder kleiner sind.

[1] Da Grundfrequenzanalyseverfahren fehlerhaft arbeiten, und Entscheidungen (speziell bei Laryngalisierungen) zum Teil Ermessenssache sind, wurden kritische Stellen akustisch überprüft.

[2] Sie traten bei den gewählten Sätzen auch kaum auf.

3.2 Daueruntersuchung

Auf eine Analyse der Sprechpausen wurde verzichtet, da nur in zwei Sätzen Sprechpausen vorkamen. Es gab auch in der Literatur keinerlei Hinweise auf ein Korrelieren freudigen Sprechausdrucks mit verstärktem oder verringertem Pausenauftreten.

3.2.1 Durchschnittliche Silbendauern

Im Hinblick auf die spätere Verwendung zur Modifikation der Laut- und Silbendauern neutraler Sprachsignale wurden als Maß für die Sprechrate die durchschnittlichen Silbendauern verwendet, die sich als Verhältnis von Gesamtsprechzeit und Silbenanzahl ergeben[3](siehe 1.3.4). Für die wort- und phrasenbetonenden Silben wurden gesondert mittlere Silbendauern berechnet.

In Tabelle 3.1 sind die Abweichungen der Silbendauern für die freudigen gegenüber den neutralen Äußerungen in Prozent und die dazugehörigen p-Werte dargestellt. In Abbildung 3.1 sieht man oben die Ergebnisse für die allgemeinen Silbendauern und unten für die phrasenbetonender Silben als Balkendiagramme. Auf eine Darstellung für wortbetonende Silben wurde verzichtet, da die Ergebnisse nicht signifikant waren. Der T-Test wurde nicht über die gemittelten Werte der Sprecher berechnet, sondern für alle Sätze.

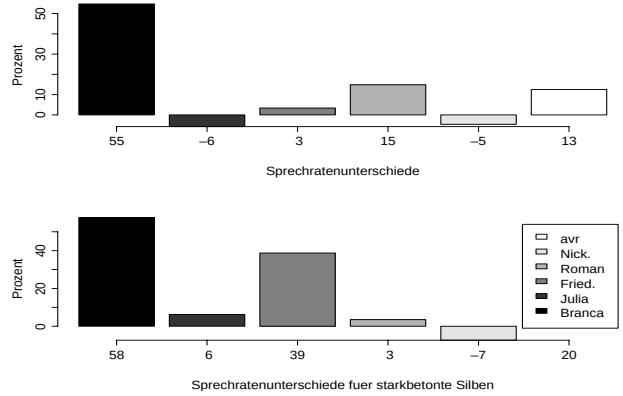

Abbildung 3.1: Abweichungen der Silbendauern

Allgemeine Silbendauern

Die Ergebnisse für die einzelnen Sprecher waren keineswegs eindeutig; während bei Sprecherin Branca die Silbendauern durchschnittlich um 55 % gestiegen waren, sind sie bei den Sprechern Julia und Nick sogar gesunken. Die Aussagen aus der Literatur (vgl. 2.2) sind hier äußerst

[3]Sprechpausen wurden in die Berechnung der Gesamtsprechzeit nicht einbezogen, da sie allgemein nicht untersucht wurden und ansonsten die Ergebnisse für alle und die betonten Silben nicht miteinander vergleichbar gewesen wären.

Sprecher	Silbendauern allg. [%]	Wortbetonungen [%]	Phrasenbetonungen [%]
Branca	54.7	35.1	57.5
Julia	-5.8	1.0	6.2
Fried.	3.4	2.5	38.7
Roman	14.9	26.4	3.5
Nick.	-4.6	-19.6	-7.4
∅	12.5	9.1	19.7
p-Wert	0.079	0.115	0.038*

Tabelle 3.1: Abweichung durchschnittlicher Silbendauern

widersprüchlich. Die meisten weisen auf ein erhöhtes Sprechtempo hin, begründet durch eine stärkere Erregung. In dieser Untersuchung gab es eine Tendenz zu langsamerer Sprechweise. Eventuell haben die Schauspieler, um besonders deutlich zu sein, eine sehr emphatische Freude simuliert. Es ist wohl so, dass die Sprechrate kein entscheidender bzw. alleiniger Parameter für die Perzeption freudiger Sprechweise ist.

Dauern der betonten Silben

Die Ergebnisse für wortbetonende Silben zeigten keinen signifikanten Unterschied. Für die phrasenbetonenden Silben ergab sich jedoch eine signifikante Daueränderung, sie wurden im Schnitt um 20 % verlängert. Einzige Ausnahme war Sprecher Nick, bei dem die Silben kürzer wurden. Man kann also auch freudig sprechen, ohne die betonten Silben zu verlängern.

Dieses Ergebnis deckt sich insofern mit denen aus der Literatur, als dass Betonungen generell stärker ausfallen[4] (vgl. 2.2), steht allerdings im Widerspruch zu Bergmann et al. (1988). Es ist allerdings zu bedenken, dass die Silben nicht direkt verglichen werden konnten, da bei freudiger Sprechweise Phrasenbetonungen auftraten, die in der neutralen gar nicht vorkamen.

3.2.2 Lautdauern

Hierbei wurden die mittleren Lautdauern für ausgewählte Lautklassen bestimmt. Die Lautklassen waren:
Plosive stimmlos/stimmhaft (Psl/Psh), Frikative stimmlos/stimmhaft (Fsl/Fsh), Liquide (L), Nasale (N) und Vokale gespannt/ungespannt (Vt/Vn). Zusätzlich wurden die Dauern von wort- und phrasenbetonenden Lauten[5] (Bw/Bs) betrachtet. In Tabelle 3.2 sowie Abbildung 3.2 sind die Ergebnisse dargestellt. Wie nach der Sprechratenmessung zu erwarten, waren die Ergebnisse auch hier nicht einheitlich. Am deutlichsten war der Trend bei den stimmhaften Frikativen. Die Dauer steigt bei allen Sprechern bei freudiger Sprechweise, der Durchschnitt liegt bei 37 %. Hier ergab sich auch der einzige signifikante p-Wert.

Das Ergebnis für phrasenbetonende Vokale ist annähernd dasselbe wie für die phrasenbetonenden Silben, eine durchschnittliche Längung von 20 % ergab sich bei einem p-Wert von 0,0996. Dies weist darauf hin, dass die Verlängerung der Silben hauptsächlich auf dem Vokal stattfindet.

[4]Da längere Silben als stärker betont wahrgenommen werden.
[5]In aller Regel handelte es sich dabei um Vokale.

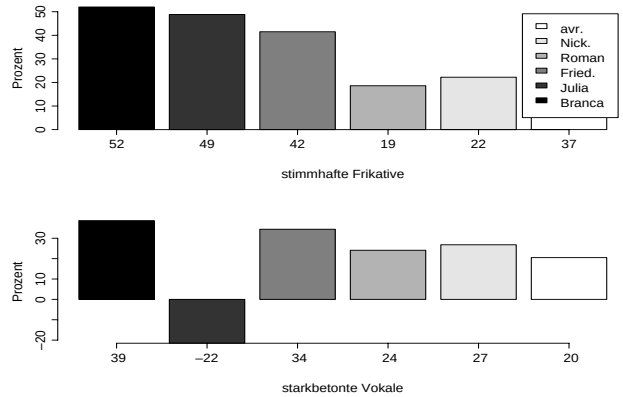

Abbildung 3.2: Lautdaueränderung für stimmhafte Frikative und phrasenbetonende Vokale

Sprecher	Psl [%]	Psh [%]	Fsl [%]	Fsh [%]	L [%]	N [%]	Vt [%]	Vn [%]	Bw [%]	Bs [%]
Branca	41	15	39	52	46	46	60	44	66	39
Julia	-15	-2	-9	19	18	-23	10	18	16	24
Fried.	9	0	5	42	13	10	4	13	-4	34
Roman	35	-30	32	22	-13	-25	24	9	52	27
Nick.	15	10	-26	49	74	-33	-22	-14	-11	-22
∅	17	-2	8	37	27	-5	15	14	24	20
p-Wert	0.14	0.8	0.5	0.00**	0.77	0.19	0.34	0.27	0.23	0.099

Tabelle 3.2: Lautdaueränderung für verschiedene Lautklassen

3.3 Grundfrequenz

3.3.1 Globale F_0-Parameter

Zunächst wurden die globalen Grundfrequenzparameter Mittelwert, Range und Standardabweichung ausgewertet. Die Werte wurden jeweils für alle Sätze eines Sprechers gemittelt. In Tabelle 3.3 sind die Abweichungen dieser Werte zwischen freudiger und neutraler Sprechweise dargestellt. Eine bildliche Darstellung ist in Abbildung 3.3 gegeben.

Wegen des logarithmischen Hörempfindens des Menschen sind die Abweichung der mittleren Grundfrequenz und der Range in Halbtönen angegeben. Zur Umrechnung eines Frequenzabstands in Halbtöne wurde die folgende Formel verwendet.

$$\text{Halbtondifferenz}(f_1, f_{bez.}) = \frac{12}{\ln(2)} \ln(\frac{f_1}{f_{bez.}})$$

Als Bezugsfrequenz wurde jeweils der Wert der neutralen Äußerung bzw. bei der Rangemessung die minimale Frequenz verwendet.

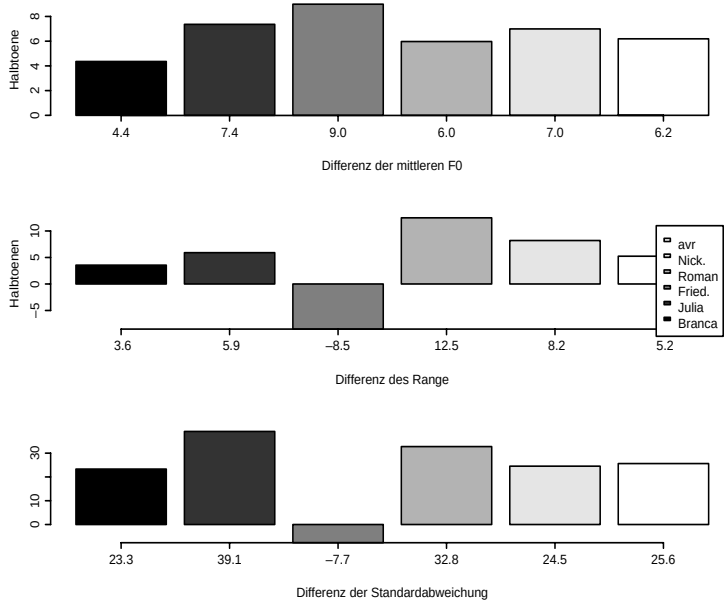

Abbildung 3.3: Abweichungen der mittleren Grundfrequenz, des Range und der Standardabweichung

Sprecher	mittl. F_0 in ST	Range in ST	Standardabweichung
Branca	4.4	3.6	23.3
Julia	7.4	5.9	39.1
Fried.	9.0	-8.5	-7.7
Roman	6.0	12.5	32.8
Nick.	7.0	8.2	24.5
∅	6.2	5.2	25.6
p-Wert	0.00**	0.075	0.00**

Tabelle 3.3: Änderung globaler F_0-Werte

Durchschnittlicher Grundfrequenzwert

Die Abweichung der mittleren F_0-Werte war erwartungsgemäß hoch signifikant. Bei allen Sprechern war die Grundfrequenz bei freudiger Sprechweise gestiegen, durchschnittlich um etwa

sechs Halbtöne. Dies deckt sich auch mit allen Ergebnissen aus der Literatur (vgl. 2.2). Es ist allerdings zu bedenken, dass die Grundfrequenz auch für andere Emotionen steigt, die einen ähnlichen Erregtheitsgrad haben, wie z.B. Ärger. Die angehobene Grundfrequenz ist zwar ein Kennzeichen freudig-erregter Sprechweise, aber vermutlich kein notwendiges Merkmal und sicherlich kein ausreichendes.

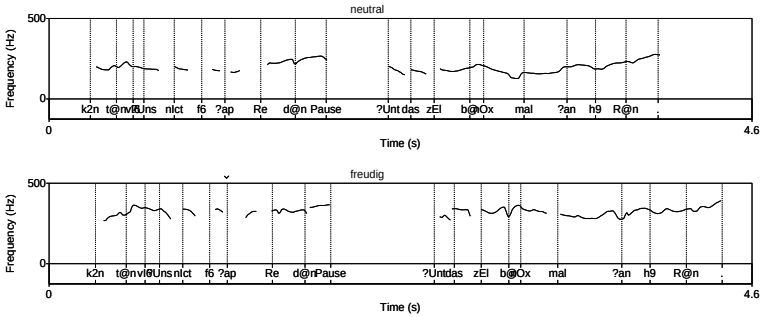

Abbildung 3.4: F_0-Konturen des Satzpaares von Sprecherin Friederike

Range

Die Abweichung des Range war nicht signifikant, was an Sprecherin Friederike lag. Ihr Wert wies eine deutliche Abweichung auf. Setzt man den Test einseitig an[6], wird auch das Merkmal Range signifikant. Eine Vergrößerung des Range gilt ebenfalls als Merkmal erregter Sprechweise, damit gelten dieselben Einschränkungen wie oben.

Da die Werte der Sprecherin Friederike deutlich von den anderen abweichen, seien die F_0-Verläufe für neutrale und freudige Sprechweise in Abbildung 3.4 dargestellt. Man sieht, dass der Range der freudigen Version tatsächlich kleiner ist, und auch nicht unbedingt mehr Bewegung der F_0-Kontur auf den Silben zu bemerken ist. Offensichtlich wird hier Freude stärker durch andere Merkmale ausgedrückt als durch die Prosodie.[7] Andererseits mag die Abweichung von den anderen Ergebnissen auch damit zusammenhängen, dass dies der einzige Fragesatz war.

Standardabweichung

Die signifikant größere Standardabweichung bei den freudigen Äußerungen kommt unter anderem daher, dass auf den Silben mehr Bewegung im F_0-Verlauf ist, wie z.B. in Abbildung 3.5 zu sehen ist. Dass dies allerdings wiederum kein notwendiges Merkmal freudiger Sprechweise ist, zeigt Sprecherin Friederike, deren geringe Standardabweichung aber natürlich auch Folge des geringen Range ist.

[6] Dies ließe sich dadurch begründen, dass in der Literatur auf eine Vergrößerung des Range hingewiesen wird.

[7] Ein ähnliches Ergebnis ist auch bei Heuft et al. (1998) beschrieben. Auch hier lag bei einem Sprecher der Range der freudigen Version 1,5 ST über der neutralen, bei einer anderen Sprecherin um 0,8 ST darunter.

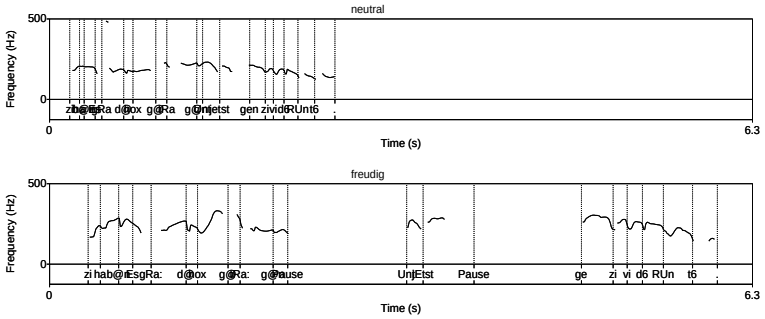

Abbildung 3.5: F_0-Kontur eines Satzpaares von Sprecherin Branca

3.3.2 F_0-Werte an betonten Silben

Um nun die Untersuchung der F_0-Kontur etwas zu differenzieren, wurden die Abweichungen des F_0-Werts auf betonten Silben zum mittleren F_0-Wert berechnet. Dafür wurde für jede betonte Silbe von den drei gemessenen Werten jeweils der maximale genommen, um die Peaks nicht durch Mittelwertbildung zu dämpfen. Zwischen diesen Werten wurde dann für jeden Sprecher über alle Betonungen gemittelt.

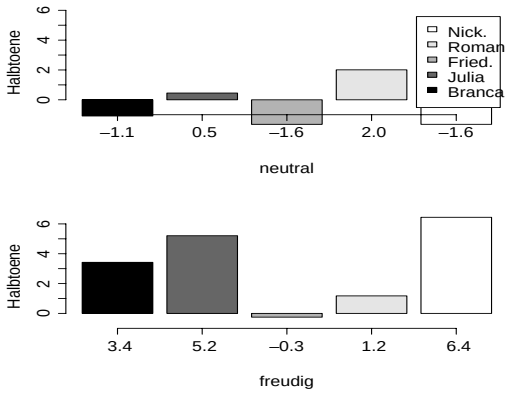

Abbildung 3.6: Abweichung der F_0 bei phrasenbetonenden Silben

Da freudige Sprechweise ja durch mehr segmentelle Bewegung gekennzeichnet sein soll (siehe 3.1), war zu erwarten, dass die Abweichungen der Maxima vom Mittelwert bei den freudigen Sätzen größer sind. Für die Wortbetonungen ist ein solcher Trend allerdings nicht vorhanden, der T-Test ergab einen p-Wert von 0,86. Lediglich Sprecher Roman zeigte eine deutliche

Tendenz, bei freudiger Sprechweise die Wortbetonungen durch stärkeres Anheben der Grundfrequenz zu realisieren.

Für die phrasenbetonenden Silben ergab sich ein signifikantes Ergebnis, der p-Wert beträgt für den beidseitigen Test 0.0126. Durchschnittlich lagen bei den freudigen Versionen die Silben 3,24 Halbtöne über der mittleren Grundfrequenz.

In Abbildung 3.6 sind für jeden Sprecher die durchschnittlichen Abweichungen in Halbtönen für neutrale und freudige Versionen dargestellt. Der Mittelwert wurde nicht angegeben, da die Sprecher in den neutralen Versionen so streuten, dass eine Interpretation nicht sinnvoll wäre.

Beachtlich ist, dass drei Sprecher in den neutralen Versionen deutlich unter dem Mittelwert liegen. In der Regel wird ja durch Anheben der Grundfrequenz betont und nicht durch Absenken. Dies mag daran liegen, dass die Festlegung einer Silbe als Satz- wie auch als Wortakzent rein willkürlich vorgenommen wurde und sicherlich von der Semantik beeinflusst war. So ist man z.B. bei dem Satz *„An den Wochenenden bin ich jetzt immer nach Hause gefahren und habe Agnes besucht“* generell geneigt, das /a/ von Agnes als betont wahrzunehmen. Wird diese Silbe aber nicht betont, wird sie wohl unter dem mittleren F_0 liegen, da sie am Ende des Satzes liegt.

3.3.3 F_0-Verlaufstypen silbischer Segmente

Vorbemerkung

Um die F_0-Analyse weiterhin zu verfeinern und Hinweise für die These zu finden, dass die F_0-Konturen auf Silbenebene sich für freudige Sprechweise signifikant von denen bei neutraler Sprechweise unterscheiden, wurden die Intonationsmuster auf den Silben analysiert. Laut Untersuchungen von Mozziconacci & Hermes (1997) sind segmentelle Intonationsmuster für unterschiedliche Emotionen charakteristisch. Die Autoren haben z.B. für Freude mehr ansteigende Konturen auf den ersten Betonungen festgestellt als bei neutraler Sprechweise.

Es wurde ein Verfahren verwendet, das an einen Algorithmus von d'Alessandro & Mertens (1995) angelehnt ist. Der Grundgedanke des Originalverfahrens wird zunächst beschrieben. Es beruht darauf, die Grundfrequenzkontur im Rahmen der Glissandoschranke zu stilisieren. Die Glissandoschranke gibt an, ab wann eine Tonhöhenänderung perzeptiv wahrnehmbar ist. Sie ergibt sich nach t'Hart et al. (1990) aus

$$G_{tr} = \frac{0,16}{T^2}$$

T ist die dabei die Dauer des Teiltons. Die Einheit der Glissandoschranke G_{tr} wird dabei in Halbtönen pro Sekunde angegeben. Damit können tonale Segmente je nachdem, ob ihre Steigung in Abhängigkeit von der Dauer die Glissandoschranke überschreitet, als statisch oder dynamisch klassifiziert werden.

Der Algorithmus erfordert zunächst eine Segmentierung des Signals auf Silbenebene. Der stimmhafte Anteil der Silbe wird auf die Anzahl von Teiltönen geprüft, indem zunächst der gesamte Intonationsverlauf mit der Glissandoschranke verglichen wird. Liegt er darunter, handelt es sich um einen statischen Ton, ansonsten wird eine Regressionsgerade zwischen Anfangs- und End-F_0-Wert gelegt und der maximal davon abweichende Wert gesucht. Liegt die Differenz unterhalb einer bestimmten Entscheidungsgrenze[8], handelt es sich bei dem Segment um einen Ton, ansonsten wird es in Teiltöne zerlegt und das Verfahren rekursiv angewendet.

[8] d'Allessandro und Mertens schlagen einen Halbton vor

Im Fall der vorliegenden Untersuchung waren die Silben bereits in zwei Teiltöne unterteilt, die durch Anfangs-, Mittel-, Endwert und Dauer festgelegt waren[9]. Die Unterteilung wurde von Hand vorgenommen mit der Intention, den Verlauf möglichst gut zu beschreiben. Damit war eine Klassifikation möglich, deren Kategorienanzahl sich durch die Kombination von zwei Teiltönen (erster und zweiter) und drei Verläufen (gerade, steigend und fallend) zu neun ergibt. Jeder Teilton wurde zunächst anhand seiner Dauer auf perzeptive Relevanz überprüft, indem er mit der Glissandoschranke verglichen wurde.

Die Umrechnung von Frequenzdifferenzen zu Halbtönen wurde wie unter 3.3.1 vorgenommen. Waren beide Teiltöne nicht perzeptiv relevant, wurde die Relevanz des gesamten Verlaufs geprüft. Die Silben wurden dann anhand der Steigungen der Teiltöne nach einer der neun möglichen Kategorien klassifiziert. Zudem wurde für jede Silbe die Steigung angegeben, falls sie nicht geraden Verlaufs war. Bei den Typen steigend/fallend und fallend/steigend wurde die Steigung angegeben, deren Absolutwert größer war. Die Häufigkeit der Konturtypen wurde für alle Sätze gezählt, die dazugehörigen Steigungen wurden für jeden Satz gemittelt.

Auf eine getrennte Analyse für wort- oder phrasenbetonende Silben wurde verzichtet, da nicht genug Daten analysiert worden waren, um statistisch abgesicherte Ergebnisse erzielen zu können[10].

Konturtypen

In Tabelle 3.4 sind Ergebnisse angegeben, die mindestens eine Tendenz zeigten. Angegeben sind die durchschnittlichen Vorkommen von Konturtypen pro Satz.

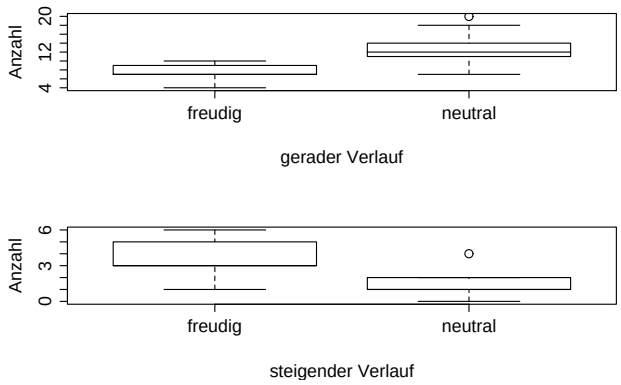

Abbildung 3.7: Boxplots für gerade und steigende Konturen

[9] Insofern gab es eine Einschränkung des oben beschriebenen Verfahrens hinsichtlich der Anzahl der maximal vorkommenden Teiltöne. Dies wurde aber als nicht so gravierend empfunden, da es ja um einen Vergleich zweier Sprechweisen ging, die beide derselben Prozedur unterzogen wurden. Die Beschränkung auf zwei Teiltöne ermöglichte eine Klassifizierung in neun Konturtypen, die mit dem Originalverfahren nicht möglich gewesen wäre.

[10] Da einige der neun neutralen Sätze einphrasig akzentuiert waren, kamen einige Konturtypen auf phrasenbetonenden Silben gar nicht oder nur einmal vor.

Es gibt hochsignifikant weniger gerade und mehr steigende Konturtypen bei freudiger Sprechweise. Tendenziell gibt es mehr fallende und steigend/fallende Typen. Es gibt signifikant mehr Silben vom Typ gerade/fallend und fallend/gerade bei den freudigen Äußerungen. Dieses Ergebnis ist allerdings dadurch zu relativieren, dass diese Typen generell sehr selten vorkommen, so kommt der Typ fallend/gerade auch bei den freudigen Sätzen im Durchschnitt nur in jedem zweiten Satz vor.

	gerade	steigend	fallend	steigend/fallend	gerade/fallend	fallend/gerade
freudig	7,6	3,5	2.8	0.44	1.1	0.55
neutral	12,8	1,4	1.8	0,11	0,22	0
p-Wert	0,0**	0,008**	0.094	0.08	0.035*	0.05*

Tabelle 3.4: Durchschn. Anzahl der Verlaufstypen pro Satz und Signifikanz

Zusammengefasst lässt sich sagen, dass es deutlich mehr Konturvariation bei freudiger Sprechweise gibt. Im Gegensatz zu den freudigen kamen bei den neutralen Äußerungen die Typen fallend/steigend und fallend/gerade gar nicht vor.

In Abbildung 3.7 sind Boxplots[11] für gerade und steigende Konturen für alle Sätze dargestellt. Die Unterschiede sind deutlich sichtbar. Während nur 25 % der neutralen Äußerungen weniger als 12 Silben mit geradem Verlauf hatten, lagen alle freudigen Äußerungen darunter. Bei den steigenden Konturen verhält es sich genau gegenläufig. 75 % der freudigen Versionen hatten mehr als drei steigende Konturen, aber nur einer der neutralen.

In Abbildung 3.8 sind die prozentualen Abweichungen der Häufigkeiten gerader und steigender Konturen für die einzelnen Sprecher dargestellt. Während es für alle Sprecher eine Abnahme gerader Konturen um durchschnittlich 40 % gibt, zeigen sich bei der Zunahme steigender Konturen sprecherabhängige Unterschiede. Es gab bei Sprecher Roman unabhängig von der Emotion gleichviel steigende Konturen und Sprecherin Friederike hatte in ihrer neutralen Äußerung gar keine steigende Kontur.

Steigungen

Die Unterschiedlichkeit der Steigungen (gemessen in Halbtönen pro Sekunde) wurde nur für steigende und fallende Konturtypen signifikant. Die Mittelwerte und die p-Werte des T-Tests sind in Tabelle 3.5 dargestellt. Im Schnitt sind die Steigungen der freudigen Äußerungen doppelt so hoch wie die der neutralen. Die korrespondierenden Boxplots sind in Abbildung 3.9 dargestellt. Wie man sieht, unterscheiden sich die Verteilungen deutlich voneinander, das untere Quartil der freudigen Werte liegt weit über dem oberen Quartil der neutralen für die Steigungen. Für die Gefälle liegen oberes Quartil der freudigen und unteres der neutralen Werte ungefähr gleich. Es gibt also nicht nur mehr Bewegung auf den freudigen Silben, sondern bei steigenden und fallenden Typen ist die Bewegung auch deutlich stärker.

3.4 Spektrale Parameter

In den folgenden beiden Abschnitten werden Parameter der Stimmqualität untersucht, da diese erwiesenermaßen ein wichtiger Träger emotionalen Ausdrucks ist (Scherer et al. (1984)).

[11] Ein Boxplot stellt auf kompakte Weise Median, unteres/oberes Quartil und Extremwerte dar

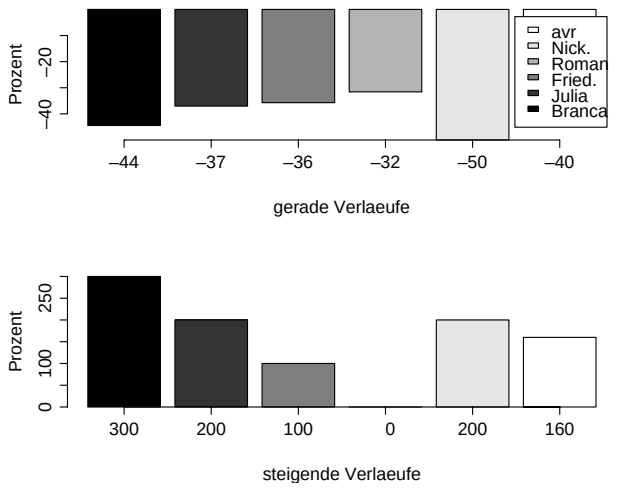

Abbildung 3.8: Abweichungen der Häufigkeiten gerader und steigender Konturen

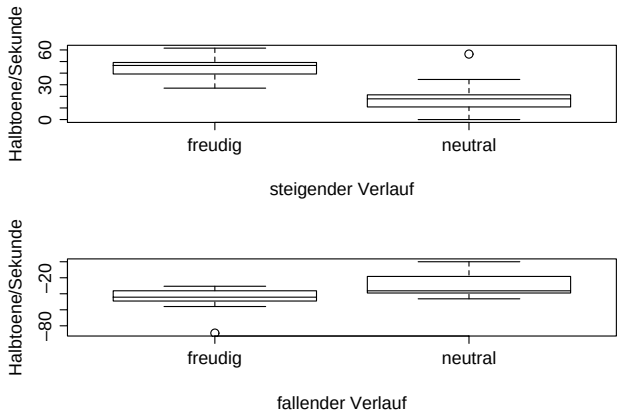

Abbildung 3.9: Boxplots der Steigungen für steigende und fallende Typen

Gerade weil in der Literatur wenig Hinweise auf den Charakter freudiger Stimmen gefunden wurden, besteht Erkenntnisbedarf. Die Untersuchungen wurden jedoch aus Zeitgründen eher stichprobenartig und vollkommen automatisiert durchgeführt. Ein gründlicheres Vorgehen muss künftigen Forschungsprojekten vorbehalten bleiben.

	steigend [ST/s]	fallend [ST/s]
freudig	43	-47
neutral	19	-26
p-Wert	0.02*	0.012*

Tabelle 3.5: Steigungen unterschiedlicher Verlaufstypen in Halbtönen pro Sekunde

3.4.1 Formanten

In der Literatur (Fant (1957, S. 19); Tartter (1980); Laver (1991, S. 219)) wird darauf hingewiesen, dass bei freudiger Sprechweise die Formanten höher liegen als bei neutraler (siehe 2.1.3). Dadurch wird das Ansatzrohr verkürzt und die Resonanzfrequenzen liegen insgesamt höher.

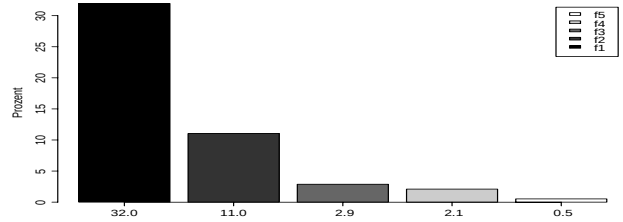

Abbildung 3.10: Änderung der Formantlagen

Um die neutralen und freudigen Versionen für die Messung vergleichbar zu machen, wurden aus allen Äußerungen 8-10 Vokale und Diphthonge gleicher Länge ausgeschnitten. Ein Liste der ausgeschnittenen Laute ist unter A.3 zu finden. Die Auswahl war hauptsächlich dadurch motiviert, bei allen Äußerungen Realisationen zu finden, die für eine Analyse lang genug waren. Die Dauern lagen zwischen 40 und 80 ms.

Die Einzellaute wurden miteinander verkettet und die Durchschnittswerte der einzelnen Formanten (F1-F5) für diese Vokalkette berechnet. Dies wurde automatisch mit dem Analyseprogramm Praat durchgeführt. Eine automatische Analyse von Formanten ist recht fehleranfällig (Scherer (1982)). Andererseits ist sie aber von Hand nur unter großem Aufwand zu bewerkstelligen. Die Mittelwerte der einzelnen Formanten lagen auf jeden Fall im Bereich der in anderen Untersuchungen (Klein (1994, S. 61)) für deutsche Sprecher gefundenen Mittelwerte. Eine getrennte Analyse der einzelnen Vokale konnte leider aus Zeitgründen nicht erfolgen.

	branca1	branca2	branca3	julia1	julia2	fried	roman1	roman2	nick
Diff F1 [%]	54.5	69.2	34.0	35.5	42.9	-0.1	14.6	16.4	36.1
Diff F2 [%]	16.1	29.1	9.5	16.4	9.0	7.5	16.8	-1.3	-0.7

Tabelle 3.6: Differenz der ersten beiden Formanten

Die Ergebnisse wurden für jeden Satz einzeln tabelliert. Es wurde nicht für jeden Sprecher gemittelt, da die Lage der Formanten natürlich von den realisierten Phonen abhängt und die

Ergebnisse damit für die einzelnen Sätzen differieren. Für die einzelnen Äußerungen wurde die Gleichheit der Realisierungen der freudigen und neutralen Versionen dabei als ausreichend angenommen.

In Abbildung 3.10 sind die durchschnittlichen Abweichungen für alle fünf Formanten freudiger gegenüber neutraler Sprechweise in Prozent dargestellt. Ein T-Test zeigte signifikante Unterschiede für die ersten beiden Formanten. Die Abweichungen für den ersten und zweiten Formanten in Prozent für alle 9 Satzpaare sind in Tabelle 3.6 dargestellt. Eine Anhebung der ersten beiden Formanten bei freudiger Sprechweise lässt sich bestätigen.

Hierbei ist allerdings zu bedenken, dass die Grundfrequenz oft über dem ersten Formanten liegt, was die überproportionale Anhebung erklärt. Interessanterweise korrelieren die Differenzen für die beiden Formanten nicht miteinander. Während beispielsweise bei Sprecherin Friederike nur der zweite Formant angehoben ist, ist beim zweiten Satz von Sprecher Roman und dem von Sprecher Nick nur der erste angehoben.

Vokal	a		e		i		ʊ		ø	
	neut.	freu.	neut.	freu.	neut.	freu.	neut.	freu.	neut.	freu.
F1 [Hz]	750	1100	400	-	400	400	300	-	400	500
F2 [Hz]	1300	1500	2100	2300	2300	2400	900	1200	1800	2000

Tabelle 3.7: Formantlagen der Vokale aus Abbildungen 3.11 und 3.12

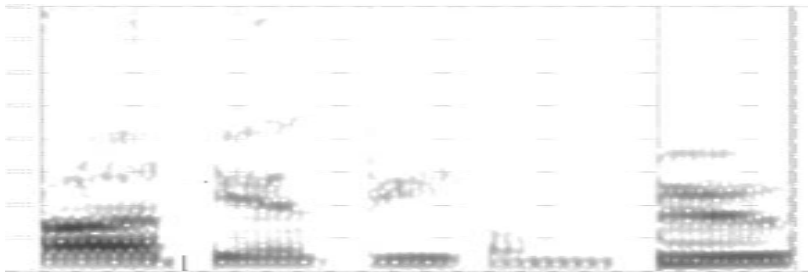

Abbildung 3.11: Wideband-Spektrogramm der Vokal a,e,i,ʊ, ø eines neutralen Satzes

In den Abbildungen 3.11 und 3.12 sind zur Veranschaulichung Sonagramme für fünf ausgeschnittenen Vokale des ersten Satzes von Sprecherin Branca dargestellt. Die geschätzten Lagen der ersten beiden Formanten sind in Tabelle 3.7 aufgeführt. Für die Vokale e und ʊ der freudigen Äußerungen waren keine erste Formanten auszumachen, da die Grundfrequenz über der geschätzten Lage lag. Dies macht auch den ersten Formanten bei i fragwürdig, der eigentlich tiefer liegen sollte. In jedem Falle liegen die Formanten bei allen freudigen Vokalen über denen der neutralen Version.

3.4.2 Energieverteilung in spektralen Bändern

Für dieselben Vokalketten wie unter 3.4.1 wurden Differenzen der Energie bestimmter Frequenzbereiche untersucht. Klasmeyer & Sendlmeier (1999) hatten eine Einteilung in drei Fre-

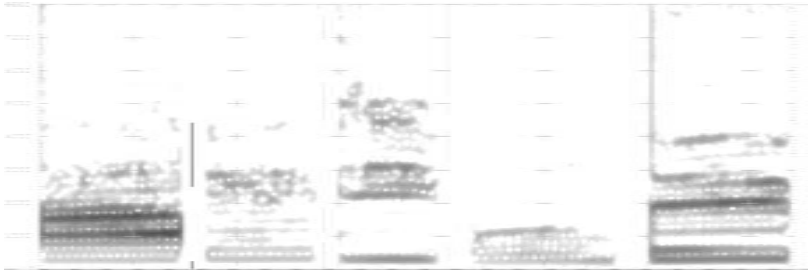

Abbildung 3.12: Wideband-Spektrogramm der Vokale a,e,i,u, ø eines freudigen Satzes

quenzbänder vorgenommen. Diese Bereiche umfassten ein Lowband von 0-1 kHz, ein Midband von 2-3 kHz und ein Highband von 4-5 kHz. Für Freude wurde dabei etwas mehr Energie im Midband und deutlich mehr im Highband gemessen.

Bei dieser Untersuchung wurde der gesamte Frequenzraum in zwei Bereiche unterteilt. Einmal wurden die Differenzen der vorhandenen Energie für eine Grenze von 1 kHz und einmal für 4 kHz bestimmt. Unter 1 kHz liegen die Grundfrequenz und die ersten ein bis zwei Formanten, unter 4 kHz die ersten für die Vokaldifferenzierung entscheidenden drei Formanten. Als untere Grenze wurden jeweils 80 Hz[12] angesetzt und als obere 8 kHz[13]. Bei der zweiten Messung lag also das untere Band zwischen 80 Hz und 4 kHz und das obere zwischen 4 und 8 kHz. Die Messung wurde ebenfalls automatisiert mit dem Analyseprogramm Praat durchgeführt.

Beide Ergebnisse, getestet über alle Äußerungen, waren hoch signifikant. Da die Werte sich für die einzelnen Sprecher ähnelten, wurden sie für jeden Sprecher gemittelt. Abbildung 3.13 a) zeigt die Differenzen der Bänder bei einer Grenze von 1 kHz der freudigen minus der neutralen Versionen für alle Sprecher. Sprecherin Julia hatte z.B. im Schnitt bei den freudigen Äußerungen 11,4 dB mehr Energie im Bereich von 1 bis 8 kHz als bei den neutralen.

In Abbildung 3.13 b) sind die Werte für die Grenze bei 4 kHz dargestellt. Hier sind die Unterschiede noch deutlicher, alle Sprecher hatten bei freudiger Sprechweise eine erheblich geringere Dämpfung der höheren Frequenzbereiche.

Sprecher	Grenze 1 kHz in dB freudig	Grenze 1 kHz in dB neutral	Grenze 4 kHz in dB freudig	Grenze 4 kHz in dB neutral
Branca	-6,7	-12,8	-26,1	-34,7
Julia	0,9	-10,5	-20,6	-28,6
Fried.	-8,3	-8,4	-26,9	-31,4
Roman	-8,1	-10,3	-26,2	-30,0
Nick	-4,1	-4,7	-27,2	-30,3
ø	-5,3	-9,3	-25,4	-31,0

Tabelle 3.8: Differenzen der Energie in verschiedenen Bändern

In Tabelle 3.8 sind die Differenzwerte für die neutralen und freudigen Äußerungen einzeln

[12] Die niedrigste vorkommende Grundfrequenz.

[13] Die halbe Samplingfrequenz und damit höchste vorkommende F_0.

dargestellt. Zum Beispiel waren im Schnitt bei den freudigen Versionen 5,6 dB ($-25,4$ dB $-$ $(-31$ dB$)$) mehr Energie im Bereich von 4-8 kHz als bei den neutralen. In den Sonagrammen

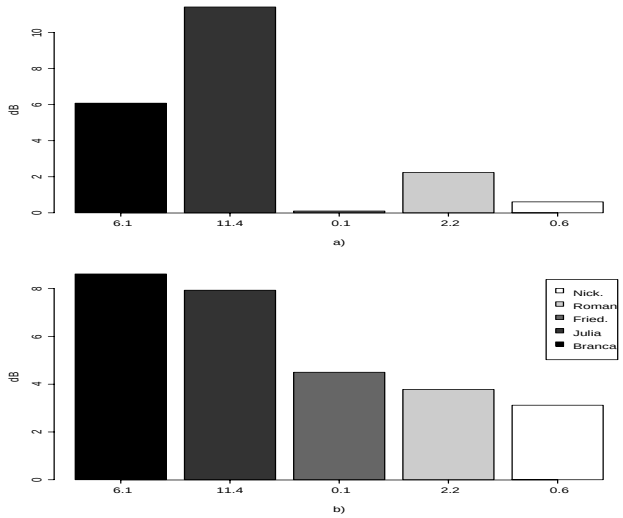

Abbildung 3.13: Differenzen der Energie zwischen verschiedenen Bändern

in Abbildungen 3.11 und 3.12 ist die stärkere Energie in den oberen Frequenzbereichen bei den freudigen Versionen gut zu sehen. Es handelt sich dabei durchaus noch um harmonische Energie. So lässt sich z.B. beim i bei 7000 Hz noch ein Formant in einer Region ausmachen, in der bei den neutralen Versionen nicht mehr genug Energie für Formantstrukturen vorhanden ist. Man kann wohl nicht mehr von modaler Phonation sprechen, deren Primärschall (Glottissignal) eine spektrale Dämpfung von -12 dB pro Oktave aufweist (Laver (1991, S. 223); Sundberg (1997, S.93)).

3.5 Zusammenfassung der Ergebnisse

Insgesamt bestätigen die Ergebnisse die Vermutung, dass nicht einzelne Parameterausprägungen für freudigen Sprechausdruck verantwortlich sind, sondern dass es auf die Kombination vieler Parameter ankommt. Es ist nochmals darauf hinzuweisen, dass die Schauspieler um möglichst deutlich zu sein eine sehr erregte Form von Freude realisiert haben. So ist das Anheben der Grundfrequenz vor allem ein Kennzeichen erregter Sprechweise, aber nicht unbedingt Voraussetzung für freudigen Sprechausdruck.

Viele Parameterausprägungen lassen sich auf die Erregung des sympathischen Nervensystems zurückführen und gelten deshalb auch für andere Emotionen (vgl. Scherer (1995)). Nicht alle Ergebnisse waren eindeutig. Viele der untersuchten Parameter waren sicherlich zu grob, um klare Trends zu erkennen, eventuell war auch die Anzahl der untersuchten Sätze zu gering.

Einige Widersprüchlichkeiten wie z.B. beim Range legen die Vermutung nahe, dass es sich hier auch um sprecherspezifische Phänomene handelt.

Andererseits legt die Tatsache, dass sich gerade die Werte des einzigen Fragesatzes hinsichtlich der Intonation von den anderen unterschieden haben, die Vermutung nahe, dass bezüglich der Intonation zwischen Frage- und Aussagesätzen unterschieden werden sollte.

Verwertbare Ergebnisse sind:

- Die Grundfrequenz steigt im Mittel um 50 %.
- Der F_0-Range ist eher größer.
- Die Äußerungen werden in mehr Phrasen unterteilt, Wortbetonungen werden zu Phrasenbetonungen.
- Phrasenbetonungen fallen stärker aus, Sie werden langsamer gesprochen und haben eine stärkere F_0-Anhebung.
- Die silbischen F_0-Verläufe sind bewegter, steigende Typen kommen häufiger vor.
- Steigungen und Gefälle silbischen F_0-Verläufe fallen bei freudiger Sprechweise stärker aus.
- Vor allem stimmhafte Frikative und starkbetonte Vokale werden verlängert.
- Die ersten beiden Formanten liegen höher.
- Die höheren Frequenzbänder enthalten deutlich mehr Energie.

Kapitel 4

Überprüfung der Relevanz ausgewählter akustischer Korrelate

In diesem Kapitel wird die Durchführung eines Hörversuchs beschrieben. Ziel war dabei, einige der bei der Analyse emotionalen Sprachmaterials gefundenen unterschiedlichen Merkmalsausprägungen auf perzeptive Relevanz zu überprüfen und weitere Hinweise für die Stärke der Ausprägung zu finden.

4.1 Vorbereitung

4.1.1 Auswahl des Satzes

Um emotionales Sprachmaterial bewerten zu lassen, ohne die Hörer durch den Inhalt der Äußerungen zu beeinflussen, bieten sich drei Wege an (Murray & Arnott (1993)): Texte mit neutralem semantischen Inhalt, sinnleere Texte[1] und durch Filterung unverständlich gemachte Texte. Beim vorliegenden Experiment wurde die erste Variante verwendet, da als Ausgangsmaterial ein Satzpaar aus der Analysedatenbank verwendet wurde. Anforderungen an den Testsatz waren:

- er soll emotional unbestimmt sein
- er soll in mehrere Phrasen unterteilt werden können
- er sollte möglichst viel vom Lautinventar des Deutschen enthalten

Mit „emotional unbestimmt“ ist hier gemeint, dass er in allen möglichen emotionalen Färbungen natürlich wirkt, also weder Sätze wie *„Das Buch steht im Regal“*, bei denen jeder emotionale Ausdruck übertrieben scheinen muss[2] noch ein Satz wie *„Bald haben wir Ferien“*, der wohl in aller Regel mit positiven Gefühlen assoziiert würde.

Es wurde das Satzpaar *„An den Wochenenden bin ich jetzt immer nach Hause gefahren und habe Agnes besucht“* von Sprecherin Julia verwendet, ein Satz, der der Einschätzung des Autors nach die oben genannten Kriterien erfüllt. Zudem war das Satzpaar bei der Analyse nicht durch Extremwerte aufgefallen.

[1] Etwa Alphabet oder Kunstwörter.

[2] Ein solcher Satz würde nach Murray & Arnott (1995) wohl als „emotional neutral“ bezeichnet.

4.1.2 Auswahl der Manipulationen

Die Auswahl der zu überprüfenden Parameter wurde durch die Ergebnisse aus Kapitel 3, durch Untersuchungen aus der Literatur und durch experimentelles Ausprobieren bestimmt. Es gab auch eine Reihe von informellen Hörtests mit insgesamt 9 Teilnehmern, durch deren Ergebnisse erste Einschätzungen darüber gewonnen wurden, welche Manipulationen sinnvoll sind.

4.2 Erstellung der Stimuli

Die Erstellung von Stimuli durch Resynthese hat gegenüber natürlichen Stimuli den Vorteil, dass die zu untersuchenden Faktoren unter weitgehender Konstanthaltung aller anderen Faktoren systematisch variiert werden können. Als Nachteile sind vor allem Artefakte zu nennen, die durch Schwächen der verwendeten Resyntheseverfahren enstehen (siehe 1.5).

Die Stimuli wurden durch Manipulationen an der freudigen und neutralen Version des Satzes gewonnen. Diese hatten eine Samplerate von 16 kHz und eine Quantisierung von 16 Bit. Diese Manipulationen waren vielfältiger Natur:

- Herausschneiden von Signalteilen zwecks Verkürzung der Dauer
- Kopieren von Signalteilen zwecks Verlängerung der Dauer
- Digitales Bandpassfiltern zwecks Verbesserung der Qualität
- LPC-Resynthese (siehe 1.5.2), um die Formantlagen zu manipulieren
- TD-PSOLA-Resynthese (siehe 1.5.1), um Dauern und Intonation stimmhafter Laute zu manipulieren

Die dabei verwendete Software ist unter A.1 beschrieben.

4.2.1 Formanten

Da die Manipulation der Formanten eine LPC-Resynthese unerlässlich machte, wurden beide Sätze zunächst einer solchen unterzogen, um eventuelle Artefakte auch in den Stimuli mit originalen Formantlagen zu erzeugen. Um die Störungen aus der LPC-Resynthese, die aus hörbarem „Zwitschern“ und einer allgemeinen „Rauhigkeit“ der Signale bestanden, zu mildern, wurden alle Stimuli bandpassgefiltert. Eine Erklärung für die Ursachen der Artefakte wird in 1.5.2 gegeben.

Das Durchlassband des Bandpassfilters hatte eine untere Grenzfrequenz von 100 Hz[3] und eine obere von 6 kHz. Als Fensterform wurde ein Kaiserfenster verwendet. Zum digitalen Filtern wurden die ESPS-Tools verwendet.

Die Formanten wurden auf zwei Arten verändert. Die Manipulationen wurden mit der Software ASL vorgenommen.

- In Version 1 wurde der erste Formant um 6 % angehoben (alle Werte im *Numerical Value Editor* von ASL mit 1.06 multipliziert), und der zweite um 4 %. Dies entspricht in etwa den Mittelwerten aus 3.4.1.

[3] Da es sich um eine Frauenstimme handelte, lag die Grundfrequenz auf jeden Fall deutlich über der unteren Grenzfrequenz.

- In der zweiten Variante wurden die Formanten 1-5 (alle von ASL angebotenen) um 5 % erhöht. Diese Manipulation wurde hinzugenommen, weil die erste sich im akustischen Eindruck nicht sehr stark bemerkbar gemacht hat. Zudem liegt die physiologische Begründung für das Anheben der Formanten in einer Verkürzung des Ansatzrohres (vgl. 2.1.3), und die betrifft das gesamte Spektrum.

  Hierbei ist allerdings zu beachten, dass der gesamte Manipulationsvorgang aufgrund von Restriktionen der verwendeten Software automatisch vorgenommen wurde und sicherlich fehlerbehaftet ist (vgl. 1.3.6). Vor allem wurden die Formanten auch für stimmhafte Frikative angehoben, bei denen man ja eigentlich nicht von Formanten sprechen kann. Es wäre unter diesem Gesichtspunkt genauer, von einer Verschiebung des gesamten Spektrums oberhalb der Grundfrequenz um etwa 5 % nach oben zu sprechen.

Die Bandbreiten wurden nicht verändert. Die zur Resynthese notwendige LPC-Analyse hatte folgende Parameter (vgl. 1.5.2):

- Sie war pitchsynchron.
- Die Koeffizienten wurden mit dem Kovarianzverfahren berechnet.
- Die Filterordnung betrug 16 (ein Richtwert, der sich aus der Samplerate geteilt durch 1000 ergibt).
- Als Analysefenster wurde ein Hammingfenster verwendet.
- Die Fensterlänge für stimmlose Abschnitte betrug 25 msek.
- Die Analyse wurde mit Preemphase durchgeführt, für die stimmhaften Abschnitte betrug sie 0.95, für die stimmlosen 0.6.

Diese Parameter haben sich nach Ausprobieren für diesen Satz als die günstigsten erwiesen. Dann wurden die Filterkoeffizienten in Formantmittenfrequenzen und -bandbreiten umgerechnet und den oben genannten Anforderungen entsprechend manipuliert.

4.2.2 Grundfrequenzverlauf

Der Grundfrequenzverlauf wurde in drei Varianten mit dem *Pitch Editor* der Sprachsoftware Praat manipuliert.

- In der ersten Variante wurde der Intonationsverlauf der freudigen Version von Hand nachgebildet. Diese Modifkation war dadurch motiviert, die Bedeutung der F_0-Kontur für freudige Eindruckswirkung durch Verwendung einer bereits im Hörversuch evaluierten Kontur zu überprüfen.

  Dazu wurden die F_0-Werte der neutralen Version jeweils silbenweise auf die Höhe der Werte der freudigen Version gebracht. Das Ergebnis wurde akustisch überprüft und gegebenenfalls nachgebessert. Klangen die einzelnen Silben von der Tonhöhe her gleich, wurden kurze Phrasen angehört und verglichen.

  Da auch die TD-PSOLA-Resynthese Artefakte bei starken F_0-Manipulationen erzeugt, wurde der gesamte Verlauf anschließend auf die ursprüngliche Stimmlage abgesenkt, so dass zwar die Kontur der freudigen Version verwendet wurde, nicht aber die tatsächlichen Werte.

- Als zweite Variante wurden in dem zugrundeliegenden Satz drei Satzbetonungen angenommen und diese Silben um 25 % angehoben. Hiermit sollten die Ergebnisse aus Kapitel 3.3.2 überprüft werden.

  Die Betonungen waren: *„An den Wochenenden bin ich jetzt immer nach Hause gefahren und habe Agnes besucht“*. Sie sind in der Abbildung durch Pfeile gekennzeichnet.

- In der dritten Variation wurden die drei Satzbetonungen um 50 % angehoben. Um unnatürlich wirkende Sprünge zu vermeiden, wurde 2-5 Silben vor und nach diesen Betonungen hinsichtlich der Grundfrequenz linear angehoben bzw. abgesenkt. Diese Variante wurde hinzugenommen, um die Stärke der Anhebung von betonten Silben überprüfen zu können.

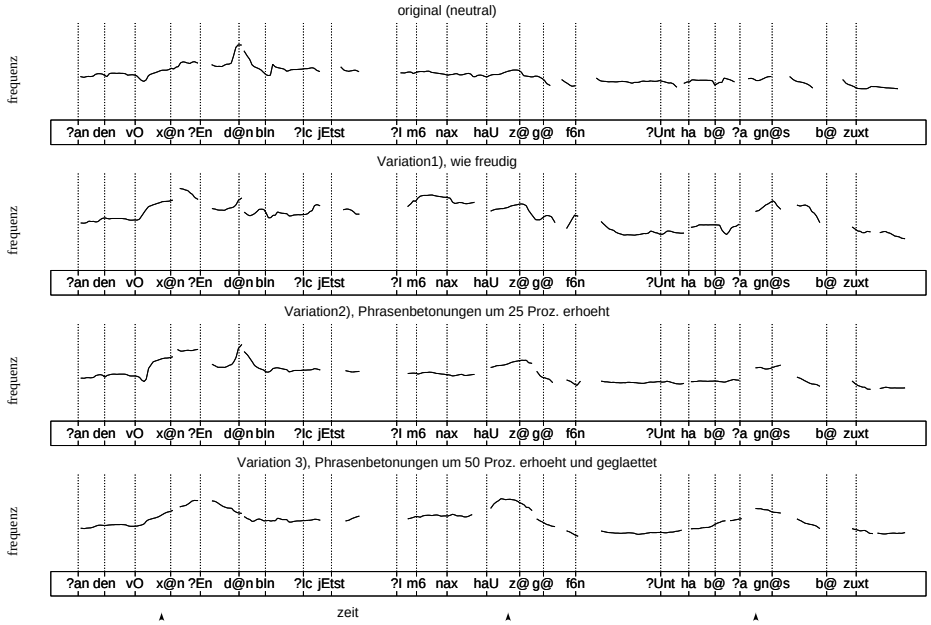

Abbildung 4.1: Die verwendeten Grundfrequenzkonturen

Die modifizierten Varianten und der neutrale Verlauf sind in Abbildung 4.1 dargestellt. Die betonten Silben bei den Variationen 2 und 3 sind dabei durch Pfeile gekennzeichnet.

4.2.3 Dauern

Die Dauern wurden auf zwei Arten verändert:

- Im einen Fall wurden die Lautdauern der freudigen Version gemessen und die Laute der neutralen Version entsprechend verlängert oder verkürzt. Diese Manipulation diente wie bei der Intonation dazu, die Prosodie der freudigen Version nachzubilden.

  Für stimmhafte Abschnitte wurde dies wieder unter dem *Pitch Editor* von Praat mit PSOLA durchgeführt, bei stimmlosen durch Ausschneiden oder Einfügen von Signalteilen.

- In der zweiten Variante wurde entsprechend den Ergebnissen aus 3.2.1 die drei phrasenbetonenden Silben um 24 % verlängert und alle anderen um 20 % verkürzt. Die informellen Voruntersuchungen hatten gezeigt, dass eine generelle Verlängerung eher als „gelangweilt“ empfunden wird.

Auf eine Verlängerung der stimmhaften Frikative, welche nach 3.2.2 ebenfalls ein Merkmal freudiger Sprechweise war, wurde verzichtet, um die Dauer des Tests nicht zu lang auszudehnen. Weiterhin musste aus diesem Grund eine Überprüfung der These, dass freudige Sprechweise durch mehr Energie in den höheren Frequenzbereichen gekennzeichnet ist, unterbleiben. Eine Veränderung dieses Merkmals hätte durch Hochpassfilterung erreicht werden können.

4.2.4 Testdesign

Insgesamt gab es also drei Varianten für die Formantstruktur, vier für die Grundfrequenz und drei für die Dauern. Hinzu kam noch eine resynthetisierte Version des freudigen Satzes, dessen Grundfrequenz und Dauern der der neutralen Version nachgebildet waren[4]. So ergab sich eine Stimulusmenge von $3 * 4 * 3 + 1 = 37$ Sätzen. Diese wurden in zufälliger Reihenfolge je zwei Mal auf ein Datband gespielt, um die Kontextabhängigkeit für die einzelnen Stimuli etwas zu dämpfen[5].

Zur Eingewöhnung der Hörer wurden als Vorlauf zwei Sätze vorangestellt, deren Bewertung nicht in der Testauswertung berücksichtigt wurde. Insgesamt waren also 76 Sätze zu bewerten. Den Sätzen wurde jeweils ihre Nummer vorangestellt[6], danach folgten 4 sek Pause. Insgesamt betrug die Dauer des Tests 15 Minuten.

4.3 Durchführung des Hörversuch

Grundsätzlich gibt es drei Arten von Tests (Murray & Arnott (1993)), die für dieses Design 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.

- **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.

[4]Diese Manipulation stellte sozusagen das Gegenexperiment zu der neutralen Version dar, deren Prosodie der der freudigen nachgebildet war.

[5]Die Bewertung für jeden Satz hängt sicherlich davon ab, wieviele und welche Sätze vorher gehört wurden

[6]Die Nummern wurden mit dem auch unter Kapitel 6 verwendeten TTS-System txt2pho/Mbrola generiert

- **Tests mit Gegensatzpaaren** bieten jeweils zwei Stimuli an, zwischen denen die Beurteiler vergleichen.

Im Prinzip wäre bei dieser Art der Fragestellung, in der lediglich der Eindruck nur einer Emotion abgefragt werden soll, ein free-response-Test am sinnvollsten, um die Beurteiler möglichst wenig durch vorgegebene Bewertungen zu beeinflussen. Die Auswertung ist jedoch äußerst problematisch, da man die beliebig abgegebenen Urteile in Kategorien fassen muss, um sie vergleichbar zu machen. Zudem ist dies ein Vorgang, der nicht objektiv durchzuführen ist. Eine Beurteilung durch Gegensatzpaare bietet sich bei der gegebenen Fragestellung nicht an, da der Test viel zu lang werden würde.

Aus diesen Gründen wurde ein forced-choice-Test gewählt. Dies entspricht auch dem gängigen Verfahren, Selbstbeobachtungen standardisiert abzufragen Scherer et al. (1984), S. 1348. Die Versuchspersonen bekamen einen Fragebogen und wurden mündlich instruiert, die Frage „*Wie freudig klingt die Sprecherin?*“ mit einem Kreuz auf einer siebenstufigen Rating-Skala zwischen „*gar nicht*“ und „*sehr*“ zu beantworten. Der Fragebogen ist unter A.4 abgebildet.

Alle Sitzungen wurden in geschlossenen Räumen in ruhiger Atmosphäre durchgeführt. Die Stimuli wurden mittels eines tragbaren DAT-Rekorders und Kopfhörern (Hersteller siehe A.1) dargeboten.

Teilnehmer waren 17 Personen im Alter zwischen 23 und 35 Jahren. Alle waren native deutsche Sprecher. Zehn davon waren weiblichen und sieben männlichen Geschlechts. Abgesehen von drei Angestellten des Instituts waren alle Laien.

4.4 Auswertung

Die Ergebnisse wurden durch eine Varianzanalyse für abhängige Gruppen ausgewertet (Diehl (1977, S. 272); Bortz (1993, S. 320); Werner (1997, S. 486)). Dieses Design ist auch unter dem Stichwort „Varianzanalyse mit kompletter Messwiederholung“ beschrieben wurden.

Die Nullhypothese war hierbei, dass die einzelnen Faktoren keine Auswirkung darauf haben, wie freudig der Satz wirkt, d.h. dass für alle Stimuli die Mittelwerte und Varianzen der Bewertungen gleich sind. Die Berechnungen wurden mit der Software SPSS (siehe A.1) durchgeführt.

Die doppelten Bewertungen der Versuchspersonen wurden dabei so gehandhabt, als wären es jeweils zwei Beurteiler gewesen. Dadurch wurde eine Dämpfung der Urteile durch Mittelwertsbildung vermieden, und der Vorteil, dass die Abhängigkeiten der Urteile vom Kontext verringert werden, blieb erhalten.

Bei der Varianzanalyse mit Messwiederholung müssen die Messwerte einer Transformation unterworfen werden, um miteinander vergleichbar zu sein. Es wurde die SPSS-Methode *deviation* gewählt, die die Abweichungen der einzelnen Werte vom Mittel der jeweiligen Versuchspersonen berechnet (Brosius (1989, S. 229)).

Alle drei Haupteffekte waren hoch signifikant (p-Wert unter 0.01). Die Interaktion zwischen F_0-Kontur und Dauern war ebenfalls signifikant (p-Wert von 0.015). Alle anderen Interaktionen sowie die Interaktion zweiter Ordnung (Interaktion aller drei Faktoren) waren nicht signifikant. Dieses war auch zu erwarten, da Dauer und F_0-Verlauf als Prosodieparameter stärker miteinander zusammenhängen als mit der spektralen Struktur.

Die Ergebnisse für die signifikanten Effekte werden im Folgenden einzeln besprochen:

- **Formantlagen** Die Mittelwerte der drei Varianten der Formantlagen sind in Abbildung

	orig.	Var. 1	Var. 2
orig.		0.04*	0.00**
Var. 1	0.04*		0.16
Var. 2	0.00**	0.16	

Tabelle 4.1: p-Werte für die Unterschiedlichkeit der einzelnen Formantmanipulationen

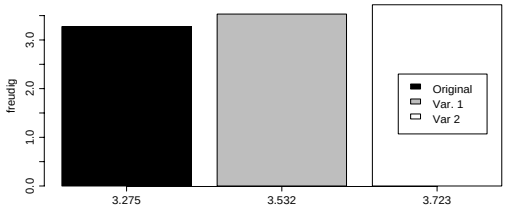

Abbildung 4.2: Mittelwerte der Formantmanipulationen

4.2 dargestellt. Aus den p-Werten der paarweisen Vergleiche (siehe Tabelle 4.1) ergibt sich, dass die beiden manipulierten Versionen signifikant besser bewertet wurden. Ein signifikanter Unterschied zwischen den Manipulationen hat sich jedoch nicht ergeben. Die Version, bei der die ersten fünf Formanten um 5 % angehoben waren, wurde allerdings tendenziell besser bewertet (bei einseitiger Fragestellung ergibt sich ein p-Wert von 0,08). Dies liegt vermutlich daran, dass freudige Sprechweise sich generell durch mehr Energie in den höheren Frequenzbereichen auszeichnet (Klasmeyer & Sendlmeier (1999)).

- **F_0-Konturen**
  Die durchschnittlichen Bewertungen für die einzelnen F_0-Modelle sind in Abbildung 4.3

	orig.	Var. 1	Var. 2	Var. 3
orig.		0,472	0,00**	0,02*
Var. 1	0,472		0,009**	0,153
Var. 2	0,00**	0,009**		0,026*
Var. 3	0,02*	0,153	0,026*	

Tabelle 4.2: p-Werte für die Unterschiedlichkeit der einzelnen F_0-Verläufe

und ihre p-Werte in Tabelle 4.2 dargestellt. Hierbei entspricht Variation 1 dem der freudigen Äußerung nachgebildeten F_0-Verlauf, Variation 2 dem Verlauf mit den um 25 % erhöhten Phrasenbetonungen und Variation 3 dem Verlauf mit um 50 % erhöhten Betonungen (vgl. 4.2.2).

Die Version mit den angehobenen Satzbetonungen (Var. 2) hat die höchste Bewertung erfahren, gefolgt von der dritten Variante. Erstaunlich ist die niedrige Bewertung der F_0-

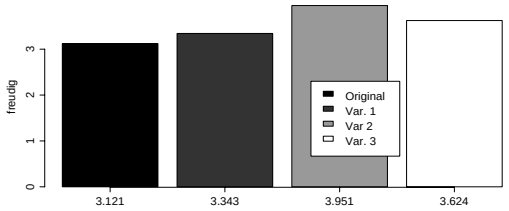

Abbildung 4.3: Mittelwerte der F_0-Konturmanipulationen

Kontur, die der freudigen Version nachgebildet wurde. Sie ist nicht einmal signifikant höher bewertet worden als der neutrale Verlauf. Dies kann mehrere Ursachen haben:

- Eventuell waren die aus der PSOLA-Manipulation entstandenen Artefakte (vgl. 1.5.1) trotz der Anpassung an das originale F_0-Niveau zu groß, als dass die Sätze noch freudig gewirkt hätten.
- Die freudige F_0-Kontur wirkt nicht isoliert freudig, sondern nur im Zusammenhang mit anderen Merkmalen, die nicht angepasst worden waren.
- Die Absenkung auf das F_0-Niveau der neutralen Äußerung hat dazu geführt, dass die Kontur nicht mehr freudig wirkt.

Die dritte Variation ist zwar nicht signifikant besser beurteilt worden als Variante 2, aber immerhin besser als der Originalverlauf.

- **Lautdauern** Abbildung 4.4 und Tabelle 4.3 zeigen die durchschnittlichen Bewertungen

	orig.	Var. 1	Var. 2
orig.		0,749	0,00**
Var. 1	0,749		0,00**
Var. 2	0,00**	0,00**	

Tabelle 4.3: p-Werte für die Unterschiedlichkeit der einzelnen Dauermodelle

der drei Dauervarianten und die p-Werte für ihre Unterschiedlichkeit. Die der freudigen Version nachempfundenen Lautdauerwerte (Variante 1) zeigen keinerlei Signifikanz hinsichtlich einer Steigerung des Eindrucks von Freude. Da durch die Veränderung der Lautdauern mittels PSOLA und Schneiden bzw. Einfügen von Signalteilen keine Störungen erzeugt wurden, läßt dies nur den Schluss zu, dass die Lautdauern der freudigen Version isoliert nicht zu einer Steigerung von freudiger Eindruckswirkung geführt haben.

Lediglich die zweite Variante, bei der die satzbetonten Vokale um 24 % gedehnt und alle anderen um 20 % verkürzt wurden, hat eine signifikant höhere Bewertung erbracht. Hier hat sich immerhin ein signifikanter Haupteffekt ergeben, der nach den doch relativ

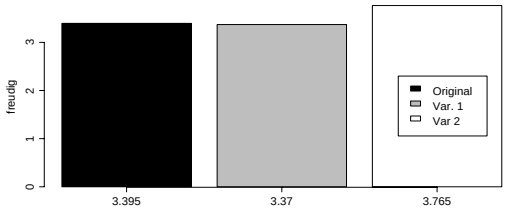

Abbildung 4.4: Mittelwerte der Dauermanipulationen

uneindeutigen Ergebnissen aus 3.2 und den widersprüchlichen Aussagen aus der Literatur nicht unbedingt zu erwarten war.

- **F_0-Konturen in Abhängigkeit von den Lautdauern**
  Die Mittelwerte für die verschiedenen F_0-Konturen in Abhängigkeit von den Lautdauern sind in Abbildung 4.5 dargestellt. Die Unterschiede wurden mittels eines T-Tests für abhängige Variablen auf Signifikanz gegeneinander getestet. Der Test ergab, dass ein Einfluss der Dauern nur für die F_0-Konturvarianten 1 und 3 existierte. Folgende Schlüsse lassen sich aus der Grafik ziehen:

  – Für die originale Grundfrequenzkontur spielt es keine Rolle, welches Dauermodell verwendet wird.

  – Die dem freudigen Satz nachempfundene Kontur wurde unter der zweiten Variation der Lautdauern signifikant höher bewertet als unter der ersten. Die Nachbildung der freudigen Intonationskontur ergab also auch unter Verwendung der Dauern der freudigen Version keine Verstärkung des Eindrucks freudiger Sprechweise.

  – Für die F_0-Kontur mit den um 25 % angehobenen Satzbetonungen spielt das Dauermodell ebenfalls keine Rolle, die Unterschiedlichkeit der Bewertungen war nicht signifikant.

  – Das Intonationsversion mit den um 50 % angehobenen Betonungen wirkt zusammen mit der zweiten Dauervariation am freudigsten.

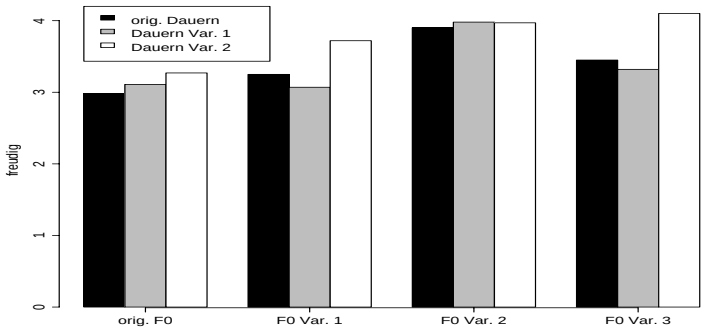

Abbildung 4.5: Mittelwerte der F_0-Konturmanipulationen in Abhängigkeit zu den Dauermodellen

Für die freudige Version, deren Dauern und Intonationskontur denen der neutralen angepasst waren, ergab ein T-Test ebenfalls einen signifikanten Unterschied der Beurteilung. Dieser Satz war als am freudigsten klingend beurteilt worden. Der durchschnittlich erreichte Freudewert betrug 4,85 gegenüber einem Gesamtmittelwert von 3,546. Dieses Ergebnis weist daraufhin, dass neben der Intonation und den Lautdauern andere Merkmale entscheidend den freudigen Ausdruck in der Äußerung erzeugen.

4.5 Zusammenfassung der Ergebnisse

Einige der durchgeführten Manipulationen haben tatsächlich zu einer Attribution der Stimuli als freudiger geführt. Dass die Bewertungen sehr dicht beieinander lagen, ist vor allem darauf zurückzuführen, dass die meisten Beurteiler die Skala nicht in ihrer vollen Breite nutzten.

Die Beurteilung des freudigen Satzes mit neutraler Prosodie als freudiger und die Beurteilung des neutralen Satzes mit freudiger Prosodie als nicht so freudig weist neben dem signifikanten Haupteffekt der Formantmanipulationen auf die Relevanz nicht-prosodischer stimmlicher und artikulatorischer Merkmale hin. Der Haupteffekt, der sich für das dritte Dauermodell ergeben hat, bestärkt die These, dass freudige Sprechweise durch stärkere Betonung gekennzeichnet ist (Trojan (1975)) Die Verlängerung der betonten und Verkürzung aller anderen Silben führte vor allem dazu, dass der Satzrhythmus prägnanter wurde.

Das relativ bessere Abschneiden der Anhebung aller Formanten gegenüber dem Anheben nur der ersten zwei bedeutet nicht unbedingt, dass Formantsynthesizer alle Formanten anheben sollten, um freudige Sprechweise zu simulieren, sondern dass dadurch tatsächlich vor allem ein Anheben der höheren Frequenzbereiche bewirkt wurde.

Weiterhin ist die eher niedrige Durchschnittsbewertung wie auch das schlechtere Abschneiden der Intonationsmodelle, die starke Eingriffe in die Originalkontur erforderten, ein Hinweis auf den störenden Einfluss der aus den Resyntheseverfahren entstandenen Artefakte. Dies ließe sich eventuell durch eine sorgfältigere Anwendung der Verfahren oder durch Verwendung von generierender Sprachsynthese zur Stimulusgewinnung verbessern.

Kapitel 5

Ein „Happiness by Rule“-System

5.1 Einleitung

Dieses Kapitel befasst sich mit einer konkreten Implementierung eines „happiness-by-rule“-Systems, das unbeschränkte Texteingabe so synthetisieren soll, dass sie als freudig wahrgenommen wird.

Ein System zur Erzeugung emotionalen Ausdrucks in synthetisierter Sprache mit unbegrenzter Texteingabe braucht als Rahmen ein Text-to-Speech System, welches wiederum eine Kombination aus mindestens zwei Subsystemen darstellt: ein Konvertierer von Text nach Lautschrift inklusive Prosodieparametern und ein System zur Erzeugung des Sprachsignals aus dem Output des ersten (vgl. 1.4). Ein weiteres Subsystem zur emotionalen Färbung müsste in beiden Generierungsprozessen Eingriffe zur Parametermanipulation vornehmen; verbale im Konvertierer, prosodische in beiden und qualitative im Signalgenerator.

Grundsätzlich sind zwei Herangehensweisen denkbar. Einerseits könnte freudige Sprache datenbasiert simuliert werden, indem als Datenbasis für die Prosodiesteuerung freudige Äußerungen verwendet werden (vgl. 1.4.2). Andererseits kann die neutrale Prosodiebeschreibung des Konvertierers durch Regeln so modifiziert werden, dass sie freudig wirkt. Dieser Ansatz wurde hier verwendet.

Um von Trägersystemen unabhängig zu sein, wurde ein Filter implementiert, der auf der Schnittstelle zwischen Konvertierer und Sprachgenerator operiert.

5.2 Aufbau des Programms

Als Programmiersprache wurde C++ (Stroustrup (1998)) verwendet, dass einerseits durch die Unterstützung von Objektorientierung eine sinnvolle Strukturierung 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.

Im Mittelpunkt des Programms steht die Klasse „Äußerung“. Diese bietet Methoden zur Analyse und Modifikation einer Liste von Lauten, die ihrerseits durch einen Namen, einen Typ[1], eine Dauer und eine F_0-Kontur-Beschreibungsliste charakterisiert sind. Sämtliche Modifikationen und Analysen beziehen sich auf die Ebene Äußerung, die eine oder mehrere Phrasen enthalten kann.

[1] Der Typ richtet sich nach der Artikulationsart.

Als Eingabe ist die phonetische Beschreibung einer Äußerung durch eine Lautliste und einer Beschreibung der Prosodie vorgesehen. Die Prosodie wird intern dadurch repräsentiert, dass jedem Laut eine Dauer und eine Liste von F_0-Targetwerten zugeordnet ist[2].

Zudem können bei Aufruf des Programms Werte für die Modifikation der unten beschriebenen Parameter angegeben werden. Um die Bedienung zu erleichtern, wurde ein in der Skriptsprache TCL/TK programmiertes interaktives Frontend entwickelt, das die Eingabe von Testsätzen und die Belegung der Parameterwerte mittels Slidern und Buttons ermöglicht. In Abbildung A.8 ist ein Screenshot davon dargestellt. Hier ist auch eine Manipulation der Phonemliste von Hand zu Testzwecken möglich.

Das Programm erhielt den Arbeitsnamen *emofilt* (ein Akronym für „Emotions-Filter“), da es ja lediglich der Manipulation prosodischer Parameter diente. Welche Emotion dadurch simuliert werden soll, wird allein durch die anzugebenden Modifikationswerte festgelegt.

5.3 Manipulierbare Parameter

Die Auswahl der manipulierbaren Parameter war einerseits durch die unter Kapitel 3 und 4 gewonnenen Erkenntnisse bestimmt, andererseits durch die Literatur (siehe 2.1.4). Grundsätzlich wurden nur prosodische Parameter implementiert. Die gegenwärtigen TTS-Systeme lassen in der Regel eine Variation der Stimmqualität innerhalb einer Äußerung nicht zu.

Allgemein wurde Wert darauf gelegt, möglichst viele für den emotionalen Ausdruck wichtige Parameter zu berücksichtigen, auch wenn sie keinen Effekt für Freude hatten, um das System später auf weitere Emotionen erweitern zu können.

Eine Manipulation einzelner Parameter kann natürlich zu Seiteneffekten für andere führen, so erhöht z.B. eine Verkürzung der stimmhaften Frikative die Sprechrate. Das Programm arbeitet die Modifikationen in der Reihenfolge „vom speziellsten zum allgemeinsten“ ab.

5.3.1 Dauerparameter

Sprechrate Eine Änderung der Sprechrate kann implizit durch Änderung der Lautdauern erreicht werden. Die Lautdauermodifkation (Pausen werden nicht modifiziert) lässt sich in Prozent angegeben. Werte unter 100 entsprechen dabei einer Verkürzung der Laute und Werte über 100 einer Verlängerung der Dauern.

Lautdauern Die Dauern der Laute lässt sich einerseits für alle und andererseits für einzelne Lautklassen getrennt 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 Dauern wie Laute gehandhabt, ein Einfügen von Pausen ist allerdings nicht implementiert, da dies eine syntaktische Analyse der Äußerung erfordert hätte.

[2] Diese Repräsentation entspricht dem Input des zur Evaluierung verwendeten Sprachgenerierungssystems (MBROLA (1999)).

5.3.2 F_0-Parameter

Die folgenden Manipulationen beziehen sich auf F_0-Targetwerte. Für alle Manipulationen gelten die Grenzwerte von 80 bzw. 500 Hz. Würden einzelne F_0-Werte aufgrund einer Manipulation außerhalb dieses Bereichs verlegt, werden sie auf die Extremwerte gesetzt.

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

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 liegen, prozentual abgesenkt oder angehoben. 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 Mittelwert.[3]

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.

Konturverläufe auf Silbenebene Vorgesehen sind drei Betonungsstufen auf Silbenebene: unbetont, wortbetonend und phrasenbetonend. Lage und Anzahl der Betonungen werden dem Programm entweder als Input gegeben oder durch eine Analyse der relativen Tonhöhen und Dauern der Silben vom Programm geschätzt.

Für die phrasenbetonenden Silben oder alle betonten Silben ist einer der drei Konturtypen steigend, fallend und gerade zu wählen. Zu fallenden oder steigenden Typen ist zusätzlich die Steigung in Halbtönen pro Sekunde anzugeben. Soll eine steigende Kontur modelliert werden, 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 Konturen verfahren.

Jitter Jitter wird durch Schwankungen der F_0-Werte simuliert. Hierbei werden alle Werte eines stimmhaften Lautes um einen anzugebenden Prozentsatz abwechselnd angehoben und abgesenkt

Phrasenintonationsverlauf Die Intonationsverläufe einzelner Phrasen der Äußerung können den Konturtypen steigend, fallend und gerade zugeordnet werden. Dafür werden die phrasenbetonenden Silben jeweils als Zielwerte angenommen. 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.

[3] Dies entspricht einer Kontrastveränderung bei der digitalen Bildverarbeitung.

5.3.3 Stimmqualitätsparameter

Die folgenden Manipulationen wurden nicht implementiert. Das zum Testen verwendete Rahmensystem (siehe 6.1) stellt leider derartige Einflussmöglichkeiten nicht zur Verfügung. Eine Erweiterung in dieser Hinsicht ist auch nicht geplant[4], obwohl dies, da es sich um eine halbparametrische Kodierung handelt, möglich wäre. Trotzdem sollen an dieser Stelle für emotionalen Ausdruck wichtige Manipulationsmöglichkeiten besprochen werden.

Formanten Sinnvoll wäre eine Möglichkeit zur Verschiebung der ersten beiden Formanten, eventuell für verschiedene Vokalklassen getrennt.

Energieverteilung in Frequenzbändern Eine Manipulation der Energieverteilung bestimmter Frequenzbänder könnte durch eine Angabe der spektralen Dämpfung beschrieben werden (in dB/Oktave).

Intensität Eine Manipulation der Lautheit einzelner Laute könnte die Unterschiede zwischen betonten und unbetonten Silben stärker hervorheben. Vorstellbar wäre etwa ein Wert, der den Lautheitsunterschied zwischen Silben verschiedener Betonungsstufen beschreibt.

[4] Auskunft vom Autor, Thierry Dutoit, per E-Mail

Kapitel 6

Hörversuch zur Evaluierung des „Happiness by Rule“-Systems

6.1 Auswahl der Trägersysteme

Für Sprachgeneratoren gibt es grundsätzlich zwei Ansätze: generierende und konkatenierende Systeme (vgl. 1.4). Nun bieten zwar generierende Systeme (wie z.B. DEC-Talk) Zugriff auf praktisch alle Stimm- wie Sprachparameter, jedoch reicht ihre Natürlichkeit noch nicht an die der konkatenierenden Synthese heran. Zudem stand kein deutschsprachiges System als Trägersystem zur Verfügung.

Als Trägersysteme wurde das frei erhältliche Diphonsystem MBROLA (1999)[1] verwendet, für das ein deutscher Konvertierer (TXT2PHO (1999)) und sowohl eine männliche als auch eine weibliche Stimme existieren. Mbrola ist ein Diphonsynthesizer, der nach dem in Dutoit & Leich (1993 a) beschriebenen Multiband-Verfahren arbeitet (siehe 1.4).

Die Schnittstelle ist das Eingabeformat von Mbrola und besteht aus einer Phonemliste. Den Phonemen ist jeweils ein Wert für die Lautdauer und eine Liste von Tupels für Dauer (in Prozent) und Tonhöhe eines Lautabschnitts zugeordnet. Ausgabeformat des Konvertierers ist eben diese Liste.

Die Prosodiesteuerung des Konvertierers arbeitet mit einem Entscheidungsbaumverfahren ähnlich dem, das unter 1.4.2 beschrieben ist.

6.2 Festlegung der Parameterwerte

Als Ansatz für die Belegung der Werte wurden die Ergebnisse aus den Kapiteln 3.5 und 4 verwendet. Um mit der Feineinstellung einzelner Parameterwerte einfacher experimentieren zu können, wurde ein graphisches Hilfsprogramm als Frontend für das Hauptprogramm entwickelt. Es wurde in der Skriptsprache TCL/TK (Welch (1996)) implementiert (siehe Abbildung A.8). Die modifizierten Parameter waren:

- Die mittlere Grundfrequenz wurde um 40 % erhöht.

  Dass freudige Sprechweise eine erhöhte Grundfrequenz aufweist, haben sowohl alle Untersuchungen aus der Literatur als auch die Messungen aus 3.3.1 ergeben.

[1] MBROLA ist ein Akronym für *multiband resynthesis overlap-add*

- Der Rangeparameter wurde auf 140 gesetzt, wodurch alle Werte um 40 % ihrer Distanz vom globalen Durchschnittswert verschoben wurden[2].

  Auch diese Veränderung ergab sich sowohl aus der Literatur wie auch aus den Messungen in 3.3.1. Es hatte sich zwar bei den Messungen keine Signifikanz für den Parameter Range ergeben, aber die meisten Sprecher hatten doch einen deutlich größeren Range.

- Der Parameter Variabilität wurde auf 200 gesetzt. Damit wurde für alle Werte eine Verdoppelung des Abstands vom Silbendurchschnittswert erreicht. Da die Variabilität der neutralen Prosodie nicht besonders hoch war, war diese Veränderung nicht so gravierend, wie sie vielleicht erscheinen mag.

  Diese Manipulation wurde durch die Ergebnisse aus 3.3.1 motiviert, bei denen die freudigen Sätze eine höhere Standardabweichung der F_0-Werte aufwiesen.

- Die F_0-Werte der phrasenbetonenden Silben wurden um 10 % angehoben.

  Dass bei freudiger Sprechweise stärker betont wird, hatte sich in 3.3 ergeben und in 4.4 (beim zweiten F_0-Kontur-Modell) bestätigt. Die Anhebung erfolgte nicht um die in 4.2.2 verwendeten 25 %, da ja durch die Konturveränderung (siehe nächster Punkt) ebenfalls eine Anhebung realisiert wurde.

- Den phrasenbetonenden Silben wurde eine linear steigende F_0-Kontur zugewiesen. Die Steigung betrug 20 ST/sek.

  Dies war dadurch motiviert, dass laut der Untersuchung in 3.3.3 bei freudiger Sprechweise mehr steigende Konturverläufe auf Silbenebene auftreten. Um eine Begrenzung auf wenige Silben zu haben, wurden die stark betonten gewählt. Die Steigung um 20 ST/sek erschien nach Experimentieren mit verschiedenen Werten als am günstigsten.

- Die Dauer der phrasenbetonenden Silben wurde um 20 % verlängert.

  Dieses Modell hatte sich zusammen mit dem nächsten Punkt in Kapitel 4.4 für freudige Eindruckswirkung als günstig erwiesen.

- Die Dauer aller nicht phrasenbetonenden Silben wurde um 20 % verkürzt.

- Die Dauer aller stimmhaften Frikative wurde um 30 % verlängert.

  Dass die stimmhaften Frikative bei freudiger Sprechweise signifikant länger dauern, ergab sich sowohl bei den Messungen aus 3.2.2 als auch bei Untersuchungen von Kienast (1998).

Um die Änderungen zu veranschaulichen, ist in Abbildung 6.1 der Intonationsverlauf der originalen Version und der modifizierten des zweiten Satzes dargestellt. Die Anhebungen auf den betonten Silben sind deutlich zu sehen.

[2] Dies entspricht nicht einer Vergrößerung des Range um 40 %, da ja nicht um 40 % des Wertes selber sondern der Distanz zum Mittelwert verschoben wurde

6.3 Erzeugung der Stimuli

Um den Output des Programms zu evaluieren, wurden zehn Sätze aus Sendlmeier et al. (1998) synthetisiert. Eine Liste der Sätze ist unter A.6 zu finden. Die Sätze waren ursprünglich alle zweiphrasig. Um den Test aber um eine Variationsebene zu erweitern, wurden fünf der Sätze auf eine Phrase gekürzt. Zwei der kurzen und einer der langen Sätze waren Fragesätze.

Da es bei diesem Test nur einen Faktor mit zwei Varianten gab, entstand das Problem, dass gutmütige Testteilnehmer alle Sätze, die nicht neutral klingen, als freudig beurteilen könnten. Daher wurden zusätzlich zu einer neutralen und einer freudigen Version jedes Satzes eine Distraktorversion generiert. Bei dieser wurden die Parameter konträr zur Belegung der freudigen gewählt. Die Grundfrequenz wurde abgesenkt, alle Lautdauern wurden um 20 % verlängert und der Range um 40 % verringert. Zusätzlich wurden die Pausendauern verlängert und ein Jitterfaktor von 1 % verwendet.

Da das Ergebnis dieser Manipulationen ausgesprochen traurig klang, werden diese Versionen künftig als „traurig“ bezeichnet. Insgesamt wurden also 30 Stimuli erzeugt: je zehn neutrale, traurige und freudige Versionen.

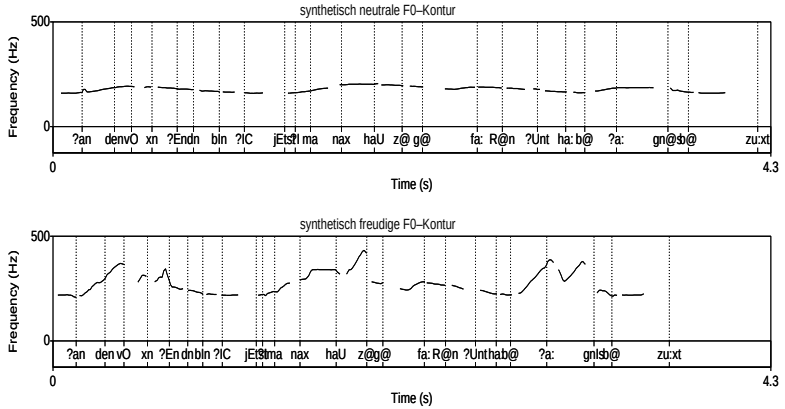

Abbildung 6.1: Neutrale und freudige F0-Kontur von Satz zwei

6.4 Versuchsdurchführung

Es kamen wiederum die unter Kapitel 4.3 erläuterten drei Testdesigns in Frage. Aus denselben Gründen wie dort wurde ein forced-choice-Test gewählt.

Da es nur zwei Varianten gab, wurde auch die Antwortmöglichkeit binär ausgelegt. Die Versuchspersonen hatten die Aufgabe, für jeden Stimulus die Frage „Wie klingt die Sprecherin?“ mit „eher freudig“ oder „eher nicht freudig“ zu beantworten. Die Frage wurde vorsichtig formuliert, da die Stimuli nach Ansicht des Autors fast alle nicht sehr freudig klangen. Der Fragebogen, der zu diesem Zweck entwickelt wurde, ist unter A.5 abgebildet.

Es nahmen 27 Versuchspersonen im Alter von 23 bis 43 Jahren teil. Das Durchschnittsalter betrug 29 Jahre. 17 waren männlichen und 10 weiblichen Geschlechts. Alle waren deutsche Muttersprachler. Die Versuche wurden in Einzelsitzungen mittels DAT-Rekorder und Kopfhörer in ruhiger Atmosphäre durchgeführt. Die verwendeten Geräte sind unter A.1 aufgeführt.

6.5 Versuchsauswertung

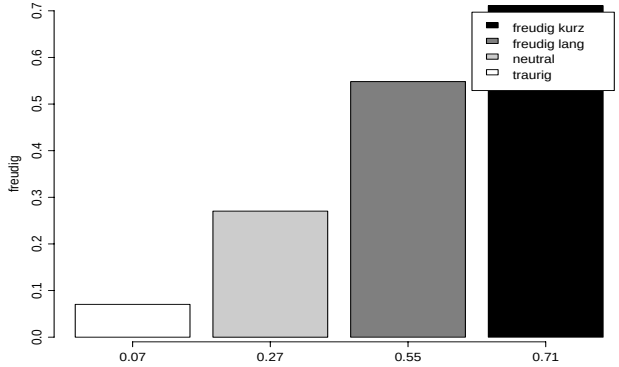

Abbildung 6.2: Durchschnittliche Bewertung der Emotionsmodelle als freudig

Die drei Gruppen wurden zunächst paarweise einem T-Test für abhängige Gruppen unterzogen. Dann wurden die langen und kurzen freudigen Sätze auf signifikante Unterschiedlichkeit überprüft. Alle Ergebnisse waren hochsignifikant. In Abbildung 6.2 sind die Mittelwerte für die Gruppen traurige Sätze, neutrale Sätze, lange freudige und kurze freudige Sätze graphisch dargestellt.

Dass die kurzen Sätze als freudiger beurteilt wurden, mag daran liegen, dass der durch das Programm erzeugte Prosodieverlauf durch die starken Betonungen doch recht unnatürlich wirkte. Bei den kurzen Sätzen kam dieser Effekt nicht so zum Tragen, da es jeweils nur eine Phrasenbetonung gab.

Dass einige der neutralen Sätze als freudig beurteilt wurden[3], liegt sicherlich am Vergleich, der unwillkürlich zu den traurigen Versionen gezogen wurde. Anscheinend war die vorsichtige Fragestellung völlig unnötig, die wenigsten Versuchshörer urteilten so kritisch wie der Autor.

Eine Darstellung der durchschnittlichen Beurteilungen für die freudigen Sätze ist in Abbildung 6.4 gegeben. Die Reihenfolge der Sätze ist die gleiche wie in A.6, die ersten fünf sind also die langen und die zweiten fünf die kurzen Sätze.

Nicht alle Sätze liegen über dem Zufallsniveau von 0,5. Zum Teil waren die Sätze wohl auch nicht semantisch unbestimmt (vgl. 4.1.1). So äußerten die Beurteiler etwa, dass Satz sieben[4]

[3] Vier der neutralen Sätze lagen in der Beurteilung über dem niedrigsten freudigen Satz

[4] *„Legen Sie mir die Unterlagen hin.“*

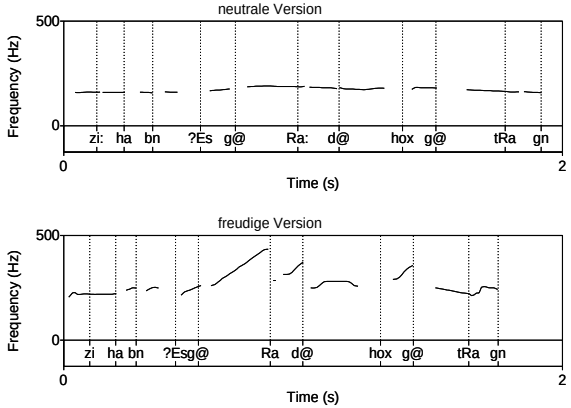

Abbildung 6.3: Neutrale und freudige F0-Kontur von Satz sechs

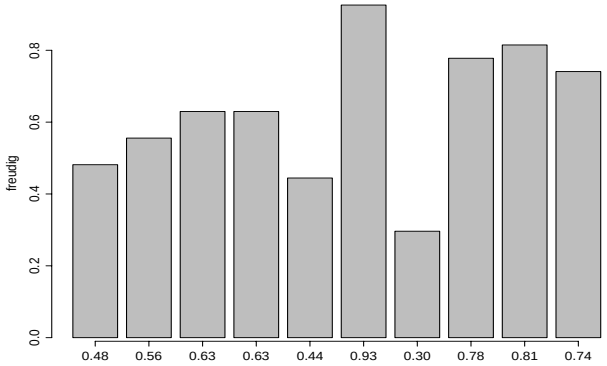

Abbildung 6.4: Durchschnittliche Bewertung der freudigen Sätze

schon vom Sinn her nicht freudig wirke. Bei Satz neun[5] hingegen wurde die Bedeutung eher mit einem freudigen Ereignis assoziiert. Satz sieben wurde von 30 % der Beurteiler als freudig empfunden, Satz neun von 81 %.

Die höchste Bewertung hat Satz sechs[6] erreicht, dessen F_0-Konturen in der neutralen und freudigen Version in Abbildung 6.3 dargestellt sind.

[5] *„Wo ist die verbogene Gabel?“*
[6] *„Sie haben es gerade hochgetragen.“*

6.6 Zusammenfassung der Ergebnisse

Es ist offensichtlich gelungen, den Output eines TTS-Systems so zu manipulieren, dass er als freudiger attribuiert wurde. Welche der durchgeführten Manipulationen dabei welchen Einfluss hatten, wäre in Folgeuntersuchungen zu klären.

Dass nicht alle Sätze von den Beurteilern als freudig beurteilt wurden, weist allerdings daraufhin, das der Output oft noch zu unnatürlich wirkt. Sinn eines TTS-Systems ist es ja gerade, dass Regeln auf jeden Input erfolgreich angewendet werden können und nicht für spezielle Eingaben optimiert sind. Dieses eher schlechte Ergebnis lässt sich sicherlich auch darauf zurückführen, dass lediglich Parameter der Prosodie berücksichtigt wurden.

Ob ein solches Verfahren tatsächlich erkennbare Emotion simulieren kann, bleibt fraglich. Dazu sollte ein Versuch mit vier bis sieben Emotionen durchgeführt werden. Die Schwierigkeit der Erkennung intendierter Emotion steigt ja mit der Anzahl der Emotionen, und bei nur einer Emotion ist die Beweiskraft nicht besonders hoch.

Basisemotionen zum Gegenstand hat.

Kapitel 7

Ausblick

In dieser Arbeit wurden viele Aspekte der Simulation freudigen Sprecherausdrucks mittels Sprachsyntheseverfahren besprochen.

Zunächst wurde von früheren Versuchen berichtet, akustische Korrelate freudiger Sprache mittels Resynthese aufzuzeigen. Ein eigenes solches Experiment wurde im Kapitel 4 beschrieben. Dabei konnten eindeutig für die Attribution von Freude günstige Parameterausprägungen nachgewiesen werden. Als wichtigste Erkenntnis erscheint hier die Verlängerung betonter Silben und das Verkürzen aller anderen, wodurch die Sprache belebter und akzentuierter wirkt.

Eine Analyse emotionaler Sprachdatenbanken wurde im Kapitel 3 durchgeführt. Dabei wurden viele Unterschiede zu neutraler Sprechweise gefunden, die sprecherübergreifende Gültigkeit haben. Als eine der wichtigsten Erkenntnisse einer solchen Analyse bleibt dennoch die sprecherspezifische Natur emotionalen Ausdrucks, die sich durch die vielen Ausreißer in den Daten ausdrückt.

Ganz eindeutig stellte sich auch die Prominenz von Parametern der Stimmqualität gegenüber solchen der Prosodie dar. Hier würden sich sicherlich genauere Untersuchungen lohnen, bei denen etwa Formanten nach Lauten getrennt manipuliert werden oder durch inverse Filterung die Wirkung verschiedener Phonationsarten untersucht wird.

Davon sind Versuche abzugrenzen, emotionale Sprechweise regelbasiert durch TTS-Systeme zu simulieren. Von zwei solchen Untersuchungen wurde berichtet und eine eigene ist in den Kapiteln 5 und 6 beschrieben. Das Problem besteht hierbei hauptsächlich darin, dass heutige TTS-Systeme bezüglich der Stimmqualität noch zu unflexibel sind. Die Stimmqualität ist für emotionalen Ausdruck immanent wichtig und sollte als dynamische Parameterklasse ebenso zur Verfügung stehen wie Intonation und Dauern.

Ein weiteres Problem, das weder in der Literatur noch in dieser Untersuchung erörtert wurde, besteht darin, dass solche doch recht einfachen Regeln bei längeren Textpassagen zu ermüdender Wiederholung immer gleicher Intonationsverläufe führen. Um hier Abhilfe zu schaffen, wären zumindest Variationsregeln nötig, die die bei natürlicher Sprache gegebene Variabilität gewährleisten.

Insofern sind solche Systeme bislang doch eher als Hilfsmittel für Syntheseexperimente zu sehen als zur Verbesserung von Sprachprothesen. Wege zur schnelleren Realisierung natürlich wirkender emotionaler Prosodie mittels datenbasierter Lernverfahren sind unter 1.4.2 dargestellt worden.

Ein Aspekt, der leider in dieser Arbeit nicht berücksichtigt werden konnte, ist der Vergleich unterschiedlicher Emotionen. Inwieweit die gefundenen Korrelate tatsächlich spezifisch für Freude gelten, kann erst eine Untersuchung zeigen, die mindestens vier der wichtigsten

Literaturverzeichnis

R. Benzmüller & W. J. Barry. Mikrosegmentsynthese - Ökonomische Prinzipien bei der Konkatenation Subphonemischer Spracheinheiten. In D. Mehnert, Hrsg., *Elektronische Sprachsignalverarbeitung*, 13, Seite 85–93. Humbold-Universität Berlin, 1996. Studientexte zue 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.

J. Bortz. *Statistik für Sozialwissenschaftler*. Springer Verlag, 4. Auflage, 1993.

G. Brosius. *SPSS, PC+ Advanced Statistics and Tables*. Mc-Graw-Hill, 1989.

J. E. Cahn. The Affect Editor. *Journal of the American Voice I/O Society*, 8:1–19, 1989.

R. Carlson, G. Granström & L. Nord. Experiments With Emotive Speech, Acted Utterances And Syntesized Replicas. *Speech Communication*, 2:347–355, 1992.

M. J. Charpentier & M. G. Stella. Diphone Synthesis using an Overlap-Add Technique for Speech Waveform Concatenation. *ICASSP*, Seite 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.

J. M. Diehl. *Varianzanalyse*. Band 3 der Reihe „Methoden in der Psychologie. Fachbuchhandlung für Psychologie, 1977.

T. Dutoit & B. Gosselin. On the Use of a Hybrid Harmonic/Stochastic Model for TTS Synthesis-by-Concatenation. *Speech Communication*, 19:119–134, 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 & H. Leich. An Analysis of the Performance of the MBE-Model when used in the Context of a TTS-System. *Proc. Eurospeech*, Seite 531–534, 1993 b.

T. Dutoit & H. Leich. High-Quality Text-To-Speech Synthesis: A Comparison of Four Candidate Algorithms. *Proc. ICASSP*, 1:565–568, 1994.

G. Fant. Modern Instruments and Methods for Acoustic Studies of Speech. *Proc. of the 8th International Congress of Linguists*, Seite 282–388, 1957.

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, Seite 51–79. Royal Swedish Academy of Music, 1981.

C. S. Fordyce & M. Ostendorf. Prosody Prediction for Speech Synthesis using Transformational Rule-Based Learning. *Proc. ICSLP*, 1998.

B. Granström. The Use Of Speech Synthesis In Exploring Different Speaking Styles. *Proc. ICSLP*, 1:671–674, 1992.

B. Heuft, T. Portele & M. Rath. Emotions in Time Domain Synthesis. *Proc. ICSLP*, 1998.

I. Karlsson. Modelling Voice Variations in Female Speech Synthesis. *Speech Communication*, 11:491–495, 1992.

KAY. Instruction Manual von ASL, 1992.

M. Kienast. Segmentelle Reduktion bei Emotionaler Sprechweise. Magisterarbeit, Institut für Kommunikationswissenschaft der TU-Berlin, 1998.

G. Klasmeyer & W. F. Sendlmeier. Voice and Emotional States. In R. Kent & M. Ball, Hrsg., *The Handbook of Voice Quality Measurement*. Singular Publishibng Group, San Diego, 1999. im Druck.

G. Klein. Zur Sprecherunabhängigkeit eines Visualisierungssystems für gesprochene Sprache. Magisterarbeit, FB Elektrotechnik der TU-Berlin, 1994.

D. R. Ladd. Phonological Features of Intonational Peaks. *Language*, 95(3):721–759, 1983.

J. Laver. *The Gift Of Speech*. Edinburgh University Press, 1991.

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*, Seite 791–831. Dekker, New York, 1992.

F. Malfrère & T. Dutoit. Speech Synthesis for TTS-Alignment and Prosodic Feature Extraction. *Proc. ICSLP*, 1998.

F. Malfrère, T. Dutoit & P. Mertens. Automatic Prosody Generation Using Suprasegmental Unit Selection. *Proc. of the 3. ESCA/IEEE Workshop in Speech Synthesis*, 1999.

MBROLA, 1999. http://tcts.fpms.ac.be/synthesis.

K. Morton. Adding Emotion To Synthetic Speech Dialogue Systems. *Proc. ICSLP*, Seite 675–678, 1992.

E. Moulines & F. Charpentier. Pitch-Synchronous WaveformProcessing 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. *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 & E. A. Rohwer. Emotional Stress in Synthetic Speech: Progress and Future Directions. *Speech Communication*, 20:85–91, 1996.

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*, Seite 79–90, 1986.

T. Portele. *Ein Phonetisch-Akustisch Motiviertes Inventar zur Sprachsynthese Deutscher Äußerungen*. Max Niemeyer Verlag, Tübingen, 1996.

K. R. Scherer & U. Scherer. Speech Behaviour and Personality. In J. K. Darby, Hrsg., *Speech Evaluation in Psychiatry*, Seite 117–135. Grune & Stratton, 1981.

K. R. Scherer, D. R. Ladd & K. E. A. Silverman. Vocal Cues to Speaker Affect: Testing Two Models. *JASA*, 76:1346–1356, 1984.

K. R. Scherer. Speech and Emotional States. In J. K. Darby, Hrsg., *Speech Evaluation in Psychiatry*, Seite 189–218. Grune & Stratton, 1981.

K. R. Scherer. Akustische Parameter der Vokalen Kommunikation. In K. R. Scherer, Hrsg., *Vokale Kommunikation*, Seite 122–137. Beltzverlag, 1982.

K. R. Scherer. Vocal Correlates of Emotionaql Arousal and Affective Disturbance. In H. Wagner & A. Manstead, Hrsg., *Handbook of Social Psychophysiology*, Seite 165–197. John Wiley & Sons Ltd., 1989.

K. R. Scherer. How Emotion is Expressed in Speech and Singing. *Proc. ICPhS*, 3:90–96, 1995.

H. Schlosberg. Three Dimensions of Emotions. *Psychological Review*, 61(2):81–88, 1954.

W. F. Sendlmeier, M. Kienast & A. Paeschke. Phonetische Reduktion und Elaboration bei Emotionaler Sprechweise, 1998. DFG-Projekt am Institut für Kommunikationswissenschaft der TU-Berlin, Kennziffer Se462/3-1.

B. Stroustrup. *Die C++ Programmiersprache*. Addison-Wesley, 1998.

J. Sundberg. *Die Singstimme*. Orpheus-Verlag, Bonn, 1997.

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.

TXT2PHO, 1999. http://www.ikp.uni-bonn.de.

H. Valbret, E. Moulines & J. E. Tubach. Voice Transformation Using PSOLA Technique. *Speech Communication*, 11:175–187, 1992.

B. Welch. *Praktisches Programmieren mit TCL/TK*. Prentice Hall, 1996.

J. Werner. *Lineare Statistik*. Psychologie Verlags Union, 1997.

C. E. Williams & K. N. Stevens. Emotions and Speech: Some Acoustical Correlates. *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.

Anhang A

Anhang

A.1 Liste der verwendeten Hard- und Software

Audiohardware		
DAT-Rekorder		Sony TCD-D7
Kopfhörer		AKG K280
Software zur Analyse und Manipulation von Sprache		
Name	Version	Beschreibung
Praat	3.8	interaktives Analyse- und Syntheseprogramm
ASL	3.0	LPC-Resynthese Programm, gehört zum CSL-System (Computerized Speech Lab) der Firma Kay Elemetrics.
ESPS-tools	5.2	Toolsammlung zur Analyse- und Manipulation von digitalen Sprachsignalen der Firma Entropic Research Laboratory.
Xwaves	5.2	interaktives Frontend zu ESPS
statistische Software		
Name	Version	Beschreibung
R	0.62.3	Skriptsprache zur statistischen Analyse und Visaualisierung (freier Clone von Splus)
SPSS	8.01	interaktives statistisches Prgrammpaket.

A.2 Die unter 3 verwendeten Satzpaare

Sprecherin Branca, weiblich, drei Sätze:
„Sie haben es gerade hochgetragen, und jetzt gehen sie wieder runter“
„Es ist möglich, das am dreißigstem März wieder geöffnet ist“
„Er hat das Heck gesehen, als sie schon umgedreht hatten“

Sprecher Nick, männlich, ein Satz:
„Sie haben es gerade hochgetragen, und jetzt gehen sie wieder runter“

Sprecherin Friederike, weiblich, ein Satz:
„Könnten wir uns nicht verabreden, und dasselbe nochmal anhoeren?“

Sprecherin Julia, weiblich, zwei Sätze:
„An den Wochenenden bin ich jetzt immer nach Hause gefahren und habe Agnes besucht“
„Es ist möglich, das am dreißigstem März wieder geöffnet ist“

Sprecher Roman, männlich, zwei Sätze:
„An den Wochenenden bin ich jetzt immer nach Hause gefahren und habe Agnes besucht“
„Die wird auf dem Platz sein, wo wir sie immer hinlegen“

A.3 Die unter 3.4 verwendeten Vokale

„Sie haben es gerade hochgetragen, und jetzt gehen sie wieder runter“
„Es ist möglich, das am dreißigstem März wieder geöffnet ist“
„Er hat das Heck gesehen, als sie schon umgedreht hatten“
„Könnten wir uns nicht verabreden, und dasselbe nochmal anhoeren?“
„An den Wochenenden bin ich jetzt immer nach Hause gefahren und habe Agnes besucht“
„Die wird auf dem Platz sein, wo wir sie immer hinlegen“

A.4 Der unter 4.5 verwendete Fragebogen

Alter: **Geschlecht:**

Sie hören im Folgenden 76 verschiedene Versionen des Satzes
“An den Wochenenden bin ich jetzt immer nach Hause gefahren und habe Agnes besucht.“
Bitte beurteilen Sie für jeden Satz, wie freudig die Sprecherin klingt

Satz 1)	gar nicht freudig ○ ○ ○ ○ ○ ○ ○ sehr freudig		Satz 2)	gar nicht freudig ○ ○ ○ ○ ○ ○ ○ sehr freudig	
Satz 3)	gar nicht freudig ○ ○ ○ ○ ○ ○ ○ sehr freudig		Satz 4)	gar nicht freudig ○ ○ ○ ○ ○ ○ ○ sehr freudig	
Satz 5)	gar nicht freudig ○ ○ ○ ○ ○ ○ ○ sehr freudig		Satz 6)	gar nicht freudig ○ ○ ○ ○ ○ ○ ○ sehr freudig	
...					
Satz 73)	gar nicht freudig ○ ○ ○ ○ ○ ○ ○ sehr freudig		Satz 74)	gar nicht freudig ○ ○ ○ ○ ○ ○ ○ sehr freudig	
Satz 75)	gar nicht freudig ○ ○ ○ ○ ○ ○ ○ sehr freudig		Satz 76)	gar nicht freudig ○ ○ ○ ○ ○ ○ ○ sehr freudig	

Vielen Dank für die Teilnahme!

A.5 Der unter 6.6 verwendete Fragebogen

Alter: **Geschlecht:**

Sie hören im Folgenden 32 verschiedene Sätze
Bitte beurteilen Sie für jeden Satz, wie freudig die Sprecherin klingt

Satz 1) eher freudig ☐ eher nicht freudig ☐ Satz 2) eher freudig ☐ eher nicht freudig ☐

Satz 3) eher freudig ☐ eher nicht freudig ☐ Satz 4) eher freudig ☐ eher nicht freudig ☐

Satz 5) eher freudig ☐ eher nicht freudig ☐ Satz 6) eher freudig ☐ eher nicht freudig ☐

...

Satz 29) eher freudig ☐ eher nicht freudig ☐ Satz 30) eher freudig ☐ eher nicht freudig ☐

Satz 31) eher freudig ☐ eher nicht freudig ☐ Satz 32) eher freudig ☐ eher nicht freudig ☐

Vielen Dank für die Teilnahme!

A.6 Die unter 6.3 verwendeten Sätze

Zweiphrasige Sätze:
Satz 1: *"Was sind denn das für Tüten, die da unter dem Tisch stehen?"*
Satz 2: *"An den Wochenenden bin ich jetzt immer nach Hause gefahren und habe Agnes besucht"*
Satz 3: *"Es ist möglich, das am dreißigstem März wieder geöffnet ist"*
Satz 4: *"Er hat das Heck gesehen, als sie schon umgedreht hatten"*
Satz 5: *"Ich will das eben wegbringen, und dann mit Karl was trinken gehen."*
Einphrasige Sätze:
Satz 6: *"Sie haben es gerade hochgetragen."*
Satz 7: *"Legen Sie mir die Unterlagen hin."*
Satz 8: *"Könnten wir uns nicht verabreden?"*
Satz 9: *"Wo ist die verbogene Gabel?"*
Satz 10: *"Die liegt immer da."*

A.7 Alle Ergebnisse der Messungen aus 3.3

	Dauer [sek]	Pausen [sek]	S.zeit [sek]	S.rate [sek/Sil]	minF_0 [Hz]	maxF_0 [Hz]	F_0 ø [Hz]	F_0-Range [Hz]
Satz 1:	*Sie haben es gerade hochgetragen, und jetzt gehen sie wieder runter.*							
branca F	5.65	2.03	3.62	0.2	149	316	237.463	167
branca N	2.37	0	2.37	0.13	105	230	183.765	125
Nick F	2.73	0	2.73	0.15	110	270	199.463	160
Nick N	3.18	0.36	2.82	0.16	88	176	131.333	88
Satz 2:	*An den Wochenenden bin ich jetzt immer nach Hause gefahren und habe Agnes besuc*							
Julia F	3.52	0	3.52	0.15	145	490	337.899	345
Julia N	3.76	0	3.76	0.16	160	311	223.611	151
Roman F	4.22	0	4.22	0.18	87	450	173.464	363
Roman N	3.39	0.05	3.34	0.14	79	160	108.899	81
Satz 3:	*Es ist möglich, das am dreißigsten März wieder geöffnet ist.*							
Branca F	4.87	0.9	3.97	0.25	160	337	242.083	177
Branca N	2.28	0	2.28	0.14	72	288	187.104	216
Julia F	2.69	0	2.69	0.17	175	490	328.562	315
Julia N	2.86	0	2.86	0.18	110	280	210.229	170
Satz 4:	*Könnten wir uns nicht verabreden, und dasselbe nochmal anhören?.*							
Fried F	4.1	0.66	3.44	0.19	263	392	328.907	129
Fried N	3.72	0.41	3.31	0.18	127	278	195.259	151
Satz 5:	*Er hat das Heck gesehen, als sie schon umgedreht hatten.*							
Branca F	3.33	0.82	2.51	0.18	177	479	251.095	302
Branca N	1.88	0	1.88	0.13	155	243	190.143	88
Satz 6:	*Die wird auf dem Platz sein, wo wir sie immer hinlegen.*							
Roman F	2.26	0	2.26	0.16	80	250	144.071	170
Roman N	2.12	0	2.12	0.15	82	157	113.548	75

A.8 Screenshot des graphischen Hilfsprogramms aus 6.2