Untersuchungen zur Auswirkung erhöhten Stimmaufwands auf Sprache unter Einbezug des Anwendungsfalls der automatischen Sprechererkennung

Inaugural-Dissertation
zur Erlangung der Doktorwürde
der
Philosophischen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität
zu Bonn

vorgelegt von

Corinna Harwardt

aus

Troisdorf

Bonn 2012

Gedruckt mit der Genehmigung der Philosophischen Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn

Zusammensetzung der Prüfungskommission:

Professor Dr. Caja Thimm
(Vorsitzende)

Professor Dr. Wolfgang Hess
(Betreuer und Gutachter)

apl. Professor Dr. Ulrich Schade
(Gutachter)

Professor Dr. Bernd Möbius
(weiteres prüfungsberechtigtes Mitglied)

Tag der mündlichen Prüfung: 03.04.2012

Danksagung

Mein besonderer Dank gilt meinem Doktorvater, Herrn Prof. Dr. Wolfgang Hess, sowie Herrn Prof. Dr. Ulrich Schade, Herrn Prof. Dr. Bernd Möbius und Herrn Dr. Matthias Hecking, die mich während der gesamten Zeit meiner Promotion mit Diskussionen und Anregungen immer wieder auf neue Ideen gebracht und unterstützt haben.

Da diese Arbeit während meiner Zeit am Fraunhofer-Institut für Kommunikation, Informationsverarbeitung und Ergonomie (FKIE) entstanden ist, gilt mein Dank auch den dortigen Kollegen. Viele unterstützten mich auf die eine oder andere Weise. Herr Dr. Matthias Hecking motivierte mich eine Dissertation zu beginnen, gab mir die Möglichkeit in seinem Team zu promovieren und unterstützte mich in jeglicher Hinsicht, sodass das Projekt „Doktorarbeit“ überhaupt möglich wurde. Besonders hervorheben möchte ich auch die Kollegen und Freunde Frederike Gottsmann, Thomas Remmersmann und Sandra Noubours.

Als letzten und wichtigsten Dank möchte ich mich an meine Familie und meine Freunde wenden, die die richtige Mischung aus Motivation, Unterstützung und Verständnis zeigten. Besonderer Dank gilt an dieser Stelle meinem Verlobten Markus Blätzinger.

Ungeachtet all dieser Unterstützung bin ich allein verantwortlich für etwaige Fehler oder Ungenauigkeiten der vorliegenden Arbeit.

Inhaltsverzeichnis

Kapitel 1

Einleitung

Die akustischen Parameter von Sprache sind abhängig von äußeren Einflüssen und Gegebenheiten. Diese können stark variieren. Die resultierenden Variationen der akustischen Parameter wirken sich auf die Leistung sprachverarbeitender Systeme aus.

Sprachverarbeitende Systeme, wie beispielsweise automatische Sprach- oder Sprechererkenner, sind mittlerweile ausreichend weit entwickelt, um für Aufnahmen hoher Qualität sehr gute Resultate zu erzielen. Die Leistung sinkt jedoch bei wechselnden äußeren Gegebenheiten stark ab. Ein Faktor, der zu einem Leistungsabfall führen kann, ist der veränderte Stimmaufwand eines Sprechers, welcher bedingt ist durch äußere Einflüsse wie beispielsweise Hintergrundgeräusche oder eine Distanz zwischen Sprecher und Hörer. Sprachverarbeitende Systeme werden vor der Nutzung im Testmodus mit Trainingsdaten, ähnlich denen des späteren Einsatzszenarios, trainiert. Unterscheidet sich der Stimmaufwand in Trainings- und Testdaten, kommt es zu einem Leistungsabfall sprachverarbeitender Systeme. In der vorliegenden Arbeit wird dieses Problem thematisiert.

Das Ziel dieser Arbeit ist die Analyse der Auswirkung erhöhten Stimmaufwands auf Sprache und sprachverarbeitende Systeme. Als Beispielszenario wird die automatische Sprechererkennung bei ungleichem Stimmaufwand in Trainings- und Testdaten ausgewählt. Die zugehörigen Problemstellungen lauten:

- Wie verändern sich die akustischen Parameter bei einer Erhöhung des Stimmaufwands?
- Welchen Einfluss haben diese Veränderungen auf Sprechererkennungssysteme und welche Verfahren können zur Verbesserung der Systemleistung bei nichtübereinstimmendem Stimmaufwand in Trainings- und Testdaten angewendet werden?

Diese Problemstellungen werden im Folgenden in einen Anwendungskontext gesetzt und detaillierter dargestellt.

1.1 Motivation

Die Veränderungen akustischer Parameter auf Grund veränderten Stimmaufwands stehen seit langer Zeit im Fokus zahlreicher Untersuchungen (Ladefoged & McKinney, 1963; Rostolland & Parant, 1974; Rostolland, 1982a, 1982b; Schulman, 1985b;

Gramming, Sundberg, Ternström, Leanderson & Perkins, 1987; Schulman, 1989). Nichtsdestotrotz sind die exakte Natur der Veränderungen der akustischen Parameter und ihre Zusammenhänge bisher nicht bekannt. Die bisherigen Studien analysieren artikulatorische (Garnier, Wolfe, Henrich & Smith, 2008), perzeptive (Dreher & O'Neill, 1957) und akustische (Gramming et al., 1987) Aspekte sowie die Zusammenhänge zwischen diesen Faktoren (van Summers, Pisoni, Bernacki, Pedlow & Stockes, 1988). Die Untersuchungen stellen verschiedene Ursachen für die Erhöhung des Stimmaufwands dar. Der Sprecher kann aus emotionalen Gründen den Stimmaufwand anheben; er kann ihn anheben, um eine Distanz zwischen Sprecher und Hörer zu überwinden; oder um sich bei Hintergrundgeräuschen verständlich zu machen (Lombardsprache). Die emotionsbedingte Erhöhung des Stimmaufwands wird in dieser Arbeit nicht betrachtet, da hier mehrere Faktoren gleichzeitig die Untersuchung beeinflussen. Der Fokus dieser Arbeit liegt auf der Analyse von Lombardsprache sowie auf der Analyse von Sprache mit erhöhtem Stimmaufwand, hervorgerufen durch eine Anweisung des Versuchsleiters. Zwischen Lombardsprache und auf Anweisung laut gesprochener Sprache bestehen zwar Unterschiede hinsichtlich der Veränderung der akustischen Parameter, die Tendenzen sind jedoch gleich (Stanton, Jamieson & Allen, 1988).

Die Motivation für die Untersuchung von erhöhtem Stimmaufwand ist vor allem im *militärischen und forensischen Kontext* zu finden. In militärischen Anwendungsbereichen von Sprachtechnologie liegt beinahe ausschließlich Lombardsprache vor. Als militärische Anwendungsbereiche für Sprachtechnologie sind *sicherheitsüberprüfende Systeme*, *Kommado-* beziehungsweise *Eingabesysteme* und *Aufklärungssysteme* zu nennen (Harwardt, 2009).

Für *sicherheitsüberprüfende Systeme* besteht kein Problem mit erhöhtem Stimmaufwand, sofern die Anwendung in einem ruhigen Umfeld angewendet wird. Werden sicherheitsüberprüfende Systeme hingegen für mobile Geräte oder Fahrzeuge eingesetzt, so sind Verschlechterungen der Erkennungsraten, unter anderem auf Grund des erhöhten Stimmaufwands, zu erwarten. Ein von Hansen und Varadarajan (2009) beschriebenes Szenario dieser Art ist ein Zugangskontrollsystem für Piloten.

Der Einsatz von *Kommando-* und *Eingabesystemen* in militärischen Anwendungen ist meistens verbunden mit Gefechtslärm und wechselnden Hintergrundgeräuschen von Geräten, Fahrzeugen oder Flugzeugen. In einem solchen Szenario ist es notwendig, die Sprachmodelle auf die äußeren Gegebenheiten anzupassen, sodass die Hintergrundgeräusche und der erhöhte Stimmaufwand bei der Systementwicklung besondere Beachtung finden.

Sprachverarbeitende Systeme in der *Aufklärung* haben mit ähnlichen Problemen zu kämpfen. Zwar gibt es selten Motorengeräusche, jedoch sind hier andere herausfordernde Geräusche und Verzerrungen vorhanden. Je nach Szenario kann es zu erhöhtem Stimmaufwand des Sprechers kommen, sodass die vorliegende Arbeit ebenfalls für die Entwicklung solcher Systeme interessant ist.

Die Leistung sprachverarbeitender Systeme im militärischen Kontext sinkt auf Grund widriger Faktoren. Die Motivation der vorliegenden Arbeit begründet sich im Wunsch zur Verbesserung der Systemleistung im Kontext erhöhten Stimmaufwands. Die Analyse der Veränderungen akustischer Parameter findet nicht nur in der Verbesserung von Sprachverarbeitungstechnologien Anwendung. Die gewonnenen Informationen können ebenfalls nutzbringend in der *forensischen Fallarbeit* eingesetzt werden. In der forensischen Phonetik besteht häufig das Problem, dass zwei

Aufnahmen verglichen werden sollen, in denen Sprache unterschiedlichen Stimmaufwands enthalten ist. In einem solchen Fall können, für die Grundfrequenzanalyse beispielsweise, nur solche Signalabschnitte verglichen werden, die den gleichen Stimmaufwand aufweisen. Die Anteile gleichen Stimmaufwands können je nach Fall aber sehr gering sein, sodass Studien zur Veränderung der akustischen Parameter helfen können, weitere Signalabschnitte für die Analyse nutzbar zu machen. Eine andere Einsatzmöglichkeit der gewonnen Informationen ist die Sortierung großer Datenmengen nach Sprache normalen, erhöhten, und niedrigen Stimmaufwands. Eine solche Sortierung kann in Abhör- oder Aufklärungsszenarien relevant sein, um potentiell relevantere Sprachsignale (laut oder leise/ geflüstert) zu identifizieren.

1.2 Herangehensweise

Die Herangehensweise an die gegebenen Problemstellungen in den definierten Anwendungsszenarien gliedert sich in verschiedene Schritte:

Wie verändern sich die akustischen Parameter bei einer Erhöhung des Stimmaufwands? Um Sprachverarbeitungstechnologien im Kontext erhöhten Stimmaufwands verbessern zu können, muss zunächst detailliertes Wissen über die akustischen Veränderungen der Stimmaufwandsvariationen vorliegen. Hierfür werden im praktischen Teil der Arbeit zunächst Untersuchungen verschiedener spektraler Parameter durchgeführt. Es wird analysiert, ob Veränderungen der Verteilungen für die einzelnen spektralen Parameter auftreten und wie groß diese sind. Anschließend werden Signifikanztests realisiert, um zu prüfen, ob die beobachteten Veränderungen signifikant sind.

Die Ergebnisse der statistischen Analysen werden mit Hilfe eines Stimmaufwandsklassifikators verifiziert. Dafür wird zunächst ein Klassifikator realisiert, welcher mit den spektralen Parametern getestet wird. Um insgesamt das beste Merkmal zur Klassifikation des Stimmaufwands zu finden, werden die spektralen Merkmale und die MFCC-Merkmale vergleichend und als Kombinationen von Merkmalen evaluiert. Über die Verifikation der Ergebnisse der statistischen Analyse hinausgehend kann der Stimmaufwandsklassifikator als unterstützendes System für andere Sprachverarbeitungssysteme verwendet werden. In der Sprach- oder Sprechererkennung bei wechselndem Stimmaufwand kann durch einen solchen Klassifikator ein geeignetes Modell für den aktuellen Stimmaufwandsgrad ausgewählt werden. Auch in der forensischen Phonetik ist die Quantifizierung des Stimmaufwands essentiell, sodass die Ergebnisse der statistischen Analyse und des Klassifikators in diesem Bereich genutzt werden können.

Nachfolgend wird F_0 hinsichtlich Veränderungen im Kontext erhöhten Stimmaufwands, mit derselben Vorgehensweise wie für die spektralen Parameter, untersucht. Weiterhin soll überprüft werden, ob über die festgestellten Veränderungen der akustischen Parameter hinaus Zusammenhänge zwischen lauter und normaler Sprache bestehen. Dies ist eine relevante Fragestellung, da starke Zusammenhänge zwischen normaler und lauter Sprache möglicherweise eine Vorhersage der Veränderungen erlauben. Deswegen wird für sämtliche analysierten akustischen Parameter eine Korrelationsanalyse durchgeführt.

Als Nächstes wird untersucht, ob Zusammenhänge zwischen F_0 und einem der spektralen Parameter für beide Stimmaufwandsgrade bestehen. Bestünden für beide Stimmaufwandsgrade sowohl starke als auch ähnliche Zusammenhänge, könnte mit Hilfe dieser Zusammenhänge ein robustes Merkmal für die Sprechererkennung entwickelt werden. Es würde geprüft, ob die Zusammenhänge für den einzelnen Sprecher gleich bleiben bei normalem und erhöhtem Stimmaufwand und ob die Zusammenhänge zwischen F_0 und einem der spektralen Parameter für verschiedene Sprecher unterschiedlich sind. In diesem Fall könnte beispielsweise das Verhältnis der beiden Parameter zueinander als Merkmal für die Sprechererkennung verwendet werden.

Die vorliegende Arbeit bietet einen **umfangreichen Vergleich spektraler Merkmale**. Ein solcher Vergleich wurde in diesem Umfang in bisherigen Untersuchungen nicht durchgeführt. Weiterhin werden die erzielten Ergebnisse auf die automatische **Stimmaufwandsklassifikation** übertragen. Die Darstellung der **Zusammenhänge zwischen den spektralen Veränderungen und F_0** ist ebenfalls in dieser Form nicht in der Literatur zu finden.

Welchen Einfluss haben diese Veränderungen auf Sprechererkennungssysteme und welche Verfahren können zur Verbesserung der Systemleistung bei nicht-übereinstimmendem Stimmaufwand in Trainings- und Testdaten angewendet werden? Die Untersuchung, welchen Einfluss der erhöhte Stimmaufwand auf sprachverarbeitende Systeme hat und wie eine verbesserte Erkennungsrate erzielt werden kann, wird im Kontext der automatischen Sprechererkennung durchgeführt. Hierzu wird ein Sprecherverifikationssystem erstellt und mit Standardmerkmalen sowie F_0-basierten Merkmalen bei übereinstimmendem und nicht-übereinstimmenden Stimmaufwand evaluiert.

Mit Hilfe der gewonnen Informationen über die akustischen Veränderungen werden neue Merkmale beziehungsweise Merkmalskombinationen zur Sprechererkennung vorgeschlagen. Diese werden ebenfalls evaluiert und bewertet.

Als weitere Verbesserungsmöglichkeit werden unterschiedliche Adaptionsverfahren vorgestellt. Auch hier werden vergleichende Evaluationen der unterschiedlichen Adaptionsverfahren auf verschiedenen Merkmalen durchgeführt. Als nächstes werden Systemkombinationen verschiedener Merkmale für die einzelnen Adaptionsverfahren getestet.

Abschließend werden die Resultate der einzelnen Systemtests einander gegenübergestellt und Vorschläge für weitere Arbeiten aufgeführt.

In dieser Arbeit werden neue Merkmale eingeführt und bestehende Methoden für die Sprechererkennung bei nicht-übereinstimmendem Stimmaufwand in Trainings- und Testdaten angepasst. Damit bietet die Arbeit eine **fundierte Grundlage für die Entwicklung sprachverarbeitender Systeme im Kontext erhöhten Stimmaufwands**.

1.3 Überblick

Das vorliegende Kapitel führt den Leser in die Thematik ein und präsentiert die Motivation für die nachfolgenden Untersuchungen. Die weitere Arbeit ist in einen Theorie- und einen Praxisteil unterteilt.

Im *Theorieteil* führt Kapitel 2 in die Problematik der Veränderung von Sprache in Abhängigkeit vom Stimmaufwand ein. Hier werden artikulatorisch-physiologische, perzeptive und akustische Aspekte diskutiert. Der Schwerpunkt der Ausführungen liegt auf den akustischen Veränderungen. In Kapitel 3 werden die Grundlagen der automatischen Sprechererkennung erläutert. Dieses Kapitel liefert die Kenntnisse für das später erstellte Sprechererkennungssystem. Es werden ausgewählte Standardtechnologien präsentiert. Kapitel 4 ist das letzte Kapitel des Theorieteils. Hier werden die Erkenntnisse aus den vorherigen Kapiteln zusammengeführt. Es wird dargestellt, wie groß der Einfluss veränderten Stimmaufwands auf sprachverarbeitende Systeme ist und welche Kompensations- und Filterverfahren zur Lösung des Problems vorgeschlagen wurden. Weiterhin wird der militärische Anwendungsbereich der Sprachverarbeitung bei verändertem Stimmaufwand dargestellt.

Ausgehend von den theoretischen Kenntnissen beginnt der *praktische Teil* der Arbeit mit einer Übersicht über die durchgeführten Tests und über die Korpora, welche für diese Tests relevant sind (Kapitel 5). Darauf folgt die statistische Analyse spektraler Parameter bei erhöhtem Stimmaufwand in Kapitel 6. Ausgehend von dieser Analyse werden die spektralen Parameter als Merkmale und Merkmalskombinationen in einem automatischen Stimmaufwandsklassifikator getestet, um zu evaluieren, ob sie zur Klassifikation des Stimmaufwands geeignet sind. In Kapitel 7 wird eine statistische Analyse der Veränderungen von F_0 präsentiert. Außerdem werden die Zusammenhänge zwischen normaler und lauter Sprache für F_0 und die spektralen Parameter untersucht. Abschließend stellt das Kapitel die Zusammenhänge zwischen F_0 und den einzelnen spektralen Parametern dar. Nach den statistischen Analysen wird in Kapitel 8 das Framework für die Sprechererkennung vorgestellt. Dieses wird mit zahlreichen Standard- und F_0-basierten Merkmalen getestet, wobei in den Trainingsdaten normaler Stimmaufwand und in den Testdaten erhöhter vorliegt. Weiterhin werden in Kapitel 8 anhand der statistischen Analysen gefundene Merkmale vorgestellt und evaluiert. Im Anschluss an die Evaluation der Merkmale werden in Kapitel 9 unterschiedliche Adaptionsverfahren dargestellt und getestet. Es wird unterschieden zwischen Adaption der Testdaten und Adaption der Modelle. Kapitel 10 fasst die Untersuchungen und ihre Ergebnisse zusammen und diskutiert mögliche Fortführungen dieser Arbeit.

Kapitel 2

Auswirkungen von Lautstärkeveränderungen der Stimme auf Sprache

Die Auswirkungen veränderten Stimmaufwands auf akustisch-phonetische und artikulatorische Merkmale sowie die Perzeption von Sprachsignalen ist seit langer Zeit Forschungsgegenstand der Phonetik. Zahlreiche Veröffentlichungen zu diesem Thema erschienen bereits in den 80er Jahren (Gramming et al., 1987; Schulman, 1989, 1985b; Rostolland, 1982a, 1982b) und früher (Ladefoged & McKinney, 1963; Rostolland & Parant, 1974). Bereits zu dieser Zeit war eine Motivation für solche Untersuchungen die Verbesserung sprachverarbeitender Systeme für Lombardsprache[1] (Bond, Moore & Gable, 1989; Rajasekaran, Doddington & Picone, 1986; Stanton et al., 1988; Moore & Bond, 1987). Wie stark die Erkennungsraten sprachverarbeitender Systeme bei verändertem Stimmaufwand noch heute absinken, verdeutlicht beispielsweise die Untersuchung von Becker (2008). Auch in der vorliegenden Arbeit wird der Einfluss veränderten Stimmaufwands in sprachverarbeitenden Systemen untersucht. Hierfür werden unterschiedliche Merkmale und Adaptionsverfahren auf ihre Eignung im Kontext automatischer Sprechererkennung mit Trainingsdaten normalen und Testdaten erhöhten Stimmaufwands analysiert (Kapitel 8 und 9). Vorab werden allgemeine Untersuchungen zur Veränderung verschiedener spektraler Parameter und zur Veränderung der Grundfrequenz vorgestellt (Kapitel 6 und 7).

Als Grundlage für diese praktischen Arbeiten wird im Folgenden zunächst der Begriff Stimmaufwand erläutert (Abschnitt 2.1), um anschließend die Veränderungen der Sprache bei erhöhtem Stimmaufwand zu beschreiben. Die Auswirkungen des Stimmaufwands werden getrennt nach artikulatorisch-physiologischen (Abschnitt 2.2), perzeptiven (Abschnitt 2.3) und akustischen Aspekten (Abschnitt 2.4) dargestellt. Der Schwerpunkt liegt hier auf der Darstellung von F_0 und den spektralen Parametern, da diese für die praktischen Aufgaben der vorliegenden Arbeit am wichtigsten sind. Die anderen akustischen Parameter sowie die artikulatorisch-physiologischen und perzeptiven Aspekte müssen aber ebenfalls betrachtet werden, um Sprache und die durch den Stimmaufwand bedingten Veränderungen als komplexes Gesamtprodukt verstehen zu können.

[1] Die Veränderung der Sprache auf Grund von Hintergrundgeräuschen wird Lombardeffekt genannt.

2.1 Stimmaufwand

Unter *Stimmaufwand* ist laut Duden der „Aufwand an Stimmkraft“ (Wissenschaftlicher Rat der Dudenredaktion, 2001) zu verstehen. Von einem großen Stimmaufwand wird gesprochen bei lauter oder geschriener Sprache, während ein geringer Stimmaufwand mit leiser Sprache assoziiert ist. In der Fachliteratur ist Stimmaufwand je nach Forschungsinteresse unterschiedlich definiert. Eine allgemeine Definition geben Jessen et al., die Stimmaufwand damit in Verbindung bringen, wie laut ein Sprecher Sprache produziert (Jessen, Köster & Gfroerer, 2005, S. 175). Eine etwas speziellere Definition wird von Traunmüller und Eriksson genutzt: „...vocal effort is the quantity that ordinary speakers vary when they adapt their speech to the demands of an increased or decreased communication distance“ (Traunmüller & Eriksson, 2000, S. 3438). Stimmaufwand wird in diesem Fall als die Quantität betrachtet, um die ein Sprecher seine Stimme verändert, sobald sich die Kommunikationsdistanz zwischen ihm und seinem Hörer verändert. Diese Definition ist stark an dem Forschungsinteresse der Autoren ausgerichtet, welches sich auf unterschiedlichen Stimmaufwand, hervorgerufen durch eine Distanzveränderung, konzentriert. Weitere Möglichkeiten, um Stimmaufwandsveränderungen hervorzurufen, wie beispielsweise durch Hintergrundgeräusche, werden in dieser Definition nicht berücksichtigt. In einer Untersuchung von Shriberg et al. werden zwar nur distanzbedingte Veränderungen untersucht, Hintergrundgeräusche werden jedoch in die Aufgabendefinition miteinbezogen: „... we sought to study raised (or high) vocal effort associated with projection over a distance, rather than over noise (as in the Lombard effect)“ (Shriberg et al., 2008, S. 609).

Ein weiteres Beispiel für eine speziell auf das Forschungsinteresse ausgerichtete Definition stammt von Cabrera und Gilfillan (2002). Ihre Untersuchung ist perzeptuell motiviert, sodass sie Stimmaufwand als perzeptuellen Schlüssel für die Distanz definieren (Cabrera & Gilfillan, 2002, S. 2): „Vocal effort is a distance perception cue, greater effort (e.g. a shout) being associated with greater distance.“

Auf Grund der zahlreichen unterschiedlichen Definitionen wird deutlich, dass es sich bei Stimmaufwand nicht um ein physikalisch eindeutig messbares Phänomen handelt. Vielmehr ist Stimmaufwand ein durch den Sprecher und Hörer subjektiv wahrnehmbares Phänomen, welches besonders bei wechselnden Aufnahmegegebenheiten nur schwer zu quantifizieren ist. Der Stimmaufwand wird in dieser Arbeit wie folgt definiert:

> Stimmaufwand ist die Quantität, welche ein Sprecher benötigt, um die Lautstärke seiner Sprache an die Kommunikationssituation anzupassen.

Mit Kommunikationssituation kann eine Distanzveränderung, Hintergrundgeräusche, Schwerhörigkeit eines Kommunikationspartners oder Weiteres gemeint sein. Wesentlich ist hierbei die Tatsache, dass unterschiedlich hervorgerufene Stimmaufwandsveränderungen auch unterschiedliche beziehungsweise unterschiedlich starke Auswirkungen auf die akustischen Parameter haben können (Gramming et al., 1987). Es können folgende Auswirkungen von erhöhtem Stimmaufwand auf Sprache genannt werden:

- höhere Intensität und Amplitude,
- höhere Grundfrequenz, welche allerdings je nach Sprecher stark variiert,

- Korrelation zwischen erhöhter Sprechrate und Lombardeffekt,
- erhöhte Wort- und Vokaldauer (nur bei manchen Studien nachgewiesen),
- Erhöhung der Mittenfrequenz des ersten Formanten, verursacht durch die niedrigere Zungenposition,
- Erhöhung des subglottalen Drucks,
- Erhöhung der Spannung der laryngalen Muskeln sowie
- Absenken des Kiefers und dadurch größere Mundöffnung.

Diese Veränderungen wurden sowohl bei Versuchen mit Hintergrundgeräuschen, welche den Lombardeffekt nutzen, beobachtet als auch bei Versuchen, die andere Methoden nutzen, um den Sprecher zu lauterem Sprechen zu bewegen (beispielsweise durch genaue Anweisungen (Barfs, 2005) oder Distanz zwischen Sprecher und Versuchsleiter (Liénard & Di Benedetto, 1999)). Ein detaillierter Überblick über den Einfluss des Stimmaufwands auf die Artikulation, die perzeptiven Eindrücke und die akustischen Parameter von Sprache ist in den Abschnitten 2.2 - 2.4 zu finden.

2.2 Artikulatorisch-physiologische Aspekte

In diesem Abschnitt werden die *artikulatorisch-physiologischen Veränderungen* bedingt durch erhöhten Stimmaufwand kurz dargestellt. Die Artikulatorische Phonetik untersucht die Gesamtheit der Sprechbewegungen (Pompino-Marschall, 2003). Folglich werden die Positionen der beteiligten Artikulatoren näher beleuchtet. In Bezug auf Veränderungen des Stimmaufwands gibt es zahlreiche Untersuchungen, welche die Vokaltraktkonfigurationen unterschiedlicher Stimmaufwandsgrade vergleichen (Schulman, 1985a, 1989; Garnier, Bailly, Dohen, Welby & Loevenbruck, 2006; Garnier et al., 2008; Garnier, Dohen, Loevenbruck, Welby & Bailly, 2006). Besonders die Parameter Kiefer- und Zungenposition sowie die Stellung der Lippen sind Gegenstand dieser Untersuchungen.

Erhöhter Stimmaufwand geht einher mit erhöhtem subglottalen Druck und erhöhter Spannung der laryngalen Muskeln (Lu, 2010). Weiterhin werden bei lauter Sprache die Artikulationsbewegungen verstärkt (Schulman, 1989; Garnier, Bailly et al., 2006). Diese Verstärkung ist nicht linear. Vielmehr handelt es sich um eine komplexe, zielorientierte Reorganisation der Artikulationsbewegungen (Schulman, 1989). Des Weiteren fand Schulman (1989) heraus, dass die Veränderungen der Vokalartikulation vorhersehbar sind und dass die abgesenkte Kieferposition bilabialer Verschlüsse ähnlich der Veränderung bei Beißblockexperimenten ist. Deswegen betrachtet er das Schreien als „natürlichen“ Beißblock. Der abgesenkte Kiefer bei lauter Sprache wird durch die Oberlippe kompensiert, was im Vergleich zu normaler Artikulation einen größeren Verschluss hervorruft. Die Kieferposition ist jedoch nicht nur für bilabiale Verschlusslaute verändert. In einer vorherigen Untersuchung fand Schulman (1985a) heraus, dass der Kiefer auch für Vokale abgesenkt ist, ebenso wie die Zungenposition. Untersuchungen zur Lippenposition wurden von Garnier et al. (Garnier, Bailly et al., 2006; Garnier et al., 2008) durchgeführt. Sie fanden heraus, dass die Amplituden der Parameter Lippenspreizung, Lippenöffnung und Öffnungsfläche der Lippen bei lauter Sprache, sowohl bei weißem Rauschen als

auch bei Cocktailpartygeräuschen, ansteigen. Zudem steigt die Geschwindigkeit der artikulatorischen Bewegungen für diese Parameter an. Dies deckt sich mit der Aussage von Schulman (1989), der einen Anstieg der Geschwindigkeit bei größeren Veränderungen der Position des Kiefers und der Zunge beobachtete.

Insgesamt zeigen sämtliche hier beschriebenen Studien **große Veränderungen der artikulatorischen Zielstellungen** für laute Sprache. Vor allem die Veränderungen der Vokale sind sehr stark und folgen einem vorhersagbaren Muster (Schulman, 1989).

2.3 Untersuchungen zur Perzeption

Perzeptionsexperimente im Kontext erhöhten Stimmaufwands untersuchen, ob durch die massiven artikulatorischen und akustischen Veränderungen *die Wahrnehmung von Sprache* verbessert oder verschlechtert wird. Auf Grund der starken Veränderungen gehen viele Autoren zunächst von einer verschlechterten Perzeption aus. Dies wurde allerdings durch unterschiedliche Studien widerlegt (Dreher & O'Neill, 1957; Schulman, 1985b, 1989; van Summers et al., 1988). Vielmehr resultierte der Lombardeffekt bei einigen Experimenten in besser verständlicher Sprache (Dreher & O'Neill, 1957; Lu & Cooke, 2009). Schulman (1985b, 1989) zeigte, dass laute Vokale nicht schlechter wahrgenommen werden als leise. Er untersuchte die Vokale im Kontext /h_t/ (Schulman, 1989). Der erhöhte Stimmaufwand wurde nicht durch Hintergrundgeräusche hervorgerufen, sondern durch eine Aufforderung an den Sprecher. Für den Erhalt des perzeptiven Eindrucks gibt der Autor zwei mögliche Erklärungen an. Bei der ersten Begründung bezieht er sich auf die Veränderungen der Formanten durch erhöhten Stimmaufwand. Für F1 beobachtete er eine Steigerung um 100 Hz. Diese Steigerung ist nicht signifikant und verursacht bei relativ stabil bleibenden F2- und F3-Werten keine größere Stimulusveränderung. Die zweite Erklärung nimmt Bezug auf das Verhältnis zwischen F_0 und F1. Da beide Werte um ca. 100 Hz steigen, bleibt die Relation gleich. Daraus resultiert die These, dass die Vokalqualität über das Verhältnis zwischen F_0 und F1 wahrgenommen wird.

Eine Untersuchung zur Perzeption von Lombardsprache wurde von Junqua und Anglade (1990) durchgeführt. Sie untersuchten verschiedene Vokabularsets, von denen vier Sets leicht verwechselbare Wörter enthielten und zwei nicht leicht verwechselbare Wörter. Es wurden zwei Experimente mit Hörern unterschiedlicher Nationalitäten durchgeführt. Die Ergebnisse beider Hörergruppen (Englisch und Französisch) sind vergleichbar. Es stellt sich heraus, dass im Gegenteil zu vorherigen Studien (Dreher & O'Neill, 1957; Lu & Cooke, 2009) ein Abfall der Perzeptionsleistung zu beobachten ist. Dies ist vor allem durch die leicht verwechselbaren Wörter erklärbar, von denen in anderen Studien keine verwendet wurden (Dreher & O'Neill, 1957).

Weitere Perzeptionsexperimente, welche die Kopplung von Distanzeinschätzungen und Sprache erhöhten Stimmaufwands untersuchen, wurden von Eriksson und Traunmüller (Eriksson & Traunmüller, 2002; Traunmüller, 1997) durchgeführt. Diese Experimente zeigen, dass kleinere Distanzen tendenziell überschätzt und größere Distanzen unterschätzt werden (Traunmüller, 1997).

Grundsätzlich lässt sich festhalten, dass **die Verständlichkeit von Sprache bei erhöhtem Stimmaufwand erhalten bleibt oder sich verbessert** (Dreher & O'Neill, 1957; Schulman, 1985b, 1989; van Summers et al., 1988). Werden jedoch zu ähnliche Wörter in den Perzeptionsexperimenten verwendet, so kann es zu einem Absinken der Perzeptionsleistung kommen (Junqua & Anglade, 1990).

2.4 Akustische Parameter

In diesem Abschnitt werden die Veränderungen unterschiedlicher *akustischer Parameter* dargestellt, welche durch eine Variation des Stimmaufwands ausgelöst werden. Zunächst werden hierzu die Auswirkungen auf die *Grundfrequenz* erläutert. Anschließend werden die *Formanten*, *zahlreiche spektrale Parameter*, *Amplitude* und *Intensität* sowie die *akustische Dauer* erörtert. Besonders ausführlich werden die Grundfrequenz und die spektralen Veränderungen dargestellt, da diese Parameter im Fokus der praktischen Untersuchungen stehen.

2.4.1 Grundfrequenz

Die *Grundfrequenz* F_0 ist die „tiefste Frequenz einer komplexen periodischen Schwingung" (Pompino-Marschall, 2003, S. 310). Sie ist das akustische Korrelat für die Anzahl der Vibrationen der Stimmbänder (Hess, 2008; Rose, 2002).

Die Beschaffenheit der Stimmbänder, besonders deren Größe, enthält wichtige sprecherspezifische Informationen. Die Intrasprechervariabilität der mittleren F_0 ist relativ gering. Aus diesem Grund gelten Merkmale, die mit der Anzahl Vibrationen der Stimmbänder, also mit F_0, zusammenhängen, als zuverlässige Parameter in der Sprechererkennung (Braun, 1992; Rose, 2002).

Problematisch wird die F_0-basierte Sprechererkennung, wenn der Sprecher seinen Stimmaufwand für ein zu vergleichendes Sprachsignal anhebt, da hierdurch die Intrasprechervariabilität stark erhöht wird. In Tabelle 2.1 wird der Einfluss des Stimmaufwands auf verschiedene F_0-Maße, wie er in unterschiedlichen Studien belegt wurde, dargestellt. Sämtliche Studien nutzten als Vergleichsmaterial Sprache normalen Stimmaufwands.

Studie	Parameter	Bedingung	Ergebnis
Schulman (1985a)	F_0	gelesene Sprache, laut (a. A.[2]), mind. 90 dB	steigt um 100-200 Hz
Stanton et al. (1988)	F_0	laut (a. A.) + Lombard + normal, jeweils mit Sauerstoffmaske	tendenziell steigend für laute und Lombardsprache
Geumann (2001)	F_0	gelesene Sprache, laut (a. A.)	steigt

[2] a. A. steht für „auf Anweisung".

Liénard und Di Benedetto (1999)	F_0	isoliert gesprochene Vokale, drei Distanzen zur Erzeugung des erhöhten Stimmaufwands	steigt signifikant
Garnier et al. (2008)	F_0	isoliert gesprochene Vokale mit erhöhtem Stimmaufwand mit den Anweisungen: 1. keine, 2. F_0 konstant halten, 3. F_0 und Artikulatoren konstant halten	steigt für alle drei Bedingungen, am meisten für 1.
Jessen et al. (2005)	Mittelwert	Lombard-, Lese- + Spontansprache	steigt signifikant
	Standard-abweichung	Lombard-, Spontansprache	bleibt ähnlich oder steigt
	Variations-koeffizient	Lombard-, Lese- + Spontansprache	bei gelesener Sprache signifikant erhöht, bei Spontansprache nicht signifikant verändert
Barfs (2005)	Mittelwert	gelesene Sprache, geschrien (a. A.)	steigt signifikant
	Standard-abweichung (Absolutwert)		steigt signifikant
	Variations-koeffizient		sinkt hoch signifikant
Gramming et al. (1987)	Mittelwert	gelesene Sprache mit drei Kategorien: laut (a. A.), lautest möglich (a. A.), Lombard	steigt (für Lombard weniger als für „lautest möglich“)
Bond et al. (1989)	Mittelwert	Lombard	steigt
		Lombard + Sauerstoffmaske	keine signifikante Änderung

Tabelle 2.1: Untersuchungen zur mittleren F_0 und zur F_0-Variation

Die in Tabelle 2.1 aufgeführten Untersuchungen zeigen eine Anhebung von F_0 beziehungsweise von der mittleren F_0 bei erhöhtem Stimmaufwand. Der Anstieg ist sowohl in Lombardsprache als auch in lauter Sprache ohne Hintergrundgeräusche signifikant. Bei der Untersuchung von Gramming et al. (1987) zeigte sich, dass F_0 unterschiedlich stark modifiziert wird, je nachdem, ob die Stimmaufwandserhöhung durch Lombardsprache oder durch lautes Sprechen auf Anweisung bedingt wird.

Es lässt sich festhalten, dass, obwohl beide Veränderungen signifikant sind, die Veränderung von F_0 bei Lombardsprache nicht so groß ist wie bei auf Anweisung laut gesprochener Sprache (Gramming et al., 1987). Weiterhin scheint die Stärke des Anstiegs sprecherabhängig zu sein (Jessen et al., 2005). Als mögliche Einflussfaktoren, welche die verschiedenartige Erhöhung von F_0 verschiedener Individuen begründen können, sind unter anderem Erfahrungswerte im Singen, Emotionen, das Rauchverhalten und Alkohol- beziehungsweise Drogeneinfluss zu nennen (Gramming et al., 1987; Barfs, 2005). Diese Faktoren werden in dieser Arbeit jedoch nicht weiter betrachtet, sodass ihr Einfluss auf die Grundfrequenz nicht in Tabelle 2.1 aufgeführt ist.

Die Variation von F_0 kann durch die Standardabweichung oder den Variationskoeffizienten ausgedrückt werden. Während die Standardabweichung ein Absolutwert ist, der die durchschnittliche Abweichung vom Mittelwert ausdrückt, gibt der Variationskoeffizient die relative Abweichung vom Mittelwert an. Der Variationskoeffizient stellt eine Normierung der Varianz dar. Bei großem Mittelwert einer Verteilung liegt im Allgemeinen eine größere Varianz vor als bei kleinem Mittelwert. Dieser Effekt wird durch den Variationskoeffizienten normiert, sodass eine Beurteilung der Varianz unabhängig von der Größe des Mittelwerts durchgeführt werden kann. Der Variationskoeffizient ist eine statistische Kenngröße und wird häufig in Prozent angegeben ($\frac{Standardabweichung \cdot 100}{Mittelwert}$), wie beispielsweise in den Untersuchungen von Künzel (1987) und Jessen et al. (2005).

Die Standardabweichung steigt oder bleibt ähnlich bei erhöhtem Stimmaufwand und geschriener Sprache. Dagegen sinkt der Variationskoeffizient für geschriene Sprache (Barfs, 2005) und ist nicht signifikant verändert für erhöhten Stimmaufwand (Jessen et al., 2005). Dies zeigt, dass die Variation zwar in absoluten Werten gemessen steigt. Relativ zum erhöhtem Mittelwert betrachtet, ist für erhöhten Stimmaufwand hingegen keine signifikante Veränderung der Variation sichtbar.

Zusammenfassend kann festgehalten werden, dass eine **Erhöhung der F_0 im Kontext erhöhten Stimmaufwands** auftritt. Dies wurde in zahlreichen Studien nachgewiesen.

2.4.2 Formanten

Formanten sind Resonanzen des Vokaltrakts, welche für die Wahrnehmung der Vokalqualität maßgeblich sind (Pompino-Marschall, 2003). Sowohl in der automatischen als auch in der forensischen Sprechererkennung spielen die Formanten eine zentrale Rolle. Aus diesem Grund wird in diesem Abschnitt der Einfluss erhöhten Stimmaufwands auf die ersten drei Formanten dargestellt. Obwohl gerade höhere Formanten sprecherspezifische Information enthalten, werden sie hier, wie in den meisten anderen Studien, nicht berücksichtigt, da Sprechererkennung meist auf Telefonsprache durchgeführt wird. Es sind nur die ersten drei Formanten in Telefonsprache relevant, da Formanten in Abständen von ca. 1 kHz auftreten, sodass Telefonsprache lediglich die ersten drei bis vier Formanten enthält (Vary, Heute & Hess, 1998).

Zunächst werden die Ergebnisse zahlreicher Studien zur Veränderung des *ersten Formanten F1* in Tabelle 2.2 dargestellt.

Studie	Parameter	Bedingung	Ergebnis
Liénard und Di Benedetto (1999)	F1	isoliert gesprochene Vokale, drei Distanzen zur Erzeugung des erhöhten Stimmaufwands	steigt signifikant
Pisoni und Yuchtman (1985); Pisoni, Bernacki, Nusbaum und Yuchtman (1985)	F1	isoliert gesprochene Worte, Lombard	Wenn eine Veränderung stattfindet, dann steigt F1.
Stanton et al. (1988)	F1	laut (a. A.) + Lombard+ normal, jeweils mit Sauerstoffmaske	steigt für laute und Lombardsprache
Schulman (1985a)	F1	gelesene Sprache, laut (a. A.), mind. 90 dB	steigt um ca. 100 Hz
Geumann (2001)	F1	gelesene Sprache, laut (a. A.)	steigt
Barfs (2005)	Mittelwert	gelesene Sprache, Vokale, a. A. geschrien	steigt signifikant
Bond et al. (1989)	Mittenfrequenz	isoliert gesprochene Worte, Lombard	steigt
		isoliert gesprochene Worte, Lombard + Sauerstoffmaske	sinkt bei Hochvokalen, steigt bei Tiefvokalen
Garnier et al. (2008)	R1, erste Vokaltraktresonanz	isoliert gesprochene Vokale erhöhten Stimmaufwands mit den Anweisungen: 1. keine, 2. F_0 konstant halten, 3. F_0 und Artikulatoren konstant halten	steigt für 1. signifikant (für untermittelhohe Vokale mehr als für offene) und geringfügig für 2. und 3.

Tabelle 2.2: Untersuchungen zum ersten Formanten

Bei dem Vergleich der Studien zeigt sich, dass der erste Formant tendenziell bei erhöhtem Stimmaufwand steigt. Nicht alle Studien finden jedoch eine signifikante Steigerung dieses Formanten. Eine Ausnahme bildet die Studie von Bond et al. (1989), in der F1 bei Hochvokalen sinkt und bei Tiefvokalen steigt. In dieser Studie

wurde zusätzlich zum Stimmaufwand der Einfluss einer Sauerstoffmaske untersucht. Das Absinken von F1 bei Hochvokalen wurde durch die Sauerstoffmaske verursacht.

Die Ergebnisse verschiedener Studien bezüglich der Veränderung des *zweiten Formanten F2* sind in Tabelle 2.3 dargestellt.

Studie	Parameter	Bedingung	Ergebnis
Pisoni und Yuchtman (1985)	F2	isoliert gesprochene Worte, Lombard	sinkt für Vorderzungenvokale, steigt für Hinterzungenvokale
Pisoni et al. (1985)	F2	isoliert gesprochene Worte, Lombard	wenn F2 signifikant verändert ist, dann sinkt F2
Stanton et al. (1988)	F2	laut (a. A.) + Lombard + normal, jeweils mit Sauerstoffmaske	individuelle Veränderungen
Schulman (1985a)	F2	gelesene Sprache, laut (a. A.), mind. 90 dB	ist relativ unverändert
Liénard und Di Benedetto (1999)	F2	isoliert gesprochene Vokale, drei Distanzen zur Erzeugung des erhöhten Stimmaufwands	keine signifikante Änderung
Geumann (2001)	F2	gelesene Sprache, laut (a. A.)	steigt für Vokal /a:/
Barfs (2005)	Mittelwert	gelesene Sprache, Vokale, geschrien (a. A.)	steigt signifikant für Hinterzungenvokale, Vorderzungenvokale sind nicht signifikant verändert, Ausnahme: Voderzungenvokal /e/: F2 sinkt signifikant ab
Bond et al. (1989)	Mittenfrequenz	isoliert gesprochene Worte, Lombard	je nach Vokal unterschiedlich: steigt für /u/, sinkt für /i/, unverändert für /a/ und /æ/
		isoliert gesprochene Worte, Lombard + Sauerstoffmaske	gleich oder leicht abgesenkt

Garnier et al. (2008)	R2, zweite Vokaltraktresonanz	isoliert gesprochene Vokale erhöhten Stimmaufwands mit den Anweisungen: 1. keine, 2. F_0 konstant halten, 3. F_0 und Artikulatoren konstant halten	keine signifikante Veränderung über alle Vokale bei sämtlichen Modi, für 1. gilt: Hinterzugenvokale steigen signifikant, Vorderzungenvokale sinken signifikant

Tabelle 2.3: Untersuchungen zum zweiten Formanten

Für F2 lässt sich kein eindeutiges Ergebnis festhalten. Einige Studien finden keinerlei signifikante Veränderungen (Liénard & Di Benedetto, 1999). Andere stellen fest, dass F2 bei Vorderzungenvokalen sinkt oder nicht verändert ist und bei Hinterzungenvokalen steigt (Barfs, 2005; Pisoni & Yuchtman, 1985). Manche geben an, dass, sofern eine signifikante Veränderung vorliegt, F2 absinkt (Pisoni et al., 1985). Obwohl die oben zusammengefassten Studien nicht alle die gleichen Ausgangsbedingungen haben, sollte, sofern es für F2 klare Ergebnisse gäbe, auch ein entsprechender Trend in den Studien absehbar sein. Da dies nicht der Fall ist, muss F2 im Einzelfall näher untersucht werden.

Die Ergebnisse für den *dritten Formanten F3* sind in Tabelle 2.4 abzulesen.

Studie	Parameter	Bedingung	Ergebnis
Stanton et al. (1988)	F3	laut (a. A.) + Lombard + normal, jeweils mit Sauerstoffmaske	individuelle Veränderungen
Schulman (1985a)	F3	gelesene Sprache, laut (a. A.), mind. 90 dB	ist relativ unverändert
Liénard und Di Benedetto (1999)	F3	isoliert gesprochene Vokale, drei Distanzen zur Erzeugung des erhöhten Stimmaufwands	keine signifikante Änderung
Geumann (2001)	F3	gelesene Sprache, laut (a. A.)	wird zentralisiert zu Werten um 2600 Hz

Tabelle 2.4: Untersuchungen zum dritten Formanten

Für F3 ist ebenfalls kein klares Muster erkennbar. F3 wird durch erhöhten Stimmaufwand nicht signifikant beeinflusst. Eine Studie (Geumann, 2001) berichtet aber von einer Zentralisierung zu 2600 Hz hin. Eine andere Interpretation wäre hier, dass F3 für Vokale mit hoher oder obermittelhoher Zungenposition gesenkt und für

tiefe Vokale erhöht wird. Zur Bestätigung dieser Interpretation müssten allerdings weitere Vokale untersucht werden.

Abschließend lässt sich festhalten, dass **die Erhöhung des Stimmaufwands Einfluss auf die Formanten hat**. Besonders der erste Formant ist stark betroffen. Er steigt, in den meisten Studien signifikant, an. F2 und F3 sind weniger beeinflusst und es lässt sich für diese zwei Formanten kein eindeutiger Trend nachweisen.

2.4.3 Spektrum

Das *Spektrum* eines Sprachsignals dient der Bestimmung der verschiedenen Komponenten des Signals. „Etwas verallgemeinert nennt man die Bestimmung geeignet gewählter Grundkomponenten aus einem Gesamtsignal Spektralanalyse, die Überlagerung der Anteile zu einem Gesamtsignal Spektralsynthese“ (Vary et al., 1998, S. 61). Die Bestimmung der einzelnen Komponenten erfolgt mit Hilfe der Fouriertransformation (Details hierzu sind in Vary et al. (1998) zu finden).

Mit einem Spektrum kann die Energie eines Signals in verschiedenen Frequenzbereichen bestimmt werden. Da sich die Energieverteilung bei verändertem Stimmaufwand ebenfalls ändert, sind die spektralen Parameter für die Bestimmung des Stimmaufwandsgrades und das Finden geeigneter stabiler Merkmale sprachverarbeitender Systeme bei veränderlichem Stimmaufwand wesentlich. In den folgenden Abschnitten werden einige spektrale Parameter eingeführt und ihre Veränderung bei erhöhtem Stimmaufwand dargestellt. Hierbei ist zu beachten, dass in der Literatur für die beschriebenen spektralen Parameter nicht immer die gleichen Definitionen genutzt werden. Diese Unterschiede werden in den nachfolgenden Abschnitten beschrieben.

Spektrale Neigung Die *spektrale Neigung (spectral tilt)* gibt Auskunft über die Energieverteilung in einem Sprachsignal. Als eine geeignete Maßeinheit kann die Steigung der Regressionsgeraden des Spektrums verwendet werden (Schneider & Möbius, 2007; van Summers et al., 1988; Hansen & Varadarajan, 2009). Hierbei wird je nach Untersuchung beispielsweise ein geeigneter Frame pro Laut oder der Mittelwert sämtlicher Frames zur Berechnung der spektralen Neigung ausgewählt.

Abgesehen von der spektralen Neigung gibt es weitere Parameter, welche ähnliche Informationen über das Spektrum liefern. Die Namen dieser Parameter werden in der Literatur häufig synonym zum Begriff spektrale Neigung verwendet, obwohl die Definition für die Berechnung der Parameter variieren kann. Solche Parameter sind beispielsweise *spektrale Balance (spectral balance)* (Sluijter & Heuven, 1996) und *spektrale Betonung (spectral emphasis)* (Heldner, Strangert & Deschamps, 1999). Diese Parameter werden in der vorliegenden Arbeit nicht berücksichtigt, da die spektrale Neigung sehr ähnliche beziehungsweise gleiche Information enthält.

Ein weiterer, der spektralen Neigung ähnlicher Parameter, ist die *Differenz der korrigierten Amplitude der ersten Harmonischen H1* und der korrigierten Amplitude des dritten Formanten A3**. Abbildung 2.1 zeigt die *nicht korrigierte Amplitude der ersten Harmonischen H1* und die *nicht korrigierte Amplitude des dritten Formanten A3* im Spektrum des Vokals /ə/. Die Differenz H1*-A3* gibt Auskunft darüber, wie abrupt die Stimmbänder geschlossen werden (Hanson & Chuang, 1999). H1*-A3* ist damit ein Maß für die *glottale Verschlussrate (rate of closure, RC)* und dient als ein

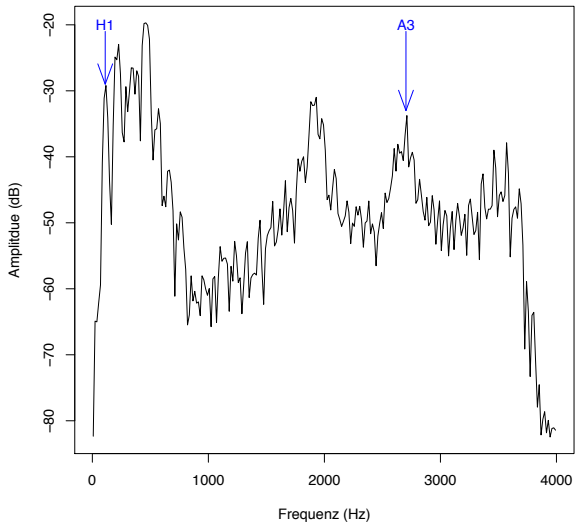

Abbildung 2.1: Spektrum des Vokals /ə/, gesprochen mit normalem Stimmaufwand (Die erste Harmonische (H1) und die Amplitude des dritten Formanten (A3) sind markiert.)

Indikator für die spektrale Neigung. Die Korrektur der Amplituden bezieht sich auf den Einfluss des Vokaltrakts. H1 wird bezüglich des Effekts des ersten Formanten korrigiert, während A3 hinsichtlich des Einflusses von F1 und F2 auf F3 korrigiert wird. Untersuchungen, welche H1*-A3* als Indikator für spektrale Neigung nutzen, sind beispielsweise (Iseli, 2007; Iseli, Shue & Alwan, 2007; Hanson & Chuang, 1999; Hanson, 1997). Wird die spektrale Neigung als Steigung der Regressionsgeraden definiert, so kann die glottale Verschlussrate als von der spektralen Neigung separater Parameter betrachtet werden (Schneider & Möbius, 2007).

Die vorliegende Arbeit definiert spektrale Neigung als die Steigung der Regressionsgeraden des Spektrums.

Eine Veränderung der spektralen Neigung kann ein Indiz für Veränderungen verschiedener Parameter, wie beispielsweise der Betonung oder des Stimmaufwands, sein. Diese Parameter sind voneinander abhängig. Es ist beobachtet worden, dass betonte Silben mit mehr Stimmaufwand produziert werden als unbetonte (Sluijter & Heuven, 1996). Bei erhöhtem Stimmaufwand flacht die spektrale Neigung ab. Dies ist in Abbildung 2.2 zu sehen. Die blaue Regressionsgerade des Spektrums mit erhöhtem Stimmaufwand ist flacher als die des Spektrums mit normalem Stimmaufwand. Bei der Berechnung der spektralen Neigung auf Telefondaten ist der Abfall des Spektrums ab 3400 Hz zu bedenken. Die Berechnung sollte sich folglich auf die von Filter unbeeinflussten Bereiche beschränken.

Ein Vorteil der spektralen Neigung gegenüber anderen Parametern der Lautstärke, wie beispielsweise der Gesamtintensität, ist, dass die spektrale Neigung nicht so leicht durch Umgebungsfaktoren beeinflusst wird. Sie ist dementsprechend robuster als die Gesamtintensität (Sluijter & Heuven, 1996, S.2472).

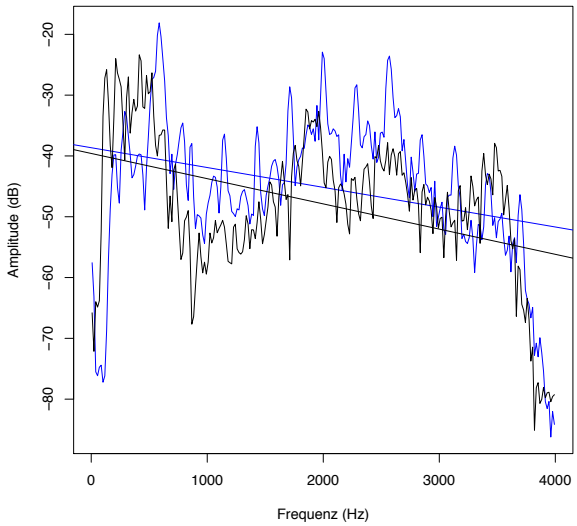

Abbildung 2.2: Spektrum des Vokals /ə/, gesprochen mit normalem (schwarz) und erhöhtem (blau) Stimmaufwand (Die zugehörigen Regressionsgeraden der beiden Spektren zur Berechnung der spektralen Neigung sind ebenfalls eingezeichnet.)

Gewichteter spektraler Schwerpunkt Der *gewichtete spektrale Schwerpunkt (center of gravity, COG)* ist, genau wie die spektrale Neigung, ein Maß für die Energieverteilung im Sprachsignal. Der spektrale Schwerpunkt kann auch als gewichtete, mittlere Frequenz des Spektrums betrachtet werden. Er wird je nach Untersuchung mit der spektralen Balance gleichgesetzt (siehe beispielsweise (van Son & van Santen, 2005)). In der vorliegenden Studie wird er dagegen unabhängig von der spektralen Balance betrachtet. Der spektrale Schwerpunkt kann entweder als Quotient zweier Integrale oder numerisch definiert werden.

Als Definition für die vorliegende Arbeit wird die numerische Definition nach van Son und Pols (1999) verwendet:

$$COG = \sum(f_i \cdot E_i) / \sum E_i. \tag{2.1}$$

f_i steht hierbei für die Frequenz in Hertz des i-ten DFT-Punkts und E_i für die spektrale Energie als Funktion der Frequenz.

Kienast und Sendlmeier (2000) verwenden dieselbe Definition für die spektrale Balance.

Der gewichtete spektrale Schwerpunkt wird, genau wie die spektrale Neigung, stark durch den Stimmaufwand eines Sprechers beeinflusst. Steigt der Stimmaufwand, so steigt auch der spektrale Schwerpunkt (Junqua, 1993). Dies ist in Abbildung 2.3 dargestellt. Der Unterschied zwischen dem spektralen Schwerpunkt für normalen und erhöhten Stimmaufwand beträgt in diesem Fall 600 Hz.

Energieverhältnis Das *Energieverhältnis* ist ein Parameter, der eingeführt wurde, um normale Phonation von geflüsterter Sprache ohne Phonation unterscheiden

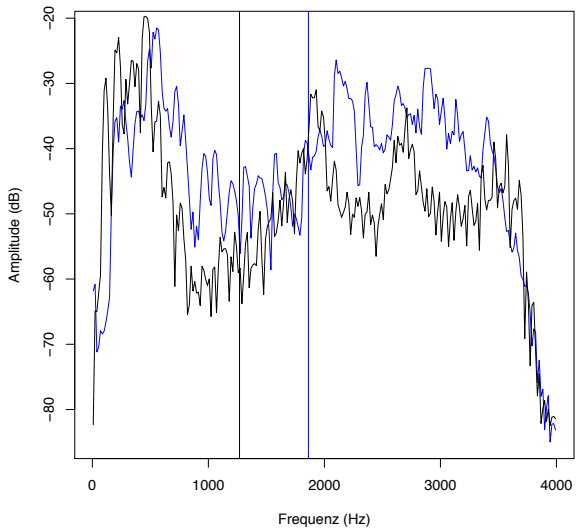

Abbildung 2.3: Spektrum des Vokals /ə/ mit normalem (schwarz) und erhöhtem (blau) Stimmaufwand artikuliert von dem gleichen Sprecher (Der gewichtete spektrale Schwerpunkt ist als vertikale Linie für beide Spektren markiert.)

zu können. Ausgehend von einer Untersuchung zur Bestimmung der mittleren Differenz im Betragsspektrum normaler und geflüsterter Sprache (Wenndt, Cupples & Floyd, 2002), wird für die Berechnung des Energieverhältnisses ein hohes Frequenzband ausgewählt (2800 bis 3000 Hz), bei dem der Unterschied zwischen den Beträgen etwa der durchschnittlichen Differenz über das gesamte Spektrum entspricht. Weiterhin wird ein niedriges Frequenzband ausgewählt (450 bis 600 Hz), welches den größten Unterschied zwischen den Beträgen normaler und geflüsterter Sprache aufweist. Das Energieverhältnis ergibt sich dann aus dem Verhältnis der spektralen Energie des hohen Frequenzbandes zu der Energie des niedrigen Frequenzbandes. Diese Methode erzielt bei der Klassifikation normaler und geflüsterter Sprache für Sprachsignale unterschiedlicher Signal-zu-Rausch-Abstände sehr gute Ergebnisse (Wenndt et al., 2002). Auch in der Klassifikation leiser, lauter und geschriener Sprache ist dieser Parameter nutzbar. Zhang und Hansen (2008a) zeigen, dass das Energieverhältnis zur Stimmaufwandswechseldetektion vier verschiedener Stimmaufwandsgrade geeignet ist.

Da das Energieverhältnis speziell für die Unterscheidung normaler und geflüsterter Sprache entwickelt wurde, empfiehlt es sich, eine ähnliche Untersuchung wie Wenndt et al. (2002) auch für laute und normale Sprache durchzuführen. Möglicherweise finden sich für diese zwei Stimmaufwandsgrade Frequenzbereiche, die zur Berechnung des Energieverhältnisses besser geeignet sind. Dies wird in Abschnitt 6.1 überprüft.

Als vorläufige Definition (bis Abschnitt 6.1) wird das Energieverhältnis als das Verhältnis der Energie zwischen den Frequenzbereichen 2800 bis 3000 Hz zu 450 bis 600 Hz festgelegt.

Spektrale Momente Ähnlich wie die bisher beschriebenen spektralen Parameter beschreiben die *spektralen Momente* die Zusammensetzung der spektralen Verteilung. Es ist üblich, zur Beschreibung der spektralen Verteilung vier Momente zu berechnen. Das *erste Moment* entspricht hierbei dem arithmetischen Mittel der Verteilung und wird häufig mit dem gewichteten spektralen Schwerpunkt gleichgesetzt. Das *zweite Moment* entspricht der Varianz, das *dritte* der Schiefe und das *vierte* der Kurtosis. Das dritte Moment wird häufig mit der spektralen Neigung gleichgesetzt. In dieser Arbeit wird hingegen explizit zwischen den Momenten und den Parametern spektraler Schwerpunkt und spektrale Neigung unterschieden (siehe oben).

Die Definition der spektralen Momente erfolgt auf Grundlage der Arbeit von Forrest, Weismer, Milenkovic und Dougall (1988). Für das erste spektrale Moment ergibt sich, bei einer 512 Punkte FFT (Fast Fourier Transform), folgende Berechnungsgrundlage:

$$L_1 = \sum_{k=1}^{256} f_k \cdot p(k). \tag{2.2}$$

Die Variable $p(k)$ entspricht dem normalisiertem Leistungsspektrum und ergibt sich aus den Werten des Leistungsspektrum $P(k)$. Es gilt:

$$p(k) = P(k) / \sum_{k=1}^{256} P(k). \tag{2.3}$$

Die Summe aller $p(k)$-Werte ergibt 1. Die Frequenz f_k in Gleichung 2.2 ergibt sich aus:

$$f_k = f_s \cdot k/512, \tag{2.4}$$

wobei f_s für die Abtastrate steht.

Das zweite bis vierte spektrale Moment berechnet sich als n-tes Moment um den Mittelwert, also um L_1. Nach Forrest et al. (1988) gilt:

$$L_n = \sum_{k=1}^{256} [(f_k - L_1)^n \cdot p(k)]. \tag{2.5}$$

Die dimensionslosen Momentenkoeffizienten der Schiefe und Kurtosis berechnen sich über $l_3 = L_3/(L_2)^{3/2}$ und $l_4 = [L_4/(L_2)^2] - 3$. In dieser Arbeit werden mit spektralen Momenten L_1, L_2, l_3 und l_4 bezeichnet.

Die Veränderung der spektralen Momente bei erhöhtem Stimmaufwand wurde ausführlich in Gottsmann (2010) sowie Gottsmann et al. (Gottsmann & Harwardt, 2011; Harwardt, Gottsmann & Noubours, 2011)[3] untersucht. Die Studien zeigen, dass das erste Moment über alle Lautklassen bei erhöhtem Stimmaufwand steigt. Ein signifikanter Anstieg ist für die Vokale, Frikative und Sonoranten, nicht jedoch für die Plosive festzustellen.

Für das zweite Moment ist ebenfalls für Vokale, Frikative und Sonoranten eine signifikante Erhöhung festzustellen, während für Plosive keine höheren Werte vorliegen.

[3]Sämtliche eigenen Veröffentlichungen zu dissertationsrelevanten Themenbereichen sind im Anhang A.3 abgedruckt.

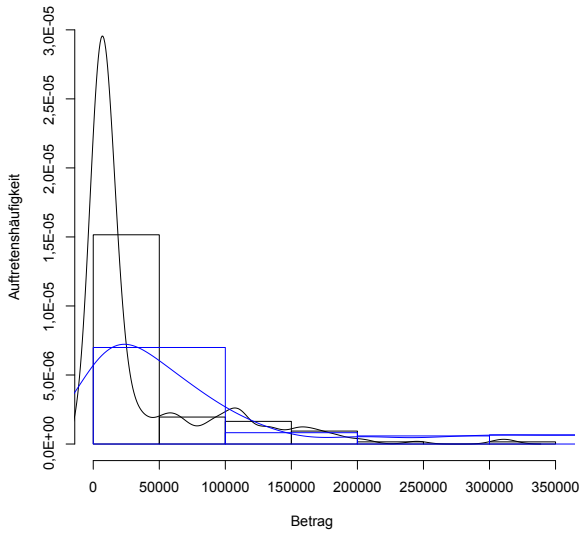

Abbildung 2.4: Spektrale Verteilung des Vokals /ə/ für normalen (schwarz) und erhöhten (blau) Stimmaufwand, artikuliert von dem selben Sprecher

Das dritte Moment ist bei normalem Stimmaufwand höher für Vokale, Frikative und Sonoranten. Signifikant ist dieser Unterschied allerdings nur für Vokale und Sonoranten. Das größere dritte Moment normaler Sprache entspricht einer stärkeren Rechtsschiefe der Verteilung für normale Sprache. Für die Plosive sind die Werte erhöhten Stimmaufwands größer. Dieser Unterschied ist allerdings nicht signifikant.

Auch das vierte Moment zeigt signifikante Veränderungen für Vokale und Sonoranten. Für beide Lautklassen liegt eine größere Steilheit der Kurven bei normalem Stimmaufwand vor. Für Frikative und Plosive kann keine signifikante Veränderung beobachtet werden.

Bei der sprecherspezifischen Untersuchung der Momente zeigt Gottsmann (2010), dass Frikative und Plosive kein einheitliches Muster aufweisen und die Veränderung des Stimmaufwands stark sprecherabhängig ist.

Ein Beispiel für die sich verändernde spektrale Verteilung bei normalem und erhöhtem Stimmaufwand ist in Abbildung 2.4 zu sehen. Bezogen auf die spektralen Momente zeigt es, dass die Verteilung für normalen Stimmaufwand eine wesentlich steilere Kurve aufweist. Außerdem ist der Mittelwert der Verteilung normalen Stimmaufwands etwas weiter links auf der x-Achse angesiedelt, sodass eine größere Rechtsschiefe vorliegt als für erhöhten Stimmaufwand.

Insgesamt zeigen die Untersuchungen zu den verschiedenen spektralen Parametern, wie auch die Illustrationen der Merkmale für normale und laute Sprache in diesem Abschnitt, dass **das Spektrum und seine Verteilung durch erhöhten Stimmaufwand verändert wird**. Grundsätzlich wird eine **Energiemigration von niedrigeren zu höheren Frequenzen** beobachtet (Stanton et al., 1988).

2.4.4 Amplitude und Intensität

Die *Amplitude* eines Sprachsignals entspricht dem maximalen Ausschlag einer Schwingung (Pompino-Marschall, 2003). Die *Intensität* wiederum ist der linguistische Parameter der Artikulationsenergie und entspricht perzeptiv der Lautstärke (Machelet). Diese beiden Parameter sind wesentlich bei der Betrachtung von Stimmaufwandsveränderungen, da sie mit der Lautstärke eines Sprachsignals assoziiert werden. Eine Erhöhung der Gesamtlautstärke eines Sprachsignals muss aber nicht zwangsläufig mit einer Erhöhung des Stimmaufwands einhergehen. Deswegen können Amplituden- oder Intensitätssteigerungen auch nicht direkt auf Stimmaufwandsveränderungen übertragen werden. Es zeigt sich jedoch in verschiedenen Studien, dass ein erhöhter Stimmaufwand immer eine Erhöhung der Amplitude und der Intensität mit sich bringt (Moore & Bond, 1987; Bond et al., 1989; Geumann, 2001; van Summers et al., 1988). Schulman (1989) ist sogar der Auffassung, dass ein Sprecher, der lauter sprechen möchte, als primäres Ziel die Erhöhung der Intensität verfolgt (S. 310).

Die Erhöhung der Intensität gilt nicht nur für erhöhten Stimmaufwand, sondern auch für geschriene Sprache. Barfs (2005) untersuchte verschiedene akustische Parameter in normaler und geschriener Sprache. Hierbei stellte sie nicht nur fest, dass sich die Gesamtintensität signifikant erhöht. Gleichzeitig verringert sich die Intensitätsspannweite hochsignifikant. Bei lauter oder geschriener Sprache ist es das Ziel des Sprechers, zu jedem Zeitpunkt eine möglichst hohe Intensität zu erlangen. Die Spannweite sinkt, da Sprecher bei einer stark erhöhten Intensität nicht mehr in der Lage sind, die Intensität stark zu variieren.

Genau wie Barfs (2005) untersuchte Geumann (2001) Vokale innerhalb gelesener Sprache. Geumann (2001) betrachtete hingegen keine geschriene, sondern auf Anweisung laut gesprochene Sprache. Auch in dieser Studie zeigte sich eine Erhöhung der Gesamtintensität bei erhöhtem Stimmaufwand.

Des Weiteren fanden van Summers et al. (1988) heraus, dass die Amplitude nicht nur zwischen normaler und lauter Sprache verändert wird. Vielmehr untersuchten sie verschiedene Geräuschbedingungen und ihre Auswirkungen auf unterschiedliche akustische Parameter zweier männlicher Sprecher. Es zeigte sich, dass bei jeder Steigerung der Geräusche (kein Geräusch, 80 dB, 90 dB, 100 dB weißes Rauschen) ein signifikanter Anstieg der Amplitude zu beobachten ist. Der größte Anstieg ist bei dem Wechsel von normaler Sprache ohne Hintergrundgeräusche auf Sprache mit 80 dB Hintergrundrauschen aufgetreten.

Abschließend kann festgehalten werden, dass **Amplitude und Intensität bei erhöhtem Stimmaufwand steigen**, jedoch nicht als Marker für Stimmaufwand verwendet werden können, da Amplitude und Intensität stark von wechselnden Signalgegebenheiten beeinflusst werden. Ein robusterer Marker für Stimmaufwand ist beispielsweise die spektrale Neigung (siehe Abschnitt 2.4.3) (Sluijter & Heuven, 1996).

2.4.5 Akustische Dauer

Zur *akustischen Dauer* der unterschiedlichen Laute und Lautklassen bei unterschiedlichen Stimmaufwandsgraden liegen ebenfalls zahlreiche Untersuchungen vor. Die Ergebnisse werden in diesem Abschnitt erläutert. Generell zeigt sich die Tendenz,

bei erhöhtem Stimmaufwand vokalische Laute zu längen und Konsonanten zu kürzen (Schulman, 1989; Junqua, 1993; Garnier, Bailly et al., 2006). Einige Ergebnisse weichen jedoch, abhängig von den Vorbedingungen der Studien, davon ab. Auch für geschriene Sprache können verlängerte Vokale und verkürzte Konsonanten nachgewiesen werden (Rostolland, 1982a). Die Wortdauer vergrößert sich bei erhöhtem Stimmaufwand ebenfalls in den meisten Untersuchungen (Schulman, 1989; van Summers et al., 1988; Junqua, 1993; Garnier, Bailly et al., 2006).

Eine Studie, welche diese Tendenzen bestätigt, ist die Untersuchung von Schulman (1989). Er untersuchte zwölf schwedische Vokale, jeweils eingebettet zwischen zwei stimmhaften bilabialen Plosiven im Kontext /i'b_b/. Er fand heraus, dass betonte Vokale gelängt und gleichzeitig intervokalische bilabiale Plosive verkürzt werden. Die Gesamtlänge der Segmente (/i'b_b/) steigt leicht an.

Ein etwas anderes Ergebnis erzielten Bond et al. (1989). Bezüglich der Wortdauer stellen sie ebenfalls eine Längung fest, welche allerdings nicht signifikant ist. Bei den untersuchten Wörtern handelte es sich um zweisilbige Wörter mit der Betonung auf der ersten Silbe. Da die Wörter isoliert gesprochen wurden, war die zweite Silbe immer die finale Phase. Bei der Auswertung der Vokallängen zeigt sich, dass Geräusche von 95 dB keinen Einfluss auf den Vokal der ersten Silbe ausüben, während der zweite Vokal gelängt wird. Eine signifikante Längung des Vokals der zweiten Silbe tritt ebenfalls beim Tragen einer Sauerstoffmaske (mit und ohne Hintergrundgeräusch) auf. Bei einer anderen Untersuchung über Sprache mit Sauerstoffmaske von Stanton et al. (1988) ergab sich kein konsistentes Muster für die Lautdauer.

In der Untersuchung von Andersson, Eriksson und Traunmüller (1996) wird die Veränderung des Stimmaufwands vom Flüstern zum Schreien beschrieben. Sie untersuchten hierfür fünf verschiedene Stimmaufwandsgrade, hervorgerufen durch fünf verschiedene Distanzen zwischen Sprecher und Hörer. Aus den Experimenten ergibt sich, dass die Dauer vokalartiger Segmente mit steigender Distanz ansteigt. Konsonanten hingegen werden bei Distanzen bis 7,5 m verkürzt. Bei größeren Distanzen werden die Konsonanten ebenfalls gelängt.

Insgesamt wird die **Längung der Vokale und der Wortdauer**, sowohl für Lombardsprache als auch für sonstige, laut artikulierte Sprache, durch zahlreiche Studien bestätigt. Die **Kürzung der Konsonanten** wird ebenfalls durch die meisten Studien bestätigt.

2.5 Zusammenhänge zwischen Artikulation, Perzeption und akustischen Merkmalen

Nachdem in den vorherigen Abschnitten der Einfluss veränderten Stimmaufwands auf artikulatorische, perzeptive und akustische Aspekte von Sprache beschrieben wurde, werden nun ihre Zusammenhänge verdeutlicht: *die drei genannten Aspekte bedingen einander*. Obwohl nur die akustischen Merkmale für den praktischen Teil dieser Arbeit relevant sind, werden auf Grund dieses Zusammenhangs hier auch die artikulatorischen und perzeptiven Veränderungen dargestellt, um die Komplexität der Sprache als Ganzes aufzuzeigen.

Eine Theorie, welche sich mit diesen Zusammenhängen befasst, ist die Quantaltheorie von Stevens (1989). Stevens beschreibt die Beziehung zwischen Artikulati-

on und Akustik als *quantal* in dem Sinne, dass die Veränderung eines akustischen Musters von einem Zustand in einen anderen durch eine Reihe artikulatorischer Veränderungen bedingt ist. Dies bedeutet, dass eine kleine akustische Veränderung durch eine große Anzahl artikulatorischer Veränderungen hervorgerufen werden kann. Die akustischen Merkmale sind in solchen Quantalbereichen relativ stabil gegenüber artikulatorischen Veränderungen. In einer Sprache werden häufig Phoneme aus unterschiedlichen Quantalbereichen ausgewählt, damit kleine Veränderungen in der Artikulation des Sprechers nicht direkt zu Problemen bei dem Hörer führen. Andersherum ist es auch möglich, dass kleine Modifikationen der Artikulation zu großen Veränderungen der akustischen Parameter führen (*Transition*). Die Zusammenhänge zwischen Artikulation und Akustik sind also nichtlinear. Auch für die Perzeption und die Akustik stellt Stevens ähnliche nichtlineare Zusammenhänge fest.

Im Kontext erhöhten Stimmaufwands bedeutet dies, dass die Sprache verändert wird, da der Sprecher sich, auf Grund von störenden Faktoren, seinem Hörer gegenüber besser verständlich machen möchte. Um die Perzeption zu verbessern, verändert der Sprecher den Artikulationsprozess. Die veränderte Artikulation führt zu Modifikationen der Akustik und somit zu einer Veränderung der Wahrnehmung. Wie in Abschnitt 2.3 gezeigt, bleibt die Güte der Perzeption trotz störender Einflüsse erhalten oder wird sogar, je nach Untersuchungsvoraussetzung, auf Grund der Anpassungen verbessert.

Der Zusammenhang zwischen den unterschiedlichen Aspekten von Sprache sowie zwischen verschiedenen akustischen Merkmalen ist in einigen Studien nachgewiesen worden. Schulman (1989) beispielsweise stellt verstärkte Artikulationsbewegungen mit einer größeren Geschwindigkeit der Artikulatoren fest. Diese erhöhte Geschwindigkeit führt zu kürzeren intervokalischen Plosiven (siehe S. 310). Die Vokale werden allerdings trotz der erhöhten Geschwindigkeit gelängt. Eine Studie, die sich ebenfalls mit artikulatorischen und akustischen Aspekten erhöhten Stimmaufwands befasst, wurde von Garnier, Bailly et al. (2006) durchgeführt. Garnier et al. stellen Korrelationen zwischen F_0 und der spektralen Energie mit der maximalen Amplitude der Lippenöffnung fest. Diese Korrelationen sind größer als die zwischen der maximalen Amplitude der Lippenöffnung und der Intensität, sodass nicht von einer Scheinkorrelation auszugehen ist.

Eine weitere Arbeit von Schulman (1985a) befasst sich mit den Zusammenhängen zwischen akustischen Merkmalen. Er stellt fest, dass das Verhältnis zwischen F_0 und F1 relativ gleich bleibt, da beide akustischen Parameter bei erhöhtem Stimmaufwand in ähnlichem Maß gesteigert werden. Der zweite und dritte Formant hingegen bleiben relativ unbeeinflusst, sodass sich das Verhältnis von F2 zu F1 oder F3 zu F1 verändert.

Die Untersuchung zu geschriener Sprache von Barfs (2005) hingegen zeigt, dass trotz der Veränderung des ersten Formanten und den unklaren Veränderungsmustern des zweiten Formanten, die Formantlagen nicht grundsätzlich verändert werden. Barfs stellt fest, dass die Formanten näher zusammenrücken; das grundlegende Muster bleibt jedoch erhalten.

Weitere Untersuchungen zu den Zusammenhängen unterschiedlicher akustischer Merkmale bei erhöhtem Stimmaufwand wurden von Liénard und Di Benedetto (1999) durchgeführt. Nachdem die Veränderungen verschiedener akustischer Parameter einzeln beschrieben wurden, werden die Differenzen F1-F_0, F2-F1, F3-F2

hinsichtlich ihrer Veränderungen bei erhöhtem Stimmaufwand untersucht. Der Hintergrund dieser Analyse ist, dass Liénard und Di Benedetto versuchen, durch diese Parameter die Konstanz in der Wahrnehmung der Vokale bei erhöhtem Stimmaufwand zu erklären. Sie gehen davon aus, dass diese Parameter, welche die Vokalhöhe und die Zungenposition repräsentieren, bei erhöhtem Stimmaufwand relativ stabil sind. Für die Differenz F1-F_0, welche die Vokalhöhe repräsentiert, werden allerdings signifikante Veränderungen bei veränderter Distanz zwischen Sprecher und Hörer festgestellt. Die Veränderungen sind jedoch nicht so groß wie für F1. In einer Korrelationsanalyse wiesen die Autoren einen großen Zusammenhang zwischen den Parametern F_0 und F1 nach. Es kann sich hierbei um eine Scheinkorrelation handeln, da beide Parameter mit der Amplitude in ähnlicher Weise korrelieren. Die anderen beiden Differenzen werden im Kontext der Repräsentativität für die Zungenposition bei der Vokalartikulation untersucht. Es stellt sich heraus, dass F3-F2 besser geeignet ist zur Repräsentation der Zungenposition, da keine signifikante Variation bei verändertem Stimmaufwand beobachtet werden konnte. Für F2-F1 hingegen wurden signifikante Variationen festgestellt. Diese Ergebnisse zeigen, dass die Differenz F3-F2 eine mögliche Erklärung für die gleichbleibende Leistung des Hörers bei verändertem Stimmaufwand ist. Insgesamt wird in dieser Studie, über die Zusammenhänge akustischer Parameter hinaus, das Zusammenspiel unterschiedlicher Aspekte von Sprache bei verändertem Stimmaufwand erarbeitet (Artikulation (Zungenposition), Akustik (F3-F2), Perzeption (gleichbleibend)).

Der Zusammenhang zwischen Perzeption und Akustik wird ebenfalls in der Studie von van Summers et al. (1988) untersucht. Hier wird festgestellt, dass die akustischen Unterschiede zwischen normaler Sprache und Lombardsprache mit der Güte der Perzeption zusammenhängen. Die Studie zeigt, dass laute Sprache, im Vergleich zu normaler Sprache mit gleichem Signal-zu-Rausch-Abstand, eine verbesserte Wahrnehmung auslöst. Diese verbesserte Wahrnehmung ist auf die Veränderungen der akustischen Parameter zurückzuführen. Van Summers et al. schlussfolgern aus ihren Untersuchungen, dass die akustischen Veränderungen bei sinkendem Signal-zu-Rausch-Abstand immer bedeutsamer werden (S. 925).

Die in diesem Abschnitt exemplarisch angeführten Studien zeigen, dass **Artikulation, Akustik und Perzeption stark zusammenhängen**. Deshalb wurde in diesem Kapitel ein Gesamtüberblick über sämtliche Aspekte der Sprache gegeben anstatt nur die Veränderung der akustischen Parameter, welche später im praktischen Teil aufgegriffen werden, zu berücksichtigen.

Kapitel 3

Einführung in die Sprechererkennung

In der vorliegenden Arbeit soll der Einfluss erhöhten Stimmaufwands auf spektrale Merkmale, F_0 und automatische Sprechererkennungssysteme untersucht werden. In diesem Kapitel werden die notwendigen Grundlagen für das später beschriebene automatische Sprechererkennungssystem vermittelt. Hierfür wird zunächst definiert, was Sprechererkennung ist und zwischen welchen Aufgaben unterschieden werden muss (Abschnitt 3.1). Nachfolgend werden die einzelnen Komponenten eines automatischen Sprechererkennungssystems vorgestellt (Abschnitt 3.2). In diesem Kontext werden der Detektor, die Vorverarbeitung und das verwendete statistische Modell dargestellt. In der Vorverarbeitung werden sowohl Standardmerkmale der Sprechererkennung als auch F_0-basierte Merkmale beschrieben. Neben diesen zwei Kategorien von Merkmalsarten ist eine Fülle anderer Merkmale und Merkmalsarten denkbar, wie beispielsweise lexikalische Merkmale oder Merkmale auf Basis akustischer Tokenisierung (Phonemerkennung). Für die vorliegende Arbeit sind diese aber nicht relevant, da lediglich unterschiedliche Standardverfahren und F_0-basierte Merkmale realisiert wurden.

3.1 Aufgabendefinition

Sprechererkennung bezeichnet die Bestimmung der Identität einer sprechenden Person (Reynolds & Campbell, 2008). Es wird grob zwischen zwei Hauptaufgaben unterschieden: der *Identifikation* und der *Verifikation*. Der Hauptunterschied dieser zwei Aufgabendefinitionen liegt in der Anzahl der jeweils gleichzeitig betrachteten Referenzsprecher. Referenzsprecher sind Sprecher, für die Audiomaterial zum Training eines Sprechermodells vorliegt. Diese Modelle werden dann gegen unbekannte Signale getestet. Bei der *Identifikationsaufgabe* wird geprüft, ob ein unbekanntes Sprachsignal von einem Sprecher aus der Menge der Referenzsprecher stammt. Die Leitfrage der Identifikationsaufgabe lautet folglich: Handelt es sich bei dem unbekannten Signal um ein Sprachsignal einer der Referenzsprecher? In der Identifikation ist weiterhin zwischen *open-set* und *closed-set* Identifikation zu differenzieren. Eine schematische Darstellung der Identifikation, inklusive der Unterscheidung zwischen open- und closed-set, findet sich in Abbildung 3.1. Die Diskrepanz zwischen closed-set und open-set Identifikation bezieht sich auf die vorab getätigte Annahme über die Identität des Sprechers gegeben ein unbekanntes Sprachsignal. *Closed-*

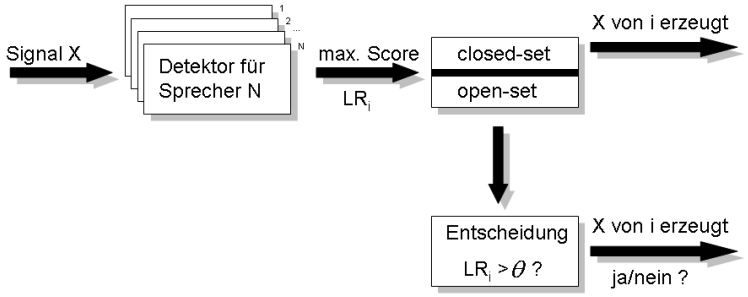

Abbildung 3.1: Identifikation

set Sprecheridentifikation wird die Untersuchungsbedingung genannt, bei der die Zuordnung der Sprecheridentität auf das Referenzsprecherset beschränkt ist. Das heißt, es ist bekannt, dass es sich bei dem Sprecher auf jeden Fall um einen der N Referenzsprecher handelt. Das System muss also lediglich den zu dem unbekannten Sprachsignal nächsten Sprecher aus dem Referenzsprecherset finden; also, wie in Abbildung 3.1 gezeigt, den Detektor[4] mit dem maximalen Score[5]. Eine Rückweisung des Signals, also die Zuordnung zu keinem der Referenzsprecher, ist in diesem Fall nicht möglich. Ist die Identität des unbekannten Sprachsignals vorab nicht auf das Referenzsprecherset festlegbar, so muss dagegen die Möglichkeit bestehen, das Signal zurückzuweisen und damit den Sprecher als keinen der Referenzsprecher zu klassifizieren. Diese Systemkonzeption wird *open-set* Sprecheridentifikation genannt, da die möglichen Zuordnungen über das Referenzsprecherset hinweg offen sind. Bei der open-set Identifikation wird, ähnlich der closed-set Identifikation, zunächst der Referenzsprecher mit der größten Ähnlichkeit zu dem unbekannten Signal gesucht. Dieser wird dann jedoch nicht automatisch als Urheber des Signals identifiziert. Viel mehr folgt eine Klassifikation anhand des Scores mit Hilfe eines vorab festgelegten Grenzwertes θ. Dieser Grenzwert ist in der open-set Aufgabe notwendig, nicht hingegen in der closed-set Aufgabe.

Die Ausgangsannahme für eine *Verifikationsaufgabe* ist die Zugehörigkeit eines Sprachsignals zu genau einem Referenzsprecher. Diese Annahme wird durch das Verifikationssystem überprüft. Das Konzept der Verifikation wird in Abbildung 3.2 dargestellt. In der Verifikation wird zunächst der Score für die Zugehörigkeit des

Abbildung 3.2: Verifikation

Signals zu dem gegebenen Referenzsprechermodell berechnet. Dann wird, wie in der open-set Identifikation, eine Klassifikation anhand des Scores mit Hilfe eines vorher festgelegten Grenzwertes θ vorgenommen. Auch in der Verifikation können

[4]Mit Detektor ist ein für einen Referenzsprecher trainierter Klassifikator gemeint, wie beispielsweise der Likelihood-Ratio-Detektor, der in Abschnitt 3.2.1 beschrieben wird.

[5]Mit Score ist der Wert gemeint, der die Zugehörigkeit von X zu S quantifiziert. Diese wird in Abschnitt 3.2.1 näher erläutert.

mehrere Referenzsprechermodelle vorhanden sein. Diese werden immer separat und nicht vergleichend im Erkennungsprozess genutzt.

Zusätzlich zu der Identifikation und Verifikation sind in der automatischen Sprechererkennung weitere, weniger häufig verwendete Aufgabendefinitionen, wie beispielsweise die Sprecherwechseldetektion vorhanden. Die Sprecherwechseldetektion wird in Abschnitt 3.2 erläutert.

3.2 Automatische Sprechererkennung

Die automatische Sprechererkennung dient der automatischen Klassifikation eines unbekannten Sprachsignals hinsichtlich des Sprechers. Sie lässt sich grob in zwei Arten von Systemen teilen: in die *textabhängigen* und die *textunabhängigen*. *Textabhängige* Systeme setzen einen kooperativen Sprecher voraus, der einen vorgegebenen Text beziehungsweise ein Passwort spricht. Diese Art von System wird in Zugangskontrollsystemen, wie beispielsweise beim Telefonbanking, verwendet. Bei textabhängigen Verfahren können Modelle verwendet werden, die den Inhalt der Äußerung auf diversen linguistischen Ebenen mit einbeziehen. Aus diesem Grund erzielen textabhängige Systeme meist bessere Erkennungsergebnisse im Vergleich zu textunabhängigen Systemkonzeptionen. Ein Nachteil der Textabhängigkeit ist die leichte Möglichkeit zum Missbrauch eines solchen Systems. Häufig besteht die Option, an Stelle eines direkt sprechenden Nutzers, gezielte Aufzeichnungen des Zielsprechers abzuspielen. Eine Möglichkeit, diesen Missbrauch zu umgehen, besteht darin, dass der Sprecher vorab nicht weiß, welchen Text er sprechen muss. Stattdessen bekommt er erst während der Nutzung die so genannten Textprompts angezeigt. *Textunabhängige* Systeme stellen keine Anforderungen an den Sprecher. Auch ein nicht-kooperativer Sprecher kann getestet werden. Dementsprechend sind solche Systeme für eine Vielzahl von Applikationen einsetzbar, wie beispielsweise:

- in der Audio-Indexierung von Besprechungen,
- in sicherheitsrelevanten Bereichen zur Überwachung,
- in der Forensik,
- als Teilkomponente in Applikationen zum Multimediaretrival oder
- in sprecherabhängigen Dialogsystemen.

Die Modelle, die in der textunabhängigen Sprechererkennung verwendet werden, unterscheiden sich von denen der textabhängigen Erkennung. Sie liefern im Vergleich meist weniger gute Ergebnisse, da keine Information über den gesprochenen Text vorliegt. Um ähnliche Modelle wie in der textabhängigen Sprechererkennung nutzen zu können, muss das Audiomaterial zunächst mit Hilfe eines Spracherkenners verschriftet werden. In diesem Fall kommt der Spracherkenner als zusätzliche Fehlerquelle hinzu. Auch bei dieser Art der Erkennung ist Missbrauch möglich, jedoch gestaltet dieser sich nicht so einfach wie in einem textabhängigen System. Der größte Vorteil eines solchen Systems ist die Vielseitigkeit hinsichtlich möglicher Einsatzszenarien. Im Rahmen dieser Arbeit wird eine textunabhängige Systemkonzeption gewählt, um einen möglichst vielseitigen Einsatz zu garantieren.

Aus den verschiedenen Applikationen der Sprechererkennung folgen verschiedene Systemaufgaben. Die Aufgaben Verifikation und Identifikation sind bereits in der allgemeinen Einführung in Abschnitt 3.1 erläutert worden. Eine weitere denkbare Aufgabendefinitionen ist die *Sprecherwechseldetektion*. Die Sprecherwechseldetektion markiert Sprecherwechsel innerhalb eines Audiosignals und annotiert außerdem Signale, die von dem gleichen Sprecher artikuliert wurden. In der vorliegenden Arbeit wird die Realisierung eines *Sprecherverifikationssystems* beschrieben. Daher werden die anderen Aufgaben hier nicht weiter erörtert.

Der typische Aufbau eines Sprecherverifikationssystems ist in Abbildung 3.3 zu sehen. Für die Nutzung eines Sprechererkennungssystems müssen zwei Phasen

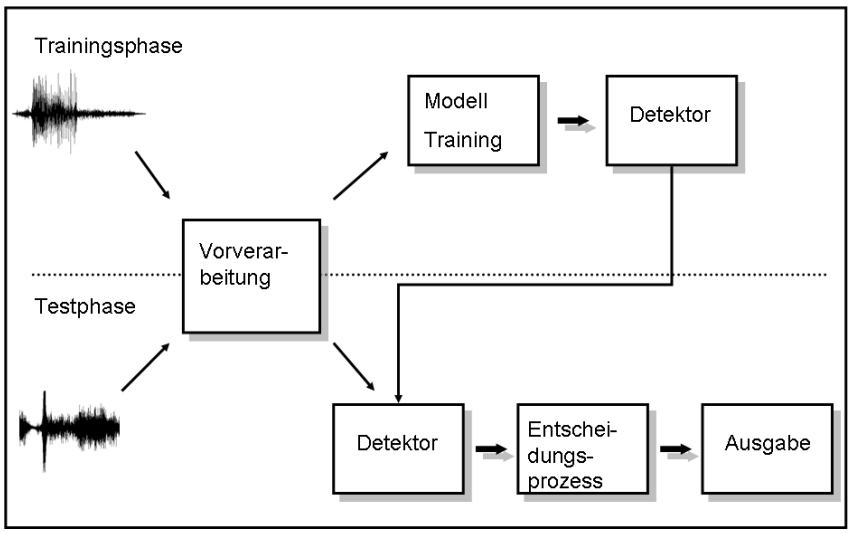

Abbildung 3.3: Systemaufbau

unterschieden werden, die *Trainings-* und die *Testphase*. Die *Gemeinsamkeit beider Phasen* ist die initiale Vorverarbeitung. In der Vorverarbeitungsphase wird zunächst eine Sprache-Pause-Detektion durchgeführt. Aus den Signalteilen, die Sprache enthalten, werden dann sprecherspezifische Merkmale extrahiert (siehe Abschnitt 3.2.2). Als nächster Schritt der *Trainingsphase* folgt das Modelltraining. Hier wird unterschieden zwischen dem Erstellen eines Hintergrundmodells und dem Training von Sprechermodellen (siehe Abschnitt 3.2.3). Darauf folgt das Zusammenstellen eines Detektors aus diesen Modellen (siehe Abschnitt 3.2.1). In der *Testphase* wird der Detektor dann genutzt, um die Merkmalsvektoren des Testsignals, welche in der Vorverarbeitungsphase generiert wurden, mit dem Referenzsprechermodell und dem Hintergrundmodell zu vergleichen. Die Ausgabe des Detektors ist ein Score, welcher für den folgenden Entscheidungsprozess benötigt wird. Bei dem Entscheidungsprozess handelt es sich meist um den Vergleich des Scores mit einem vorher festgelegten Grenzwert. Der Score und die zugehörige Entscheidung bilden für Systeme mit Entscheidungsprozess die Ausgabe. In den nachfolgenden Abschnitten wird zunächst der Likelihood-Ratio-Detektor als Framework der einzelnen Komponenten eines textunabhängigen Sprechererkennungssystems vorgestellt, um dann die einzelnen Komponenten sowie die statistischen Grundlagen näher zu erläutern. Die Entscheidungsprozedur wird nicht behandelt, da sie nicht in dem System, welches für diese Arbeit implementiert wurde, enthalten ist.

3.2.1 Likelihood-Ratio-Detektor

Der *Likelihood-Ratio-Detektor* verdeutlicht das Grundkonzept der automatischen Sprecherverifikation. Deswegen wird diese Systemkomponente zuerst dargestellt. Zur Erläuterung des Detektors werden die Begriffe *Likelihood* und *Wahrscheinlichkeit* verwendet. Diese Begriffe sind voneinander zu differenzieren.

Für die Wahrscheinlichkeit gibt es in der Literatur unterschiedliche Konzepte. Die Ausführungen dieses Abschnitts orientieren sich an den Erläuterungen zur Likelihood von Edwards (1972, S.9). Die Grundlage für statistisches Schlussfolgern ergibt sich aus dem Triplett bestehend aus dem Wahrscheinlichkeitsmodell, einem Set statistischer Hypothesen und den Daten. Die *Wahrscheinlichkeit* $p(R|H)$ist definiert als die Wahrscheinlichkeit, dass das Resultat R erzielt wird unter der Bedingung der Hypothese H und dem gegebenen statistischen Modell. $p(R|H)$ kann als Funktion beider Variablen betrachtet werden, meistens wird es jedoch als Funktion von R genutzt. R ist hierbei als Variable zu verstehen und H als Konstante.

Die Likelihood $L(H|R)$ der Hypothese H für gegebene Daten R und ein gegebenes Modell ist proportional zur Wahrscheinlichkeit $p(R|H)$. (Definition nach Edwards (1972, S. 9)).

Der grundlegende Unterschied zur Wahrscheinlichkeit besteht darin, dass die Hypothese H die Variable ist und die Daten R die Konstante. Für die Wahrscheinlichkeit verhält es sich genau anders herum. Dies bedeutet, dass bei der Wahrscheinlichkeit die Hypothese H festgelegt ist und die Beobachtungen R variieren, während bei der Likelihood die Beobachtungen R festgelegt sind und die Hypothese H gesucht wird. Deswegen tritt in Parameterschätzverfahren, wie beispielsweise der Maximum-Likelihood-Methode, der Begriff Likelihood anstatt dem Begriff der Wahrscheinlichkeit auf. Weiterhin muss die Summe der Likelihoods verschiedener Hypothesen bei gleichen Daten und gleichem Modell nicht Eins ergeben.

Bei der Sprechererkennung soll die Wahrscheinlichkeit dafür bestimmt werden, dass ein gegebenes Sprachsignal Y durch einen Referenzsprecher S artikuliert wurde. Dabei müssen zwei Hypothesen in Betracht gezogen werden (Reynolds, Quatieri & Dunn, 2000):

- H_0: Y wurde von Sprecher S erzeugt,
- H_1: Y wurde nicht von Sprecher S erzeugt.

Die Wahrscheinlichkeitsdichtefunktionen der zwei Hypothesen ($p(Y|H_i), i = 0, 1$) werden zur Berechnung der *Likelihood-Ratio LR* verwendet. Die Wahrscheinlichkeitsdichtefunktion einer Hypothese H_i wird auch Likelihood der Hypothese H_i, gegeben das Sprachsignal Y, genannt.[6] Die Likelihood-Ratio ist definiert durch:

$$LR = \frac{p(Y|H_0)}{p(Y|H_1)}. \tag{3.1}$$

Ist der Wert der Likelihood-Ratio größer oder gleich dem Entscheidungsgrenzwert θ, so wird die Hypothese H_0 angenommen. Ist die Likelihood-Ratio kleiner, so wird

[6] Die Notation der Likelihood erfolgt in Reynolds et al. (2000) anders ($p(A|B)$) als in Edwards (1972) ($L(B|A)$). Nachfolgend wird die Notation von Reynolds et al. verwendet, da sich die weiteren Ausführungen auf die Studie von Reynolds et al. beziehen.

die H_0 zurückgewiesen. Der Entscheidungsgrenzwert muss für jede spezifische Applikation experimentell festgelegt werden. Bei der Festlegung eines Grenzwertes muss beachtet werden, dass zwischen den beiden möglichen Fehlern *fehlender Alarm* und *falscher Alarm* einer Verifikationsaufgabe eine Austauschbeziehung besteht. Dass heißt, bei der Zunahme des einen Fehlers nimmt der andere ab und umgekehrt. Diese Austauschbeziehung ist für die Darstellung und Auswertung von Evaluationsergebnissen wesentlich und wird, ebenso wie die Beschreibung der verschiedenen Fehler, in Abschnitt 8.2 aufgegriffen.

Die Likelihood-Ratio wird oft auch als *logarithmische Größe (Log-Likelihood-Ratio, LLR)* ausgedrückt. Die Hypothese H_0 kann mathematisch auch als Modell λ_{hyp} dargestellt werden. H_1 wird dann repräsentiert durch das Modell $\lambda_{\overline{hyp}}$, welches auch Hintergrundmodell genannt wird. Die mathematische Repräsentation der Merkmale des Sprachsignals Y wird durch die Vorverarbeitung erzielt, welche das Signal Y in eine Sequenz von Merkmalsvektoren X transformiert (siehe Abbildung 3.4). Für die Log-Likelihood-Ratio gilt dann:

$$\Lambda(X) = \log p(X|H_0) - \log p(X|H_1) = \log p(X|\lambda_{hyp}) - \log p(X|\lambda_{\overline{hyp}}). \qquad (3.2)$$

Allgemein lässt sich festhalten, dass die Unterscheidung zwischen den beiden Likelihoods der Modelle λ_{hyp} und $\lambda_{\overline{hyp}}$ elementar ist, da die Likelihoods den Unterschied zwischen Intra- und Intersprechervariabilität widerspiegeln. Das Modell λ_{hyp} prüft, ob das Sprachsignal, unter Berücksichtigung der Intrasprechervariabilität, von Sprecher S erzeugt wurde. Das Modell $\lambda_{\overline{hyp}}$ verifiziert die Möglichkeit, ob das gegebene Sprachsignal einem beliebigen anderen Sprecher, unter Einbezug der Intersprechervariabilität, zugeordnet werden kann. Eine geeignete Modellierung, welche das Finden adäquater Likelihoods impliziert, ist eine der bedeutsamsten Aufgaben in der automatischen Sprechererkennung. Sie wird in Abschnitt 3.2.3 näher erläutert.

Die verschiedenen *Komponenten eines Likelihood-Ratio-Detektors* werden in Abbildung 3.4 so dargestellt, wie sie in den meisten Sprecherverifikationsapplikationen realisiert werden. Die Modelle λ_{hyp} und $\lambda_{\overline{hyp}}$ werden in dieser Darstellung als Sprechermodell und Hintergrundmodell umgesetzt, sie können jedoch je nach Systemkonzeption anders realisiert werden (siehe unten). Die Abbildung zeigt, dass vor dem Einsatz des Detektors zunächst die Vorverarbeitung erfolgt, welche die Extraktion sprecherspezifischer Information aus dem Sprachsignal sowie ihre Repräsentation in Form eines Merkmalsvektors umfasst (siehe Abschnitt 3.2.2). Diese Merkmalsvektoren werden zum Training der Modelle verwendet. Das *Sprechermodell* λ_{hyp} repräsentiert die Sprache des Sprechers S mit den zugehörigen Intrasprechervariationen. Die Zusammenstellung eines Sprechermodells ist verhältnismäßig einfach, da der Sprecher und das zugehörige Trainingsset genau definiert sind. Die Konzeption eines Modells $\lambda_{\overline{hyp}}$, welches die H_1 repräsentiert, ist dagegen komplexer.

Es gibt zwei Hauptansätze zur Erstellung von $\lambda_{\overline{hyp}}$ (Reynolds, 2002). Eine Möglichkeit ist die Verwendung von so genannten *Hintergrundsprechern* (auch *Likelihood-Ratio-Sets* oder *Cohorts* genannt). Bei den Hintergrundsprechern handelt es sich um eine Anzahl verschiedener Sprecher, die dem Referenzsprecher ähnlich, jedoch nicht mit ihm identisch sind. Für jeden dieser Hintergrundsprecher wird jeweils ein Hintergrundsprechermodell erstellt. Zur Berechnung der Likelihood $p(X|\lambda_{\overline{hyp}})$ für die H_1-Hypothese werden die Likelihoods für jedes einzelne Hintergrundsprecher-

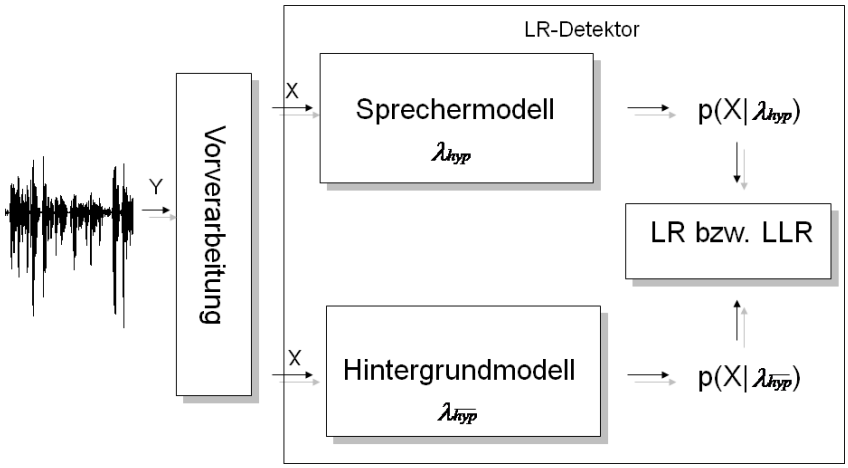

Abbildung 3.4: Likelihood-Ratio-Detektor

modell berechnet. Dann werden diese über eine Funktion, welche die Likelihoods der einzelnen Hintergrundsprechermodelle beispielsweise mittelt oder das Maximum ausgibt, zu einem Wert zusammengefasst. Dieser Ansatz ist besonders dann geeignet, wenn großer Aufwand für die Auswahl der Hintergrundsprecher betrieben werden kann. Durch die gezielte Auswahl von Hintergrundsprechern kann ein sprecherspezifisches Hintergrundsprecherset für jeden Referenzsprecher festgelegt werden. Dieser Prozess ist aufwändig und deswegen für viele Szenarien zu zeit- und kostenintensiv.

Der zweite Ansatz, die Verwendung eines *Hintergrundmodells* (*Universal Background Model*, *UBM*) beziehungsweise *Weltmodells*, ist für das gegebene Szenario wesentlich besser geeignet. Ein solches Hintergrundmodell enthält Sprache von möglichst vielen unterschiedlichen Sprechern in einem Modell. Auch in diesem Sprecherset darf der Referenzsprecher nicht enthalten sein. Die verwendeten Sprachdaten sollen eine bestimmte Sprecherpopulation repräsentieren. Dementsprechend erfolgt die Auswahl der Audiodaten an die Sprechererkennungsaufgabe angepasst. Dies bedeutet, dass für jedes Sprachsignal bedeutsame Charakteristika wie beispielsweise Kanalqualität, Sprechstil, Landessprache und Sprechsituation ähnlich beziehungsweise gleich denen des Trainingsmaterial des Referenzsprechers sind. Der Vorteil dieses Ansatzes ist, dass nur ein Hintergrundmodell für sämtliche Sprecher einer Erkennungsaufgabe trainiert werden muss. Trotzdem besteht, ähnlich dem ersten Ansatz, die Möglichkeit für verschiedene Sprecherpopulationen unterschiedliche Hintergrundmodelle zu trainieren.

Sowohl das Sprechermodell als auch das Hintergrundmodell werden detailliert in Abschnitt 3.2.3 beschrieben.

3.2.2 Vorverarbeitung

Die *Vorverarbeitung* in der Sprechererkennung umfasst den Prozess der *Sprache-Pause-Detektion*, die Anwendung von *Normalisierungsverfahren*, wie beispielsweise der Kanalnormalisierung, und die *Merkmalsextraktion*. Die *Sprache-Pause-Detektion* schließt Pausensegmente aus dem Audiomaterial vor der weiteren Verarbeitung aus. Sie kann beispielsweise mit Hilfe energiebasierter Verfahren durchgeführt werden

(Euler, 2006). Ein kurzer Überblick über aktuelle Verfahren sowie die Nutzung Gauß'scher Wahrscheinlichkeitsdichtefunktionen für die Sprache-Pause-Detektion ist in Górriz, Ramìrez, Lang und Puntonet (2008) nachzulesen. Die *Normalisierungsverfahren* in der Vorverarbeitungsphase dienen der Minimierung der Störeffekte über die gegebenen Sprachsignale. Die *Merkmalsextraktion* dient der Gewinnung der relevanten Merkmale aus dem Sprachsignal. Das Ergebnis ist eine Sequenz von Merkmalsvektoren für jedes Sprachsignal. Die Reduktion der Vektordimension ist ebenfalls im Prozess der Merkmalsextraktion enthalten. Zur Unterscheidung der zwei Prozesse wird häufig eine Unterscheidung zwischen *Merkmalsgewinnung* (Generierung der Merkmalsvektoren) und *Merkmalsextraktion* (Reduktion der Vektordimension) vorgenommen. In dieser Arbeit werden beide Prozesse unter dem Begriff Merkmalsextraktion zusammengefasst.

Um geeignete Merkmale aus dem Sprachsignal extrahieren zu können, müssen zunächst die optimalen Eigenschaften eines Merkmals definiert werden. Ein Merkmal sollte:

- eine hohe Intersprechervariation und
- eine niedrige Intrasprechervariation aufweisen,
- sollte leicht messbar sein,
- möglichst wenig Aufwand in der Berechnung benötigen,
- robust gegenüber Verstellung, Nachahmung und Krankheit sein,
- robust gegenüber Verzerrung und Hintergrundgeräusch sein sowie
- maximal unabhängig von den anderen Merkmalen sein.

Es wurde bisher noch kein Merkmal gefunden, dass alle diese Eigenschaften optimal erfüllt. Die bisher bekannten Merkmale lassen sich, gemäß der zur Gewinnung der Merkmale verwendeten *Informationsebenen*, in zwei Gruppen unterteilen; in die *Merkmale höherer Informationsebenen* (*High-Level Merkmale*) und die *Merkmale niedrigerer Informationsebenen* (*Low-Level Merkmale*). Der Unterschied zwischen diesen zwei Arten von Merkmalen ist in Abbildung 3.5 dargestellt. Die High-Level Merkmale und ihre Eigenschaften sind auf der rechten Seite dargestellt, während die Low-Level Merkmale und ihre Eigenschaften auf der linken Seite zu finden sind. Die *Merkmale niedrigerer Informationsebenen* beruhen auf akustischen Informationen aus dem Sprachsignal. Ihnen liegen physikalische Eigenschaften des Sprechers, nämlich anatomische Gegebenheiten, wie beispielsweise Größe und Länge des Vokaltrakts, des Nasaltrakts und der Stimmlippen zu Grunde. Die Segmentlänge, welche zur Berechnung der Merkmale verwendet wird, ist sehr kurz (meistens circa 20 bis 25 ms).

High-Level Merkmale hingegen beziehen sich häufig auf linguistische Informationen (Semantik, Analyse der Dialogstruktur). Je mehr linguistische Informationen zur Gewinnung der Merkmale verwendet werden, desto schwieriger wird die Extraktion der Merkmale. Die High-Level Merkmale repräsentieren persönliche Charakteristika des Sprechers, wie beispielsweise die regionale Herkunft oder den sozialen Status. Die Berechnung dieser Merkmale erfolgt für gewöhnlich über längere Signalabschnitte, teilweise sogar über das komplette Signal. Für High-Level Merkmale

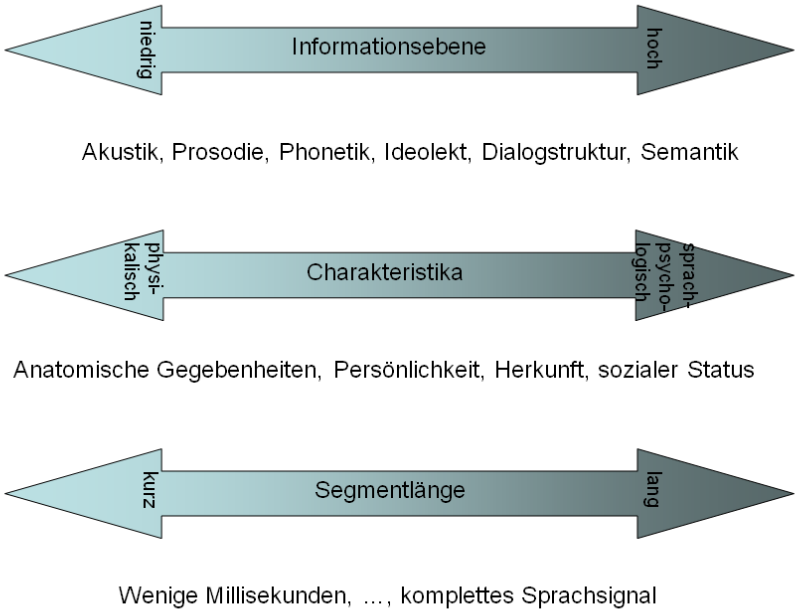

Abbildung 3.5: Darstellung der Informationsebenen für die Merkmalsextraktion mit den assoziierten Charakteristika und Segmentlängen.

können nicht nur linguistische, sondern auch akustische Eigenschaften des Signals verwendet werden. Voraussetzung ist hierbei, dass die betrachtete Segmentlänge höher ist als bei Low-Level Merkmalen und dass dynamische Eigenschaften erfasst werden. Shriberg (2007)) definiert High-Level Merkmale beziehungsweise Merkmale höherer Informationsebenen als solche Merkmale, die entweder linguistische Informationen nutzen oder sich auf Segmentlängen beziehen, die länger als die für Low-Level Merkmale üblichen 20 bis 25 ms sind. Die Pfeile in Abbildung 3.5 zeigen, dass die Übergänge zwischen den zwei Merkmalstypen fließend sind. Grundsätzlich gilt: je höher die genutzte Informationsebene ist, desto schwieriger wird die korrekte Extraktion der Merkmale. Eine Kombination von Low-Level und High-Level Merkmalen ist häufig hilfreich, da beide Merkmalsarten komplementäre Information enthalten.

Die Merkmalsextraktion geht der Nutzung des Likelihood-Ratio-Detektors voran. Hier gilt für die Sprecherverifikation die Voraussetzung, dass das betrachtete Sprachsignal, und damit die extrahierten Merkmalsvektoren, nur Sprache von einem Sprecher enthält. Der Fall, dass mehrere Sprecher in einem Signal enthalten sind, gehört in den Aufgabenbereich der Sprecherwechseldetektion und wird hier nicht weiter diskutiert.

In den folgenden Abschnitten werden hauptsächlich Low-Level Merkmale beschrieben. Hierbei werden zunächst die Standardmerkmale der Sprach- und Sprechererkennung vorgestellt; um anschließend F_0-basierte Merkmale einzuführen. Bei der Betrachtung der akustischen Messungen der Grundfrequenz zu einzelnen kurzen Zeitpunkten handelt es sich ebenfalls um Low-Level Merkmale. Im Rahmen F_0-basierter Merkmale geht Abschnitt 3.2.2.2 sowohl auf solche Low-Level Merkmale als auch auf High-Level Merkmale ein.

3.2.2.1 Standardmerkmale

Die gängigsten *Standardmerkmale* der Sprach- und Sprechererkennung sind die *Mel-Cepstrum-Koeffizienten* (*Mel-Frequency Cepstrum Coefficients*, *MFCC*). Außerdem werden häufig die Merkmale der *Linearen Prädiktiven Codierung* (*LPC*) und der *Perzeptuellen Linearen Prädiktion* (*PLP*) genutzt. Die *lineare Prädiktion* (*LP*) stellt eine Alternative zur *Diskreten Fourier Transformation (DFT)* dar und beruht auf dem Quelle-Filter-Modell (Fant, 1960). Bei der linearen Prädiktion wird von den vorherigen Signalwerten auf die nachfolgenden geschlossen. Die perzeptuelle lineare Prädiktion verknüpft Methoden der linearen Prädiktion mit der diskreten Fourier Transformation und ist damit das Verbindungsstück zwischen MFCC- und LPC-Merkmalen. Eine detaillierte Darstellung der LPC- und PLP-Merkmale ist in Makhoul (1975) und Hermansky (1990) zu finden. Die MFCC-Merkmale sollen im Folgenden etwas genauer beschrieben werden, da diese Merkmale für das Basissystem verwendet werden, während die beiden anderen nur in einem Vergleichstest (siehe Abschnitt 8.3) angewendet werden.

Die *MFCC-Merkmale* gehören, ebenso wie die LPC- und PLP-Merkmale, zu den Low-Level Merkmalen und sind die am häufigsten genutzten Merkmale in der Sprach- und Sprechererkennung. Die zu Grunde liegende Idee, das Sprachsignal nicht im Zeitbereich, sondern im Frequenzbereich zu betrachten, ist physiologisch motiviert. Zur Transformation des Sprachsignals vom Zeitbereich in den Frequenzbereich wird die diskrete Fourier Transformation verwendet. Sie ist nur für periodische Signale anwendbar. Deswegen werden die Sprachsignale in kleine, überlappende Fenster eingeteilt, sodass das Signal in dem betrachteten Fenster als stationär angenommen werden kann. An den Enden dieser Fenster können Abschneidefehler (Sprungstellen) entstehen. Um diese Abschneidefehler zu vermeiden, werden Fensterfunktionen verwendet, welche eine Gewichtung vornehmen. Eine häufig verwendete Fensterfunktion ist das Hamming-Fenster, welches die Frequenzen zu den Enden der Fenster hin dämpft. Eine detailliertere Beschreibung der diskreten Fourier Transformation ist in Vary et al. (1998) zu finden.

Auch bei der *Generierung der MFCC-Merkmale* wird zunächst eine *diskrete Fourier Transformation* durchgeführt. Die resultierenden Spektren werden *logarithmiert*. Ausgehend von dem Quelle-Filter Modell von Fant kann das Sprachsignal im Zeitbereich durch eine Multiplikation von Anregungssignal und Impulsantwort dargestellt werden. Die Extraktion der Impulsantwort ist das Ziel der Berechnung der MFCC-Vektoren, da die Impulsantwort für die Formung des Vokaltrakts und damit für die Formantstruktur steht. Die Anregung, welche für die von der Glottis erzeugte Grundfrequenz steht, soll nicht für die MFCC-Merkmale verwendet werden. Die oben genannte Logarithmierung führt zu einem additiven, statt dem vorher vorliegenden multiplikativen Zusammenhang und vereinfacht damit die Trennung zwischen den verschiedenen, im Signal enthaltenen Anteilen. Durch diesen additiven Zusammenhang können beispielsweise leicht Einflüsse der Aufnahmegeräte entfernt werden, wie in der cepstralen Subtraktion. Die Anwendung der *inversen diskreten Fourier Transformation* (*IDFT*) auf das logarithmierte Spektrum führt zu einer Erhaltung des additiven Zusammenhangs. Diese Transformation ist eigentlich eine Rücktransformation in den Zeitbereich. Zur Abgrenzung vom ursprünglichen Sprachsignal im Zeitbereich wird dieser neue Raum cepstraler Bereich genannt. Der Begriff *Cepstrum* und die zugehörige Einheit *Quefrency* entsprechen den Wörtern *Spectrum* und *Frequency*, nur mit einer entsprechenden Vertauschung der Buchsta-

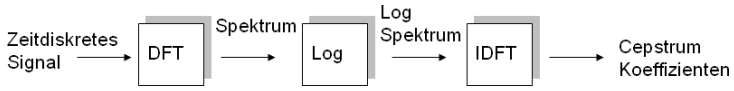

Abbildung 3.6: Cepstrumanalyse

ben (Bogert, Healy & Tukey, 1963). Dies verbildlicht, dass die Transformation in den neuen Raum aus dem spektralen Bereich erfolgt ist und deswegen auch die Benennungen aus den Bezeichnungen des spektralen Bereichs hervorgehen. Das Resultat ist ein *komplexes Cepstrum*. Für die Sprachsignalverarbeitung ist hingegen *das reelle Cepstrum* relevant, welches durch die Logarithmierung der *Betragsspektren* gewonnen wird. Weiterhin wird statt des komplexen Logarithmus der *reelle Logarithmus* gebildet. Dann erst wird die inverse diskrete Fourier Transformation durchführt. Das Cepstrum ist somit das Spektrum des logarithmierten Spektrums (Noll, 1964). Bei der ausschließlichen Betrachtung der Betragsspektren kann statt der inversen diskreten Fourier Transformation die *diskrete Kosinustransformation (DCT)* zur Gewinnung der Cepstrum-Koeffizienten eingesetzt werden.

Eine Übersicht über die notwendigen Schritte zur Generierung von Cepstrum-Koeffizienten wird in Abbildung 3.6 gegeben. Das ursprüngliche Ziel, nämlich die Trennung von Anregung und Impulsantwort, ist im Cepstrum gut sichtbar, da die Anregung stimmhafter Signale als Gipfel (Maxima) erscheinen. Die Anwendung eines Tiefpassfilters auf das Cepstrum wird *Liftering* genannt. Beim Liftering wird der Bereich unterhalb des ersten cepstralen Gipfels ausgeschnitten, sodass nur noch ein Gipfel der Anregungsfunktion in diesem Ausschnitt vorhanden ist. Dieser Gipfel erscheint nach der Transformation in den Spektralbereich als Konstante und ist für die weitere Verarbeitung nicht störend. Das Liftering erfolgt nach der Berechnung des Cepstrums durch die inverse diskrete Fourier Transformation.

Sollen statt der Cepstrum-Koeffizienten die Mel-Cepstrum-Koeffizienten berechnet werden, so muss das Cepstrum aus den Spektren der Mel-gewichteten Signalausschnitte bestimmt werden. Die *Mel-Gewichtung* erfolgt mit Hilfe der Mel-Skala, welche die menschliche Hörempfindung nachbildet.[7] Anschließend folgt wie bei den einfachen Cepstrum-Koeffizienten eine diskrete Kosinustransformation.

Eine ausführliche Beschreibung der oben beschriebenen Verfahren zur Gewinnung der cepstralen Koeffizienten und der Mel-Cepstrum-Koeffizienten ist in Vary et al. (1998) und Schukat-Talamazzini (1995) zu finden. Zusammengefasst sind die folgenden Schritte notwendig zur Generierung der Mel-Cepstrum-Koeffizienten:

- Diskrete Fourier Transformation zur Transformation in den Frequenzbereich,
- Gewichtung der Frequenzachse nach der Mel-Skala,
- Logarithmierung des Betragsspektrums zur Auflösung des multiplikativen Zusammenhangs von Anregung und Impulsantwort,
- inverse diskrete Fourier Transformation zur Transformation in den cepstralen Bereich.

[7] Das menschliche Gehör bildet so genannte Frequenzgruppen, welche größere Frequenzbereiche zusammenfassen. Diese Frequenzgruppen, beziehungsweise Frequenzbänder haben unter 500 Hz dieselbe Breite (ca. 100 Hz). Höhere Frequenzen werden nicht so gut wahrgenommen, sodass die Breite der Frequenzbänder ungefähr proportional mit der Frequenz steigt. Details hierzu sind in Zwicker (1982) nachzulesen.

3.2.2.2 F_0-basierte Merkmale

Die eben vorgestellten MFCC bilden die Charakteristika der Vokaltrakts ab. Die Information über die Grundfrequenz F_0 wird absichtlich entfernt. Diese Information kann jedoch für die Sprechererkennung zusätzlichen Nutzen zu den MFCC liefern. Es gibt zahlreiche Ansätze und Möglichkeiten die *Grundfrequenz als Merkmal in der Sprechererkennung* einzusetzen. Eine Auswahl davon wird im Folgenden dargestellt. Hierbei ist zwischen Merkmalen *für die automatische Sprechererkennung* und der Nutzung der Grundfrequenz *im forensischen Kontext* sowie zwischen *Low-Level* und *High-Level Merkmalen* zu unterscheiden. Die beschriebenen Low-Level Merkmale sind im Rahmen dieser Arbeit verwendet worden. Die Realisierung wird in Abschnitt 8.4 erläutert. Um dem Leser einen Überblick über die verschiedenen F_0-basierten Merkmale zu liefern, werden nachfolgend auch High-Level Merkmale beschrieben, welche jedoch nicht umgesetzt wurden.

Eine Standardlösung zur Realisierung eines *F_0-basierten Low-Level Sprechererkennungssystems* ist die Messung der *logarithmierten F_0* (*logF_0*) pro Frame. Die logF_0 wird dann mit der *logarithmierten Energie* (*logE*) kombiniert. Von beiden Werten wird die erste Ableitung gebildet, sodass ein vierdimensionaler Merkmalsvektor entsteht. Die Nutzung eines solchen Systems wird von Reynolds et al. (2002) vorgestellt.

Eine in der *forensischen Phonetik* angewendete Standardmethode ist die Berechnung von *F_0-Statistiken* für ein gegebenes Sprachsignal. Von Rose (2002) werden als Momente der F_0-Verteilung, die für den Vergleich zweier Sprachproben in Betracht gezogen werden, der *Mittelwert*, die *Standardabweichung*, die *Schiefe*, die *Kurtosis* und der *Modalwert* genannt. Der Modalwert ist der in der Verteilung am häufigsten vorkommende Wert. Dies bedeutet, dass der Modalwert nicht in jeder Verteilung vorhanden sein muss, da es auch mehrere gleichhäufig vorkommende Werte in einer Verteilung geben kann. Aus diesem Grund ist der Modalwert nicht für die Nutzung in automatischen Systemen geeignet. Die anderen Merkmale können hingegen zur automatischen Klassifikation verwendet werden. Laut Shriberg (2007) zählt die F_0-Statistik nicht zu den High-Level Merkmalen, da sie sich zwar auf die komplette Sprachprobe bezieht, jedoch keine dynamische Information enthält. Untersuchungen zu unterschiedlich zusammengesetzten F_0-Statistiken in der Sprechererkennung wurden beispielsweise von Kinnunen und González Hautamäki (2005), Labutin, Koval und Raev (2007) sowie Becker und Kreuzer (2008) beschrieben.

Eine Möglichkeit zur Nutzung von *F_0 als High-Level Merkmal* ist die Erfassung prosodischer Verläufe zur Verifikation des Sprechers. Einen guten Überblick über prosodische Merkmale vermittelt Shriberg (2007). Die stückweise lineare Stilisierung der Grundfrequenzkontur bietet eine Möglichkeit zur Beschreibung prosodischer Characteristika. Sie wird von Sönmez, Shriberg, Heck und Weintraub (1998) beschrieben. Die Autoren führen hierbei zunächst eine Medianfilterung durch und entfernen verdoppelte beziehungsweise halbierte F_0-Werte, um Probleme des F_0-Trackers zu beheben. Anschließend wird ein stückweise lineares Modell an die F_0-Kontur angepasst. Die stilisierten Konturen werden dann verwendet, um die Merkmale für die Sprechererkennung zu extrahieren. Aus den einzelnen Abschnitten der Konturen wird pro Segment der Median, die Steigung und die Dauer extrahiert. Dies führt zu einer weitreichenden Datenreduktion, da nicht pro Frame ein Vektor vorhanden ist. Des Weiteren wird über die stimmhaften Regionen und die Pausenabschnitte jeweils pro Segment die Dauer bestimmt. Die Merkmalsvektoren werden dann durch unterschiedliche Verteilungen modelliert, welche anschließend im Testmodus zum

Vergleich genutzt werden. Eine andere Variante der Modellierung der F_0-Dynamik wird in Reynolds et al. (2002) präsentiert. Auch hier werden stilisierte Grundfrequenzkonturen angewendet. Es werden jedoch nicht die oben genannten Parameter extrahiert, sondern lediglich bestimmt, ob die betrachtete Gerade steigt oder fällt (+,-) beziehungsweise ob der betrachtete Abschnitt der Sprachprobe stimmlos ist (uv — unvoiced). Das gleiche Verfahren wird für die Energiekontur angewendet. Die Modellierung erfolgt mit einem einfachen n-gram Modell. In einem weiteren Experiment wurde dieses System um die Information der Segmentlänge (S — short, M — medium, L — long) erweitert. Durch die Nutzung der Segmentlänge konnte eine gute Verbesserung erzielt werden. Diese Verbesserung konnte durch Einbezug des Phonemkontexts weiter gesteigert werden (Reynolds et al., 2002). Die Fusion des GMM-logF_0/E-Systems mit dem n-gram-basierten System inklusive Steigung, Segmentdauer und Phonemkontext erzielte eine weitere Steigerung der Systemleistung (Reynolds et al., 2002).

3.2.3 Klassifikation mit Hilfe Gauß'scher Mischverteilungsmodelle

In Abschnitt 3.2.1 haben wir den statistischen Rahmen für ein Sprecherverifikationssystem kennengelernt. Die Modellierung der H_0- und H_1-Hypothesen ist hingegen noch offen geblieben und soll in diesem Abschnitt allgemein für das Modell λ und die Likelihood $p(X|\lambda)$ dargestellt werden. Hierfür werden zunächst grundlegende Begriffe der Statistik erläutert, um dann die einzelnen Komponenten eines *Sprecherverifikationssystems, welches Gauß'sche Mischverteilungsmodelle und ein Hintergrundmodell nutzt (GMM-UBM-basiertes System)*, darzustellen.

Kovarianz und Kovarianzmatrix Die *Kovarianz (Cov)* ist eine Maßzahl zur Beschreibung linearer Zusammenhänge zweier statistischer Zufallsvariablen X und Y. Sie wird berechnet durch:

$$Cov(X,Y) = E((X - E(X))(Y - E(Y))) = \sigma_{xy} = \sigma_{yx} = Cov(Y,X), \qquad (3.3)$$

wobei E für Erwartungswert ($E(X) = \mu_X$) steht (Sachs, 2002). Allgemein kann festgehalten werden, dass eine positive Kovarianz auf einen gleichsinnigen linearen Zusammenhang der betrachteten Variablen hinweist (hohe X-Werte führen zu hohen Y-Werten, beziehungsweise niedrige X- zu niedrigen Y-Werten), während eine negative Kovarianz auf einen gegensinnigen linearen Zusammenhang (hohe X-Werte führen zu niedrigen Y-Werten, beziehungsweise niedrige X- zu hohen Y-Werten) hindeutet. Für Cov(X,Y) = 0 besteht kein linearer Zusammenhang zwischen X und Y. Für die Kovarianz Cov(X,X) gilt:

$$Cov(X,X) = Var(X). \qquad (3.4)$$

Die Kovarianz gibt an, ob ein linearer Zusammenhang zwischen den Variablen vorliegt, und welcher Art (positiv oder negativ) dieser gegebenenfalls ist. Eine Aussage über die Stärke des Zusammenhangs wird nicht gemacht. Hierfür ist der *Korrelationskoeffizient* nutzbar, welcher als Normierung der Kovarianz angesehen werden kann (Bortz, 1993).

Die *Kovarianzmatrix (**Cov**)* enthält paarweise Kovarianzen der Elemente einer Variable Z, welche aus einer Sequenz statistischer Zufallsvariablen besteht ($Z = \{Z_1, ..., Z_n\}$). Z ist wie folgt aufgebaut:

$$\begin{aligned}\mathbf{Cov}(Z) &= \begin{pmatrix} Cov(Z_1, Z_1) & ... & Cov(Z_1, Z_n) \\ ... & Cov(Z_i, Z_i) & ... \\ Cov(Z_n, Z_1) & ... & Cov(Z_n, Z_n) \end{pmatrix} \\ &= \begin{pmatrix} Var(Z_1) & ... & Cov(Z_1, Z_n) \\ ... & Var(Z_i) & ... \\ Cov(Z_n, Z_1) & ... & Var(Z_n) \end{pmatrix}\end{aligned} \tag{3.5}$$

Wie aus der obigen Formel ersichtlich, ist auf der Diagonalen die Varianz der jeweils betrachteten Zufallsvariable eingetragen. Die Matrix ist an dieser Diagonalen gespiegelt. Ein Sonderfall der Kovarianzmatrix ist die diagonale Kovarianzmatrix, welche lediglich die n Elemente der Diagonalen, also die Varianzen der n Zufallsvariablen enthält.

Der Expectation-Maximation-Algorithmus Der *Expectation-Maximation-Algorithmus (EM-Algorithmus)* wird verwendet, wenn die vorliegenden Daten eines Zufallsexperiments als unvollständig angesehen werden. Hierbei ist es irrelevant, ob die Daten tatsächlich fehlen oder ob die fehlenden Daten lediglich als Erweiterung der vorhandenen Daten dienen. Wären sämtliche Daten bekannt, so könnte die *Maximum-Likelihood-Methode (ML-Methode)* angewendet werden. Um das Problem der unvollständigen Daten auf die ML-Methode herunter zu brechen, wird der EM-Algorithmus verwendet. Gesucht sind die Parameter, welche die optimale Verteilungsfunktion für die vorliegenden und die gesuchten Daten charakterisieren. Handelt es sich bei der Verteilung beispielsweise um eine Normalverteilung, so sind die gesuchten Parameter der Mittelwert μ und die Varianz σ^2. Es ergeben sich folgende Schritte:

1. Initialisierung der gesuchten Parameter.
2. E-Schritt: Schätzung der gesuchten Parameter anhand der vorliegenden Daten und der Endparameter aus Schritt 3 (beziehungsweise der Initialisierung).
3. M-Schritt: Maximierung der geschätzten Parameter hinsichtlich der Wahrscheinlichkeit für das Eintreffen des Stichprobenergebnisses der vorliegenden Daten. Hier wird der ML-Algorithmus verwendet.

Schritt 2 und 3 werden so oft wiederholt, bis die geschätzten Parameter konvergieren, sich also nicht mehr wesentlich verändern.

Gauß'sche Mischverteilungsmodelle Zur *Modellierung der Likelihood* $p(X|\lambda)$ für eine Sequenz von Merkmalsvektoren X können *Gauß'sche Mischverteilungsmodelle* (*Gaussian Mixture Model - GMM*) verwendet werden (Reynolds et al., 2000). Eine Mischverteilung ist eine zusammengesetzte Verteilung, welche in diesem Fall aus M eingipfligen Gauß'schen Dichtefunktionen $p_i(\mathbf{x})$ besteht. Die Mischverteilung für einen Vektor $\mathbf{x}$ ist gegeben durch

$$p(\mathbf{x}|\lambda) = \sum_{i=1}^{M} w_i p_i(\mathbf{x}), \tag{3.6}$$

mit w_i als Gewicht der Dichtefunktion $p_i(\mathbf{x})$ und $\sum_{i=1}^{M} w_i = 1$. Für *eindimensionale Vektoren* $\mathbf{x}$ kann die Wahrscheinlichkeitsdichte der Normalverteilung für $p_i(\mathbf{x})$ verwendet werden:

$$p_i(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)\sigma^2}} e^{(-\frac{1}{2}\frac{(\mathbf{x}-\mu)^2}{\sigma^2})}. \tag{3.7}$$

Sie ist komplett charakterisiert durch den Mittelwert μ und die Varianz σ^2. Die Wahrscheinlichkeitsdichtefunktion für *Merkmalsvektoren der Dimension D* ist durch die Dichtefunktion der multivariaten Normalverteilung gegeben:

$$p_i(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^D |\mathbf{Cov}(i)|}} e^{(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T(\mathbf{Cov}(i))^{-1}(\mathbf{x}-\boldsymbol{\mu}_i))}. \tag{3.8}$$

Der Mittelwertvektor $\boldsymbol{\mu}_i$ hat die Dimension $D \times 1$, während die Kovarianzmatrix $\mathbf{Cov}(i)$ die Dimension $D \times D$ aufweist. Die Unterschiede zwischen den beiden Formeln 3.7 und 3.8 bestehen hauptsächlich in der Nutzung der Kovarianzmatrix statt der Varianz in Formel 3.8. Die Kovarianzmatrix wird deswegen verwendet, weil sie eine Verallgemeinerung der Varianz darstellt, welche nicht nur eindimensionale, sondern auch mehrdimensionale Variablen zulässt. Weiterhin ist in Formel 3.8 die Dimension D eingefügt. Die mehrdimensionale Gauß'sche Wahrscheinlichkeitsdichtefunktion wird komplett beschrieben durch das Modell

$$\lambda = \{w_i, \boldsymbol{\mu}_i, \mathbf{Cov}(i)\}. \tag{3.9}$$

In diesem Modell wird eine volle Kovarianzmatrix verwendet. In der Sprechererkennung ist dagegen die Verwendung einer diagonalen Kovarianzmatrix üblich, da die Rechenzeit wesentlich geringer ist und sie trotzdem bessere Ergebnisse liefert als die komplette Matrix (Reynolds et al., 2000). Zum Training eines GMMs mit einer Menge vorgegebener Trainingsvektoren wird der EM-Algorithmus verwendet (siehe oben). Für die Berechnung der *Log-Likelihood* eines Modells λ und die Reihe von Merkmalsvektoren $X = \{\mathbf{x}_1, ..., \mathbf{x}_T\}$ ergibt sich folgende Gleichung:

$$log p(X|\lambda) = \sum_{t=1}^{T} log p(\mathbf{x}_t|\lambda), \tag{3.10}$$

wobei T die Anzahl der Vektoren in X angibt und $p(\mathbf{x}_t|\lambda)$ gemäß Formel 3.6 berechnet wird. Eine Normalisierung der Log-Likelihood kann mittels der Division von $log p(X|\lambda)$ durch die Anzahl der Trainingsvektoren T erfolgen. Dieser normalisierte Wert wird *mittlere Log-Likelihood* (*average log-likelihood*) genannt.

Die Vorteile bei der Nutzung eines GMMs zur Modellierung von $log p(X|\lambda)$ ist die schnelle Rechenzeit sowie die Bekanntheit des vorliegenden statistischen Modells. In einem GMM werden keine temporalen Aspekte des akustischen Signals modelliert, sondern nur die vorliegende Verteilung der akustischen Ereignisse zu verschiedenen Zeitpunkten (vorgegeben durch die Frames und deren Überlappung). Diese temporalen Informationen können durch Merkmale höherer Informationsebenen erfasst werden. Aus diesem Grund ist die Kombination eines GMM-basierten Systems mit akustischen Merkmalen (beispielsweise MFCC-Merkmalen) mit einem oder mehreren High-Level Systemen sinnvoll (Reynolds et al., 2002).

Vorverarbeitung Die *Vorverarbeitung* umfasst in einem GMM-UBM-basierten Sprechererkennungssystem für gewöhnlich, wie in Abschnitt 3.2.2 bereits beschrieben, die Sprache-Pause-Detektion, die Normalisierung und die Merkmalsextraktion. Für die Merkmalsextraktion werden häufig die in Abschnitt 3.2.2.1 beschriebenen MFCC-Merkmale verwendet. Es sind aber grundsätzlich auch viele andere Merkmale, wie später gezeigt wird, denkbar.

Das Hintergrundmodell Das *Hintergrundmodell* ist die Komponente eines Sprecherverifikationssystems, welche für die Berechnung der Likelihood $p(X|\lambda)$ benötigt wird (siehe Abbildung 3.4). Wie bereits in Abschnitt 3.2.1 erläutert wurde, gibt es verschiedene Realisierungsmöglichkeiten des Hintergrundmodells. Die Ausführungen in diesem Abschnitt beschränken sich auf die Beschreibung der zweiten Version, also die Beschreibung der Realisierung eines *Hintergrund-* beziehungsweise *Weltmodells (Universal Background Model, UBM)*, welches die Sprache mehrerer Sprecher in einem Modell zusammenführt. Ein solches UBM modelliert sprecherunabhängig die Verteilung der Eigenschaften von Sprache im Allgemeinen oder die Verteilung der Sprache einer bestimmten Sprecherpopulation. Unter Sprecherpopulation ist eine Gruppe von Sprechern zu verstehen, die hinsichtlich bestimmter Parameter (wie beispielsweise Geschlecht, Übertragungskanal, Sprechsituation...) homogen ist.

Modelliert das UBM Sprache im Allgemeinen, so werden die Daten von verschiedenen Sprechern unterschiedlicher Sprecherpopulationen zusammengeführt und für das Training des UBMs verwendet. Die Modellierung des UBMs erfolgt mit einem GMM und das Training durch den EM-Algorithmus (siehe oben). Die Zusammensetzung der Trainingsdaten ergibt sich aus der Beschaffenheit der Testdaten, da beide Datensets möglichst ähnlich sein sollten.

Ist bekannt, dass die Testdaten aus einer bestimmten Sprecherpopulation stammen, so sollten nur Daten dieser Sprecherpopulation zum Training des UBMs verwendet werden. Die späteren Testdaten dürfen nicht im Trainingsset enthalten sein. Umfasst das Trainingsmaterial Daten verschiedener Sprecherpopulationen, so ist zu prüfen, ob von allen Populationen gleich viel Trainingsmaterial vorhanden ist. Ist dies nicht der Fall, kann das UBM unausgewogen sein. Um dies zu vermeiden, kann für jede Sprecherpopulation ein eigenes UBM trainiert werden. Die resultierenden UBMs werden dann zu einem Modell kombiniert. Hierbei ist zu beachten, dass die Anzahl Mischungskomponenten in den einzelnen Modellen der Sprecherpopulationen niedriger sein muss als die des Gesamt-UBMs. Die Summe der Anzahlen sämtlicher einzelner Modelle ergibt die Anzahl der Mischungskomponenten für das Gesamt-UBM. Weiterhin muss beachtet werden, dass die einzelnen Modelle nicht zu einem Modell zusammengeführt werden können, ohne vorab eine Normalisierung der Gewichte innerhalb der einzelnen Modelle vorzunehmen, sodass die Summe der Gewichte im Gesamt-UBM Eins ergibt.

Das Sprechermodell Das *Sprechermodell* soll die H_0 repräsentieren; also die Likelihood dafür, dass der Referenzsprecher das gegebene Signal Y erzeugt hat. Als Basis für das Sprechermodell wird das UBM verwendet, da dieses eine Vielzahl akustischer Ereignisse modelliert und mit ausreichend Daten trainiert wurde. Für das Sprechermodell liegen hingegen häufig nur wenig Daten vor. Das UBM wird mit Hilfe des *Maximum-A-Posteriori-Algorithmus (MAP-Algorithmus)* adaptiert. Während der Adaption werden die Merkmale des Referenzsprechers für akustischen

Ereignisse, welche im Trainingsmaterial auftauchen, im UBM aktualisiert. Der MAP-Algorithmus ist dem EM-Algorithmus sehr ähnlich. Er umfasst ebenfalls zwei Schritte. Der erste Schritt ist derselbe wie im EM-Algorithmus: Die gesuchten Parameter einer Verteilung werden geschätzt. Statt der Maximierung folgt darauf die Bestimmung der Endparameter durch eine Kombination der alten und neuen Parameter. Welche Parameter für die Endparameter verwendet werden, wird durch einen datenabhängigen Mischungskoeffizienten bestimmt. Dieser Koeffizient erlaubt eine separate Gewichtung der alten und neuen Parameter für jede Mischungskomponente. Die Gewichtung hängt ab von der Anzahl der Elemente des Trainingssignals für die betrachtete Mischungskomponente. Tritt ein akustisches Ereignis nicht im Trainingssignal des Referenzsprechers auf, so werden die alten Parameter beibehalten. Liegen hingegen ausreichend Daten für die betrachtete Mischungskomponente vor, so werden die neuen Parameter verwendet beziehungsweise höher gewichtet. Während der MAP-Adaption werden ausschließlich die Erwartungswerte adaptiert, die restlichen Parameter werden beibehalten. Es wäre durchaus möglich, auch die anderen Parameter zu adaptieren; es hat sich jedoch erwiesen, dass die Adaption des Erwartungswerts ausreicht und zu besseren Ergebnissen führt (Reynolds & Campbell, 2008). Eine ausführliche Beschreibung des MAP-Algorithmus sowie dessen Nutzung kann in Gauvain und Lee (1991, 1994) nachgelesen werden.

Kapitel 4

Erhöhter Stimmaufwand im Kontext sprachverarbeitender Systeme

Sprachverarbeitende Systeme zeigen mittlerweile sehr gute Erkennungsraten für Mikrophon- oder Telefonsprache. Auch bei Hintergrundgeräuschen, die in einem Auto zustandekommen, existieren bereits zufriedenstellende Lösungen. Die Leistung für Szenarien mit nicht-alltäglichen Hintergrundgeräuschen, unter Stress stehenden oder laut redenden Sprechern ist hingegen noch schlecht. Derartige Szenarien sind jedoch häufig im militärischen aber auch im forensischen Kontext gegeben.

Eine Arbeit, die die negativen Auswirkungen veränderten Stimmaufwands auf die Erkennungsleistung eines Spracherkenners demonstriert, wurde von Rajasekaran et al. (1986) durchgeführt. Sie zeigen die Auswirkungen des Lombardeffekts sowie verschiedener Stimmaufwandsstufen auf die Erkennungsleistung eines ASR-Systems, welches mit normaler Sprache trainiert wurde. Die schlechtesten Ergebnisse erzielt der Worterkenner bei geschriener Sprache. Die Untersuchung, ob Hintergrundgeräusche oder der Lombardeffekt die ASR stärker beeinflussen, ergab eindeutig schlechtere Erkennungsergebnisse für dem Lombardeffekt, auch im Vergleich zu einer SNR von nur 10 dB.

Aktuellere Untersuchungen zum Einfluss erhöhten Stimmaufwands auf die Sprechererkennung (Becker, 2008; Shriberg et al., 2008) bestätigen die Ergebnisse von Rajasekaran et al. (1986). Becker (2008) berichtet über den Einfluss verschiedener Intra-Sprecher-Variabilitäten auf ein F_0-basiertes Sprechererkennungssystem. Bei diesen Variabilitäten handelt es sich um jeweils zwei verschiedene Sprechstile (Lesen vs. Spontansprache) und Stufen des Stimmaufwands (normal vs. erhöht). In seinem Test kombiniert Becker die vier Variabilitäten miteinander und stellt fest, dass sofern im Training beide Bedingungen anders sind als im Testmaterial, die schlechtesten Ergebnisse erzielt werden. Dies bestätigt Beckers vorherige Untersuchung (Becker, 2007) mit einem GMM-basierten Sprechererkennungssystem, welches tendenziell die gleichen Verhaltensweisen auf dem gleichen Korpus zeigte.

Nachdem in Kapitel 2 bereits der Einfluss erhöhten Stimmaufwands auf unterschiedliche Parameter gesprochener Sprache vorgestellt wurde, soll nun die verringerte Leistung sprachverarbeitender Systeme im Kontext erhöhten Stimmaufwands näher beleuchtet werden. Hierfür werden zunächst unterschiedliche Möglichkeiten der Stimmaufwandsklassifikation vorgestellt (Abschnitt 4.1), welche als Unterstützung für andere Sprachverarbeitungstechnologien genutzt werden können. Als Nächstes werden gängige Kompensationsverfahren der Sprach- (Abschnitt 4.2) und

Sprechererkennung (Abschnitt 4.3) dargestellt. Abschließend wird der militärische Anwendungsbereich samt zugehöriger Studien vorgestellt (Abschnitt 4.4).

4.1 Stimmaufwandsklassifikation

Ein *Stimmaufwandsklassifikator* klassifiziert Sprachsignale hinsichtlich des Stimmaufwands. Dies bedeutet, dass zwei oder mehr Stimmaufwandsgrade im Training definiert werden. Die Zuordnung der Sprachsignale zu einem dieser Stimmaufwandsgrade oder die Zuordnung einzelner Abschnitte des Sprachsignals zu einem der Stimmaufwandsgrade soll in diesem Abschnitt dargestellt werden.

Die meisten aktuellen Arbeiten nutzen zur Modellierung der Stimmaufwandsgrade GMM-Modelle. Ein einfacher Klassifikator zur Unterscheidung neutraler Sprache von Lombardsprache wurde von Bořil und Hansen (2009a) realisiert. Hier wurde ein *GMM-basierter Klassifikator* mit unterschiedlichen Merkmalen getestet. Die besten Ergebnisse erzielten die *Standard-MFCC-Merkmale*. Ebenfalls gute Ergebnisse erzielten die Merkmale *20BandsLPC0-3200*. Diese Merkmale wurden von der PLP-Analyse abgeleitet, allerdings wurden statt einer trapezoiden Filterbank 20 nicht-überlappende Rechtecke im Codebuch verwendet. Da der Rechenaufwand für diese Merkmale geringer ist, bei vergleichbarer Leistung zu den MFCC-Merkmalen, wurden die Merkmale 20BandsLPC0-3200 für die weitere Anwendung genutzt. Die Stimmaufwandsklassifikation wurde in der Arbeit von Bořil und Hansen zur Verbesserung eines Spracherkennungssystems verwendet. Details zu diesen und den weiteren verwendeten Merkmalen sind in Bořil und Hansen (2009a) nachzulesen.

Ein solcher Standardklassifikator kann auch im Kontext der automatischen Sprechererkennung nutzbringend eingesetzt werden (siehe (Hansen & Varadarajan, 2009)). Hansen und Varadarajan (2009) verwenden hierfür GMM-Modelle mit 64 Mischungskomponenten und 23-dimensionalen Merkmalsvektoren. Die Merkmale setzen sich zusammen aus 19 MFCC-Koeffizienten (ohne den nullten Koeffizient) und vier Koeffizienten des spektralen Schwerpunktes. Getestet wurde dieser Klassifikator zur Klassifikation von Lombard- und neutraler Sprache, zur Klassifikation des Geräuschtyps bei Lombardsprache und zur Klassifikation von Lombardsprache unterschiedlicher Geräuschpegel. Die besten Ergebnisse (81,5% Erkennungsrate) werden für die einfachste Klassifikation, nämlich die Unterscheidung zwischen normaler und Lombardsprache erzielt. Die Klassifikation des Geräuschtyps fällt wesentlich schwerer, sodass hier nur Erkennungsraten zwischen 46% und 59% erzielt werden. Bei der Klassifikation des Geräuschpegels zeigt sich für eine der Klassen eine Erkennungsrate, die einer zufälligen Zuordnung gleich kommt. Die beiden anderen Klassen liefern bessere Ergebnisse.

Die bisher beschriebenen Klassifikatoren wurden auf normaler und lauter Sprache getestet. Tests zur Klassifikation normaler und *geflüsterter Sprache* sind von Wenndt et al. (2002) sowie von Zhang und Hansen (2008b, 2008a) durchgeführt worden. Wenndt et al. schlagen drei Methoden vor. Als erstes kann das Betragsverhältnis angewendet werden. Es berechnet sich als Verhältnis der hohen Energie zur niedrigen Energie im Frequenzband 450-650 Hz. Für sinkende Signal-zu-Rausch-Abstände sinkt auch die Erkennungsrate dieser Methode stark. Um eine robustere Klassifikation zu erzielen, schlagen Wenndt et al. das Verhältnis zwischen der Energie eines hohen Frequenzbands (2800-3000 Hz) zu der Energie eines niedrigen Frequenzbands

(450-650 Hz) vor. Dieses Verhältnis liefert insgesamt eine zuverlässige Klassifikation, auch bei starken Geräuschen (weißes Rauschen). Die dritte Methode entspricht den oben beschriebenen Klassifikatoren, bei denen für jeden Stimmaufwandsgrad ein Modell trainiert wird. Die verwendeten Merkmale und Modelle werden nicht spezifiziert. Für diese Art von Klassifikator werden ähnlich gute Ergebnisse wie für das zweite Verhältnis erzielt. Tests mit additivem Rauschen wurden für diese Methode nicht durchgeführt.

Einen alternativen Ansatz zeigen Zhang und Hansen (2008a). Sie nutzten statt der üblichen GMM-Modelle den *T^2-BIC-Algorithmus*. Dieser Algorithmus nutzt das *Bayes'sche Informationskriterium (BIC)* kombiniert mit der T^2-Statistik. Der Vorteil dieses Algorithmus ist, dass er sowohl für Sprachsignale unter fünf Sekunden Länge bei als auch für längere Signale zuverlässige Ergebnisse liefert. Der T^2-BIC-Algorithmus wurde von Zhou und Hansen (2005) eingeführt. Zhang und Hansen (2008a) vergleichen in ihrer Arbeit folgende Merkmale für die Nutzung mit dem T^2-BIC-Algorithmus: *MFCC, ZEPS (Zero Cross Rate, Energy, Pitch, Energy slope - Nulldurchgangsrate, Energie, Grundfrequenz, Steigung der Energie), das Energieverhältnis, die spektrale Informationsentropie und die spektrale Neigung*. Die besten Ergebnisse lieferte das Energieverhältnis und die MFCC-Merkmale. Das Energieverhältnis entspricht dem Verhältnis zwischen einem hohen und einem niedrigen Frequenzband nach Wenndt et al. (2002) (siehe oben). Die Beschreibung der anderen Merkmale ist in Zhang und Hansen (2008a) zu finden. Eine weitere Verbesserung der Klassifikation konnte durch die Entwicklung eines neuen Entropie-basierten Merkmals erzielt werden (siehe (Zhang & Hansen, 2008b)). Bei diesem neuen Merkmal handelt es sich um eine Kombination des Energieverhältnisses mit Merkmalen der spektralen Informationsentropie. Diese Merkmalskombination zeigt über sämtliche Tests bessere Ergebnisse als das Energieverhältnis allein.

Für die **Entwicklung eines eigenen Stimmaufwandsklassifikators** werden einige der beschriebenen Merkmale und Modelle aufgegriffen (siehe Abschnitt 6.3). Als Ausgangspunkt wird ein GMM-basierter Klassifikator mit MFCC-Merkmalen verwendet. Dieser Klassifikator wird mit unterschiedlichen spektralen Merkmalen kombiniert. Eines dieser Merkmale ist das Energieverhältnis. Es wurde ausgewählt, da es, wie bereits vorgestellt, in mehreren Publikationen zur Klassifikation leiser Sprache gute Ergebnisse lieferte. Daher soll es auch zur Klassifikation lauter Sprache getestet werden (siehe Abschnitt 6.3).

4.2 Erhöhter Stimmaufwand in der Spracherkennung

Frühe Arbeiten zur *Spracherkennung bei erhöhtem Stimmaufwand* wurden unter anderem von Rajasekaran et al. (1986), Moore und Bond (1987), Hansen (1988), Anglade, Fohr und Junqua (1992) sowie Junqua (1993) durchgeführt. Diese Studien sind teilweise militärisch motiviert und werden deswegen im nachfolgenden Abschnitt (4.4) beziehungsweise bereits in der Einleitung dieses Kapitels beschrieben. Ein sehr umfassendes Werk ist die Dissertation von Hansen (1988). Diese Arbeit steht in Zusammenhang mit dem später erschienen NATO-Bericht (Verlinde et al., 2000), an dem Hansen selbst auch mitwirkte. In seiner Dissertation setzt sich der Autor mit der Analyse von Sprache unter Stress und Sprache bei Hintergrundgeräuschen

auseinander. Er untersucht verschiedene Stressbedingungen, wie zum Beispiel unterschiedliche Stimmaufwandsgrade oder Sprechraten, hinsichtlich ihrer akustischen Korrelate. Weiterhin folgen Ausführungen zum Stand der Technik von Sprachverbesserungsverfahren für geräuschbehaftete Sprache und Kompensationsverfahren für Sprache unter Stress. Abschließend werden Möglichkeiten zur Spracherkennung gestresster und geräuschbehafteter Sprache vorgestellt. Hierbei werden nicht die Erkennungsalgorithmen selbst angepasst. Stattdessen verwendet Hansen auf das Szenario angepasste Präprozessoren. Bei dem untersuchten Spracherkenner handelt es sich um ein HMM-basiertes System zur sprecherabhängigen Erkennung isoliert gesprochener Worte. Die Untersuchung der Kompensationsverfahren für geräuschbehaftete Sprache sind für diese Arbeit nicht relevant, da es sich dabei um additive Geräusche handelt und somit keine Veränderungen der Sprache als Anpassung an die Hintergrundgeräusche durch den Sprecher erfolgt. Die Evaluation des Spracherkennungssystems auf unterschiedlichen Stressbedingungen ist hingegen sehr relevant, da unter anderem normale, laute, leise und Lombardsprache im Testset enthalten sind. Die Untersuchungen zeigen, dass die Geräuschkompensationsverfahren für geräuschbehaftete Sprache kombiniert mit Formantkompensationsverfahren (Anpassung der Lokation und/ oder der Bandweite), besonders für laute, ärgerliche und Lombardsprache, zu großen Verbesserungen führen. Hier werden Verbesserungen zwischen 35% und 43% erzielt.

Eine weitere Studie, die sich mit Sprache erhöhten Stimmaufwands im Kontext automatischer Spracherkennung auseinander setzt, ist die Untersuchung von Anglade et al. (1992). Der Schwerpunkt dieser Untersuchung liegt jedoch nicht auf der Verbesserung der Spracherkennung bei erhöhtem Stimmaufwand, sondern auf der Verbesserung der Spracherkennung bei leicht verwechselbaren Wörtern. Anglade et al. schlagen vor, die Unterscheidung zwischen ähnlichen Wörtern anhand von diskriminativen Frames durchzuführen. Für jedes der Wortpaare wird eine phonetische Analyse zur Bestimmung der diskriminativen Frames durchgeführt. Die automatische Detektion dieser Frames erfolgt mit Hilfe von Energiemessungen. Für Training und Test werden anschließend nur diese diskriminativen Frames verwendet. Die vorgeschlagene Methode nutzt ein neuronales Netz. Der Vergleich erfolgt zu einem HMM-basierten System, welches das Wort als Einheit beim Training und Test verwendet. Der Vergleich normaler Sprache zu Lombardsprache ohne Hintergrundgeräusche zeigt, dass die Erkennungsraten stark abfallen. Für normale Sprache verbessert sich die Systemleistung mit der neuen Methode oder bleibt gleich. Die Leistung bleibt auch bei Lombardsprache ohne Hintergrundgeräusche erhalten. Beim Hinzufügen unterschiedlicher additiver Geräusche führt das neue Verfahren zu teilweise erheblichen Verbesserungen. Allerdings ist nicht klar, ob die Veränderung der Erkennungsergebnisse tatsächlich nur auf der Nutzung der diskriminativen Frames basiert, da im Vergleichssystem ein anderer Klassifikator und andere Merkmale verwendet werden. Insgesamt lässt sich festhalten, dass der negative Einfluss des Lombardeffekts durch das neue Gesamtsystem für viele Wortkombinationen etwas abgeschwächt werden kann.

Ähnliche Ansätze zur Verbesserung der Spracherkennung bei Lombardsprache werden von Junqua (1993) vorgeschlagen. Eine Methode, die er vorschlägt ist die Nutzung von Clustering-Verfahren zur Auswahl eines Musters aus zahlreichen Mustern eines Wortes. Diese zahlreichen Muster repräsentieren Variabilitäten, wie beispielsweise unterschiedlichen Stimmaufwand oder das Geschlecht. Das Geschlecht ist bei

Lombardsprache ein wesentlicher Einflussfaktor, da auf Grund der unterschiedlichen akustischen Eigenschaften verschieden starke Veränderungen auftreten. Der zweite Verbesserungsvorschlag bezieht sich auf den Sprecher selbst. Handelt es sich bei der Spracherkennungsaufgabe um ein Szenario mit kooperativem Sprecher, so kann der Sprecher trainiert werden, seine Sprache bei Hintergrundgeräuschen an den Spracherkenner anzupassen. Der dritte Vorschlag steht in Zusammenhang mit der bereits erwähnten Studie von Anglade et al. (1992). Junqua schlägt vor eine wissensbasierte Erkennung durchzuführen. Dies bedeutet, dass sich ändernde phonetische Charakteristika in den Erkennungsprozess mit einbezogen werden. Denkbar seien auch Merkmale, die sich auf das Verhältnis zweier akustischer Eigenschaften beziehen, wie beispielsweise das Verhältnis von Konsonantendauer zu Vokaldauer oder F1-F_0.

Aktuellere Studien zu diesem Themenbereich wurden von Bořil und Hansen (Bořil, 2008; Bořil & Hansen, 2009a, 2009b, 2011) durchgeführt. Einen ausführlichen Überblick über den Einfluss des Lombardeffekts auf die Artikulation, die akustischen Merkmale und die automatische Spracherkennung gibt Bořil in seiner Dissertation (Bořil, 2008). In (Bořil & Hansen, 2009b) stellen die Autoren dann eine nicht-überwachte Methode zur Dämpfung des Lombardeffekts vor. Die nicht-überwachte Natur der Methode erlaubt es, auch eine Anpassung bei wechselndem Stimmaufwand innerhalb einer Äußerung vorzunehmen. Dadurch ist keine Vorabklassifikation des Stimmaufwands oder der Geräuschumgebung nötig. Die Methode umfasst Transformationen auf Frequenz- und cepstraler Ebene für Lombardsprache, hin zur normalen Sprache. Die Autoren schlagen die *Maximum-Likelihood-Frequenztransfomation* vor, welche an die *Vokaltraktlängennormalisierung (VTLN)* angelehnt ist. Es handelt sich um eine lineare Transformation, welche die Vokaltraktlängennormalisierung und eine weitere Variable kombiniert. Diese Variable enthält Information über das Verhalten der Formanten bei Lombardsprache. Weiterhin wird die *Quantil-basierte Normalisierung der cepstralen Dynamik* vorgestellt. Diese Form der Normalisierung ist robust gegen Ausreißer. Die Experimente zeigen, dass die gleichzeitige Verwendung beider Verfahren zu einer großen Verbesserung der Spracherkennung auf Lombardsprache führt. Allerdings ist der Rechenaufwand sehr groß, da die Merkmale während der Laufzeit berechnet werden. Die nachfolgende Studie der Autoren (Bořil & Hansen, 2009a) greift diese Arbeit auf und entwickelt eine Lösung, die weniger rechenintensiv ist. Außerdem wird das System um einen Detektor zur Klassifkation von Umgebungsgeräuschen erweitert, damit ein Codebuch verwendet werden kann, welches auf Daten zahlreicher Umgebungsgeräusche trainiert wurde. In (Bořil & Hansen, 2011) zeigen die Autoren, wie die automatische Spracherkennung mit großem Vokabular bei Lombardsprache unterschiedlicher Hintergrundgeräusche verbessert werden kann.

Die gezeigten Ansätze zur **Verbesserung der Spracherkennung** umfassen **Normalisierungsverfahren, Transformationen oder Filterverfahren**.

4.3 Erhöhter Stimmaufwand in der Sprechererkennung

Ähnlich wie bei der automatischen Spracherkennung führt erhöhter Stimmaufwand zu Verschlechterungen der Erkennungsleistung in der *Sprechererkennung*. Dies gilt

jedoch nicht nur für *automatische Sprechererkennung*. Auch die *Erkennungsleistung menschlicher Hörer* verschlechtert sich. Laut einer Studie von Kajarekar, Bratt, Shriberg und Leon (2006) sinkt die Erkennungsleistung der menschlichen Hörer bei absichtlich veränderter Stimme sogar mehr ab als die Leistung eines automatischen Systems. In dieser Studie veränderten die Sprecher ihre Stimme unter anderem hin zu geflüsterter Sprache sowie zu einer hohen beziehungsweise niedrigen Grundfrequenz. Es wurden auch weitere für diese Arbeit nicht relevante Variationen vorgenommen, wie beispielsweise die Imitation anderer Dialekte oder Sprachen. Diese Verschlechterung der Erkennungsleistung menschlicher Hörer steht im Gegensatz zu den Ergebnissen der Perzeptionsexperimente, welche sich mit der Verständlichkeit von Sprache erhöhten Stimmaufwands beschäftigen (siehe Abschnitt 2.3). Hier wurde eine gleich gute oder sogar bessere Verständlichkeit beobachtet. Dies zeigt, dass der Sprecher seine Sprache so anpasst, dass der Hörer das Gesagte besser versteht; die sprecherspezifischen Charakteristika, welche in normaler Sprache zu beobachten sind, nehmen dabei jedoch ab.

Eine weitere Untersuchung von Shriberg et al. (2008) analysiert unterschiedliche Sprechstile und Stimmaufwandsgrade im Kontext sprecherunabhängiger Sprecherverifikation. Es wird leise, normale und laute Sprache in Trainings- und Testdaten variiert. Die Kombination von lauter Sprache im Training und leiser Sprache im Test führt zu den schlechtesten Erkennungsleistungen. Besonders die leise Sprache, welche nicht geflüstert war, wird als herausfordernd für die automatische Sprechererkennung herausgestellt. Um die Erkennung zu verbessern, wird eine Anpassung der Sprache-Pause-Detektion an den Stimmaufwandsgrad vorgeschlagen. Die Sprache-Pause-Detektion ist bei leiser Sprache problematisch, da Sprache geringer Energie häufig nicht als solche erkannt wird.

Für weitere Verbesserungen der Sprechererkennung bei verändertem Stimmaufwand könnten die oben beschriebenen Ansätze aus der Spracherkennung geprüft werden. Die Anwendung von Techniken der Spracherkennung auf die Sprechererkennung ist üblich. Allerdings muss bei dem Einsatz von Kompensations- und Filterverfahren, wie sie in Abschnitt 4.2 vorgestellt wurden, geprüft werden, ob nicht genau die sprecherspezifischen Charakteristika des Sprachsignals entfernt werden.

Eine Studie, die sich mit der Kompensation von Lombardsprache für unterschiedliche Geräuscharten und -pegel im Kontext der automatischen Sprechererkennung auseinandersetzt, ist die Studie von Hansen und Varadarajan (2009). Die Autoren zeigen zunächst eine Analyse von Lombardsprache hinsichtlich der Satzdauer, der Energie und der spektralen Neigung. Anschließend beschreiben sie Tests mit einem Stimmaufwandsklassifikator (siehe Abschnitt 4.1). Dann folgen Tests mit einem Sprechererkennungssystem. Es wurde ein sogenanntes „in-set/out-of-set“ Sprecheridentifikationssystem genutzt. Bei dieser Art der Sprecheridentifikation geht es darum zu identifizieren, ob es sich bei der Testaufnahme um einen Sprecher aus dem Referenzsprecherset handelt oder nicht. Im Gegenteil zu der „open-set“ Sprecheridentifikation muss der Referenzsprecher nicht identifiziert werden. Ein denkbares Szenario für ein „in-set/out-of-set“ Sprecheridentifikationssystem ist ein Zugangskontrollsystem. Das Sprecheridentifikationssystems wurde als GMM-UBM-System mit MFCC-Merkmalen realisiert. Für das Training verwendeten die Autoren immer neutrale Sprache. Als Testdaten wurden neutrale Sprache normalen Stimmaufwands und Lombardsprache sowie geräuschbehaftete Sprache normalen Stimmaufwands und Lombardsprache angewendet. Es zeigt sich, dass die Kompensation

der Geräusche allein nicht ausreicht, da auch Lombardsprache ohne Hintergrundgeräusche zu starken Verschlechterungen der Systemleistung führt. Aus diesem Grund schlagen die Autoren zur Kompensation des Lombardeffekts die Adaption der Sprechermodelle mit Lombardsprache vor. Hierbei werden die neutralen Sprechermodelle mit wenig Lombardsprache (ca. 15 Sekunden) adaptiert. Durch eine solche Adaption, welche nur wenig Adaptionsdaten pro Sprecher benötigt, wurden große Verbesserungen erzielt. Diese Adaptionsmethode ist allerdings nur in sehr wenigen Anwendungsbereichen nutzbar, da für viele Szenarien keine zusätzlichen Sprecherdaten erhöhten Stimmaufwands vorliegen. Bei einigen militärisch-relevanten Szenarien, wie beispielsweise einer Zugangskontrolle für Piloten, ist diese Adaptionstechnik aber anwendbar. Die Kombination des Sprechererkennungssystem mit dem Stimmaufwandsklassifikator eröffnet die Möglichkeit Identifikationsprozesse sowohl mit den Modellen normalen Stimmaufwands als auch mit den Modellen der Lombardsprache durchzuführen. Der Einsatz des Sprechererkennungssystems in Kombination mit dem Stimmaufwandsklassifikator erzeugt für sämtliche Lombardbedingungen eine Verbesserung. Diese ist nicht so groß, wie bei der Nutzung der Lombardmodelle mit manueller Klassifikation. Jedoch werden je nach Lombardbedingung Verbesserungen der EER zwischen 27,51% und 78,77% beobachtet, bei einem Vergleich mit der Nutzung der neutralen Modelle für sämtliche Lombardbedingungen. Die Nutzung der Lombardmodelle für die neutralen Testdaten führt fast zu einer Verdreifachung der EER im Vergleich zu der Nutzung der neutralen Modelle. Die Nutzung des Klassifikators reduziert diese Verschlechterung der EER auf 22,46%. Insgesamt zeigt das Konzept der Sprechererkennung mit vorgeschaltetem Stimmaufwandsklassifikator also gute Ergebnisse.

Eine Studie zur Verbesserung der textabhängigen Sprechererkennung für geschriene Sprache wurde von Shahin (2006) durchgeführt. Shahin schlägt eine Veränderung der HMM-Topologie vor. Statt den sonst für die textabhängige Sprechererkennung üblichen links-rechts Modellen erster oder zweiter Ordnung beziehungsweise statt zirkulär angeordneter HMMs erster Ordnung verwendet Shahin zirkulär angeordnete HMMs zweiter Ordnung. So erzielt er bessere Erkennungsraten bei geschriener Sprache. Das Verfahren ist für die vorliegende Arbeit jedoch nicht relevant, da nur textunabhängige Sprechererkennung betrachtet wird.

Sprecheridentifikation bei geflüsterter Sprache ist in Studien von Fan und Hansen (2008, 2009) betrachtet worden. In (Fan & Hansen, 2008) präsentieren die Autoren eine zweistufige Frequenzverzerrung. Die erste Stufe der Frequenzverzerrung nutzt die Ähnlichkeit geflüsterter und normaler Sprache im höheren Frequenzbereich. Für F3 und F4 wird beinahe keine Veränderung festgestellt. Dementsprechend wird die Frequenzskala so abgeändert, dass die hohen Frequenzen betont werden anstatt wie bei der Mel-Skala die niedrigen. Die zweite Stufe der Frequenzverzerrung nutzt die Information, dass F1 von Vokalen geflüsterter Sprache 1,3 bis 1,6fach höher ist. Um wieviel höher F1 liegt, ist sprecherabhängig und wird durch die Variable α ausgedrückt. Die Position der ersten drei Filterbänke wird um α verschoben. Anstatt α für jeden Sprecher zu bestimmen, werden 5 festgelegte Werte für α angenommen, sodass für jeden Testframe 5 Merkmalsvektoren entstehen. Die zweite Stufe der Frequenzverzerrung wird nur in der Testphase angewendet. Dementsprechend liegt für jeden der k Sprecher ein GMM vor, welches mit Merkmalsvektoren trainiert wurde, auf die die erste Frequenzverzerrung angewendet wurde. Da pro Sprecher 5 Scores berechnet werden, liegen nach dem Testdurchlauf $k \cdot 5$ Scores vor. Hieraus wird der

höchste Score ausgewählt. Eine Rückweisung wird in dieser Arbeit nicht betrachtet, da es sich um eine „closed-set“ Anwendung handelt. Zusätzlich zu diesen Verfahren wurde eine Detektion stimmloser Frikative durchgeführt, durch welche weitere Verbesserungen erzielt werden konnten. Als Erweiterung dieser Arbeit führen Fan und Hansen (2009) die LFCC-Merkmale (Linear Frequency Cepstrum Coefficients) als Merkmale für die Sprechererkennung geflüsterter Sprache ein. Bei dieser Art der Merkmalsextraktion wird die Filterbank so geändert, dass nicht mehr die niedrigen Frequenzen betont werden. Es wird eine lineare Filterbank eingesetzt, bei der die Informationen des Frequenzbereichs 0-1000 Hz entfernt werden. Außerdem wird eine weitere Frequenzverzerrung durchgeführt. Die Kombination der angepassten LFCC-Merkmale mit der Frequenzverzerrung führt zu einer Verbesserung von 20% im Vergleich zu dem MFCC-basierten Standardsystem. Ein Vergleich zwischen den Merkmalen der zwei verschiedenen Studien wurde nicht durchgeführt.

Durch die dargestellten Studien angeregt, erfolgen **im praktischen Teil** dieser Arbeit sowohl **Untersuchungen zur Eignung verschiedener Parameter für die Sprechererkennung bei erhöhtem Stimmaufwand** als auch **Tests unterschiedlicher Adaptionsverfahren**. Die in der Literatur beschriebenen Merkmale und Adaptionsverfahren können nicht direkt auf das hier zutreffende Szenario übertragen werden, da die vorgestellten Filter in der Merkmalsextraktion auf geflüsterte Sprache angepasst wurden und das Adaptionsverfahren zusätzliche Sprache erhöhten Stimmaufwands erfordert. Stattdessen werden Standardmerkmale, F_0-basierte Merkmale und speziell für das Szenario erstellte Merkmale evaluiert (siehe Kapitel 8). Als Adaptionsverfahren werden in der vorliegenden Arbeit eine Adaption der Testdaten und verschiedene Adaptionen der Modelle geprüft (siehe Kapitel 9).

4.4 Untersuchungen im militärischen Kontext

Nachdem in den vorherigen Abschnitten der Einfluss erhöhten Stimmaufwands auf Sprache und sprachverarbeitende Systeme erläutert wurde, soll nun ein konkreter Arbeitsbereich beschrieben werden, bei dem ebendiese Problematik relevant ist. Im Bereich der Sprachverarbeitung für militärische Anwendungsbereiche sind unterschiedliche Szenarien denkbar, bei dem Sprache erhöhten Stimmaufwands in sprachverarbeitenden Systemen auftritt. Allgemein lassen sich die Aufgabenbereiche der Sprachverarbeitung gesprochener Sprache im militärischen Kontext in die drei Bereiche *sicherheitsüberprüfende Systeme*, *Kommando-* beziehungsweise *Eingabesysteme* und *Aufklärungssysteme* unterteilen (Harwardt, 2009). Für *sicherheitsüberprüfende Systeme* ist die Untersuchung von Sprache erhöhten Stimmaufwands in der Regel irrelevant, da solche Zugangskontrollsysteme meistens von kooperativen Sprechern in geschützter, also nicht lauter, Umgebung genutzt werden. Eine Ausnahme bildet hierbei das im vorherigen Abschnitt beschriebene Szenario einer Zugangskontrolle für Piloten. Bei den anderen beiden Bereichen kann laut gesprochene Sprache auftreten und zu einer Absenkung der Leistung führen. *Eingabesysteme* sind im militärischen Kontext zum Beispiel denkbar für teilautonome Roboter oder zur Unterstützung eines Piloten im Cockpit. In einer Studie von Ahluwalia (2008) wurde beispielsweise eine Sprachsteuerung für teilautonome unbemannte Roboter erstellt. Diese wurde umfangreich evaluiert, um zu testen, ob das System auch bei Gefechts-

lärm oder anderen Hintergrundgeräuschen funktionsfähig ist. Allerdings wurden die Geräusche nachträglich zugemischt, sodass gerade der Aspekt der Veränderung der Sprache durch erhöhten Stimmaufwand nicht Teil der Untersuchung war.

Studien, die gerade diese akustisch-phonetischen Veränderungen von Sprache untersuchten, wurden beispielsweise von Moore und Bond (1987) sowie Stanton et al. (1988) durchgeführt. Hierbei wurde Sprache aus dem Cockpit beziehungsweise einer Cockpitsimulation verwendet. Das Ziel eines sprachverarbeitenden Systems im Cockpit ist die Unterstützung des Pilots während des Fluges. Gerade bei solchen Einsatzbereichen sind zusätzlich zu dem erhöhten Stimmaufwand, bedingt durch Hintergrundgeräusche, auch Aspekte wie Stress, G-Belastung und das Tragen einer Sauerstoffmaske mit in die Analyse einzubeziehen. Das Tragen einer Sauerstoffmaske kann durch die künstliche Verlängerung des Vokaltrakts zu starken Veränderungen der akustischen Eigenschaften der Sprache führen. Bond et al. (1989) fanden heraus, dass besonders bei F_0 und F1, die bei lauter Sprache ohne Sauerstoffmaske normalerweise erhöht werden, bei lauter Sprache mit Sauerstoffmaske entweder keine signifikanten oder andere Veränderungen bewirkt werden. Dies zeigt, wie unerlässlich die Durchführung einer separaten Untersuchung für ein neues Szenario mit veränderten Gegebenheiten, wie zum Beispiel der Art der Hintergrundgeräusche oder wechselnde Situation des Sprechers, ist, damit sprachverarbeitende Systeme angepasst werden können. In der Untersuchung von Stanton et al. (1988) wurde ebenfalls Sprache untersucht, die während des Tragens einer Sauerstoffmaske produziert wurde. Die Aufnahmen wurden in einem schalltoten Raum durchgeführt, sodass hier simulierte Cockpitbedingungen vorliegen und Aspekte wie die G-Belastung und der damit verbundene Stress ausgeklammert sind. Bei dem Stimmaufwandsgrad wird unterschieden zwischen normaler, lauter und Lombardsprache. Die laute Sprache zeichnet sich dadurch aus, dass sie ungefähr 10 dB lauter ist als die normale Sprache. Zur Generierung der Lombardsprache wurde dem Sprecher rosa Rauschen (90 dB) per Kopfhörer zugeführt. Nach der Analyse von 18 Merkmalen über ungefähr 11.000 Phoneme berichten Stanton et al. über Energiemigrationen im Frequenzbereich für laute Sprache und Lombardsprache. Diese Energiemigration führt zu einer Verschiebung der Frequenzen hin zu den sensibelsten Frequenzbereichen des menschlichen Gehörs. Insgesamt beobachten Stanton et al. nur geringe Unterschiede zwischen lauter und Lombardsprache (Stanton et al., 1988).

Da lautes Sprechen im militärischen Kontext oft nicht nur durch Hintergrundgeräusche, sondern auch durch Stress bedingt ist, sind viele militärisch relevante Studien unter dem Arbeitsbereich „Sprache unter Stress“ zu finden. Beispiele hierfür sind die NATO Berichte (Trancoso, 1996; Verlinde et al., 2000). Diese Berichte beschreiben Untersuchungen zur Lombardsprache und zu sprachverarbeitenden Systemen mit Lombardsprache als Eingabe, ebenso wie Datenbanken gestresster Sprache (SUSC-0, SUSC-1, SUSAS). Die Datenbanken enthalten sowohl Sprache aus tatsächlichen Stresssituationen als auch aus simulierten Situationen. Das Korpus SUSC-0 enthält beispielsweise Audioaufzeichnungen eines Flugzeugunfalls, angefangen von der Nachricht, dass der Funk nicht funktioniert, bis zum (sicheren) Ausstieg der Piloten per Schleudersitz. Simulierte Stresssituationen sind in der SUSAS Datenbank Domäne „Talking Styles“ zu finden. In dieser Domäne des Korpus wurden neun männliche Sprecher mit unterschiedlichen Sprechstilen aufgezeichnet (langsam, schnell, ärgerlich, fragend, leise, laut, klar). Diese und zahlreiche weitere

Datenbanken sind geeignet, um Studien zu Sprache erhöhten Stimmaufwands im militärischen Kontext durchzuführen. Bei der Betrachtung von Sprache im militärischen Kontext sind Einflussfaktoren, wie beispielsweise Stress und Emotionen, in einigen Szenarien relevant. In der vorliegenden Arbeit wurden diese Aspekte jedoch bewusst ausgeklammert, da bei der Betrachtung mehrerer Faktoren gleichzeitig die Ursache für Veränderungen der Sprache nicht definitiv klärbar ist.

Zusätzlich zu den Problemen bei Eingabesystemen können Probleme für sprachverarbeitende Systeme der *Aufklärung* auftreten. Hier kann der Sprecher gezwungen sein, auf Grund einer schlechten Verbindung zu seinem Kommunikationspartner oder auf Grund von Hintergrundgeräuschen laut zu reden. Zu diesem Themenbereich sind keine Veröffentlichungen zu finden. Allerdings ist dieser Aspekt ähnlich der Situation in der forensischen Phonetik, sodass Untersuchungen aus diesem Bereich herangezogen werden können. Andere Aspekte, wie die Verbindung an sich, können nicht verglichen werden und müssen daher separat untersucht werden.

Insgesamt kann festgehalten werden, dass **die Untersuchung erhöhten Stimmaufwands im militärischen Kontext essentiell ist** und zur Verbesserung sprachverarbeitender Systeme beitragen kann, wenn diese im militärischen Kontext eingesetzt werden sollen. Die Untersuchung sollte hierbei speziell auf das Szenario angepasst sein, da wir in den oben beschriebenen Studien sehen, dass die Voraussetzungen sehr unterschiedlich sein können (beispielsweise mit oder ohne Sauerstoffmaske).

Kapitel 5

Methodik

Nach der Darstellung der theoretischen Grundlagen folgt nun die Beschreibung der praktischen Tätigkeiten. Aus den theoretischen Problemstellungen dieser Arbeit, welche bereits in der Einleitung definiert wurden, ergeben sich für den praktischen Teil folgende Aufgabenstellungen:

- Untersuchungen akustischer Veränderungen bei erhöhtem Stimmaufwand
    * Statistische Analyse spektraler Veränderungen bei Erhöhung des Stimmaufwands
    * Evaluation spektraler Parameter zur automatischen Klassifikation des Stimmaufwands
    * Zusammenhänge zwischen den spektralen Parametern und F_0 bei normalem und erhöhtem Stimmaufwand
- Realisierung eines Sprecherverifikationssystems für nicht-übereinstimmenden Stimmaufwand
    * Erstellen eines Frameworks zur automatischen Sprechererkennung sowie Evaluation des Sprechererkennungssystems hinsichtlich des Einflusses erhöhten Stimmaufwands
    * Entwicklung von Strategien zur Verbesserung der Erkennungsraten des Sprechererkennungssytems für das gegebene Szenario

Die für die Lösung dieser Aufgabenstellungen notwendigen Untersuchungen und Tests werden im Folgenden erläutert (Abschnitt 5.1). Anschließend werden die verwendeten Korpora (Abschnitt 5.2) vorgestellt.

5.1 Tests

Dieser Abschnitt präsentiert einen Überblick über die Untersuchungen des praktischen Teils. Die Gliederung erfolgt angepasst an die oben dargestellten Aufgabenstellungen.

5.1.1 Untersuchungen akustischer Veränderungen bei erhöhtem Stimmaufwand

Statistische Analyse spektraler Veränderungen bei Erhöhung des Stimmaufwands Für alle Untersuchungen des Spektrums wird eine 512 Punkte FFT mit einer Fensterlänge von ebenfalls 512 Punkten mit Hilfe des Snack Toolkits (Sjölander & Beskow, 2000) durchgeführt.

Zur Bestimmung der optimalen Frequenzbereiche für das Energieverhältnis wird vor der statistischen Analyse eine *Untersuchung der Betragsdifferenzen* zwischen normaler und lauter Sprache durchgeführt (Abschnitt 6.1). Für diese Voruntersuchung wird das OLLO-*Korpus* (Abschnitt 5.2.2) verwendet.

Die *statistische Analyse* der spektralen Veränderungen abhängig vom Stimmaufwand (SA) wird für die folgenden spektralen *Parameter* durchgeführt: spektrale Neigung (SN), gewichteter spektraler Schwerpunkt (center of gravity, COG), Energieverhältnis (EV), die ersten vier spektralen Momente (Mom1 bis Mom4 für einzelne Momente; Momente, für die Kombination aller vier Momente).

Die statistische Analyse (Abschnitt 6.2) nutzt ebenfalls das OLLO-*Korpus*. Es ist auf Phonemebene annotiert. Die Untersuchungen werden mit dem *Tool* Gnu R (R Development Core Team, 2010) durchgeführt. Die Testdaten werden in vier unterschiedliche *Testsets* unterteilt: Gesamtdaten über alle Sprecher und Phoneme (Gesamt); Sortierung nach Lautklasse (LK): Obstruenten (Obstr), Sonoranten (Son) und Vokale (Vok); Sortierung nach Phonemen (Phonem); Sortierung nach Sprecher (Spk). Für jeden spektralen Parameter werden die in Tabelle 5.1 dargestellten *Tests* durchlaufen. Die Testergebnisse werden in Kapitel 6 präsentiert.

		Dateneinteilung			
		Gesamt	LK	Phonem	Spk
Tests	Analyse der Verteilung	X	X	X	X
	Test auf Normalverteilung	X	X	-	-
	Signifikanztest	X	X	-	-

Tabelle 5.1: Durchgeführte Untersuchungen zur Analyse der spektralen Veränderungen bei Erhöhung des Stimmaufwands

Evaluation spektraler Parameter zur automatischen Klassifikation des Stimmaufwands Für diesen Aufgabenpunkt wird ein *System zur Klassifikation des Stimmaufwands erstellt*. Als *Trainings- und Testkorpus* wird das „Pool 2010"-Korpus (Abschnitt 5.2.1) verwendet. Es liegt keine Annotation auf Wort- oder Phonemebene vor, sodass komplette spontansprachliche Äußerungen für Training und Test verwendet werden. Der Klassifikator wird mit Hilfe des *Tools* Hidden Markov Toolkit (HTK) (Young et al., 2006) realisiert. Es erfolgt eine Unterscheidung zwischen den zwei Klassen: normaler und erhöhter Stimmaufwand. Der Klassifikator wird mit unterschiedlichen *Merkmalen und Merkmalskombinationen* getestet:

- Tests mit jedem spektralen Parameter (siehe oben) separat als Merkmal,
- Tests mit Kombinationen unterschiedlicher spektraler Parameter,
- Test mit MFCC-Merkmalen,

- Tests mit MFCC-Merkmalen und jeweils einem spektralen Parameter und
- Test mit MFCC-Merkmalen und Kombinationen unterschiedlicher spektraler Parameter.

Die Klassifikationsergebnisse sowie ein Vergleich der Leistung der verschiedenen Merkmale und Merkmalskombinationen werden in Abschnitt 6.3 dargestellt.

Zusammenhänge zwischen den spektralen Parametern und F_0 bei normalem und erhöhtem Stimmaufwand Zur Darstellung der Zusammenhänge zwischen den spektralen Parametern und F_0 im Kontext erhöhten Stimmaufwands wird zunächst eine *statistische Analyse von F_0* durchgeführt (Abschnitt 7.1). Diese erfolgt so, wie sie bereits oben für die spektralen Parameter beschrieben wurde. F_0 wird mit der ESPS Funktion des Snack Toolkits berechnet. Es werden die Standardeinstellungen des Tools gewählt.

Um *Zusammenhänge zwischen normaler und lauter Sprache* festzustellen, werden für die oben genannten spektralen Parameter und F_0 *Korrelationsanalysen* durchgeführt (Abschnitt 7.3). Als *Korpus* wird das OLLO-Korpus verwendet. Für jeden Parameter wird eine Spearman-Rangkorrelation mit dem *Tool* Gnu R durchgeführt.

Abschließend werden *Korrelationsanalysen zwischen F_0 und jedem spektralen Parameter* jeweils für normalen und erhöhten Stimmaufwand durchgeführt (Abschnitt 7.4). Auch hier werden Spearman-Rangkorrelationen mit dem *Tool* Gnu R durchgeführt und das OLLO-*Korpus* verwendet.

5.1.2 Realisierung eines Sprecherverifikationssystems für nicht-übereinstimmenden Stimmaufwand

Erstellen eines Frameworks zur automatischen Sprechererkennung sowie Evaluation des Sprechererkennungssystems hinsichtlich des Einflusses erhöhten Stimmaufwands Das *Framework* zur Sprecherverifikation wird mit Hilfe des *Tools* HTK erstellt (Abschnitt 8.1). Es handelt sich hierbei um ein GMM-UBM-basiertes System mit 1024 Mischungskomponenten. Als *Korpus* für die Trainings- und Testdaten wird das „Pool 2010“-Korpus verwendet. Das UBM besteht aus Daten des Kiel-Korpus sowie in der erweiterten Form zusätzlich aus Daten des OLLO-Korpus.

Zur *Festlegung der besten Merkmale* für das Basissystem werden einige Standardmerkmale (MFCC, PLP, LPC) evaluiert (Abschnitt 8.3). Es werden folgende *Tests* je Merkmal durchgeführt:

- ein Vergleichstest, der normalen Stimmaufwand in Trainings- und Testdaten aufweist (Vgl);
- ein Basistest, welcher normalen Stimmaufwand in den Trainingsdaten nutzt und erhöhten Stimmaufwand im Testset (Basis);
- ein Basistest inklusive Telefonfilter, der zusätzlich zu den Gegebenheit des Basistests ein Telefonfilter anwendet (Basis(TF));
- ein Test mit erweitertem UBM (UBM+);
- ein Test mit erweitertem UBM und Telefonfilter (UBM+(TF)).

Für die nachfolgenden Tests werden die MFCC-Merkmale als *Standardmerkmale für das Basissystem* festgelegt.

Entwicklung von Strategien zur Verbesserung der Erkennungsraten des Sprechererkennungssytems für das gegebene Szenario Es werden zwei Strategien zur Verbesserung der Erkennungsraten vorgeschlagen: Die *Analyse unterschiedlicher Merkmale* auf ihre Eignung im gegebenen Szenario, allein oder in Kombination mit den MFCC-Merkmalen, und die *Untersuchung verschiedener Adaptionsverfahren*.

Im Rahmen der *Analyse unterschiedlicher Merkmale* werden nach dem Vergleich der Standardmerkmale *F_0-basierte Merkmale* evaluiert (Abschnitt 8.4). Die getesteten Merkmale sind die F_0-Statistik, logF_0 und logF_0/E. Die Extraktion der F_0 erfolgt mit dem *Tool* Snack. Die *Korpora* werden in derselben Form angewendet wie für die Standardmerkmale. Die durchgeführten *Tests* je Merkmale sind teilweise mit den oben beschriebenen für die Standardmerkmale identisch:

- ein Vergleichstest, der normalen Stimmaufwand in Trainings- und Testdaten aufweist (Vgl);
- ein Basistest, welcher normalen Stimmaufwand in den Trainingsdaten nutzt und erhöhten Stimmaufwand im Testset (Basis);
- ein Test mit erweitertem UBM (UBM+);
- eine Fusion der Ergebnisse jeweils des F_0-basierten Merkmals mit den MFCC-Merkmalen (MFCC(TF)+Basis).

Als nächstes folgen *Untersuchungen zu Merkmalen, speziell für die Verwendung bei erhöhtem Stimmaufwand* (Abschnitt 8.5). Es wird das COG-Verhältnis (COG-V. mit/ohne VAD) vorgestellt sowie die Kombination ausgewählter spektraler Parameter (Momente Kombi). Die *Korpora* werden in derselben Form angewendet wie für die Standardmerkmale. Die *Tests* sind die gleichen wie bei den F_0-basierten Merkmalen, unter Hinzunahme eines weiteren Tests. Bei diesem zusätzlichen Test handelt es sich um eine Fusion der Ergebnisse des UBM+-Tests für jeweils eins der speziellen Merkmale mit den MFCC-Merkmalen (MFCC(TF)+UBM+). Dieser Test wird eingefügt, da die Ergebnisse der UBM+-Tests einen solchen Test als vielversprechend ausweisen.

Die *Untersuchung verschiedener Adaptionsverfahren* analysiert eine Form der *Adaption der Testdaten* sowie Möglichkeiten zur *Adaption der Modelle*. Diese Untersuchungen werden in Kapitel 9 beschrieben. Für sämtliche Adaptionsverfahren werden die besten *Merkmale* der vorherigen Tests herausgesucht (MFCC(TF); logF_0; logF_0/E; COG-Verhältnis ohne VAD) und für die unterschiedlichen Adaptionsverfahren evaluiert. Als *Adaptionsverfahren* der Testdatenadaption wird die Methode von Goldenberg et al. verwendet (Abschnitt 9.1). Zur Modelladaption werden die MAP- und die MLLR-Adaption getestet (Abschnitt 9.2). Die *Korpora* sind wie bei den Tests der Standardmerkmale zusammengestellt. Zur Modelladaption wird das *Tool* HTK verwendet. Die Adaption der Testdaten wird in Java selbst ausprogrammiert, da kein Tool vorliegt. Als *Tests* werden je Adaptionsverfahren ein Basistest (Goldenberg-Basis; MAP-Basis; MLLR-Basis) und ein Test mit erweitertem UBM (Goldenberg-UBM+; MAP-UBM+; MLLR-UBM+) ausgeführt. Abschließend werden je Adaptionsverfahren 10 Fusionen der Ergebnisse durchgeführt.

5.2 Korpora

Im Rahmen dieser Arbeit werden verschiedene Korpora zur Entwicklung, zum Training und zum Test der verschiedenen Systeme beziehungsweise Systemkonfigurationen genutzt. Die folgenden Abschnitte erläutern kurz die drei verwendeten Korpora „Pool 2010“-, OLLO- und Kiel-Korpus sowie ihre Einteilung für die Entwicklungs-, Trainings- und Testszenarien. Alle Korpora liegen als 8 kHz, 16 Bit Daten vor.

5.2.1 „Pool 2010“-Korpus

Das „Pool 2010“-Korpus wurde vom Bundeskriminalamt erstellt, um die Forschung an verschiedenen Sprechstilen, Stimmaufwandsgraden und in GSM übertragener Sprache zu fördern. Eine detaillierte Beschreibung des „Pool 2010“-Korpus findet sich in Jessen et al. (2005). Für das Korpus wurden 106 männliche Sprecher aufgenommen. Die meisten Sprecher stammen aus der Region um Wiesbaden und zeigen teilweise dialektale Eigenschaften des Hessischen. Für jeden Sprecher sind je zwei Aufnahmen, eine in gelesener Sprache und eine in Spontansprache, vorhanden. Zusätzlich wurden Telefonkonversationen aufgezeichnet, welche in dieser Arbeit jedoch nicht verwendet werden. Je Sprechstil wurde eine Aufnahme ohne besondere Aufnahmebedingungen aufgezeichnet, was zu normaler Sprache mit normalem Stimmaufwand führte. Für die zweite Aufnahme hörten die Sprecher 80 dB lautes, weißes Rauschen per Kopfhörer. Auf diese Weise wurde der Lombardeffekt hervorgerufen, sodass die Sprecher lauter sprachen und einen erhöhten Stimmaufwand aufwiesen. Die so erzeugten Aufnahmen liegen sowohl in Studioqualität als auch als nachträglich per GSM übertragene Daten vor. Für die gelesene Sprache wurde der Text „Nordwind und Sonne“ verwendet, während die Spontansprache durch Beschreibung von Bildern unter Auslassung bestimmter Wörter produziert wurde. Es liegen keine Annotationen auf Wort- oder Phonemebene vor, sondern nur Informationen zu dem Stimmaufwandsgrad und dem Sprecher.

Um das „Pool 2010“-Korpus *für die Sprechererkennung* nutzen zu können, wird der spontansprachliche Anteil des Korpus (Studiosprache) in dieser Arbeit in drei Teile eingeteilt: Trainings-, Test- und Entwicklungsdaten. Die zwei Aufnahmen pro Sprecher werden jeweils in zwei gleich lange Signale geteilt. Der erste Teil des normal lauten Sprachsignals der ersten 50 Sprecher wird als Trainingskorpus verwendet. Die zwei Teile des lauten Signals der ersten 50 Sprecher werden als Testdaten genutzt. Die Daten der restlichen 56 Sprecher werden zu Entwicklungszwecken, für die Adaption, zurückgehalten.

Für die Nutzung des „Pool 2010“-Korpus *zur Stimmaufwandsklassifikation* wird das Korpus für die nachfolgenden Tests in Trainings- und Testdaten unterteilt. Als Testdaten werden die normalen und lauten Audiodateien von 50 Sprechern verwendet. Die restlichen 56 Sprecher werden für das Training der Modelle eingesetzt. Anders als in der Sprechererkennung wird sowohl für das Training als auch für die Tests jeweils die komplette Datei verwendet.

5.2.2 „Oldenburger Logatom“-Korpus

Das OLLO-Korpus („Oldenburger Logatom“-Korpus) enthält Sprachdaten von 50 verschiedenen Sprechern beider Geschlechter. Bei den Sprechern handelt es sich

um 40 deutsche und 10 französische Sprecher. Die 40 deutschen Sprecher stammen aus verschiedenen Dialektregionen. Es wurde Standarddeutsch in Oldenburg (Universitätspopulation), Bayerisch in München (Bewohner der ländlichen Umgebung um München), Ostfriesisch in Oldenburg (ländliche Bevölkerung) und Ostfälisch in Magdeburg (Bewohner der Vororte) aufgenommen. Für jeden Dialekt gibt es 10 Sprecher. Die Logatome wurden auf verschiedene Arten gesprochen: Sprechrate (langsam, normal, schnell), Stimmaufwand (leise, normal, laut) und Sprechstil (Aussage oder Frage). Es gibt insgesamt 150 verschiedene Logatome. Diese wurden in jeder dieser Varianten pro Sprecher drei Mal gesprochen. Zusätzlich zu den Logatomen existieren weitere Aufnahmen mit vorgelesenen Sätzen. Diese wurden in den folgenden Tests jedoch nicht verwendet. Ebenso wurden die weiblichen und die französischen Sprecher aussortiert. Übrig blieben demnach 19 deutschsprachige, männliche Sprecher. Von diesen 19 Sprechern wurden die normalen und die lauten Aufnahmen verwendet. Für das OLLO Korpus liegen automatische Annotationen auf Phonemebene vor. Da vorab bekannt war, welche Phoneme gesprochen wurden, musste lediglich die zeitliche Alignierung automatisiert werden. Um etwaige Ungenauigkeiten auf Grund der automatischen Annotation auszuschließen, werden in den nachfolgenden Tests die ersten und letzten zwei Frames je Laut nicht verwendet.

Das OLLO-Korpus wird im Folgenden *für die Untersuchung der spektralen Parameter* und *für die Erweiterung der UBM-Daten* verwendet. Für das erweiterte Hintergrundmodell werden die Logatome normalen Stimmaufwands der männlichen deutschen Sprecher verwendet. Eine detaillierte Beschreibung des Korpus und unterschiedlicher Einsatzmöglichkeiten ist in Wesker et al. (2005) zu finden.

5.2.3 Kiel-Korpus

Das Kiel-Korpus enthält Spontansprache von 52 deutschen Sprechern beider Geschlechter. Es wurde im Rahmen des Projekts Verbmobil zum Training eines Reiseauskunftssystems erstellt. Das Kiel-Korpus wird in dieser Arbeit ausschließlich *zum Training des Hintergrundmodells* für das Sprechererkennungssystem verwendet. Es werden ausschließlich die männlichen Sprecher genutzt. Details zur Konzeption des Kiel-Korpus sind in Kohler, Pätzold und Simpson (1995) zu finden.

Kapitel 6

Quantifizierung des Stimmaufwands

Nachdem Kapitel 2 den Stimmaufwand definiert und seine Auswirkung auf Sprache erläutert, wird nun das Problem der Quantifizierung des Stimmaufwands ausführlicher behandelt. Die Quantifizierung des Stimmaufwands ist problematisch, weil es sich beim Stimmaufwand nicht um eine physikalisch messbare Größe handelt. Vielmehr stellt der Stimmaufwand eine subjektiv durch Sprecher und Hörer wahrgenommene Größe dar. Diese zu quantifizieren ist vor allem bei wechselnden Aufnahmegegebenheiten problematisch.

Die Quantifizierung oder Klassifikation des Stimmaufwands ist sowohl in der forensischen Phonetik als auch in der automatischen Sprachverarbeitung von großem Nutzen. In der forensischen Phonetik kann eine Quantifizierung des Stimmaufwands genutzt werden, um Abschnitte gleichen Stimmaufwands zu bestimmen und nur solche Abschnitte miteinander zu vergleichen. In der automatischen Sprachverarbeitung kann ein Stimmaufwandsklassifikator zur Vorabklassifikation des Stimmaufwands verwendet werden. Mit Hilfe dieser Klassifikation kann dann für das folgende sprachverarbeitende System ein Modell ähnlichen Stimmaufwands ausgewählt werden, welches in der Regel bessere Ergebnisse erzielt als ein Standardmodell normalen Stimmaufwands.

Da bekannt ist, dass sich das Spektrum der Sprache bei Erhöhung des Stimmaufwands verändert (siehe Abschnitt 2.4.3), werden im Folgenden verschiedene spektrale Merkmale als Parameter zur Quantifizierung des Stimmaufwands vergleichend untersucht. Hierfür wird zunächst eine Vorabuntersuchung zur Definition des Energieverhältnisses präsentiert (Abschnitt 6.1). Nach der Darstellung der Veränderungen der spektralen Parameter und dem Vergleich der Ergebnisse (Abschnitt 6.2) werden die Parameter in einem Stimmaufwandsklassifikator angewendet, um die Ergebnisse aus Abschnitt 6.2 zu verifizieren (Abschnitt 6.3).

6.1 Spektrale Veränderungen bei erhöhtem Stimmaufwand

In Kapitel 2 wurden die Auswirkungen von Lautstärkeveränderungen auf Sprache, wie sie in der Literatur zu finden sind, detailliert beschrieben. Außerdem wurden verschiedene spektrale Parameter vorgestellt, welche zur Quantifizierung des Stimmaufwands denkbar sind (siehe Abschnitt 2.4.3). Ein solcher Parameter ist beispielsweise das *Energieverhältnis*. Zur Festlegung der Energiebereiche, die zur

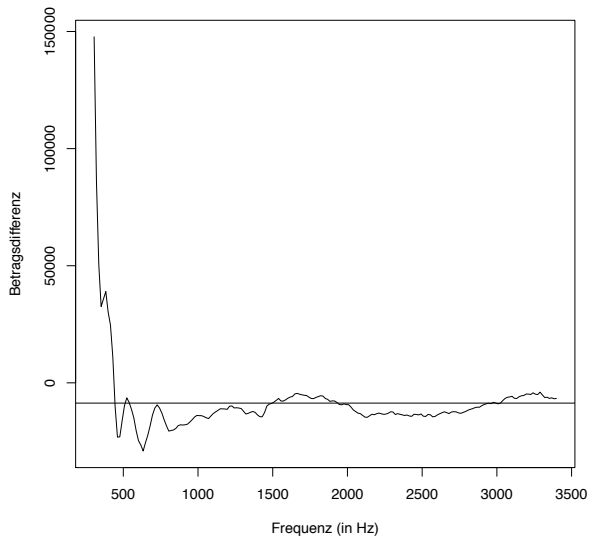

Abbildung 6.1: Betragsdifferenz zwischen normaler und lauter Sprache (Die durchschnittliche Betragsdifferenz ist als horizontale Linie gekennzeichnet.)

Berechnung des Energieverhältnisses genutzt werden sollen, haben Wenndt et al. (2002) eine Voruntersuchung zur Veränderung des Spektrums in verschiedenen Frequenzbereichen für normale und geflüsterte Sprache durchgeführt. Andere Autoren, wie Zhang und Hansen (2008a), haben diesen Parameter auch erfolgreich zur Klassifikation lauter Sprache genutzt. Es ist allerdings fraglich, ob die Energiebereiche, die für geflüsterte und normale Sprache gewählt wurden, ebenfalls für normale und laute Sprache optimal sind. Aus diesem Grund wird im Folgenden eine Untersuchung ähnlich der Untersuchung von Wenndt et al. (2002) für laute Sprache vorgestellt.

Wenndt et al. haben in ihrer Untersuchung festgestellt, dass die größte Betragsdifferenz (9dB) nahe 400 Hz liegt. Die durchschnittliche Betragsdifferenz über den Frequenzbereich zwischen 0 und 4000 Hz liegt bei 4 dB. Für die Berechnung des Energieverhältnisses wurde das Frequenzband zwischen 450 und 600 Hz ausgewählt, also der Bereich mit sehr starken Betragsdifferenzen. Als hoher Frequenzbereich wurde 2800-3000 Hz ausgewählt. In diesem hohen Bereich entspricht die Betragsdifferenz ungefähr dem Durchschnittswert von 4 dB. Außerdem sind die Frequenzbereiche so gewählt, dass sie auch bei einer Telefonbandbegrenzung noch nutzbar sind.

In dieser Arbeit wird die Betragsdifferenz für normale und laute Sprache auf Daten von 19 deutschen Sprechern des OLLO-Korpus (siehe Abschnitt 5.2) für den Frequenzbereich 300-3400 Hz vorgestellt. Eine Darstellung der unterschiedlichen Differenzen im Leistungsspektrum ist in Abbildung 6.1 zu finden. Auch hier zeigt sich, dass die größte Abweichung des Betrags bei ca. 400 Hz liegt und dass um 3000 Hz ungefähr der Durchschnittswert über das Gesamtspektrum erreicht wird. Dies entspricht den Beobachtungen für normale und geflüsterte Sprache in Wenndt et al. (2002). Ein weiterer Bereich, in dem ungefähr der Durchschnittswert erreicht ist, liegt bei 1500-2000 Hz.

Insgesamt zeigt sich, dass **die für geflüsterte und normale Sprache verwendeten Energiebereiche auch für laute Sprache anwendbar sind**. Als alternativer hoher Frequenzbereich könnte der Bereich zwischen 1500 und 2000 Hz gewählt werden, allerdings liegt der Bereich 2800-3000 Hz näher am Durchschnittswert, also an der horizontalen Linie in Abbildung 6.1.

Die Definition für das Energieverhältnis aus Abschnitt 2.4.3 wird beibehalten.

6.2 Spektrale Parameter

Spektrale Parameter liefern eine gute Ausgangsposition für die Klassifikation des Stimmaufwands (Zhang & Hansen, 2008a). Um zu untersuchen, welcher Parameter am besten für die Klassifikation geeignet ist und welcher Parameter möglicherweise robust gegenüber Stimmaufwandsveränderungen ist, wird in diesem Abschnitt ein Vergleich verschiedener spektraler Parameter durchgeführt. Als Korpus werden Aufnahmen lauter und normaler Sprache des OLLO-Korpus verwendet. Da das OLLO-Korpus nicht manuell annotiert wurde, werden die ersten und letzten zwei Frames pro Laut nicht verwendet, um Annotationsungenauigkeiten auszuschließen. Dementsprechend werden nur Laute mit mehr als 4 Frames verwendet. Für die Berechnung des Spektrums wird eine 512 Punkte FFT mit einer Fensterlänge von ebenfalls 512 Punkten mit Hilfe des Snack Toolkits (Sjölander & Beskow, 2000) durchgeführt. In den folgenden Abschnitten wird zunächst die Methodik der statistischen Auswertung beschrieben, um anschließend die Auswertung für die einzelnen spektralen Parameter durchzuführen. Abschließend wird ein Vergleich der Ergebnisse sämtlicher spektraler Parameter durchgeführt.

6.2.1 Statistische Auswertung der spektralen Parameter

Die Untersuchung jedes spektralen Parameters beginnt zunächst mit einer *statistischen Auswertung der Verteilung*. Hierfür werden der Mittelwert (MW), der maximale Wert (Max), der minimale Wert (Min), die mittlere Abweichung (Mittl.AW), die Standardabweichung (St.AW), die Varianz (Var), die Schiefe (Sch) und die Wölbung (Wöl) der Verteilung analysiert. Diese statistische Auswertung erfolgt für die Gesamtdaten sowie unterteilt nach Lautklassen (LK), Lauten (Phonem) und Sprechern (Spk).

Die *Lautklassen*, zwischen denen unterschieden wird, sind: Obstruenten, Sonoranten und Vokale. Obstruenten werden auch Geräuschlaute genannt und sind dadurch charakterisiert, dass ein Hemmnis im Ansatzrohr durch die Luft überwunden wird (Bußmann, 2002). Dies führt entweder zu einer Verwirbelung der Luft oder zu einem Verschluss, sodass sowohl Frikative als auch Plosive zu den Obstruenten zählen. Auch die Affrikate [t͡s], welche aus einem Plosiv besteht, der in einen homorganen Frikativ gelöst wird (Pompino-Marschall, 2003), gehört zu der Klasse der Obstruenten. Die anderen Laute werden den Sonoranten zugeordnet. Diese sind immer stimmhaft. Die Gruppe der Sonoranten wird in der vorliegenden Arbeit nicht komplett als Lautklasse verwendet, sondern unterteilt in die Klassen Vokale und Sonoranten (jegliche Sonoranten außer den Vokalen). Bei der Interpretation der

LK	$\alpha = 0,05$	$\alpha = 0,02$	$\alpha = 0,01$
Gesamt	0,0075	0,0083	0,0090
Obstr	0,0126	0,0140	0,0151
Son	0,0683	0,0763	0,0819
Vokale	0,0094	0,0105	0,0112

Tabelle 6.1: Schranken für den D-Wert der verschiedenen Lautklassen

Ergebnisse der Obstruenten muss beachtet werden, dass die Auswahl der Frames automatisch erfolgt. Hierfür werden, wie bei den anderen Lautklassen auch, die ersten und letzten zwei Frames pro Laut aussortiert und nur die mittleren Frames verwendet. Bei Plosiven kann dies dazu führen, dass die Verschlusspause in diesen Frames enthalten ist und nicht die Verschlusslösung. Dies kann zu inkonsistenten Ergebnissen führen.

Nach der Betrachtung der Momente der Verteilung werden die Gesamtverteilung und die Verteilungen der Lautklassen auf *Normalverteilung* getestet. Hierzu wird der *Lilliefors-Test* verwendet, welcher eine Modifikation des Kolmogoroff-Smirnoff-Tests (Sachs, 2002) ist. Das Testen auf Normalverteilung ist nötig, um entscheiden zu können, ob parametrische oder nicht-parametrische Tests für die Untersuchung der Differenzen zwischen normalem und erhöhtem Stimmaufwand eingesetzt werden sollten. Der Lilliefors-Test berechnet die Differenz zwischen der gegebenen Verteilung und der Normalverteilung. Für größere Differenzen ergibt sich ein kleinerer p-Wert. Ein kleiner p-Wert ($< 0,05$) führt demnach zur Verwerfung der H_0-Hypothese, welche besagt, dass die gegebene Verteilung normalverteilt ist. Die Annahme der H_0-Hypothese hingegen ist kein Nachweis, dass eine Kurve normalverteilt ist. Durch den Lilliefors-Test kann lediglich nachgewiesen werden, dass eine Kurve nicht-normalverteilt ist. Die Differenz ist im D-Wert dargestellt. Dieser berechnet sich nach Gross (2009) als maximaler Wert des Betrags von D^+ und D^-:

$$D^+ = \max_{i=1,\ldots,n} \{i/n - p(i)\}, D^- = \max_{i=1,\ldots,n} \{p(i) - (i-1)/n\} \tag{6.1}$$

n steht hierbei für die Anzahl der Stichprobenelemente und $p(i) = \Phi([x(i) - \overline{x}]/s)$, mit Φ als kumulative Verteilungsfunktion der Standardnormalverteilung sowie $\overline{x}$ als Mittelwert und s als Standardabweichung der gegebenen Stichprobe. Wie der D-Wert zu interpretieren ist, hängt von der Größe der Stichprobe und dem Signifikanzniveau ab. Für $n > 30$ und ein Signifikanzniveau von beispielsweise $\alpha = 0,05$ gilt als Schranke für $D : 1,358/\sqrt{n}$. Schranken für andere Signifikanzniveaus sind in Sachs (2002, S. 428) nachzulesen. Die Stichprobengrößen der im Folgenden durchgeführten Tests sind: $n = 33082$ für die Gesamtverteilung, $n = 11703$ für die Obstruenten, $n = 395$ für die Sonoranten und $n = 20984$ für die Vokale. Die Schranken für die D-Werte unterschiedlicher Signifikanzniveaus sind in Tabelle 6.1 abzulesen.

Bei den nachfolgenden Lilliefors-Tests treten p-Werte auf, die je nach Ausgabe des Tools Nortest (Gross, 2009) entweder mit $p < 2,2E-16$ oder mit $p = 0$ dargestellt werden. Für diese p-Werte scheint der maximale negative Exponent des Tools überschritten zu sein. In der gesamten Auswertung werden, je nach Einstellung des Tools, einige Werte kleiner $2,2E-16$ gemessen, für die mit dem Tool eine Darstellung möglich war. Deswegen ist davon auszugehen, dass die nicht-darstellbaren Werte zwischen 0 und dem kleinsten gemessenen Wert ($3,32E-296$) liegen. Aus diesem

Grund wird der p-Wert für diese nicht-darstellbaren p-Werte der Lilliefors-Tests mit $p < 3,32E{-}296$ festgelegt.

Da sich bei der Prüfung auf Normalverteilung herausstellt (siehe nachfolgende Abschnitte), dass nicht alle Verteilungen normalverteilt sind, wird ein nichtparametrischer Test zur Überprüfung der Unterschiede zwischen normalem und erhöhtem Stimmaufwand ausgewählt. Der *U-Test nach Wilcoxon, Mann und Whitney* bietet hierzu die Möglichkeit. Er überprüft, ob zwei unabhängige Stichproben zur selben Grundgesamtheit gehören. Für diesen Test ist keine Annahme der Normalverteilung notwendig. Es handelt sich um einen Rangsummentest, bei dem die Stichprobenelemente der beiden gegebenen Verteilungen der Größe nach sortiert werden. Für beide Verteilungen wird aus diesen Rängen separat die Rangsumme berechnet. Die niedrigere der beiden Rangsummen wird auch Testwert W genannt. Für Stichprobengrößen mit $n \geq 30$ wird statt dem Testwert W der approximative kritische Wert Z berechnet, welcher sich aus dem Testwert W und den Stichprobengrößen ergibt. Die genaue Berechnung ist in Hollander und Wolfe (1999, S. 109) zu finden, wobei Z hier mit W^* bezeichnet wird. Bei einem Signifikanzniveau von $\alpha = 0,05$ (beziehngsweise $\alpha = 0,02$, $\alpha = 0,01$) ergeben sich für Z folgende Grenzen: $z = 0 \pm 1,95996$ (beziehungsweise $z = 0 \pm 2,32635$, $z = 0 \pm 2,57583$). Für Z-Werte außerhalb dieses Bereichs wird die H_0-Hypothese zurückgewiesen. Die im Folgenden beschriebenen Mann-Whitney Rangsummentests wurden mit dem Tool Gnu R (R Development Core Team, 2010) in Kombination mit dem Paket „coin“ (Horthorn, Hornik, Wiel & Zeileis, 2010) durchgeführt. Eine Darstellung der Werte $p < 2,2E{-}16$ ist bei diesem Tool nicht möglich, sodass in den folgenden Tests häufig dieser Wert auftaucht. Eine feinere Darstellung ist nicht nötig, da die H_0-Hypothese sowohl für $p = 2,2E{-}16$ als auch für kleinere Werte zurückgewiesen wird.

6.2.2 Spektrale Neigung

In diesem Abschnitt werden die Veränderungen der *Spektralen Neigung (SN)* bedingt durch eine Erhöhung des Stimmaufwands beschrieben. Hierbei wird zunächst die Veränderung der spektralen Regressionsgerade *auf den Gesamtdaten* normalen Stimmaufwands des OLLO-Korpus (siehe Abschnitt 5.2.2) mit sämtlichen Daten erhöhten Stimmaufwands verglichen. Es zeigt sich, dass sämtliche Parameter der Verteilung um 10% oder mehr verändert werden (siehe Tabelle A.1). Der Mittelwert und die Wölbung steigen, während die anderen Werte sinken. Diese starken Veränderungen sind hingegen nicht in Abbildung 6.2 zu sehen. Hier scheinen die Unterschiede zwischen normaler und lauter Sprache nicht so groß zu sein. Es zeigt sich hingegen, dass der Anteil der niedrigeren Werte der spektralen Neigung (-150 bis -50) bei erhöhtem Stimmaufwand geringer wird und stattdessen die Anzahl der Werte um 0 sowie zwischen -50 und 0 ansteigt. Diese Beobachtung geht konform mit der in der Literatur häufig gefundenen These, dass die spektrale Neigung bei erhöhtem Stimmaufwand abflacht (siehe Abschnitt 2.4.3).

Weitere Details zur Veränderung der spektralen Neigung durch erhöhten Stimmaufwand sollen durch eine Auswertung nach *Lautklassen* erarbeitet werden. Betrachtet werden die Lautklassen Obstruenten, Sonoranten und Vokale. Die Werte der spektralen Neigung lauter und normaler Sprache sind, genau wie die der Gesamtauswertung, in Tabelle A.1 zu finden. Die Verteilungen sind in Abbildung 6.3 dargestellt. Es zeigt sich, dass der Mittelwert sowie der Wert der Wölbung um

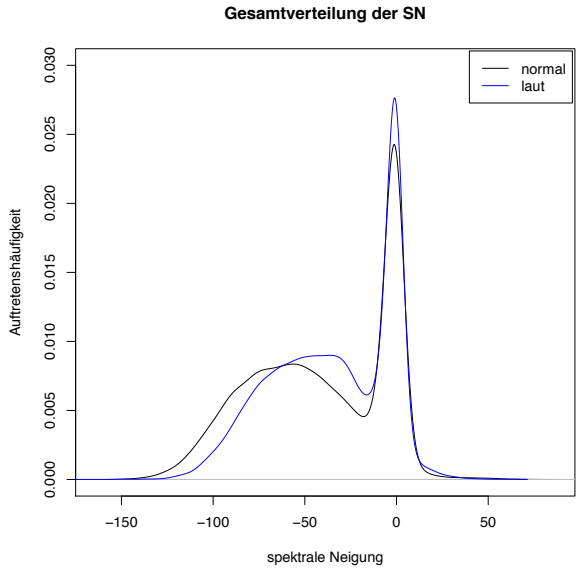

Abbildung 6.2: Verteilung der spektralen Neigung für normale und laute Sprache über sämtliche Laute und Sprecher

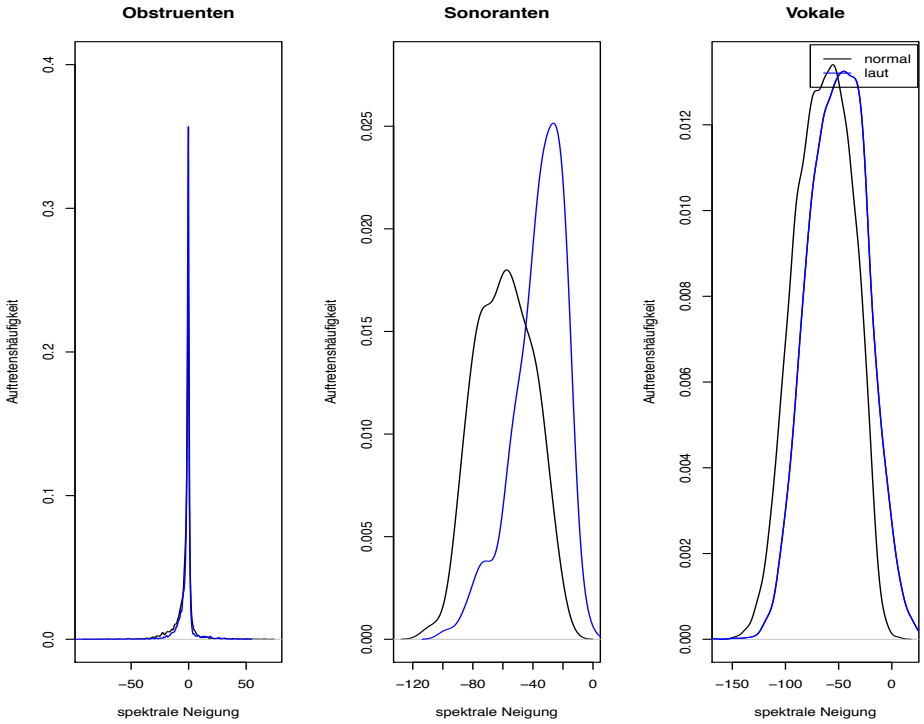

Abbildung 6.3: Verteilung der spektralen Neigung für normale und laute Sprache der drei Lautklassen Obstruenten, Sonoranten und Vokale

mehr als 10% für die drei Lautklassen steigen. Die Standardabweichung wird kaum verändert durch die Erhöhung des Stimmaufwands. Die steigende Wölbung ist in Abbildung 6.3 an den spitzeren blauen Kurven, besonders bei den Obstruenten und Sonoranten, zu erkennen. Ferner sind die Kurven der Obstruenten für laute und normale Sprache nahezu identisch. Für Sonoranten und Vokale bleibt die Form der Verteilung ähnlich; die Kurve für laute Sprache der Sonoranten ist hingegen leicht nach rechts verschoben und spitzer. Die Sonoranten lauter Sprache weisen zudem eine Linksschiefe auf, während fast keine Neigung für normale Sprache vorliegt.

Um das Verhalten der einzelnen Lautklassen besser aufschlüsseln zu können, werden die Veränderungen der einzelnen Phoneme ebenfalls betrachtet. Zunächst werden die Ergebnisse für die *Obstruenten* dargestellt (siehe Tabelle A.2). Bei den Plosiven zeigt sich der Laut [p] als Laut mit den meisten Abweichungen vom allgemeinen Trend. Der Mittelwert für [p] sinkt, während er für die anderen Plosive steigt. Die mittlere Abweichung, die Standardabweichung und die Varianz dagegen steigen für [p] und sinken für die anderen Plosive. Für die Schiefe besteht die allgemeine Tendenz zum Absinken des Wertes und für die Wölbung zur Steigerung. Auch hier gibt es jeweils einen Plosiv, der nicht diesem Trend folgt. Auffällig ist weiterhin, dass sämtliche Werte der Schiefe negativ sind. Dies bedeutet, dass die Kurven linksschief sind und für laute Sprache eine stärkere Neigung aufweisen. Die Werte der Wölbung sind positiv, sodass die Kurven leicht spitz sind. Der Vergleich der Frikative miteinander legt sinkende Mittelwerte für stimmlose Frikative und steigende für stimmhafte dar. Der Maximalwert sinkt für sämtliche Frikative, während der Minimalwert für alle Frikative außer für [f] steigt. Das heißt, dass die Werte der Verteilung insgesamt in einem kleineren Wertebereich liegen. Die Streuungsparameter mittlere Abweichung, Standardabweichung und Varianz sinken für alle Frikative außer für [f]. Dies passt zu dem verengten Wertebereich. Die Werte der Schiefe haben, wie bei den Plosiven, die Tendenz zum Sinken und die der Wölbung zum Steigen. Insgesamt bestehen ähnliche Tendenzen zur Veränderung der spektralen Neigung für Plosive und Frikative. Die Affrikate [t͡s] zeigt ebenfalls ein ähnliches Muster.

Die *Sonoranten* weisen ebenfalls untereinander ähnliche Veränderungen auf (siehe Tabelle A.3). Die Werte Mittelwert, Maximalwert und die Wölbung steigen um mehr als 10% an, während der Wert der Schiefe sinkt. Die zwei Nasale weisen auch bei den übrigen Werten das gleiche Muster auf; der Minimalwert steigt, während die anderen Werte um mehr als 10% sinken. Für den Liquid [l] gibt es für die anderen Werte keine Veränderungen größer oder gleich 10%.

Die *Vokale* zeigen kein eindeutiges Muster (siehe Tabelle A.4). Allerdings gibt es für die meisten Werte geringfügig sichtbare Tendenzen. Der Mittelwert beispielsweise wird immer erhöht, meistens sogar um mehr als 10%. Auch der Minimalwert steigt für Vorderzungenvokale um mehr als 10%. Bei Zentral- und Hinterzungenvokalen steigt der Minimalwert ebenfalls für die meisten Laute, jedoch sind die Veränderungen häufig kleiner 10%. Eine Ausnahme bildet hierbei der Vokal [uː], für den der Minimalwert um mehr als 10% sinkt. Auch bei den anderen Werten gibt es, sofern eine Tendenz erkennbar ist, stets Ausnahmen. Dies gilt für die Schiefe, welche meistens abgesenkt wird, für die Wölbung, welche meistens erhöht wird, aber auch für die Varianz, welche häufig niedriger ist bei erhöhtem Stimmaufwand.

Die Auswertung der Daten *nach Sprechern sortiert* (siehe Tabelle A.5) bestätigt die Beobachtungen aus der Gesamtauswertung für die Werte Mittelwert, mittlere

LK	SA	D-Wert	p-Wert
Gesamt	N	$0,144$	$< 3,32E-296$
	L	$0,134$	$< 3,32E-296$
Obstr	N	$0,225$	$< 3,32E-296$
	L	$0,282$	$< 3,32E-296$
Son	N	$0,0515$	$0,198$
	L	$0,09$	$8,07E-04$
Vok	N	$0,0331$	$7,44E-28$
	L	$0,0219$	$8,2E-13$

Tabelle 6.2: Ergebnisse des Lilliefors-Tests auf Normalverteilung für die spektrale Neigung

Abweichung, Standardabweichung und Varianz für sämtliche Sprecher, auch wenn die Veränderungen nicht für jeden Sprecher größer 10% sind. Für die anderen Werte gibt es keine klaren Muster, sodass hier von sprecherspezifischen Unterschieden auszugehen ist.

Nach dieser Beschreibung der Unterschiede zwischen den Verteilungen lauter und normaler Sprache folgt nun eine Analyse der Verteilungen bezüglich des *Verteilungstyps*. Es wird geprüft, ob die Gesamtverteilung sowie die Verteilungen der einzelnen Lautklassen normalverteilt sind. Hierzu wird der *Lilliefors-Test* durchgeführt (siehe Abschnitt 6.2.1). Die Ergebnisse des Lilliefors-Test sind in Tabelle 6.2 dargestellt. Für die Gesamtverteilung ist ein hoher D-Wert und ein sehr kleiner p-Wert, sowohl für normale als auch für laute Sprache, zu beobachten. Der D-Wert liegt für beide Stimmaufwandsgrade klar über den Schranken für die drei Signifikanzniveaus, die in Tabelle 6.1 dargestellt sind. Dies bedeutet, dass die H_0-Hypothese, welche eine Normalverteilung annimmt, mit einem Signifikanzniveau von $\alpha = 0,01$ zu verwerfen ist. Dies wird durch den sehr kleinen p-Wert (<0,01) bestätigt. Für die Obstruenten und Vokale wiederholen sich diese Ergebnisse, obwohl die Kurven der Vokale in Abbildung 6.3 normalverteilt erscheinen. Für die Signifikanztests ist jedoch das Ergebnis des Lilliefors-Tests maßgeblich. Der Unterschied zwischen normaler und lauter Sprache für die Gesamtverteilung und die Verteilung der Vokale besteht darin, dass der D-Wert für laute Sprache etwas geringer ist als der für normale Sprache. Für die Obstruenten ist der D-Wert dagegen etwas höher für laute Sprache, im Vergleich zur normalen Sprache. Die Sonoranten unterscheiden sich von den bisher betrachteten Verteilungen. Hier kann für normale Sprache die H_0-Hypothese nicht verworfen werden, da der D-Wert kleiner der Grenze 0,0683 für $\alpha = 0,05$ und auch der p-Wert größer $0,05$ ist. Dies bedeutet, dass die Stichprobe der Sonoranten normaler Sprache möglicherweise normalverteilt ist. Dies ist auch in Abbildung 6.3 sichtbar. Für laute Sprache hingegen kann die H_0-Hypothese verworfen werden. Jedoch weichen sowohl der D-Wert als auch der p-Wert nicht derartig deutlich von den Grenzen ab wie bei den vorher beschriebenen Verteilungen.

Da nur für eine der Verteilungen die Normalverteilung nicht mit Hilfe der Signifikanztests ausgeschlossen werden kann, wird ein nicht-parametrischer Test für die weitere Untersuchung der Verteilungen verwendet. Wie in Abschnitt 6.2.1 dargestellt, wird der *Mann-Whitney-Test* genutzt. Die Ergebnisse werden in Tabelle 6.3 präsentiert. Für sämtliche Verteilungen ist ein hoch signifikanter Unterschied zwischen lauter und normaler Sprache ersichtlich. Die Signifikanztests bestätigen demnach

LK	Z-Wert	p-Wert
Gesamt	15,9159	$< 2,2E-16$
Obstr	9,4189	$< 2,2E-16$
Son	10,9377	$< 2,2E-16$
Vokale	35,1518	$< 2,2E-16$

Tabelle 6.3: Ergebnisse des Mann-Whitney-Test für die spektrale Neigung

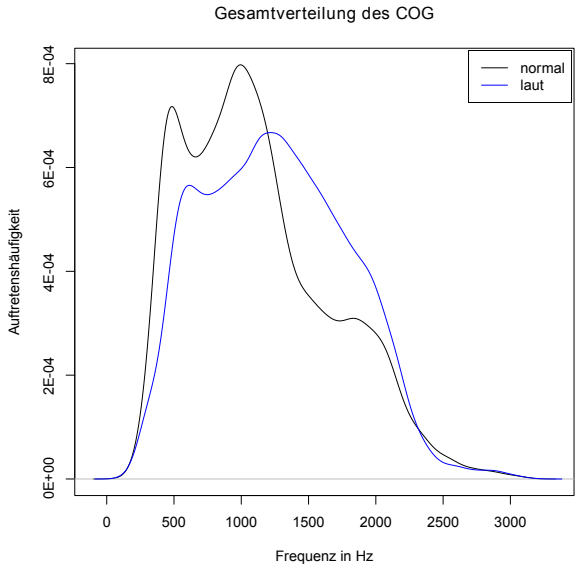

Abbildung 6.4: Verteilung des spektralen Schwerpunktes für normale und laute Sprache über sämtliche Laute und Sprecher

die Beobachtungen, die oben bezüglich der Momente der Verteilung beschrieben wurden. Für die Gesamtverteilung sowie die Verteilungen aller Lautklassen werden die meisten Momente der Verteilung um mehr als 10% verändert (siehe Tabelle A.1). Diese Veränderungen spiegeln die signifikanten Veränderungen der spektralen Neigung bei der Erhöhung des Stimmaufwandes wider.

Die Ergebnisse lassen den Rückschluss zu, dass die **spektrale Neigung als Parameter zur Unterscheidung zwischen normalem und erhöhtem Stimmaufwand**, und damit zur Quantifizierung des Stimmaufwands, geeignet ist. Die Eignung wird im Rahmen eines Stimmaufwandsklassifikators weiter geprüft (siehe Abschnitt 6.3).

6.2.3 Gewichteter spektraler Schwerpunkt

Im folgenden Abschnitt wird die Verteilung des *gewichteten spektralen Schwerpunkts (COG)*, berechnet über jeden einzelnen Frame, untersucht. Die Betrachtung der Momente der Verteilung für die Gesamtmenge aller Laute und Sprecher in Tabelle A.6 und der zugehörigen Abbildung 6.4 verdeutlicht einen Unterschied zwischen normaler und lauter Sprache. Dieser ist hauptsächlich auf die Parameter Mittelwert, Schiefe und Wölbung zurückzuführen. Insgesamt kann festgehalten werden, dass

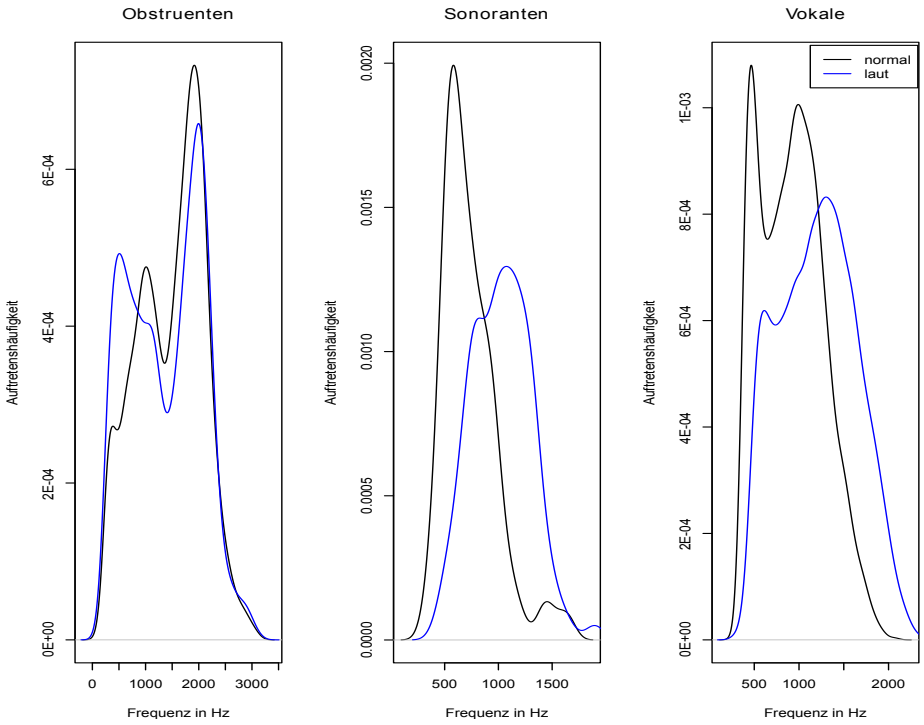

Abbildung 6.5: Verteilung des spektralen Schwerpunktes für normale und laute Sprache der drei Lautklassen Obstruenten, Sonoranten und Vokale

der Mittelwert bei einer Erhöhung des Stimmaufwands ansteigt. Für die Schiefe der Verteilung kann eine Veränderung von einer rechtsschiefen Verteilung hin zur symmetrischen Form beobachtet werden. Weiterhin ist die Verteilungskurve (siehe Abbildung 6.4) für laute Sprache flacher, sodass sich hier niedrigere Werte für die Wölbung ergeben als für normale Sprache.

Um die Unterschiede zwischen normaler und lauter Sprache besser herauszuarbeiten, wurden die Statistiken der drei *Lautklassen* Obstruenten, Sonoranten und Vokale untersucht. Diese sind ebenfalls in Tabelle A.6 zu sehen. Die zugehörige Darstellung der Verteilungen ist in Abbildung 6.5 zu finden. Für normale Sprache haben die Obstruenten und die Vokale jeweils eine bimodale Verteilung, wobei die der Obstruenten linksschief ist und die der Vokale rechtsschief. Die Verteilung der Sonoranten für normale Sprache ist ebenfalls rechtsschief, jedoch nicht bimodal. Für sämtliche Lautklassen ist die Tendenz beobachtbar, dass die Schiefe bei erhöhtem Stimmaufwand zur symmetrischen Form hin reduziert wird. Für die Obstruenten, die auch bei normaler Sprache nur eine geringe Schiefe aufweisen, ist die Schiefe für laute Sprache gleich Null. Unterscheidungen der drei Lautklassen sind bezüglich der Veränderung des Mittelwertes zu finden. Während der Mittelwert für Sonoranten und Vokale bei erhöhtem Stimmaufwand steigt, sinkt er leicht für Obstruenten. Der Maximal- und der Minimalwert bleiben bei Obstruenten ungefähr gleich, während beide Werte für Vokale und Sonoranten steigen. Auffällig ist vor allem, dass für Vokale und Sonoranten, bis auf jeweils einen Wert, die Werte um mehr als 10% verändert werden, wobei nur die Werte für Schiefe und Wölbung sinken und die anderen steigen. Auch die Werte, deren Veränderung unter 10% liegen, steigen bei erhöhtem Stimmaufwand an.

Um die Differenzen und Ähnlichkeiten zwischen den Lautklassen besser beschreiben zu können, werden im Folgenden die Statistiken der *einzelnen Laute* näher betrachtet. Hierfür werden zunächst die Statistiken der *Obstruenten* des OLLO-Korpus in Tabelle A.7 zusammengefasst. Die Betrachtung der Plosive zeigt die Tendenz, den Mittelwert abzusenken, auch wenn die Veränderung nicht immer größer 10% ist. Eine Ausnahme bildet hierbei der Plosiv [d], bei dem der Mittelwert leicht steigt. Ein weiterer Trend ist die Anhebung der mittleren Abweichung, der Standardabweichung und der Varianz für stimmlose Plosive. Für stimmhafte Plosive bleiben die mittlere Abweichung und die Standardabweichung relativ unverändert oder sinken ab ([b]). Insgesamt zeigt sich der stimmhafte bilabiale Plosiv [b] als Ausnahme unter den Plosiven, da er sehr viele Werte aufweist, welche um mehr als 10% sinken. Für sämtlich Plosive gilt, dass der Wert der Schiefe ansteigt. Da die meisten Werte auch für normale Sprache bereits positiv sind, bedeutet dies, dass die Kurven rechtsschief sind und diese Rechtsschiefe für laute Sprache stärker ausgeprägt ist. Eine Ausnahme bildet hierbei der stimmlose Plosiv [t], welcher einen negativen Wert für normale Sprache aufweist, der dann bei erhöhtem Stimmaufwand steigt. In diesem Fall wird die leichte Linksschiefe weiter zur symmetrischen Form hin verschoben. Für Frikative zeigt sich die Tendenz, dass die mittlere Abweichung, die Standardabweichung und die Varianz sinken. Eine Ausnahme bildet hierbei der Laut [f], für den diese Werte steigen. Umgekehrt verhält es sich bei den Minimalwerten, welche tendenziell steigen; erneut mit Ausnahme des Lautes [f]. Bezüglich der Schiefe und der Wölbung schließt sich [f] den allgemeinen Tendenzen an: die Schiefe wird abgesenkt und der Wert der Wölbung erhöht. Auch hier gibt es jeweils einen Frikativ, der sich nicht gemäß dieser Tendenzen verhält. Die Affrikate [t͡s] ähnelt in ihren Veränderungen den Plosiven. Bezüglich der Schiefe hingegen orientieren sich die Veränderungen der Affrikate an denen des Frikativs [s].

Insgesamt kann für die Klasse der Obstruenten aus den Einzeluntersuchungen der Laute kein klares Muster gewonnen werden. Im Gegenteil dazu stehen die Veränderungen der *Sonoranten*, welche in Tabelle A.8 dargestellt sind. Sämtliche Momente sind um 10% oder mehr verändert. Es lässt sich festhalten, dass der Mittelwert sowie der Maximal- und der Minimalwert bei erhöhtem Stimmaufwand für alle Sonoranten ansteigen. Der Liquid [l] ist insgesamt gut abgrenzbar zu den zwei untersuchten Nasalen [m] und [n]. Die Nasale weisen für jegliche Werte der Verteilung eine Steigerung auf, abgesehen von der Schiefe und der Wölbung, welche sinken. Der Lateral hat einen positiven Wert für die Wölbung der Verteilung normaler Sprache. Dies bedeutet, dass es sich um eine leptokurtische, spitze Kurve handelt. Bei erhöhtem Stimmaufwand steigt die Steilheit weiter an. Sämtliche Werte der Wölbung der Nasale sind negativ, was bedeutet, dass die Kurve flacher ist und bei erhöhtem Stimmaufwand noch weiter abflacht. Die mittlere Abweichung, die Standardabweichung und die Varianz sinken für den Lateral, was ebenfalls im Kontrast zu den Veränderungen der Nasale steht.

Die Veränderung der Verteilung des gewichteten spektralen Schwerpunktes der *Vokale* ist in Tabelle A.9 dargestellt. Für die Vokale sind klare Tendenzen zu erkennen. Der Mittelwert und der Maximalwert steigen. Die Schiefe sinkt, wobei für normale Sprache immer ein positiver Wert oder ein Wert nahe Null vorhanden ist. Die Wölbung steigt für Vorderzungenvokale mit nicht-tiefer Zungenposition an und sinkt für die anderen Vokale um mehr als 10%. Eine Ausnahme ist hier für den Vokal [ɪ] zu beobachten, dessen Wölbung nur geringfügig verändert wird. Der

LK	SA	D-Wert	p-Wert
Gesamt	N	0,068	$5,38E-205$
	L	0,0407	$3,66E-72$
Obstr	N	0,079	$7,04E-105$
	L	0,0891	$5,91E-118$
Son	N	0,105	$9,28E-06$
	L	0,0644	0,0553
Vok	N	0,0602	$6,92E-97$
	L	0,0431	$3,29E-53$

Tabelle 6.4: Ergebnisse des Lilliefors-Tests auf Normalverteilung für den gewichteten spektralen Schwerpunkt

Minimalwert sinkt für die gelängten hohen und ober-mittelhohen Vokale, während er für die anderen Vokale steigt. Der hohe Vokal [u:] bildet hierbei eine Ausnahme, da er insgesamt leicht ansteigt und keine Veränderung über 10% zeigt. Für die mittlere Abweichung, die Standardabweichung und die Varianz kann ein Anstieg über 10% für Zentral- und Hinterzungenvokale festgestellt werden. Für Vorderzungenvokale ergibt sich kein klares Muster.

Durch eine Auswertung der Daten *nach Sprechern sortiert* und einen Vergleich der Ergebnisse (Tabelle A.10) mit denen der Gesamtauswertung (Tabelle A.6), werden die Tendenzen aus der Gesamtauswertung für die meisten Sprecher bestätigt: der Mittelwert wird für sämtliche Sprecher erhöht. Diese Erhöhung ist jedoch für viele Sprecher geringer als 10%. Die Schiefe sinkt für sämtliche Sprecher um 10% oder mehr. Auch für die Wölbung wird ein Absinken des Wertes um 10% oder mehr für die meisten Sprecher festgestellt. Für zwei Sprecher ist hingegen eine Steigerung über 10% festzustellen. Ein weiterer Sprecher zeigt keine größeren Veränderungen. Bei der Varianz ist die generelle Tendenz zum Absinken des Wertes, welche für die Gesamtpopulation festgestellt wurde, für zahlreiche Sprecher zu beobachten. Jedoch gibt es auch hier Sprecher, für die die Varianz steigt, teilweise sogar über 10%. Die Veränderung der anderen Werte scheint stark sprecherabhängig zu sein, da hier kein konsistenter Trend festgestellt werden kann. Besonders auffällig ist hierbei der Minimalwert, bei dem ca. die Hälfte der Sprecher eine Steigerung über 10% und die andere Hälfte eine Senkung um mehr als 10% zeigt.

Um die oben beschriebenen Veränderungen der Momente der Verteilungen des spektralen Schwerpunktes näher zu untersuchen, werden die Gesamtverteilung sowie die Verteilungen der Lautklassen auf Normalverteilung getestet. Dies wird mit Hilfe des *Lilliefors-Tests* gemacht (siehe Abschnitt 6.2.1). Die Ergebnisse des Lilliefors-Test sind in Tabelle 6.4 nachzulesen. Es zeigt sich insgesamt ein ähnliches Ergebnis wie für die spektrale Neigung. Für die Gesamtverteilung, für die Obstruenten und für die Vokale kann die H_0-Hypothese, dass die jeweilige Verteilung normalverteilt ist, für normale und laute Sprache zurückgewiesen werden. Die p-Werte der Verteilungen sind jeweils sehr gering, sodass eine klare Rückweisung der H_0 für die drei in Abschnitt 6.2.1 beschriebenen Signifikanzniveaus möglich ist. Die D-Werte der Gesamtverteilung, der Obstruenten und der Vokale sind größer als die vorgegebenen Schranken, jedoch ist der Unterschied nicht so groß wie für die spektrale Neigung (siehe Abschnitt 6.2.2). Die Verteilung der Sonoranten normalen Stimmaufwands ist, im Gegenteil zur selbigen für die spektrale Neigung, ebenfalls

LK	Z-Wert	p-Wert
Gesamt	25,3868	$< 2,2E-16$
Obstr	-6,5763	$4,824E-11$
Son	10,3844	$< 2,2E-16$
Vokale	48,2942	$< 2,2E-16$

Tabelle 6.5: Ergebnisse des Mann-Whitney-Test für den gewichteten spektralen Schwerpunkt

nicht-normalverteilt. Der p-Wert ist allerdings wesentlich größer als der der anderen bisher beschriebenen Verteilungen des gewichteten spektralen Schwerpunkts. Die Verteilung der Sonoranten lauter Sprache ist möglicherweise normalverteilt. Die H_0-Hypothese kann nicht zurückgewiesen werden. Der D-Wert liegt knapp unter der vorgegebenen Schranke und der p-Wert ist größer 0,05. Dies illustriert Abbildung 6.5. Nur die blaue Kurve der Sonoranten sieht annähernd normalverteilt aus, während die anderen zweigipflig, stark schief oder spitz sind.

Da auch hier die meisten Verteilungen nicht-normalverteilt sind, wird der nichtparametrische *Mann-Whitney-Test* zur Überprüfung der Unterschiede zwischen normaler und lauter Sprache verwendet. Die Ergebnisse sind in Tabelle 6.5 dargestellt. Genau wie für die spektrale Neigung zeigt sich auch hier für die Gesamtverteilung und die Verteilungen aller Lautklassen ein signifikanter Unterschied zwischen normalem und erhöhtem Stimmaufwand. Für die Obstruenten ist die Differenz zwischen normaler und lauter Sprache nicht so groß wie für die anderen Lautklassen und für die Gesamtverteilung. Auch dieses Ergebnis ist in Abbildung 6.5 gut sichtbar. Die Obstruenten haben sowohl für laute als auch für normale Sprache jeweils eine zweigipflige Verteilung. Die Form der Verteilung ist folglich erhalten geblieben. Der Unterschied zwischen normaler und lauter Sprache ist insgesamt sichtbar weniger als die Unterschiede für Sonoranten und Vokale. Die weniger große Differenz zwischen normaler und lauter Sprache des gewichteten spektralen Schwerpunktes der Obstruenten stimmt überein mit den Ergebnissen aus Gottsmann (2010). Hier wurde für Frikative ein signifikanter Unterschied zwischen normaler und lauter Sprache für das erste spektrale Moment, welches ebenfalls den gewichteten Schwerpunkt des Spektrums beschreibt, beobachtet. Für Plosive wurde zwar ein Anstieg der Werte beschrieben, der Unterschied war jedoch nicht signifikant. Durch eine Veränderung der Plosive, welche nicht signifikant ist, könnte folglich die geringere Differenz zwischen normaler und lauter Sprache für den spektralen Schwerpunkt der Obstruenten erklärt werden.

Insgesamt liefert der gewichtete spektrale Schwerpunkt eine gute Unterscheidung zwischen normaler und lauter Sprache, auch wenn der Unterschied für die Obstruenten nicht so groß ist wie für die anderen Lautklassen. Die Eignung des gewichteten spektralen Schwerpunktes zur Quantifizierung des Stimmaufwands wird mit Hilfe eines Stimmaufwandsklassifikators weiter geprüft (siehe Abschnitt 6.3).

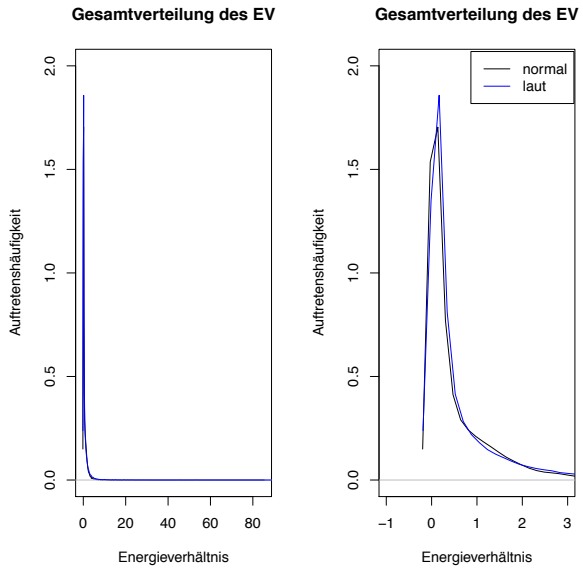

Abbildung 6.6: Verteilung des Energieverhältnisses für normale und laute Sprache über sämtlicher Laute und Sprecher als Gesamtdarstellung (links) und Vergrößerungen des wichtigsten Bereichs der Kurve (rechts)

6.2.4 Energieverhältnis

Nachfolgend wird die Veränderung des *Energieverhältnisses (EV)* bei erhöhtem Stimmaufwand untersucht. Zunächst wird die Verteilung sämtlicher Audiodaten lauter und normaler Sprache *insgesamt* miteinander verglichen (siehe Tabelle A.11). Es wird ersichtlich, dass nur der Mittelwert und die mittlere Abweichung um mehr als 10% verändert sind. Die anderen Werte sind nicht oder nur wenig verändert. Dies wird auch in Abbildung 6.6 deutlich. Auch bei Vergrößerung der Kurve im Wertebereich um 0 zeigt sich, dass die Kurven (rechts) nahezu identisch sind. Lediglich der Mittelwert und die mittlere Abweichung steigen bei erhöhtem Stimmaufwand. Das Energieverhältnis scheint dementsprechend nicht als Parameter zur Messung von Veränderungen des Stimmaufwands bei lauter und normaler Sprache geeignet zu sein. Stattdessen sollte in Betracht gezogen werden, das Energieverhältnis auf seine Robustheit in der Sprechererkennung normalen und lauten Stimmaufwands zu testen. Dafür werden die Audiodaten weiter aufgesplittet und genauer untersucht.

Zunächst werden die *Lautklassen* der Obstruenten, Sonoranten und Vokale detaillierter untersucht. Die Ergebnisse sind in Tabelle A.11 und Abbildung 6.7 zu sehen. Die Obstruenten sind, ähnlich der Gesamtauswertung, nur für zwei Werte um mehr als 10% verändert (Minimalwert und Wölbung). Eine absolute Erhöhung des Minimalwertes von 0,001 führt zu einer Veränderung über 10%. Auch die Verteilung der Obstruenten scheint dementsprechend relativ robust gegenüber einer Erhöhung des Stimmaufwands zu sein. Dies wird durch die Kurven links oben und unten in Abbildung 6.7 bestätigt. Bei den Sonoranten hingegen sind jegliche Werte um mindestens 10% verändert. Alle Werte außer der Schiefe und der Wölbung steigen. Für die Vokale ergibt sich ein ähnliches Bild. Sämtliche Werte außer dem Minimalwert steigen um 10% oder mehr. Der Minimalwert bleibt unverändert.

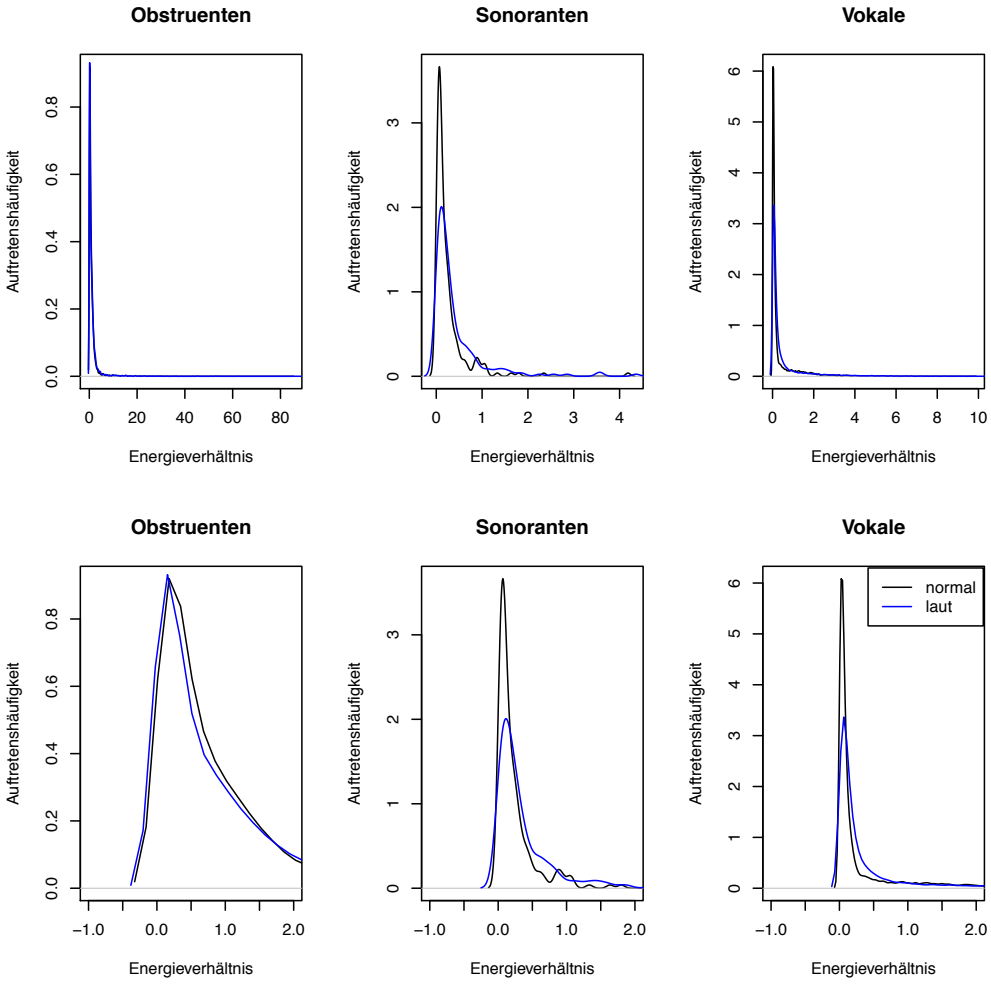

Abbildung 6.7: Verteilung des Energieverhältnisses für normale und laute Sprache der drei Lautklassen Obstruenten, Sonoranten und Vokale als Gesamtdarstellung (oben) und Vergrößerungen des wichtigsten Bereichs jeder Kurve (unten)

Die genaue Betrachtung der Klasse der *Obstruenten* zeigt, dass bei der Untersuchung der einzelnen Laute keine klaren Tendenzen bezüglich der Veränderungen bei erhöhtem Stimmaufwand ablesbar sind. Für viele Laute sind Veränderungen über 10% vorhanden (siehe Tabelle A.12). Die durchschnittlich geringe Veränderung für die Gesamtklasse der Obstruenten ergibt sich aus der starken Variation der Veränderungen für die einzelnen Obstruenten. Dies deutet daraufhin, dass das Energieverhältnis nicht, wie oben angenommen, robust ist gegenüber Stimmaufwandsveränderungen. Die Art der Veränderung wechselt lediglich stärker. Wertet man nur die Plosive oder ausschließlich die Frikative aus, so sind ebenfalls keine eindeutigen Trends sichtbar. Nichtsdestotrotz steigt die Schiefe und die Wölbung bei Plosiven, mit Ausnahme des Lautes [t], an. Der Mittelwert sinkt für vorne im Mundraum gebildete Laute (labiale), bleibt relativ unverändert für alveolare Plosive und steigt bei den velaren Plosiven. Für die anderen Parameter kann kein eindeutiger Trend beobachtet werden. Hinsichtlich der Frikative kann für keinen Parameter ein klarer Trend herausgearbeitet werden.

LK	SA	D-Wert	p-Wert
Gesamt	N	$0,374$	$< 3,32E{-}296$
	L	$0,356$	$< 3,32E{-}296$
Obstr	N	$0,359$	$< 3,32E{-}296$
	L	$0,351$	$< 3,32E{-}296$
Son	N	$0,279$	$4,92E{-}45$
	L	$0,271$	$1,67E{-}38$
Vok	N	$0,319$	$< 3,32E{-}296$
	L	$0,326$	$< 3,32E{-}296$

Tabelle 6.6: Ergebnisse des Lilliefors-Tests auf Normalverteilung für das Energieverhältnis

LK	Z-Wert	p-Wert
Gesamt	11,7371	$< 2,2E{-}16$
Obstr	-2,7387	0.006169
Son	3,7133	0.0002046
Vokale	25,007	$< 2,2E{-}16$

Tabelle 6.7: Ergebnisse des Mann-Whitney-Test für das Energieverhältnis

Die Klasse der *Sonoranten* ist, wie auch schon bei den anderen spektralen Parametern beobachtet, relativ homogen in ihren Veränderungen über die einzelnen Laute (siehe Tabelle A.13). Es besteht die Tendenz, alle Parameter außer der Schiefe und der Wölbung um mehr als 10% zu erhöhen. Die anderen zwei Parameter sinken. Eine Ausnahme wird für das Energieverhältnis für den Nasal [m] beobachtet, welcher für drei Parameter keine Veränderungen größer 10% aufweist.

Die *Vokale* zeigen für die meisten Werte eine Steigerung bei erhöhtem Stimmaufwand (siehe Tabelle A.14). Die einzigen Werte, die häufiger sinken als steigen, sind die Schiefe und die Wölbung.

Bei der *sprecherspezifischen Auswertung* zeigt sich kein stringentes Muster (siehe Tabelle A.15). Für manche Sprecher sinken sämtliche Werte, für andere steigen alle Werte und wiederum andere haben sowohl steigende als auch sinkende und unveränderte Werte. Das Energieverhältnis scheint dementsprechend von den bisher betrachteten Merkmalen das am stärksten sprecherspezifische zu sein.

Um die bisher beschriebenen Veränderungen der Verteilung näher analysieren und auf signifikante Unterschiede testen zu können, werden zunächst die Gesamtverteilung sowie die Verteilungen der Lautklassen auf Normalverteilung untersucht. Hierfür wird der *Lilliefors-Test* durchgeführt. Die Ergebnisse sind in Tabelle 6.6 nachzulesen. Es ist eindeutig erkennbar, dass keine der Verteilungen des Energieverhältnisses normalverteilt ist. Für die drei Lautklassen und die Gesamtverteilung kann die H_0-Hypothese, dass es sich um eine Normalverteilung handelt, zurückgewiesen werden. Dies ist auch in den Abbildungen 6.6 und 6.7 sichtbar.

Basierend auf den Ergebnissen des Lilliefors-Tests wird, zur Überprüfung der Unterschiede zwischen normaler und lauter Sprache, ein nicht-parametrischer Test, der *Mann-Whitney-Test*, angewendet. Die Ergebnisse des Mann-Whitney-Tests stehen in Tabelle 6.7. Die Unterschiede zwischen normaler und lauter Sprache sind für die Gesamtverteilung ebenso wie für die Verteilung der Vokale hoch signifikant. Für die

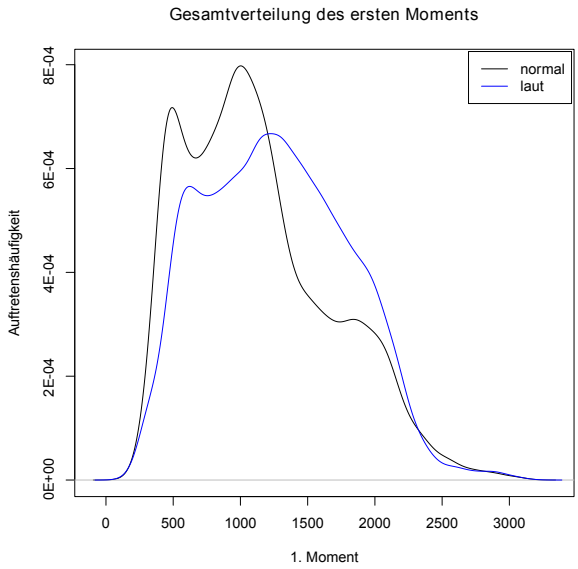

Abbildung 6.8: Verteilung des ersten Moments für normale und laute Sprache über sämtliche Laute und Sprecher

Klasse der Obstruenten und Sonoranten ist auch eine signifikante Veränderung zu beobachten. Diese ist hingegen nicht so groß wie die der anderen zwei Verteilungen.

Anhand der Ergebnisse der Signifikanztests zeigt sich, dass **das Energieverhältnis zur Quantifizierung des Stimmaufwands geeignet sein könnte**. Allerdings ist diese Eignung **vor allem für die Gesamtverteilung und die Verteilung der Vokale** gegeben. Für die anderen zwei Lautklassen sind andere Parameter möglicherweise besser zur Klassifikation geeignet. Um die Eignung im Gesamtkontext weiter untersuchen zu können, wird das Energieverhältnis in Abschnitt 6.3 als Merkmal in einem Stimmaufwandsklassifikator getestet.

6.2.5 Spektrale Momente

In den folgenden Abschnitten werden die Veränderungen der *ersten vier spektralen Momente (Momente)* nach Forrest et al. (1988) bei erhöhtem Stimmaufwand beschrieben.

6.2.5.1 1. Moment

Die Verteilung des *ersten Moments (Mom1)* für die *Gesamtdaten* zeigt für die Werte Schiefe und Wölbung eine Absenkung größer 10% und für den Mittelwert eine Steigerung größer 10% (siehe Tabelle A.16). Die anderen Werte sind nicht oder nur wenig verändert. Abbildung 6.8 veranschaulicht die Verteilung der Gesamtdaten für normale und laute Sprache. Die Veränderung der Wölbung zeigt sich in der Abflachung der blauen gegenüber der schwarzen Kurve. Auch die Veränderung der Schiefe ist sichtbar. Für normale Sprache ist die Kurve rechtsschief. Der Wert der Schiefe sinkt und bewegt sich mehr zur Null hin. Dies äußert sich in einer weniger

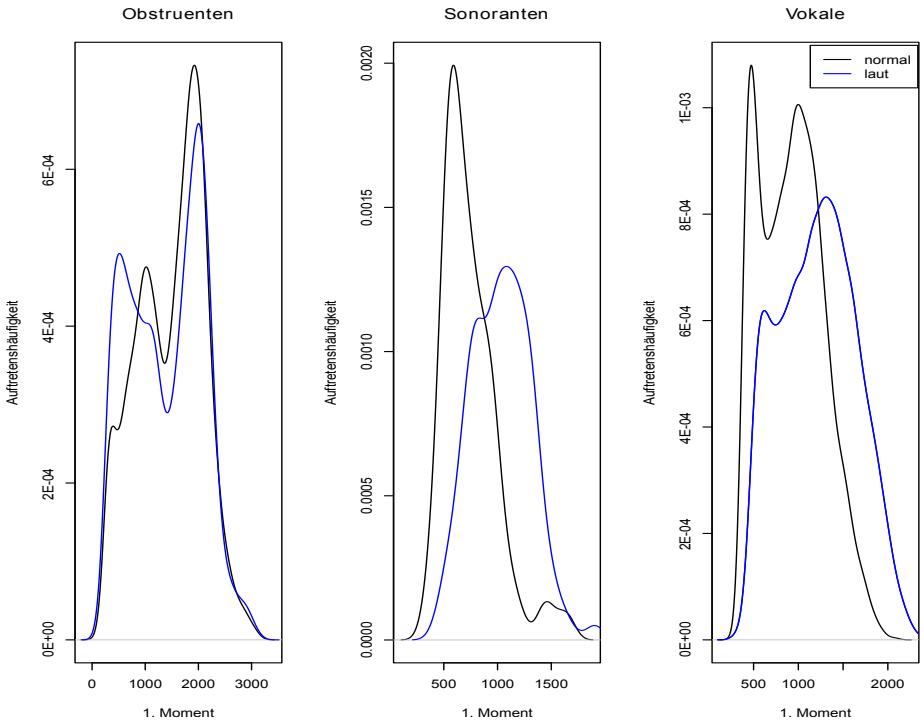

Abbildung 6.9: Verteilung des ersten Moments für normale und laute Sprache der drei Lautklassen Obstruenten, Sonoranten und Vokale

rechtsschiefen Kurve für laute Sprache. Die Werte für Abweichung und Varianz sind weitgehend unverändert, was sich in den Kurven in einer ähnlichen Breite niederschlägt.

Zur genauen Analyse der Differenzen werden die Daten nach *Lautklassen* getrennt betrachtet. Die Ergebnisse sind in Tabelle A.16 und Abbildung 6.9 ersichtlich. Es zeigt sich, dass das erste spektrale Moment der Sonoranten und Vokale wesentlich stärker durch eine Erhöhung des Stimmaufwands beeinflusst ist als das der Obstruenten. Die Obstruenten weisen eine Erhöhung über 10% für die mittlere Abweichung, die Varianz und die Schiefe auf. Die Wölbung wird um mehr als 10% abgesenkt. Die anderen Werte zeigen nur Veränderungen kleiner 10%. Insgesamt bleibt die Form der Verteilung für laute Sprache nahezu unverändert. Für die Sonoranten gibt es dagegen nur einen Wert, der keine Veränderung größer oder gleich 10% aufweist. Die Schiefe und die Wölbung sind abgesenkt, die anderen Werte sind erhöht. Diese Veränderungen sind in Abbildung 6.9 deutlich erkennbar. Die Formen der Verteilungen für normale und laute Sprache sind vollkommen verändert. Auch für die Vokale ist nur ein Wert nicht um mindestens 10% verändert. Die übrigen Parameter verändern sich genau wie die der Sonoranten. Auch die Form der Verteilung für laute Sprache weicht in Abbildung 6.9 von der für normale Sprache ab. Folglich steigen für Sonoranten und Vokale sämtliche Werte außer der Schiefe und der Wölbung stark an. Schiefe und Wölbung sinken stark. Die Obstruenten weisen ein anderes Veränderungsmuster auf.

Diese Veränderungsmuster werden durch die Analyse einzelner Laute detaillierter betrachtet. Zunächst werden die Laute der Klasse *Obstruenten* analysiert (siehe

Tabelle A.17). Die Momente der Verteilung der einzelnen Plosive lassen kein klares Muster erkennen. Der einzige Wert, der konsistent die gleiche Veränderung erfährt, ist die Schiefe. Sie steigt für sämtliche Plosive an, wobei diese Steigerung für einen Plosiv unter 10% liegt. Die Wölbung ist ebenfalls für beinahe alle Plosive um mehr als 10% verändert. Hier existieren sowohl Steigerungen als auch Absenkungen. Für die anderen Parameter ist kein eindeutiger Trend erkennbar. Für die Laute [p], [d] und [g] sind die meisten Parameter unverändert. Die Frikative weisen insgesamt größere Veränderungen auf als die Plosive. Für die mittlere Abweichung, die Standardabweichung, die Varianz und die Schiefe kann eine Absenkung der Werte beobachtet werden, wobei für die ersten drei Werte der Laut [f] eine Ausnahme bildet und für die Schiefe der Laut [z]. Auch die Wölbung ist stets um mehr als 10% verändert, mit der Tendenz zur Steigerung. Die Affrikate [t͡s] zeigt für die meisten Werte starke Veränderungen. Diese sind aber auf Grund der fehlenden Muster für Plosive und Frikative keiner dieser Lautgruppen zuzuordnen.

Bei den *Sonoranten* sind die Werte jeglicher Laute um mehr als 10% erhöht (siehe Tabelle A.18). Für die Nasale [m] und [n] sind die gleichen Veränderungen eingetreten. Sämtliche Werte außer der Schiefe und der Wölbung steigen. Der Liquid [l] dagegen hat je zur Hälfte steigende und fallende Werte. Auch für das erste spektrale Moment bestehen folglich Unterschiede bezüglich der Veränderungen der Nasale und dem Liquid.

Die separat betrachteten *Vokale* zeigen erneut einige Gemeinsamkeiten miteinander (siehe Tabelle A.19). Der Mittelwert und der Maximalwert steigen für sämtliche Vokale, während die Schiefe sinkt. Die Werte der Schiefe sind für alle Vokale positiv, sodass die Verteilung rechtsschief ist. Das Absenken dieses Wertes führt nun zu einer Veränderung der Kurve hin zur symmetrischen Form. Für zwei Laute ergibt sich bei erhöhtem Stimmaufwand eine negative Schiefe, sodass die Kurven leicht linksschief sind. Weiterhin kann für Zentral- und Hinterzungenvokale eine Steigerung der mittleren Abweichung, der Standardabweichung und der Varianz beobachtet werden. Dies führt zu einer breiteren Verteilung. Für die Zentral- und Hinterzungenvokale sinkt der Wert der Wölbung, ebenso wie für die offenen Vorderzungenvokale [a] und [aː].

Bei der *sprecherspezifischen Auswertung* zeigt sich für den Mittelwert eine Steigerung des Wertes für sämtliche Sprecher, auch wenn diese Steigerung nicht durchgehend über 10% liegt (siehe Tabelle A.20). Die Schiefe sinkt für alle Sprecher um mehr als 10%. Die Wölbung sinkt ebenfalls für nahezu alle Sprecher. Zwei Sprecher weisen hier jedoch eine Steigerung größer 10% auf. Weiterhin interessant sind der Maximalwert und die mittlere Abweichung. Beide Werte scheinen für die einzelnen Sprecher relativ robust gegenüber Veränderungen zu sein. Nur wenige Sprecher zeigen Veränderungen größer 10%. Es ist zu prüfen, ob diese Werte zusätzlich eine große Intersprechervariabilität aufweisen, da sie in diesem Fall als Merkmale für die Sprechererkennung relevant sein könnten.

Nach der detaillierten Auswertung der Momente der Verteilungen über einzelne Laute und Sprecher folgt nun eine genauere Analyse der Gesamtverteilung und der Verteilungen der Lautklassen hinsichtlich ihres *Verteilungstyps*. Es wird mit Hilfe des *Lilliefors-Tests* geprüft, ob es sich bei diesen Verteilungen um Normalverteilungen handelt. Die Ergebnisse des Tests sind in Tabelle 6.8 zu finden. Das erste spektrale Moment entspricht dem gewichteten Schwerpunkt der Verteilung und ist damit eine andere Variante des in Abschnitt 2.4.3 beschriebenen gewichteten spektralen Schwer-

LK	SA	D-Wert	p-Wert
Gesamt	N	0,068	$6,49E{-}205$
	L	0,0407	$4,62E{-}72$
Obstr	N	0,079	$6,44E{-}105$
	L	0,0891	$5,24E{-}118$
Son	N	0,105	$9,28E{-}06$
	L	0,0644	0,0553
Vok	N	0,0602	$7,15E{-}97$
	L	0,0431	$3,33E{-}53$

Tabelle 6.8: Ergebnisse des Lilliefors-Tests auf Normalverteilung für das erste spektrale Moment

LK	Z-Wert	p-Wert
Gesamt	25,3836	$< 2,2E{-}16$
Obstr	-6,5796	$4,716E{-}11$
Son	10,3844	$< 2,2E{-}16$
Vokale	48,2937	$< 2,2E{-}16$

Tabelle 6.9: Ergebnisse des Mann-Whitney-Test für das erste spektrale Moment

punktes. Dementsprechend sind die Ergebnisse des Lilliefors-Tests ähnlich denen aus Abschnitt 6.2.3. Die Werte normaler und lauter Sprache der Gesamtverteilung, der Obstruenten und der Vokale sind nicht normalverteilt. Die H_0-Hypothese kann für die sechs Verteilungen mit sehr geringen p-Werten zurückgewiesen werden. Die D-Werte weichen ebenfalls von den in Tabelle 6.1 gegebenen Schranken ab. Diese Abweichung ist ähnlich wie für den gewichteten spektralen Schwerpunkt (siehe Abschnitt 6.2.3) und damit wesentlich geringer als die Abweichung von den Schranken für die spektrale Neigung (siehe Abschnitt 6.2.2). Die Verteilung für Sonoranten normalen Stimmaufwands ist ebenfalls nicht normalverteilt. Die Werte weichen hier, genau wie beim spektralen Schwerpunkt, nicht so stark von den Grenzwerten ab wie die der vorher beschriebenen Verteilungen des ersten Moments. Für laute Sprache kann die H_0-Hypothese nicht zurückgewiesen werden. Hier liegt p knapp über 0,05 und der D-Wert knapp unter der Schranke von 0,683.

Nachdem auch für das erste Moment die meisten Verteilungen nicht normalverteilt sind, wird im Folgenden ebenfalls der nicht-parametrische *Mann-Whitney-Test* angewendet. Die Ergebnisse des Mann-Whitney-Tests sind in Tabelle 6.9 dargestellt. Auch hier sind die Ergebnisse ähnlich denen des gewichteten spektralen Schwerpunktes. Für sämtliche Verteilungen ergeben sich signifikante Unterschiede zwischen lauter und normaler Sprache. Für die Obstruenten ist die Differenz nicht so groß wie für die anderen Verteilungen.

Infolgedessen ist auch **das erste spektrale Moment zur Differenzierung verschiedener Stimmaufwandsgrade nutzbar**. Es ist zu erwarten, dass eine Klassifizierung oder Quantifizierung des Stimmaufwands für Vokale und Sonoranten bessere Ergebnisse erzielt als für Obstruenten. Ob eine Klassifikation des Stimmaufwands mit den spektralen Momenten möglich ist, wird in Abschnitt 6.3 näher untersucht.

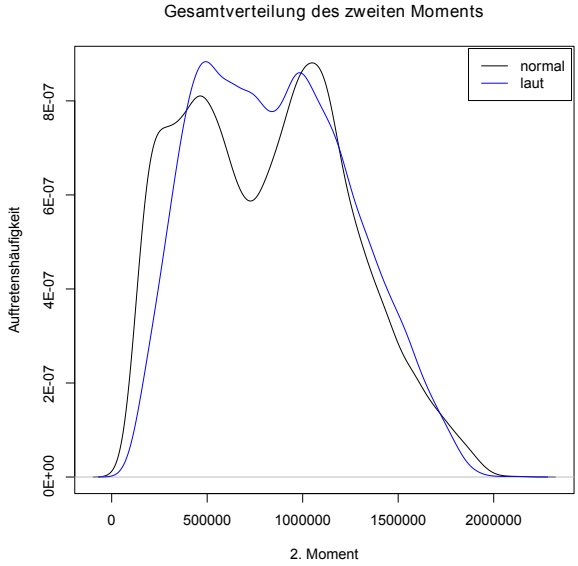

Abbildung 6.10: Verteilung des zweiten Moments für normale und laute Sprache über sämtliche Laute und Sprecher

6.2.5.2 2. Moment

In diesem Abschnitt werden die Veränderungen des *zweiten spektralen Moments (Mom2)* beschrieben. Zunächst wird die Verteilung der *Gesamtdaten* untersucht. Die Ergebnisse werden in Tabelle A.21 und Abbildung 6.10 präsentiert. Auch für das zweite Moment sind insgesamt wenige Veränderungen zu beobachten. Die Varianz sinkt um mehr als 10% und der Minimalwert steigt um mehr als 10%. Die anderen Werte bleiben relativ konstant. Dies wird durch Abbildung 6.10 bestätigt. Die Kurven für laute und normale Sprache sind insgesamt ähnlich, jedoch ist die Kurve für laute Sprache auf Grund der sinkenden Varianz weniger breit. Außerdem sind die in der schwarzen Kurve gut ausgeprägten zwei Maxima in der blauen Kurve für laute Sprache weniger ausgeprägt. Die blaue Kurve ist mehr an die Normalverteilung angenähert.

Die Auswertung nach *Lautklassen* zeigt eine Steigerung des Minimalwertes um mehr als 10% und ein Absinken der Wölbung für die drei Lautklassen (siehe Tabelle A.21). Dies ist in den blauen Kurven aus Abbildung 6.11 sichtbar, welche im Vergleich zu den schwarzen Kurven abgeflacht sind. Darüber hinaus sind die Tendenzen der Sonoranten und Vokale erneut für alle Werte außer einem gleich. Allerdings sind die Veränderungen nicht immer gleich groß. Die Obstruenten haben, abgesehen von dem Wert der Wölbung, immer andere Veränderungen als die Sonoranten und Vokale. Obwohl bei den anderen spektralen Parametern die Veränderungen für die Obstruenten meist eher gering waren, sehen wir für das zweite spektrale Moment eine etwas größere Differenz zwischen lauter und normaler Sprache in Abbildung 6.11.

Diese Differenzen werden nun weiter aufgeschlüsselt, indem die einzelnen Laute näher betrachtet werden. Zunächst wird die Klasse der *Obstruenten* detaillierter beleuchtet (siehe Tabelle A.22). Bei den Plosiven existieren zahlreiche Parameter, die

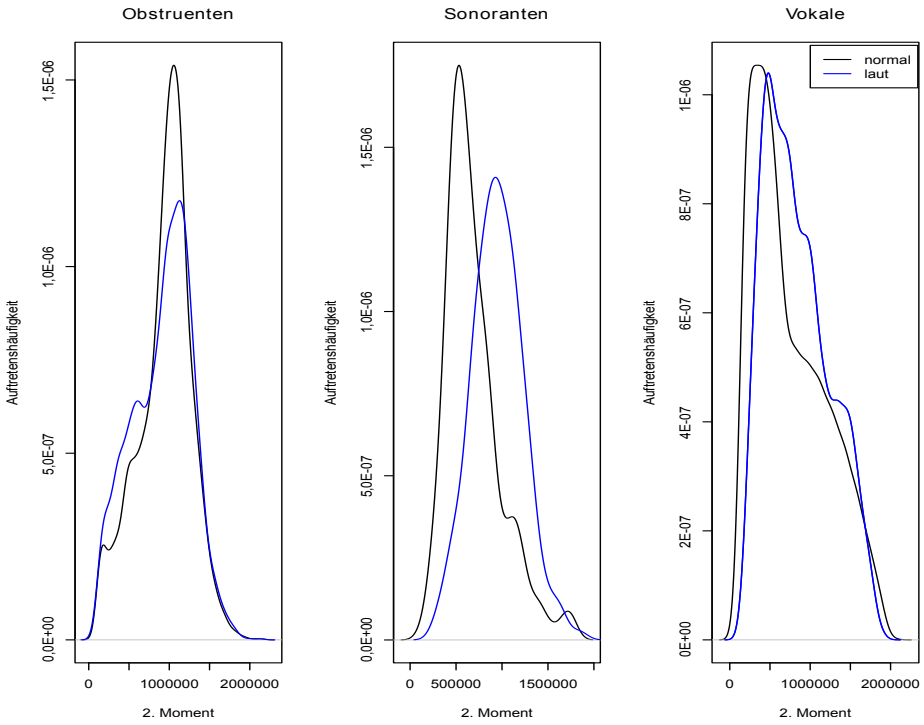

Abbildung 6.11: Verteilung des zweiten Moments für normale und laute Sprache der drei Lautklassen Obstruenten, Sonoranten und Vokale

sich für einzelne Laute nicht beziehungsweise nicht stark verändern. Die einzigen Werte, die sich meistens um mehr als 10% verändern, sind Schiefe und Wölbung. Die Schiefe steigt für sämtliche Plosive, während für die Wölbung kein klarer Trend festzustellen ist. Auch sonst ist für keinen Parameter der Plosive ein Trend feststellbar. Besonders auffällig ist der velare Plosiv [k], welcher nur für die Schiefe und Wölbung größere Veränderungen aufweist. Die anderen Parameter sind nur wenig beeinflusst. Für die Frikative sind ebenfalls Schiefe und Wölbung um mehr als 10% verändert. Die Schiefe steigt, genau wie für die Plosive, für die meisten Laute. Die Wölbung sinkt für labio-dentale Frikative und steigt für alveolare und post-alveolare Frikative sowie für die Affrikate [t͡s]. Der Minimalwert steigt für sämtliche Frikative, wobei [f] der einzige Frikativ ist, für den die Veränderung kleiner 10% ist. Die Laute [f] und [s] sind insgesamt die Frikative mit der geringsten Differenz zwischen normaler und lauter Sprache. Die Affrikate [t͡s] hingegen zeigt einige Veränderungen über 10%. Die Erhöhung der Schiefe, welche die einzige Veränderung ist, die sowohl für Plosive als auch für Frikative beobachtet werden konnte, gilt nicht für die Affrikate.

Die Werte der *Sonoranten* für das zweite spektrale Moment sind, wie auch für die anderen spektralen Merkmale, meistens um mehr als 10% verändert (siehe Tabelle A.23). Die Werte der Nasallaute [m] und [n] werden gesenkt für die Schiefe und die Wölbung. Die anderen Werte werden erhöht. Der Liquid [l] unterscheidet sich von den Nasalen bezüglich der Wölbung, der mittleren Abweichung, der Standardabweichung und der Varianz. Zusätzlich ist der Maximalwert zwar erhöht, jedoch weniger als 10%.

LK	SA	D-Wert	p-Wert
Gesamt	N	0,0654	$5,39E-189$
	L	0,053	$2,49E-124$
Obstr	N	0,0755	$1,37E-95$
	L	0,0641	$9,56E-60$
Son	N	0,114	$8,04E-07$
	L	0,0281	0,973
Vok	N	0,110	$< 3,32E-296$
	L	0,0711	$2,58E-149$

Tabelle 6.10: Ergebnisse des Lilliefors-Tests auf Normalverteilung für das zweite spektrale Moment

Bei den *Vokalen* ist nur ein Parameter durchgängig für jegliche Laute um mehr als 10% abgesenkt: die Schiefe (siehe Tabelle A.24). Bei einigen anderen Parametern sind klare Tendenzen erkennbar. Der Mittelwert beispielsweise steigt für Zentral- und Hinterzungenvokale. Die mittlere Abweichung, die Standardabweichung und die Varianz sinken für fast alle Vorderzungen- und Zentralvokale und steigen für Hinterzungenvokale. Diese Veränderungen der Streuungsparameter sind für die meisten Laute größer 10%.

Bei der *sprecherspezifischen Auswertung* zeigt sich kein klares Muster (siehe Tabelle A.25). Zeigt sich eine Veränderung größer 10% für den Minimalwert, so steigt er meistens. Diese Steigerung geht konform mit der Gesamtauswertung zu Beginn dieses Abschnittes. Allerdings gibt es auch Sprecher, für die der Minimalwert um mehr als 10% sinkt. Die Veränderung des zweiten spektralen Moments scheint dementsprechend stark von dem einzelnen Sprecher abhängig zu sein.

Zur genaueren Analyse der Gesamtverteilung sowie der Verteilungen der Lautklassen werden diese nun *auf Normalverteilung getestet*. Hierfür wird der *Lilliefors-Test* durchgeführt. Die Ergebnisse für laute und normale Sprache sind in Tabelle 6.10 ersichtlich. Ähnlich wie beim ersten Moment sind die Verteilungen normaler und lauter Sprache für die Gesamtdaten, die Obstruenten und die Vokale nicht-normalverteilt. Die p-Werte sind sehr gering für diese Verteilungen. Die Sonoranten bilden auch beim zweiten Moment eine Ausnahme. Die Stichprobe der normalen Sprache ist, wie die der anderen Verteilungen, nicht-normalverteilt. Der p-Wert ist hingegen nicht so gering wie die Werte der vorher beschriebenen Verteilungen. Für die Verteilung lauter Sprache kann die H_0-Hypothese, dass es sich um eine Normalverteilung handelt, nicht zurückgewiesen werden. Mit 0,97 ist der p-Wert sehr hoch und auch der D-Wert ist weit unter den in Tabelle 6.1 angegebenen Schranken der drei Signifikanzniveaus 0,05, 0,02 und 0,01. Somit gilt für die Sonoranten lauter Sprache die H_1-Hypothese.

Die Ergebnisse des Lilliefors-Tests zeigen, dass die meisten Verteilungen nicht normalverteilt sind. Aus diesem Grund wird im Folgenden der nicht-parametrische *Mann-Whitney-Test* für die weiteren Untersuchungen verwendet. Die Ergebnisse sind in Tabelle 6.11 dargestellt. Wie beim ersten Moment ähneln die Ergebnisse der Mann-Whitney-Tests tendenziell den Testergebnissen des spektralen Schwerpunkts. Die Unterschiede sind für die Gesamtverteilung, die Sonoranten und die Vokale hoch signifikant. Auch für die Obstruenten sind die Unterschiede signifikant. Hier ist der p-Wert hingegen wesentlich größer als für die anderen Verteilungen. Die

LK	Z-Wert	p-Wert
Gesamt	9,8721	$< 2,2E-16$
Obstr	-4,2063	$2,596E-05$
Son	9,1243	$< 2,2E-16$
Vokale	19,4614	$< 2,2E-16$

Tabelle 6.11: Ergebnisse des Mann-Whitney-Test für das zweite spektrale Moment

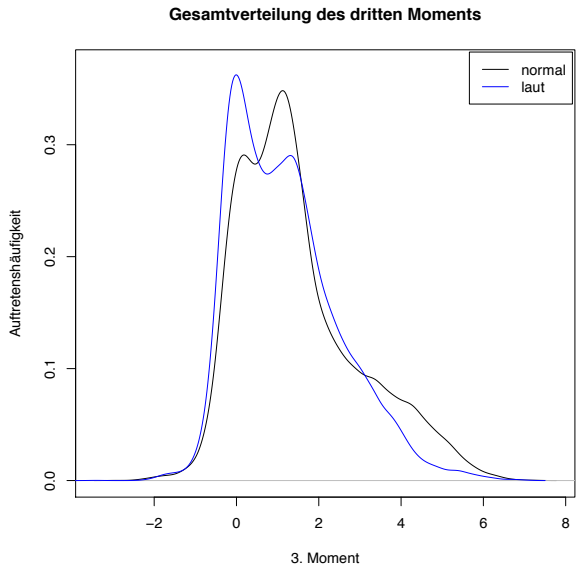

Abbildung 6.12: Verteilung des dritten Moments für normale und laute Sprache über sämtliche Laute und Sprecher

Begründung für die nicht so starke Signifikanz könnte, wie beim gewichteten spektralen Schwerpunkt (siehe Abschnitt 6.2.3), in der nicht signifikanten Veränderung der Plosive liegen (Gottsmann, 2010).

Insgesamt kann gefolgert werden, dass auch **das zweite spektrale Moment zur Quantifizierung des Stimmaufwands geeignet ist**. Besonders für Sonoranten und Vokale scheint das zweite Moment eine zuverlässige Unterscheidung zwischen normaler und lauter Sprache zu liefern. Die Obstruenten weisen nicht so große Unterschiede auf. Diese Tendenzen gleichen denen des gewichteten spektralen Schwerpunktes. Ob das zweite Moment insgesamt für die Klassifikation des Stimmaufwands geeignet ist, wird in Abschnitt 6.3 näher untersucht.

6.2.5.3 3. Moment

Im Folgenden wird die Veränderung des *dritten spektralen Moments (Mom3)* durch erhöhten Stimmaufwand beschrieben. Hierfür wird zunächst die Verteilung der *Gesamtdaten* untersucht. Die Momente der Verteilung sind in Tabelle A.26 dargestellt. Abbildung 6.12 ist die zugehörige Visualisierung der Verteilung. Für das dritte spektrale Moment werden alle Werte außer der Wölbung gesenkt. Nur zwei der

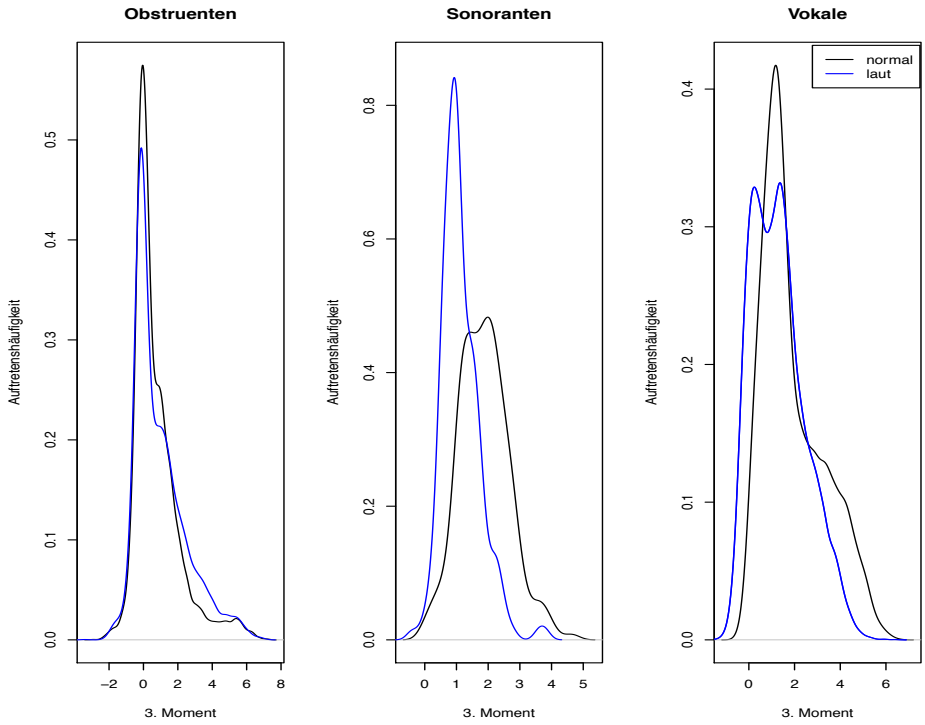

Abbildung 6.13: Verteilung des dritten Moments für normale und laute Sprache der drei Lautklassen Obstruenten, Sonoranten und Vokale

abgesenkten Werte sind weniger als 10% verändert. Die übrigen Werte sind um mindestens 10% verändert. Dies ist ein starker Unterschied zu den ersten zwei Momenten, bei denen nur zwei beziehungsweise drei Parameter verändert waren.

Um diese starken Veränderungen für das dritte spektrale Moment besser analysieren zu können, werden die Daten nach *Lautklassen* getrennt ausgewertet (siehe Tabelle A.26). Die Sonoranten und Vokale zeigen erneut ein ähnliches Veränderungsmuster. Die Wölbung steigt für beide Lautklassen mehr als 10%, während die anderen Werte sinken. Lediglich der Wert der Schiefe wird bei den Sonoranten gesteigert und für Vokale abgesenkt. Für die Obstruenten steigen die Momente Mittelwert, mittlere Abweichung, Standardabweichung und Varianz, jeweils um mehr als 10%. Der Minimalwert sinkt um mehr als 10%, ebenso wie die Schiefe und die Wölbung. Die Betrachtung der Veränderungen der drei Lautklassen in Abbildung 6.13 macht ähnliche Tendenzen wie bei anderen spektralen Parametern sichtbar. Im Vergleich zu der Verteilung der Obstruenten sind die Kurven der Sonoranten und Vokale stärker verändert, da auch die Form der Verteilung betroffen ist. Die Form der Verteilung bei den Obstruenten bleibt hingegen weitestgehend erhalten.

Zur detaillierteren Analyse werden im Folgenden die einzelnen Laute der Klasse *Obstruenten* näher untersucht (siehe Tabelle A.27). Für Plosive zeigt sich, dass der Mittelwert steigt, sofern er sich verändert. Der Wert der Schiefe sinkt für sämtliche Plosive. Die anderen Parameter sind für die einzelnen Laute unterschiedlich, sodass kein klarer Trend erkennbar ist. Bezüglich der Frikative finden viele Absenkungen statt, vor allem für den Maximalwert sowie die Streuungsparameter mittlere Abweichung, Standardabweichung und Varianz. Eine Ausnahme bildet der Laut

LK	SA	D-Wert	p-Wert
Gesamt	N	$0,103$	$< 3,32E{-}296$
	L	$0,0646$	$3,38E{-}186$
Obstr	N	$0,131$	$3,32E{-}296$
	L	$0,127$	$6,8E{-}245$
Son	N	$0,0469$	$0,324$
	L	$0,121$	$4,33E{-}07$
Vok	N	$0,127$	$< 3,32E{-}296$
	L	$0,0539$	$1,95E{-}84$

Tabelle 6.12: Ergebnisse des Lilliefors-Tests auf Normalverteilung für das dritte spektrale Moment

[f], dessen Werte, abgesehen von dem Mittelwert, steigen. Weiterhin lässt sich für sämtliche Frikative, außer dem post-alveolaren [ʃ], die Tendenz zur Steigerung des Minimalwertes beobachten. Interessant ist besonders die Affrikate [t͡s], da hier sämtliche Werte um mehr als 10% steigen. Eine solch eindeutige Veränderung ist zuvor bei keinem Laut für keinen spektralen Parameter beobachtet worden.

Die *Sonoranten* zeigen erneut ein relativ eindeutiges Bild (siehe Tabelle A.28). Für den Nasal [n] sinken sämtliche Werte um mehr als 10%. Bei den anderen zwei Lauten sinken alle Werte mit Ausnahme der Werte der Schiefe und der Wölbung, welche um mehr als 10% steigen. Interessant ist, dass die Veränderungen des Nasals [m] denen des Liquids [l] ähnlicher sind als denen des Nasals [n].

Auch die *Vokale* zeigen für einige Parameter klare Muster (siehe Tabelle A.29). Der Mittelwert und der Minimalwert sinken für sämtliche Werte um mehr als 10%. Der Maximalwert sinkt ebenfalls für die meisten Vokale, für [iː] und [eː] steigt er hingegen um mehr als 10%. Die Streuungsparameter sind für die meisten Vokale unverändert. Tritt jedoch eine Veränderung um mehr als 10% auf, so werden die Werte abgesenkt. Diese Absenkung tritt meistens für gelängte Vokale auf. Für die Schiefe und Wölbung sind sämtliche Werte um mindestens 10% verändert. Es ergibt sich allerdings kein einheitliches Muster bezüglich der Art der Veränderung.

Bei der *sprecherspezifischen Auswertung* ist der Mittelwert für alle Sprecher abgesenkt (siehe Tabelle A.30). Nur für einen Sprecher ist diese Senkung kleiner 10%. Die Streuungsparameter verhalten sich ähnlich wie bei den Vokalen. Viele sind unverändert, aber die Werte, die um mehr als 10% verändert sind, sinken, mit einer Ausnahme, ab. Auch der Maximalwert ist für die meisten Sprecher abgesenkt. Die anderen Parameter sind je nach Sprecher unterschiedlich verändert.

Zur detaillierten Analyse wird als Nächstes eine *Untersuchung auf Normalverteilung* für die Gesamtverteilung und die Verteilungen der Lautklassen durchgeführt. Hierfür wird der *Lilliefors-Test* verwendet. Die Ergebnisse der Tests auf Normalverteilung sind in Tabelle 6.12 abgebildet. Auch für das dritte Moment sind die Gesamtverteilung sowie die Verteilungen der Obstruenten und Vokale sowohl für laute als auch für normale Sprache nicht-normalverteilt. Die Sonoranten bilden erneut eine Ausnahme. Wie bei der spektralen Neigung ist die H_0-Hypothese, dass es sich bei der betrachteten Verteilung um eine Normalverteilung handelt, für normalen Stimmaufwand nicht zurückweisbar. Diese Homogenität hängt damit zusammen, dass das dritte spektrale Moment ein mögliches Maß für die spektrale Neigung ist. Die Sonoranten lauter Sprache sind dagegen abermals nicht-normalverteilt. Der p-Wert

LK	Z-Wert	p-Wert
Gesamt	-17,0732	$< 2,2E-16$
Obstr	5,2644	$1,406E-07$
Son	-9,952	$< 2,2E-16$
Vokale	-32,1066	$< 2,2E-16$

Tabelle 6.13: Ergebnisse des Mann-Whitney-Test für das dritte spektrale Moment

ist jedoch wesentlich größer als die p-Werte der anderen nicht-normalverteilten Verteilungen für das dritte Moment.

Da auch für das dritte spektrale Moment die meisten Stichproben nicht-normalverteilt sind, wird im Folgenden der *Mann-Whitney-Test* zur Untersuchung der Differenz zwischen normaler und lauter Sprache angewendet. Die Ergebnisse der Mann-Whitney-Tests sind in Tabelle 6.13 nachzulesen. Die Differenz zwischen normaler und lauter Sprache ist für sämtliche untersuchten Verteilungen signifikant. Es zeigt sich weiterhin, dass die Obstruenten, wie auch bei einigen anderen spektralen Parametern, erneut eine Ausnahme bilden, da hier der p-Wert größer ist, im Vergleich zu den anderen Verteilungen. Das bedeutet, dass die Differenz zwischen normaler und lauter Sprache kleiner ist für Obstruenten. Für die spektrale Neigung wurde dies nicht beobachtet.

Die Betrachtung der Gesamtauswertung dieses Abschnitts verdeutlicht, dass **das dritte spektrale Moment als möglicher Parameter zur Quantifizierung des Stimmaufwands in Frage kommt**. Für die Obstruenten kann voraussichtlich keine so gute Unterscheidung des Stimmaufwands erfolgen wie für die Sonoranten und Vokale. Die Leistung der spektralen Momente als Parameter zur Stimmaufwandsklassifikation wird in Abschnitt 6.3 überprüft.

6.2.5.4 4. Moment

In diesem Abschnitt wird der Einfluss von erhöhtem Stimmaufwand auf das *vierte spektrale Moment (Mom4)* untersucht. Zunächst werden die *Gesamtverteilungen* aller Daten normaler und lauter Sprache betrachtet (siehe Tabelle A.31). Die Schiefe und die Wölbung steigen um mehr als 10%. Das Ansteigen der Wölbung führt zu einer größeren Steilheit für laute Sprache. Dies ist in Abbildung 6.14 zu sehen. Weiterhin sinken der Mittelwert, die mittlere Abweichung, die Standardabweichung und die Varianz um mehr als 10%. Demnach ist insgesamt eine klare Veränderung zwischen den Parametern der Verteilung für normale und laute Sprache festzustellen.

Diese Differenzen sollen für verschiedene *Lautklassen* untersucht werden. Sie sind in Tabelle A.31 und Abbildung 6.15 dargestellt. Die Obstruenten verändern sich bei erhöhtem Stimmaufwand konträr zu der Gesamtauswertung. Der Mittelwert, die mittlere Abweichung und die Varianz steigen um mehr als 10%, während die Werte der Schiefe und Wölbung um mehr als 10% sinken. In Abbildung 6.15 zeigt sich der niedrigere Wert der Wölbung für laute Sprache durch das höhere Maximum für Obstruenten normaler Sprache. Die Sonoranten und Vokale verhalten sich ähnlich wie in der Gesamtauswertung beobachtet. Die Werte der Schiefe und Wölbung steigen, während die anderen Werte, mit Ausnahme des Minimalwertes, sinken. Dies gilt sowohl für Sonoranten als auch für Vokale.

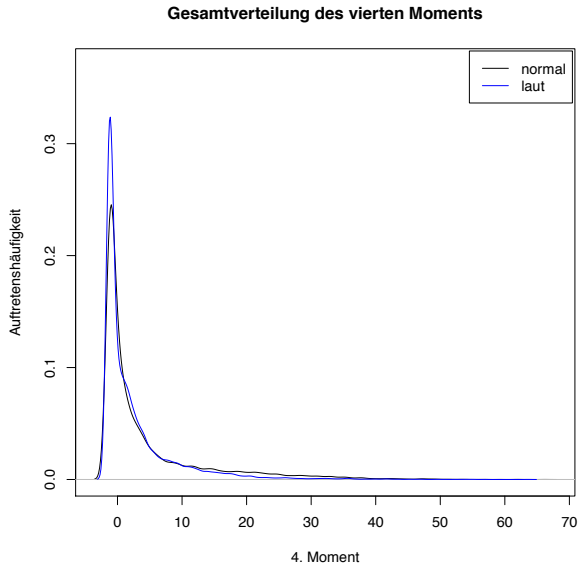

Abbildung 6.14: Verteilung des vierten Moments für normale und laute Sprache über sämtliche Laute und Sprecher

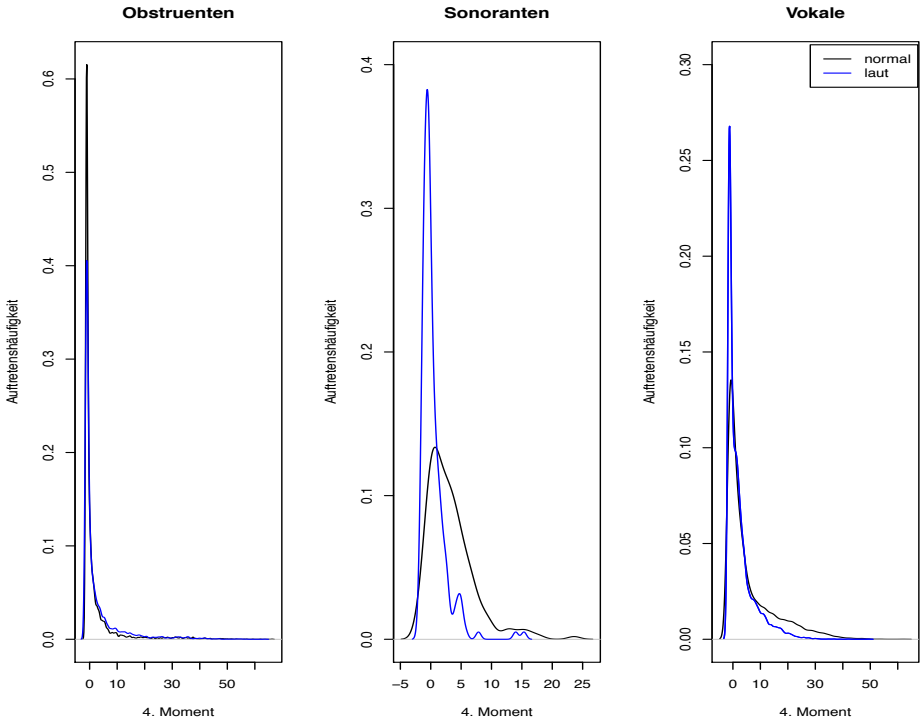

Abbildung 6.15: Verteilung des vierten Moments für normale und laute Sprache der drei Lautklassen Obstruenten, Sonoranten und Vokale

Für die einzelnen Laute der *Obstruenten* kann kein eindeutiges Muster bezüglich der Veränderungen bedingt durch erhöhten Stimmaufwand festgestellt werden (siehe Tabelle A.32). Auffällig ist, dass die Minimalwerte sowohl für die Plosive als auch für die Frikative selten über 10% verändert werden. Weiterhin interessant ist, dass der Laut [f], wie beim dritten spektralen Moment, abgesehen von dem Mittelwert nur steigende Werte aufweist. Der Laut [f] ist der einzige Frikativ, für den die Streuungsparameter mittlere Abweichung, Standardabweichung und Varianz nicht sinken. Weitere einheitliche Veränderungsmuster sind weder bei den Plosiven noch bei den Frikativen feststellbar.

Die *Sonoranten* dagegen weisen abermals mehr Ähnlichkeiten auf (siehe Tabelle A.33). Die Veränderungsmuster ähneln denen des dritten spektralen Moments. Für die drei Laute sind die Parameter Mittelwert, Maximalwert, mittlere Abweichung, Standardabweichung und Varianz um mehr als 10% abgesenkt. Der Minimalwert ist ebenfalls abgesenkt für die Nasale, nicht aber für den Liquid [l]. Die Werte der Schiefe und Wölbung sind für [l] und [m] um mindestens 10% erhöht. Für den Nasal [n] sind diese Werte abgesenkt. Dies bedeutet, dass sämtliche Parameter des Nasals [n] um mehr als 10% abgesenkt werden.

Die *Vokale* zeigen für den Mittelwert und die mittlere Abweichung klare Ergebnisse (siehe Tabelle A.34). Die Werte sinken für sämtliche Laute um mindestens 10%. Auch die anderen zwei Streuungsparameter zeigen die Tendenz einer Verringerung um mehr als 10%. Die Ausnahme bildet hierbei der Laut [e:], für den beide Werte um mehr als 10% steigen. Der Minimalwert ist für die meisten Vokale relativ unbeeinflusst. Dies konnte bereits bei anderen spektralen Parametern beobachten werden (siehe oben). Für Hinterzungenvokale ist der Minimalwert hingegen um mehr als 10% abgesenkt.

Die *sprecherbezogene Auswertung* der Daten zeigt erneut ein Absinken des Mittelwertes und der mittleren Abweichung für sämtliche Sprecher (siehe Tabelle A.35). Außer für einen Sprecher ist diese Absenkung für alle Sprecher größer 10%. Die anderen zwei Streuungsparameter sowie der Maximalwert haben ebenfalls die Tendenz, stark abgesenkt zu werden, hier gibt es jedoch einige Ausnahmefälle für verschiedene Sprecher. Besonders auffällig ist bei dieser Untersuchung der Minimalwert. Dieser ist für die meisten Sprecher beinahe unverändert. Ähnlich wie für die mittlere Abweichung und den Maximalwert des 1. Moments in Abschnitt 6.2.5.1 gilt es zu prüfen, ob dieser Parameter eine ausreichende Intersprechervariabilität zeigt, sodass er als Merkmal für die Sprechererkennung nutzbar ist.

Die oben beschriebenen Ergebnisse für die Parameter der Verteilungen für das vierte Moment sollen auf Signifikanz geprüft werden. Vorab müssen *Tests auf Normalverteilung* für die Gesamtverteilung und die Verteilungen der Lautklassen mit Hilfe *Lilliefors-Test* durchgeführt werden. Die Ergebnisse dieser Tests sind in Tabelle 6.14 dargestellt. Für sämtliche Verteilungen, sowohl für laute als auch für normale Sprache, kann die H_0-Hypothese zurückgewiesen werden. Dies bedeutet, dass keine der Verteilungen normalverteilt ist und ein nicht-parametrischer Test zur Überprüfung der Unterschiede zwischen normaler und lauter Sprache benutzt werden muss.

Nachfolgend wird der *Mann-Whitney-Test* als nicht-parametrischer Signifikanztest eingesetzt. Tabelle 6.15 zeigt die Ergebnisse der Signifikanztests. Die Unterschiede zwischen normaler und lauter Sprache sind für alle untersuchten Verteilungen signifikant. Die p-Werte sind für sämtliche Verteilungen sehr gering. Für die Obstruenten

LK	SA	D-Wert	p-Wert
Gesamt	N	$0,241$	$< 3,32E{-}296$
	L	$0,242$	$< 3,32E{-}296$
Obstr	N	$0,317$	$< 3,32E{-}296$
	L	$0,286$	$< 3,32E{-}296$
Son	N	$0,117$	$2,55E{-}07$
	L	$0,190$	$3,32E{-}18$
Vok	N	$0,202$	$< 3,32E{-}296$
	L	$0,210$	$< 3,32E{-}296$

Tabelle 6.14: Ergebnisse des Lilliefors-Tests auf Normalverteilung für das vierte Moment

LK	Z-Wert	p-Wert
Gesamt	-16,7922	$< 2,2E{-}16$
Obstr	7,3907	$1,461E{-}13$
Son	-9,8523	$< 2,2E{-}16$
Vokale	-27,3897	$< 2,2E{-}16$

Tabelle 6.15: Ergebnisse des Mann-Whitney-Test für das vierte spektrale Moment

ist der p-Wert nicht so klein wie für die anderen Verteilungen. Die D-Werte sind für sämtliche Verteilungen stark von den in Tabelle 6.1 vorgegebenen Schranken entfernt, sodass von hoch signifikanten Unterschieden zwischen den zwei Stimmaufwandsklassen ausgegangen werden kann.

Diese signifikanten Unterschiede zwischen normalem und erhöhtem Stimmaufwand führen zu der These, dass **das vierte spektrale Moment zur Quantifizierung und Klassifikation des Stimmaufwands geeignet ist**. Diese These wird in Abschnitt 6.3 getestet, indem die spektralen Momente als Parameter in einem Stimmaufwandsklassifikator angewendet werden.

6.2.6 Vergleich der Parameter

Dieser Abschnitt vergleicht die statistischen Auswertungen der spektralen Parameter der vorherigen Abschnitte. Zunächst werden die Ergebnisse der *Gesamtverteilung* verglichen. Hierbei fällt auf, dass viele Momente der Verteilung der spektralen Merkmale spektrale Neigung, drittes und viertes Moment um mehr als 10% verändert sind. Die Verteilungen der anderen spektralen Parameter dagegen haben nur wenige stark veränderte Momente. Die Signifikanztests dagegen ergeben signifikante Veränderungen bei erhöhtem Stimmaufwand der Gesamtverteilungen aller spektraler Parameter. Der Vergleich des ersten spektralen Moments mit dem spektralen Schwerpunkt bestätigt tendenziell gleiche Veränderungen. Für das dritte Moment und die spektrale Neigung ist dies nicht der Fall. Bei der Betrachtung der Veränderung des Stimmaufwands unabhängig von den gesprochenen Lauten und unabhängig vom Sprecher kann festgestellt werden, dass sämtliche untersuchten Parameter zur Unterscheidung von normaler und lauter Sprache genutzt werden können. Die spektrale Neigung sowie das dritte und vierte Moment scheinen bei dem Vergleich der Momente der Gesamtverteilung (siehe Anhang A.1) insgesamt

besser geeignet zu sein als beispielsweise das Energieverhältnis. Die Z-Werte der Gesamtverteilungen verdeutlichen hingegen, dass der COG und das erste spektrale Moment die größte Differenz zwischen normalem und erhöhtem Stimmaufwand aufweisen, während das Energieverhältnis sowie das zweite spektrale Moment die geringste Differenz zeigen.

Bei dem Vergleich der Verteilungen der spektralen Parameter für die *Obstruenten* gibt es die wenigsten veränderten Momente für das Energieverhältnis. Für die spektrale Neigung und das dritte spektrale Moment sind die meisten Veränderungen sichtbar. Beim Vergleich der Veränderungen des ersten spektralen Moments mit denen des gewichteten spektralen Schwerpunktes sind auch hier gleiche Tendenzen zu finden. Für das dritte Moment und die spektrale Neigung ist dies, ebenso wie für die Gesamtverteilung, nicht der Fall. Die Signifikanztests der Obstruenten zeigen für alle Parameter einen signifikanten Unterschied zwischen normaler und lauter Sprache. Für die spektrale Neigung ist der p-Wert so niedrig wie für die anderen Lautklassen und die Gesamtverteilung. Die anderen Verteilungen haben größere p-Werte für die Obstruenten als für die anderen Lautklassen und die Gesamtverteilung. Hieraus lässt sich ableiten, dass das Spektrum der Obstruenten nicht so stark durch die Erhöhung des Stimmaufwands beeinflusst wird wie die Spektren der anderen Lautklassen. Das Energieverhältnis zeigt insgesamt den größten p-Wert der Obstruenten. Die Obstruenten sind außerdem für alle spektralen Parameter nicht-normalverteilt.

Die Momente der Verteilungen der *Sonoranten* erfahren ausgeprägte Veränderungen für sämtliche spektralen Parameter. Für die meisten Parameter sind alle oder zumindestens die meisten Momente um mehr als 10% verändert. Die Ausnahme bildet hierbei das zweite spektrale Moment, bei dem drei der Momente der Verteilung unverändert beziehungsweise wenig verändert sind. Ein Vergleich der Veränderungen des ersten Moments mit denen des gewichteten spektralen Moments zeigt die gleichen Tendenzen für beide Parameter. Die gleichen Momente werden für beide Parameter jeweils um mehr als 10% erhöht oder abgesenkt und nur ein Moment, die Standardabweichung, wird nicht um mehr als 10% verändert. Die Tendenzen der spektralen Neigung sind hingegen nicht vergleichbar mit denen des dritten spektralen Moments. Bei der Untersuchung der Signifikanztests zeigen sich signifikante Veränderungen für sämtliche spektralen Parameter. Für das Energieverhältnis ist diese Veränderung geringer als für die anderen Parametern, da der p-Wert relativ groß ist im Vergleich zu denen der anderen Parameter. Weiterhin ist interessant, dass die Sonoranten für sämtliche Parameter eine Verteilung aufweisen, entweder laut oder normal, welche möglicherweise normalverteilt ist. Für die anderen Lautklassen und die Gesamtverteilung kann eine Normalverteilung für alle spektralen Parameter ausgeschlossen werden.

Die Momente der Verteilungen der *Vokale* sind für sämtliche untersuchten spektralen Parameter stark verändert, auch für das Energieverhältnis. Hier besteht ebenfalls eine Übereinstimmung bezüglich der Veränderungstendenzen für den spektralen Schwerpunkt und das erste spektrale Moment. Für das dritte Moment und die spektrale Neigung kann keine solche Übereinstimmung gefunden werden. Die starken Veränderungen der Momente für alle Parameter spiegeln sich in den Signifikanztests durch sehr niedrige p-Werte wider. Für sämtliche Parameter konnten signifikante Unterschiede zwischen normalem und erhöhtem Stimmaufwand für Vokale gefunden werden. Die Verteilungen lauter und normaler Sprache waren für

jegliche Parameter nicht normalverteilt.

Die *sprecherspezifische Auswertung* zeigt für jeden Parameter einige stark sprecherspezifische Momente. Einige Parameter zeigen gar keine klaren Trends. Hierzu zählen das zweite spektrale Moment und die Energieverteilung. Für diese Parameter scheint kein Moment der Verteilung über alle Sprecher ähnlich verändert zu werden. Für die anderen Parameter existieren hingegen einige Momente, welche solche Tendenzen zeigen. Besonders hervorzuheben sind im Rahmen der sprecherspezifischen Auswertung der Maximalwert und die mittlere Abweichung des ersten spektralen Moments und der Minimalwert des vierten spektralen Moments. Diese Parameter werden für fast alle Sprecher sehr geringfügig verändert. Es ist möglich, dass es sich bei diesen Parametern um robuste Merkmale für sprachverarbeitende Systeme handelt, die beispielsweise für die Sprechererkennung genutzt werden könnten. Hierfür muss geprüft werden, ob die Intersprechervariabilität groß genug und die Intrasprechervariabilität klein genug ist.

6.3 Der Stimmaufwandsklassifikator

Der Stimmaufwandsklassifikator soll die oben ausgeführten Untersuchungen zur Quantifizierung des Stimmaufwands vertiefen. Es wird geprüft, ob die oben dargestellten spektralen Parameter zur Unterscheidung normaler und lauter Sprache in einem Stimmaufwandklassifikator verwendet werden können und ob sie einen Mehrwert liefern in Kombination mit den Standard-MFCC-Merkmalen. Hierfür wird ein System erstellt, welches je Merkmal oder Merkmalskombination für beide Stimmaufwandsgrade (normal und erhöht) jeweils ein GMM mit 64 Mischungskomponenten trainiert. Angelehnt an Hansen und Varadarajan (2009) werden für die Kombination der spektralen Parameter mit den MFCC-Merkmalen die 19 ersten Koeffizienten ohne den nullten Koeffizienten verwendet. Eine Aufteilung nach Lauten oder Lautklassen wird nicht vorgenommen. Die Klassifikation wird auf spontansprachlichen Äußerungen des „Pool 2010“-Korpus durchgeführt (siehe Abschnitt 5.2.1). Das „Pool 2010“-Korpus ist so eingeteilt, dass insgesamt 56 Trainingsdateien und 50 Testdateien je Stimmaufwandsgrad zur Verfügung stehen.

Zuerst werden die spektralen Parameter separat als Merkmale für die Stimmaufwandsklassifikation getestet. Die Ergebnisse sind in Tabelle 6.16 dargestellt. Der Vergleich der ersten sieben Merkmale aus Tabelle 6.16 zeigt, dass der COG und das erste spektrale Moment von den einzelnen spektralen Merkmalen die besten Ergebnisse liefern. Außerdem ist ersichtlich, dass beide Parameter, genau wie in der statistischen Analyse in Abschnitt 6.2, gleiche Ergebnisse erzielen. Daher kann geschlussfolgert werden, dass der COG und das erste spektrale Moment, berechnet nach Forrest et al. (1988), tatsächlich gleichwertig für die Klassifikation von lauter und normaler Sprache sind. Für die spektrale Neigung und das dritte spektrale Moment gilt eine solche Gleichsetzung nach den Daten nicht. Auch wenn das dritte Moment in der Literatur häufig gleich gesetzt wird mit der spektralen Neigung, sind die Ergebnisse der Klassifikation nicht identisch. Dies bestätigt die Ergebnisse der statistischen Analyse, die bereits in Abschnitt 6.2 Unterschiede zwischen den beiden Parametern aufzeigte. Insgesamt stimmt die gute Klassifikation des COG-Wertes und des ersten spektralen Moments mit den hohen Z-Werten der Mann-Whitney-Tests aus Abschnitt 6.2 überein. Auch die schlechten Klassifikations-

Merkmal	Richtige Detektionen	Vertauschungen (normal → laut)	Vertauschungen (laut → normal)
SN	85	4	11
COG	90	5	5
EV	63	17	20
Mom1	90	5	5
Mom2	77	13	10
Mom3	82	10	8
Mom4	81	10	9
Momente	92	5	3
COG+SN+EV	94	4	2
Momente+COG+SN+EV	92	5	3

Tabelle 6.16: Ergebnisse des Stimmaufwandsklassifikators für die spektralen Merkmale

ergebnisse spiegeln sich in den Z-Werten des Energieverhältnisses und des zweiten spektralen Moments aus Abschnitt 6.2 wider.

Um eine Verbesserung der Stimmaufwandsklassifikation zu erzielen, wurden unterschiedliche spektrale Parameter miteinander kombiniert. Die Ergebnisse dieser Merkmalskombinationen sind ebenfalls in Tabelle 6.16 abzulesen. Für die drei Merkmalskombinationen werden bessere Ergebnisse erreicht als für die einzelnen Merkmale. Die Kombination COG+SN+EV liefert insgesamt die größte Anzahl richtiger Detektionen. Die Nutzung sämtlicher Parameter gleichzeitig erzielt keine Verbesserung, sodass davon auszugehen ist, dass die spektralen Momente keine komplementäre Information zu der Merkmalskombination COG+SN+EV enthalten.

Bei der Analyse der Vertauschungen zeigt sich, dass die spektrale Neigung wesentlich häufiger laute Sprachsignale als normale Sprache erkennt (laut → normal). Für die anderen Merkmale ist das Verhältnis der Vertauschungen normal als laut detektiert (normal → laut) und laut als normal detektiert relativ ausgeglichen.

Als Nächstes wird die Leistung der spektralen Parameter mit denen der Standard-MFCC-Merkmale verglichen. Weiterhin werden die spektralen Merkmale mit den MFCC-Merkmalen kombiniert, um zu testen, ob eine weitere Verbesserung der Systemleistung erzielt werden kann. Die Ergebnisse sind in Tabelle 6.17 dargestellt. Das Ergebnis der MFCC-Merkmale ist mit 94 richtigen Erkennungen so gut, wie das beste Ergebnis der spektralen Parameter aus Tabelle 6.16. Hier wurden die 94 richtigen Detektionen aber mit nur drei Elementen im Merkmalsvektor erzielt, während die MFCC-Merkmale 19 Koeffizienten nutzen. Die spektralen Merkmale COG+ST+EV haben dementsprechend den Vorteil, dass bei gleichem Ergebnis weniger Rechenkapazität benötigt wird.

Ausgehend von den bisherigen Ergebnissen ist in Tabelle 6.17 erkennbar, dass eine Kombination der MFCC-Merkmale mit spektralen Merkmalen keine Verbesserung gegenüber den MFCC-Merkmalen allein oder der Merkmalskombination COG+SN+EV erzielt. Teilweise führt die Kombination sogar zu leichten Verschlechterungen. Eine Ausnahme bildet hierbei die Kombination MFCC+SN. Mit Hilfe dieser Merkmalskombination kann eine verbesserte Leistung erzielt werden. Folglich handelt es sich um die besten Merkmale zur Unterscheidung normaler und lauter Sprache im gegebenen Szenario.

Merkmal	Richtige Detektionen	Vertauschungen (normal → laut)	Vertauschungen (laut → normal)
MFCC	94	5	1
MFCC+SN	96	4	0
MFCC+COG	93	5	2
MFCC+EV	94	5	1
MFCC+Mom1	94	5	1
MFCC+Mom2	94	5	1
MFCC+Mom3	93	6	1
MFCC+Mom4	94	5	1
MFCC+Momente	94	5	1
MFCC+COG+SN+EV	92	6	2
MFCC+Momente+ COG+SN+EV	94	5	1

Tabelle 6.17: Ergebnisse des Stimmaufwandsklassifikators für die MFCC-Merkmale allein sowie für die spektralen Merkmale kombiniert mit den MFCC-Merkmalen

Auffällig ist weiterhin, dass mit der Merkmalskombination MFCC+SN keine inkorrekte Detektion lauter Sprache als normale Sprache erfolgte. Dies stimmt mit den Tendenzen der anderen Tests mit MFCC-Merkmalen überein, für die kaum Fehler dieser Art vorhanden sind (siehe Tabelle 6.17). Stattdessen sind mehr Fehldetektionen vorhanden, bei denen normale als laute Sprache erkannt wurde. Bei den spektralen Merkmalskombinationen aus Tabelle 6.16, die ähnlich gute Detektionsergebnisse erzielen wie die MFCC-Merkmale (Momente, COG+SN+EV, Momente+COG+SN+EV), ist der gleiche Trend bei den Vertauschungen zu beobachten.

Zusammenfassend lässt sich festhalten, dass **die Stimmaufwandsklassifikation die Ergebnisse der statistischen Tests bestätigt**. Bei dem Vergleich der einzelnen spektralen Merkmale liefert der **COG und das erste spektrale Moment die besten Resultate**. Insgesamt erreicht die **Merkmalskombination MFCC+SN die beste Leistung**. Auszüge der Forschungsergebnisse dieses Kapitels sind in den Veröffentlichungen (Harwardt, 2011a, 2011b)[8] zu finden.

[8]Ein Abdruck befindet sich im Anhang A.3.

Kapitel 7

Zusammenhänge zwischen den spektralen Parametern und F_0

Nachdem im vorherigen Kapitel die Veränderungen spektraler Parameter bei erhöhtem Stimmaufwand beschrieben wurden, soll nun F_0 näher untersucht werden, da F_0 bekannt ist für seine Veränderungen bei erhöhtem Stimmaufwand. Hierfür wird zunächst eine statistische Analyse der F_0-Messungen für normalen und erhöhten Stimmaufwand durchgeführt (Abschnitt 7.1), welche an den Ausführungen für die spektralen Parameter aus Kapitel 6 angelehnt ist und dasselbe Korpus nutzt. Abschnitt 7.2 beschreibt die für die nachfolgenden Abschnitte relevanten Grundlagen der Korrelationsanalyse. Anschließend werden die Zusammenhänge zwischen normaler und lauter Sprache für die spektralen Parameter und F_0 dargestellt (Abschnitt 7.3). In den darauffolgenden Abschnitten wird geprüft, ob ein Zusammenhang zwischen F_0 und einem der spektralen Parameter besteht (Abschnitt 7.4). Die Überprüfung erfolgt sowohl für normale als auch für laute Sprache. Sie dient dem besseren Verständnis der Modifikationen durch veränderten Stimmaufwand. Interessant ist hierbei vor allem, ob gegebenfalls vorhandene Zusammenhänge normalen Stimmaufwands bei einer Stimmaufwandserhöhung erhalten bleiben. In diesem Fall könnte möglicherweise ein robustes Merkmal für die Sprechererkennung bei wechselndem Stimmaufwand in Trainings- und Testdaten entwickelt werden.

7.1 Veränderungen von F_0

Der Untersuchung der statistischen Eigenschaften unterschiedlicher spektraler Parameter im vorherigen Kapitel folgt hier die Analyse der *Grundfrequenz* F_0. Es werden dabei aber nur die stimmhaften Laute mit in die Analyse einbezogen. F_0 wird mit der ESPS Funktion des Snack Toolkits berechnet. Es werden die Standardeinstellungen gewählt. Darüber hinaus erfolgt die Vorgehensweise, wie in Abschnitt 6.2.1 für die spektralen Parameter beschrieben. Daher wird zunächst die *Gesamtverteilung* aller Daten über sämtliche Phoneme und Sprecher ausgewertet. Die Ergebnisse sind in Tabelle A.36 im Anhang und in Abbildung 7.1 zu sehen. Tabelle A.36 zeigt, dass nur zwei der untersuchten Momente der Verteilung um weniger als 10% verändert sind. Hierbei handelt es sich um den Minimal- und den Maximalwert. Die Maximalwerte sind jedoch so groß, dass es hierbei um Ausreißer oder Messfehler handelt und diese Werte nicht repräsentativ sind. Die Schiefe und die Wölbung sinken, während die anderen Werte, die eine Mindestveränderung von 10% aufweisen, steigen. Sowohl

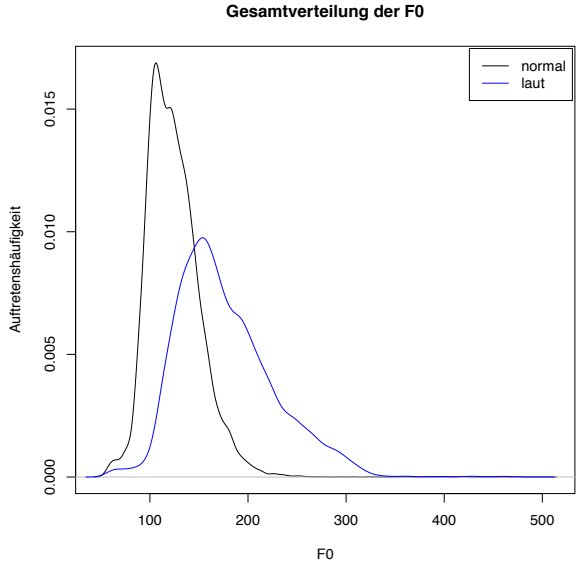

Abbildung 7.1: Verteilung von F_0 für normale und laute Sprache über sämtliche Laute und Sprecher

für die Wölbung als auch für die Schiefe sind jegliche Werte positiv. Die Werte der normalen Sprache sind hingegen höher als die der lauten, sodass für normale Sprache eine weiter ausgeprägte rechtsschiefe und eine spitzere Kurve in Abbildung 7.1 zu sehen ist. Ebenfalls ist der größere, also nach rechts verschobene, Mittelwert der Kurve für laute Sprache klar zu erkennen. Die steigenden Streuungsparameter bei erhöhtem Stimmaufwand zeigen sich in der insgesamt breiteren Kurve der Verteilung lauter Sprache.

Zur weiteren Analyse der Effekte des erhöhten Stimmaufwands werden die Verteilungen der *Lautklassen* Obstruenten, Sonoranten und Vokale separat voneinander untersucht. Die Momente der Verteilungen sind in Tabelle A.36 zufinden. Die zugehörigen Kurven sind in Abbildung 7.2 dargestellt.

Bei der Analyse der Obstruenten für F_0 ist, genau wie bei der Analyse der Obstruenten der spektralen Parameter zu beachten, dass es zu Messungen der Verschlusspause kommen kann (siehe Abschnitt 6.2.1). Weiterhin ist zu beachten, dass F_0 für Obstruenten häufig gar nicht existiert. So kommt es zu einer wesentlich geringeren Stichprobe für die Tests des vorliegenden Kapitel (siehe Tabelle 7.3). Um eine komplette Analyse vorzunehmen, folgt nun die Beschreibung der Ergebnisse der Obstruenten. Es zeigt sich, dass der Mittelwert und die Streuungsparameter bei erhöhtem Stimmaufwand um mehr als 10% steigen. Die Schiefe und Wölbung hingegen sinken. Die Veränderungen der Obstruenten zeigen demnach die gleichen Tendenzen wie die Gesamtverteilung. Für die Lautklasse der Vokale sind ebenfalls dieselben Tendenzen sichtbar wie für die Gesamtverteilung. Die Sonoranten zeigen ein anderes Muster. Für die Klasse der Sonoranten steigen sämtliche Werte um mindestens 10%. Abbildung 7.2 verdeutlicht die Unterschiede zwischen normaler und lauter Sprache für die drei Lautklassen. Hier zeigen vor allem die Kurven der Sonoranten eine große Veränderung durch den erhöhten Stimmaufwand.

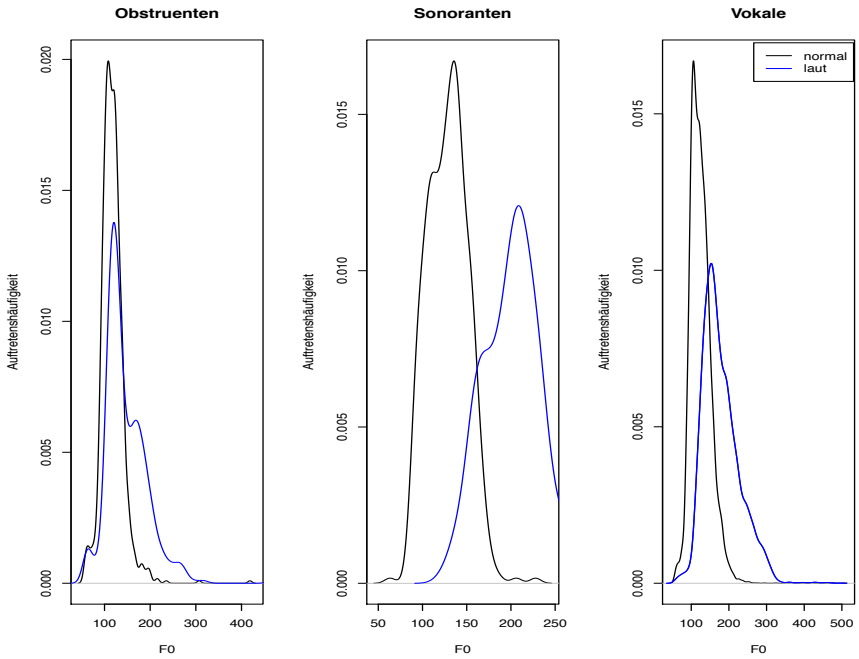

Abbildung 7.2: Verteilung von F_0 für normale und laute Sprache der drei Lautklassen Obstruenten, Sonoranten und Vokale

Nach der Untersuchung der Lautklassen sollen nun die einzelnen Laute betrachtet werden. Zunächst werden die Laute der Klasse der *Obstruenten* analysiert (siehe Tabelle A.37). Die Plosive haben für die meisten Momente der Verteilung steigende Werte. Nur die Wölbung sämtlicher Plosive und die Schiefe des Lauts [g] werden um mehr als 10% abgesenkt. Der Minimalwert wird für keinen der Plosive um mehr als 10% verändert. Für die Auswertung stehen nur zwei Frikative zur Verfügung, da nicht mehr stimmhafte Frikative im Korpus enthalten sind. Der Frikativ [v] folgt, mit leichten Abweichungen, dem Trend der Plosive. Für den Frikativ [z] hingegen sinken die meisten Werte um mehr als 10%. Eine Erhöhung um mehr als 10% ist für diesen Laut nicht zu beobachten.

Die Betrachtung der *Sonoranten* (siehe Tabelle A.38) verdeutlicht, dass die meisten Werte der drei Laute um mehr als 10% erhöht sind. Eine Ausnahme bilden die Werte Schiefe und Wölbung des Nasals [n]. Beide sind abgesenkt, wobei die Wölbung um mehr als 10% fällt.

Für die *Vokale* werden die meisten Momente der Verteilung um mehr als 10% erhöht (siehe Tabelle A.39). Die Wölbung bildet eine Ausnahme, da sie viele Absenkungen aufweist. Für die Schiefe sind wechselnde Veränderungen für die unterschiedlichen Vokale sichtbar. Teilweise werden keine großen Modifikationen der Schiefe beobachtet. Es gibt einige Laute, für die jegliche Werte um mehr als 10% erhöht sind. Dies sind durchweg gelängte Laute. Dagegen kann andersherum nicht bestätigt werden, dass sämtliche gelängten Vokale ein solches Muster zeigen.

Die *sprecherspezifische Auswertung* (siehe Tabelle A.40) verdeutlicht, dass keine großen Unterschiede hinsichtlich der Tendenzen der F_0-Veränderung für die 19 Sprecher bestehen. Die Muster der Sprecher stimmen mit dem Veränderungsmuster der Gesamtverteilung überein. Der Mittelwert und die Streuungsparameter werden

LK	SA	D-Wert	p-Wert
Gesamt	N	$0,0518$	$1,88E{-}88$
	L	$0,0754$	$2,41E{-}202$
Obstr	N	$0,0804$	$3,79E{-}16$
	L	$0,121$	$8,87E{-}40$
Son	N	$0,0506$	$0,0162$
	L	$0,0503$	$0,0364$
Vok	N	$0,0521$	$9,7E{-}80$
	L	$0,0816$	$2,7E{-}213$

Tabelle 7.1: Ergebnisse des Lilliefors-Tests auf Normalverteilung für F_0

für sämtliche Sprecher um mehr als 10% erhöht. Die Schiefe und Wölbung werden, mit einigen Ausnahmen, um mehr als 10% abgesenkt. Für die Maximal- und Minimalwerte ist kein eindeutiges Muster erkennbar. Tritt eine Veränderung um mehr als 10% auf, so handelt es sich meistens um eine Erhöhung.

Die Tatsache, dass für die einzelnen Sprecher dieselben Tendenzen beobachtet werden, scheint im Widerspruch zu der Untersuchung von Jessen et al. (2005) zu stehen. Jessen et al. (2005) stellen eine sprecherspezifische Veränderung von F_0 fest. Allerdings sind die hier vorgestellten Ergebnisse nur eine Beschreibung der Tendenzen und keine genaue Analyse dazu, wie groß die Veränderungen von F_0 für die einzelnen Sprecher ist. Die allgemeine Tendenz, dass F_0 steigt, ist auch in Jessen et al. (2005) zu finden.

Nachdem die Veränderungen der F_0-Verteilung beschrieben wurde, soll nun ein Signifikanztest durchgeführt werden. Wie bei der statistischen Analyse der spektralen Parameter wird zunächst ein *Test auf Normalverteilung* durchgeführt. Hierfür wird der *Lilliefors-Test* verwendet. Die Ergebnisse sind in Tabelle 7.1 dargestellt. Die Auswertung der p-Werte zeigt nicht-normalverteilte Kurven für die Gesamtdaten, die Obstruenten und die Vokale in normaler und lauter Sprache. Die D-Werte bestätigen diese Rückweisung der H_0-Hypothese. Die Klasse der Sonoranten normaler Sprache zeigt einen p-Wert, der für ein Signifikanzniveau von $\alpha = 0,01$ nicht zurückgewiesen werden kann, wohl aber für $\alpha = 0,03$ und $\alpha = 0,05$. Für laute Sprache kann die H_0-Hypothese nur für $\alpha = 0,05$ zurückgewiesen werden. Die D-Werte der Sonoranten bestätigen diese Ergebnisse nicht, da sie kleiner sind als die in Abschnitt 6.2.1 dargestellten Schranken. Aus diesem Grund wird die Schrankenberechnung für die Sonoranten abgeändert. Als zweite Möglichkeit der Schrankenberechnung für $n > 30$ schlägt Sachs (2002, S. 429) für $\alpha = 0,05$ vor: $D = 0,895/d_n$ und für $\alpha = 0,01$: $D = 1,035/d_n$. Es gilt:

$$d_n = \sqrt{n} - 0,01 + 0,83/\sqrt{n}. \tag{7.1}$$

Hieraus ergibt sich, bei n=395 Sonoranten, für $\alpha = 0,05$ die Grenze $D = 0,045$ und für $\alpha = 0,01$ die Grenze $D = 0,052$. Bei dieser Art der Schrankenberechnung stimmt die Aussage von p-Wert und D-Wert überein. Sowohl für laute als auch für normale Sprache kann die H_0-Hypothese nur für $\alpha = 0,05$ zurückgewiesen werden. Die Veränderung der Schrankenberechnung ist möglicherweise dadurch bedingt, dass das benutzte Toolkit Nortest (Gross, 2009) bei $p > 0,1$ nicht die sonst übliche Berechnung durchführt, sondern die modifizierte Z-Statistik $Z = \sqrt{n} - 0,01 + 0,85/\sqrt{n}$ nutzt. Durch diese verschiedenen Arten der Schrankenberechnung kann es zu den beobachteten Unterschieden kommen.

LK	Z-Wert	p-Wert
Gesamt	95,3437	$< 2,2E-16$
Obstr	17,0947	$< 2,2E-16$
Son	22,1594	$< 2,2E-16$
Vokale	92,4342	$< 2,2E-16$

Tabelle 7.2: Ergebnisse des Mann-Whitney-Test für F_0

Nachdem die Lilliefors-Tests auch für die Verteilungen der Grundfrequenz nicht-normalverteilte Kurven zeigen, werden nicht-parametrische Signifikanztests durchgeführt. Die Ergebnisse der nicht-parametrischen *Mann-Whitney-Tests* sind in Tabelle 7.2 dargestellt. Für jegliche Vergleiche sind die p-Werte kleiner $2,2E-16$, sodass für sämtliche Lautklassen und die Gesamtverteilung jeweils ein signifikanter Unterschied zwischen normaler und lauter Sprache besteht. Dies wird durch die hohen Z-Werte, die stark über den vorgegebenen Grenzen aus Abschnitt 6.2.1 liegen, bestätigt.

Zusammenfassend kann festgehalten werden, dass **F_0 für normale und laute Sprache unterschiedlich ist**. Die Untersuchung der Momente der Lautklassen hat besonders für die Sonoranten große Unterschiede ergeben. Die Signifikanztests in Tabelle 7.2 hingegen zeigen sehr große Z-Werte und damit **ausgeprägte Unterschiede** für die **Gesamtverteilung** und die **Vokale**. In den folgenden Abschnitten soll nun überprüft werden, ob die Veränderungen von F_0 mit den Veränderungen der spektralen Parameter aus Kapitel 6 zusammenhängen. Zunächst wird jedoch die Untersuchung des Zusammenhangs für normale und laute Sprache für die spektralen Parameter und F_0 vertieft.

7.2 Korrelationsanalyse

In den nachfolgenden Abschnitten werden die Zusammenhänge zwischen normaler und lauter Sprache sowie zwischen F_0 und den spektralen Parametern untersucht. Um zu testen, ob ein Zusammenhang besteht, wird eine Korrelationsanalyse durchgeführt. Die gängigste Methode ist die *Pearson-Korrelation*. Zur Durchführung von Signifikanztests über diesen Korrelationskoeffizienten wird vorausgesetzt, dass die Daten annähernd normalverteilt sind. Da dies bei den gegebene Daten nicht zutrifft, wird die *Spearman-Rangkorrelation* durchgeführt. Bei diesem Verfahren werden die Variablen sortiert und Ränge vergeben. Diese Ränge werden statt der Variablenwerte für die Berechnung des Korrelationskoeffizienten genutzt. Der *Korrelationskoeffizient* r_S wird nach Sachs (2002, S. 512) folgendermaßen berechnet:

$$r_S = 1 - \frac{6 \sum D^2}{n(n^2 - 1)}, \tag{7.2}$$

wobei D für die Differenzen der Rangplätze der einzelnen Paare steht und n für die Anzahl der Stichprobenelemente. Vorausgesetzt wird $n \geq 6$. Der Korrelationskoeffizient kann Werte zwischen 1 und -1 annehmen. Für 1 ist die Rangfolge der verglichenen Mengen identisch, während sie für -1 genau umgekehrt ist. Es besteht

ein monoton steigender oder fallender Zusammenhang. Nimmt der Koeffizient den Wert 0 an, so ist kein Zusammenhang feststellbar.

Für die *Signifikanztests über den Korrelationskoeffizienten* besagt die H_0-Hypothese, dass kein Zusammenhang zwischen den Variablen besteht ($r_S = 0$). Bei p-Werten kleiner dem Signifikanzniveau $\alpha = 0,05$ kann diese zurückgewiesen werden. In diesem Fall besteht ein echter Zusammenhang zwischen den Verteilungen. Andernfalls kann der durch den Korrelationskoeffizienten berechnete Zusammenhang zufällig sein. Die Signifikanz ist abhängig von der Stichprobengröße und der Höhe des Korrelationskoeffizienten. Aus kleinen Stichproben resultiert ein größerer p-Wert. Die Korrelation muss ausreichend groß sein, um eine Signifikanz feststellen zu können. Bei sehr großen Stichproben wird die H_0-Hypothese meistens zurückgewiesen. Trotz der signifikanten Korrelation kann der Korrelationskoeffizient sehr klein sein. Die Signifikanz macht keine Aussage über die Stärke des Zusammenhangs zweier Verteilungen. Hierfür ist ausschließlich der Korrelationskoeffizient nutzbar.

Die Untersuchung des *Zusammenhangs zwischen normaler und lauter Sprache* erfolgt für jeden untersuchten Parameter separat (siehe Abschnitt 7.3). Für die einzelnen Parameter muss die gleiche Anzahl Elemente lauter und normaler Sprache je Phonem vorliegen. Über die verschiedenen Parameter kann die Anzahl jedoch variieren. Es werden je Parameter Wertepaare zusammengestellt, die jeweils das gleiche Phonem, vom selben Sprecher gesprochen im gleichen Kontext mit unterschiedlichem Stimmaufwand enthalten.

In Abschnitt 7.4 wird der *Zusammenhang zwischen F_0 und jeweils einem spektralen Parameter* untersucht. Die Korrelationsanalyse wird für normale und laute Sprache separat durchgeführt. Die Daten werden paarweise angeordnet. Hierbei werden nur Realisierungen des selben Lauts aus dem selben Logatom für den selben Sprecher mit dem selben Stimmaufwand miteinander verglichen. Phoneme, für die nur einer der Werte (F_0 oder spektraler Parameter) vorliegt, werden aussortiert.

Die Korrelationsanalyse wird in dieser Arbeit mit dem *Tool* Gnu R (R Development Core Team, 2010) durchgeführt. Die Ausgabe umfasst den Korrelationskoeffizienten, den p-Wert und die Teststatistik. Die *Teststatistik* ist eine Prüfgröße, die, wie der p-Wert, zur Rückweisung der H_0-Hypothese genutzt werden kann. Die Verwendung einer Prüfgröße ist nur bei häufig auftretenden Bindungen sinnvoll (siehe Sachs (2002, S. 514)). Eine Bindung tritt auf, wenn zwei oder mehr Elemente der untersuchten Verteilung den gleichen Wert und damit den gleichen Rang annehmen. Die Auftretenswahrscheinlichkeit solcher Bindungen ist in den vorliegenden Daten sehr gering, da es sich um akustische Messungen mit mehreren Nachkommastellen pro Wert handelt. Aus diesem Grund wird die Teststatistik nicht mit in die Auswertung einbezogen. Weitere Details zur Spearman'schen Rangkorrelation und deren Umsetzung im Tool Gnu R sind in Duller (2008) zu finden.

7.3 Korrelation zwischen normaler und lauter Sprache

Nachdem in Kapitel 6 die Unterschiede zwischen normaler und lauter Sprache herausgearbeitet wurden, um den Stimmaufwand quantifizieren zu können, sollen nun mögliche Zusammenhänge zwischen den zwei Stimmaufwandsgraden gefunden werden. Anhand starker Zusammenhänge zwischen den zwei Stimmaufwandsgraden kann möglicherweise von lauter Sprache auf Ausprägungen der Parameter

LK	F_0	spektrale Parameter
Gesamt	9686	11104
Obstr	306	1724
Son	87	87
Vok	9293	9293

Tabelle 7.3: Stichprobengrößen für die Korrelationsanalysen normaler zu lauter Sprache für F_0 und die spektralen Parameter

LK	p-Wert	Korrelations-koeffizient
Gesamt	$< 7,9E{-}323$	0,736
Obstr	$9,98E{-}218$	0,662
Son	$1,34E{-}08$	0,572
Vok	$< 7,9E{-}323$	0,596

Tabelle 7.4: Korrelationsanalyse zwischen normalem und erhöhtem Stimmaufwand für die spektrale Neigung

in normaler Sprache geschlossen werden, und umgekehrt. Nachfolgend werden die spektralen Parameter aus Kapitel 6 und F_0 untersucht. Je Parameter wird die Gesamtverteilung und die Verteilungen der Lautklassen untersucht.

Die Anzahl der Wertepaare für die Korrelationsanalyse ist in Tabelle 7.3 dargestellt. Die unterschiedliche Anzahl der Stichproben für F_0 und die spektralen Merkmale resultiert daraus, dass für F_0 nur solche Laute, die stimmhafte Anteile enthalten, mit in die Analyse einbezogen werden können, während für die spektralen Parameter auch stimmlose Laute untersucht werden können. Aus diesem Grund ist die Anzahl der analysierten Obstruenten für die spektralen Parameter wesentlich größer im Vergleich zur F_0. Weiterhin wurden sowohl für die spektralen Parameter als auch für F_0 nur solche Audiodateien zur Analyse herangezogen, für die zuverlässige Messungen für laute und normale Sprache vorlagen.

Als erstes Merkmal wird die *spektrale Neigung* analysiert. Die Ergebnisse sind in Tabelle 7.4 dargestellt. Für alle Verteilungen wird ein signifikanter Zusammenhang zwischen normaler und lauter Sprache festgestellt. Der höchste Korrelationskoeffizient wird für die Gesamtverteilung erzielt, der geringste für die Sonoranten. Folglich ist der Zusammenhang für die Sonoranten der kleinste. Interessanterweise sind die Obstruenten die Lautklasse mit dem größten Korrelationskoeffizienten zwischen normaler und lauter Sprache. Dies bestätigt die Beobachtungen aus Kapitel 6, in dem nachgewiesen wurde, dass Obstruenten weniger geeignet sind zur Quantifizierung des Stimmaufwands als Vokale und Sonoranten. Dies begründet sich darin, dass sie für die meisten spektralen Parameter, wie auch für die spektrale Neigung, nicht so große Unterschiede zwischen normaler und lauter Sprache aufweisen. Die Korrelation ist folglich stärker, weil die Unterschiede zwischen normaler und lauter Sprache geringer sind. Die geringste Korrelation für die Sonoranten lässt sich in Abbildung 6.3 aus Abschnitt 6.2.2 ablesen. Hier zeigen die Kurven der Sonoranten den größten Unterschied zwischen normaler und lauter Sprache und den geringsten für Obstruenten. Insgesamt ist der Unterschied für sämtliche Lautklassen gering.

LK	p-Wert	Korrelations-koeffizient
Gesamt	$< 7,9E{-}323$	$0,801$
Obstr	$7,9E{-}323$	$0,759$
Son	$8,25E{-}08$	$0,545$
Vok	$< 7,9E{-}323$	$0,861$

Tabelle 7.5: Korrelationsanalyse zwischen normalem und erhöhtem Stimmaufwand für den gewichteten spektralen Schwerpunkt

LK	p-Wert	Korrelations-koeffizient
Gesamt	$< 7,9E{-}323$	$0,776$
Obstr	$2,86E{-}243$	$0,689$
Son	$3,66E{-}11$	$0,607$
Vok	$< 7,9E{-}323$	$0,773$

Tabelle 7.6: Korrelationsanalyse zwischen normalem und erhöhtem Stimmaufwand für das Energieverhältnis

Die Betrachtung der p-Werte zeigt, dass auch der p-Wert der Sonoranten nicht so niedrig ist wie die p-Werte der anderen Verteilungen. Dies ist nicht ausschließlich durch den kleineren Korrelationskoeffizienten bedingt, sondern auch durch die kleinere Stichprobengröße (siehe Tabelle 7.3).

Die Untersuchung des *gewichteten spektralen Schwerpunkts* ergibt ähnliche Trends (siehe Tabelle 7.5). Auch hier sind für die Gesamtverteilung, die Obstruenten und die Vokale hoch signifikante Zusammenhänge sichtbar. Für die Sonoranten besteht ebenfalls ein signifikanter Zusammenhang, der p-Wert ist jedoch auf Grund der Stichprobengröße und der Höhe des Korrelationskoeffizienten größer.

Der Korrelationskoeffizient der Sonoranten ist der geringste. Er zeigt einen ähnlichen Wert wie die Verteilung der Sonoranten der spektralen Neigung. Die anderen Verteilungen weisen wesentlich höhere Korrelationskoeffizienten auf als die spektrale Neigung, sodass ein stärkerer Zusammenhang vorliegt. Für den gewichteten spektralen Schwerpunkt ergibt sich die höchste Korrelation für die Vokale. In Abschnitt 6.2.3 wurde für die Vokale gezeigt, dass ein Unterschied zwischen normaler und lauter Sprache besteht für den spektralen Schwerpunkt (siehe Abbildung 6.5). Der spektrale Schwerpunkt hat sich außerdem als geeignet zur Klassifikation des Stimmaufwands herausgestellt (siehe Abschnitt 6.3). Die sehr hohe Korrelation für Vokale lässt sich dementsprechend so interpretieren, dass ein großer Zusammenhang zwischen lauter und normaler Sprache besteht, obwohl starke Veränderungen bei lauter Sprache vorliegen. Durch den großen Zusammenhang können die spektralen Veränderungen möglicherweise aus den Werten normalen Stimmaufwands abgeleitet oder vorhergesagt werden. Auch die Gesamtverteilung kann hierfür in Betracht gezogen werden. Dies ist in weiteren Untersuchungen zu prüfen.

Die Analyse des *Energieverhältnisses* ist in Tabelle 7.6 abgebildet. Sämtliche Verteilungen haben Korrelationskoeffizienten zwischen 0,6 und 0,8 und zeigen signifikante Korrelationen zwischen normaler und lauter Sprache. Die Gesamtverteilung sowie die Verteilung der Vokale zeigen die stärksten Zusammenhänge. Auch hier ist die

LK	p-Wert	Korrelations-koeffizient
Gesamt	$< 7,9E-323$	0,8
Obstr	$8,9E-323$	0,759
Son	$8,25E-08$	0,545
Vok	$< 7,9E-323$	0,861

Tabelle 7.7: Korrelationsanalyse zwischen normalem und erhöhtem Stimmaufwand für das erste spektrale Moment

LK	p-Wert	Korrelations-koeffizient
Gesamt	$< 7,9E-323$	0,83
Obstr	$1,16E-244$	0,69
Son	$1,38E-05$	0,452
Vok	$< 7,9E-323$	0,852

Tabelle 7.8: Korrelationsanalyse zwischen normalem und erhöhtem Stimmaufwand für das zweite spektrale Moment

Korrelation so hoch, dass Möglichkeiten der Prädiktion untersucht werden könnten. Da bei dem Energieverhältnis die Unterschiede zwischen normaler und lauter Sprache nicht so groß sind wie für die anderen Parameter (siehe Abschnitt 6.3), wird der spektrale Schwerpunkt als geeigneter zur Repräsentation der spektralen Veränderungen angesehen. Die Sonoranten weisen auch beim Energieverhältnis die geringste Korrelation auf. Die Unterschiede zwischen den Lautklassen sind geringer als bei den bisher untersuchten Parametern. Die p-Werte des Energieverhältnis verhalten sich ähnlich wie die der zuvor betrachteten Parameter.

Als Nächstes werden die *spektralen Momente* untersucht. Die Ergebnisse des *ersten spektralen Moments* sind in Tabelle 7.7 zu sehen. Da das erste Moment und der gewichtete spektrale Schwerpunkt sehr ähnliche Parameter sind, zeigt sich auch bei der Korrelationsanalyse ein nahezu identisches Bild. Für alle Verteilungen außer der Verteilung der Sonoranten sind normale und laute Sprache hoch signifikant korreliert. Es ergeben sich Korrelationskoeffizienten um 0,8 für diese Tests. Die normale und laute Sprache der Sonoranten ist ebenfalls signifikant korreliert. Der Korrelationskoeffizient ist jedoch wesentlich geringer, im Vergleich zu den Koeffizienten der anderen Verteilungen. Auf Grund der hohen Korrelationen der Vokale und der Gesamtverteilung sowie der Unterschiede dieser zwei Verteilungen bei normaler und lauter Sprache (siehe Abschnitt 6.2.5.1) könnten die Gesamtverteilung und die Vokale zur Beschreibung und Vorhersage der spektralen Veränderungen bei erhöhtem Stimmaufwand geeignet sein. Dies bietet, wie bei dem gewichteten spektralen Schwerpunkt, einen Ansatz für weiterführende Arbeiten.

Die Analyse des *zweiten spektralen Moments* führt zu ähnlichen Ergebnissen (siehe Tabelle 7.8). Die Signifikanztests zeigen signifikante beziehungsweise hoch signifikante Korrelationen. Der größte Zusammenhang ist für die Gesamtverteilung und die Vokale zu beobachten. Die Klasse der Obstruenten weist ebenfalls einen hohen Korrelationskoeffizienten auf. Der Unterschied zwischen den zwei höchsten Koeffizienten und dem der Obstruenten ist dagegen größer als bei dem ersten spektralen

LK	p-Wert	Korrelations-koeffizient
Gesamt	$< 7,9E{-}323$	$0,842$
Obstr	$4,81E{-}300$	$0,741$
Son	$3,77E{-}06$	$0,478$
Vok	$< 7,9E{-}323$	$0,893$

Tabelle 7.9: Korrelationsanalyse zwischen normalem und erhöhtem Stimmaufwand für das dritte spektrale Moment

LK	p-Wert	Korrelations-koeffizient
Gesamt	$< 7,9E{-}323$	$0,84$
Obstr	$2,31E{-}259$	$0,705$
Son	$7,83E{-}06$	$0,464$
Vok	$< 7,9E{-}323$	$0,886$

Tabelle 7.10: Korrelationsanalyse zwischen normalem und erhöhtem Stimmaufwand für das vierte spektrale Moment

Moment. Die Sonoranten zeigen nur einen kleineren Korrelationskoeffizienten. Er ist geringer als bei den bisher untersuchten Verteilungen. Eine Nutzung der Korrelationen der Gesamtverteilung und der Vokale zur Prädiktion der Veränderungen im Spektrum ist auch für das zweite spektrale Moment denkbar, allerdings scheinen andere Parameter geeigneter, da der Unterschied zwischen lauter und normaler Sprache nicht so groß ist wie bei anderen spektralen Parametern (siehe Tabelle 6.16, Abschnitt 6.3).

Das *dritte spektrale Moment*, welches ähnlich der spektralen Neigung ist, folgt den bisher beschriebenen Mustern (siehe Tabelle 7.9). Die Korrelationen sind signifikant oder hoch signifikant. Die Korrelationsanalyse des dritten spektralen Moments bestätigt den Trend, dass der größte Zusammenhang zwischen normalem und erhöhtem Stimmaufwand für die Vokale und die Gesamtverteilung vorhanden ist. Für die Klasse der Vokale des dritten spektralen Moments wird, im Gesamtvergleich mit den anderen Parametern, der höchste Korrelationskoeffizient erzielt. Die Obstruenten erzielen ebenfalls einen hohen Korrelationskoeffizienten. Die Sonoranten hingegen zeigen einen eher geringen Koeffizienten und damit geringere Zusammenhänge. Obwohl das dritte Moment und die spektrale Neigung ähnliche Eigenschaften des Spektrums beschreiben, sind die Ergebnisse der Korrelationsanalyse nicht vergleichbar. Dies bestätigt die Beobachtungen aus Kapitel 6. Hier wurde festgestellt, dass der spektrale Schwerpunkt und das erste Moment ähnliche Resultate erzielen, nicht aber das dritte Moment und die spektrale Neigung. Die Ergebnisse der Korrelationsanalyse ähneln viel mehr den Analysen des ersten und zweiten spektralen Moments.

Die Untersuchung des *vierten spektralen Moments* (siehe Tabelle 7.10) zeigt das gleiche Muster wie das dritte Moment. Es sind keine weiteren Besonderheiten zu beobachten.

Die Ergebnisse der Korrelationsanalyse für F_0 sind in Tabelle 7.11 dargestellt. Die Korrelationen sind für sämtliche untersuchten Verteilungen signifikant oder

LK	p-Wert	Korrelations-koeffizient
Gesamt	$< 7,9E-323$	0,537
Obstr	$< 7,9E-323$	0,493
Son	$7,45E-06$	0,465
Vok	$< 7,9E-323$	0,538

Tabelle 7.11: Korrelationsanalyse zwischen normalem und erhöhtem Stimmaufwand für F_0

hoch signifikant. Die Korrelationskoeffizienten sind im Vergleich zu den spektralen Parametern wesentlich geringer. Sie liegen ungefähr bei 0,5. Auch bei F_0 weisen insgesamt die Sonoranten den geringsten Zusammenhang zwischen normaler und lauter Sprache auf. Die Unterschiede zwischen den einzelnen Verteilungen sind hingegen gering. Für F_0 scheint kein ausreichender Zusammenhang zu bestehen, welcher eine Prädiktion von F_0 aus den Werten normalen Stimmaufwands ermöglichen würde.

Zur Illustration der Zusammenhänge zwischen normaler und lauter Sprache sind die Streudiagramme für die Gesamtverteilungen jeglicher untersuchter Parameter in Abbildung 7.3 dargestellt. Auffällig sind die Darstellungen von F_0 und die des Energieverhältnisses, welche die relativ niedrigen Korrelationskoeffizienten visualisieren. Für die anderen Parameter wird der Zusammenhang zwischen normaler und lauter Sprache durch die Abbildungen bestätigt. Allerdings scheint für keine der Gesamtverteilungen der Zusammenhang so groß zu sein, dass eine zuverlässige Vorhersage der Veränderungen möglich scheint. Dies müsste jedoch im Einzelfall überprüft werden.

Insgesamt lässt sich festhalten, dass **ein Zusammenhang zwischen normaler und lauter Sprache besteht**, der bei der **Gesamtverteilung und den Vokalen stärker ausgeprägt** ist als bei den anderen Lautklassen. Die Sonoranten weisen bei sämtlichen akustischen Parametern den geringsten Zusammenhang auf. Der Parameter F_0 zeigt im Vergleich mit den spektralen Parametern die geringsten Zusammenhänge zwischen normaler und lauter Sprache. Eine detailliertere Analyse des Zusammenhangs der spektralen Parameter mit dem größten Korrelationskoeffizienten (Vokale des dritten und vierten Moments) könnte zu einer allgemeingültigen Beschreibung der Veränderung des Spektrums bei Erhöhung des Stimmaufwands führen.

7.4 Korrelation von F_0 mit den spektralen Parametern

Im vorherigen Abschnitt wurden die Zusammenhänge zwischen normaler und lauter Sprache dargestellt. Es wurde ein Zusammenhang zwischen den beiden Stimmaufwandsgraden festgestellt. Nun soll untersucht werden, ob für diese beiden Stimmaufwandsgrade, separat betrachtet, Korrelationen zwischen F_0 und jeweils einem der spektralen Parameter existieren. Bestehen Zusammenhänge, so wird geprüft, ob diese bei beiden Stimmaufwandsgraden in der selben Intensität auftreten und ob ein robustes Merkmal für die Sprechererkennung mit Hilfe dieser Zusammenhänge entwickelt werden kann.

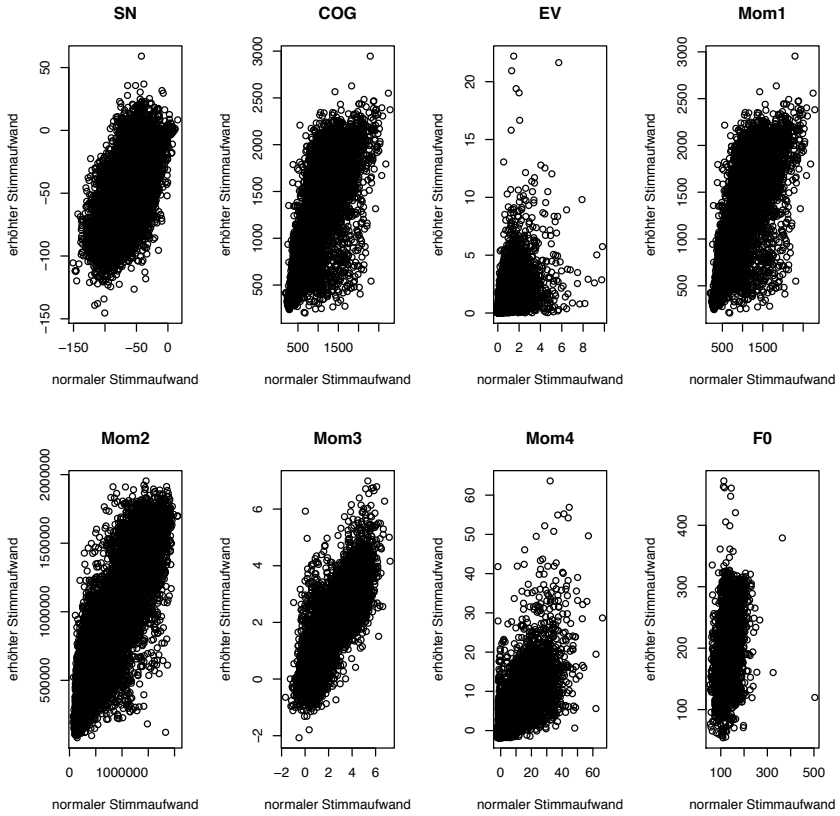

Abbildung 7.3: Streudiagramme zum Vergleich normaler und lauter Sprache für unterschiedliche akustische Parameter

Die Stichprobengröße für die nachfolgenden Korrelationsanalysen ist in Tabelle 7.12 dargestellt. Die Anzahl der Stichproben unterscheidet sich von denen der vorherigen Korrelationsanalysen, da hier nicht zwischen normaler und lauter Sprache abgeglichen wird, ob zuverlässige Messungen vorhanden sind, sondern zwischen F_0 und den spektralen Parametern. Dies erfolgt für die zwei Stimmaufwandsgrade separat. In den nachfolgenden Analysen sind nur stimmhafte Laute enthalten. Auch für diese Korrelationsanalyse wird die Spearman-Rangkorrelation genutzt.

Zunächst werden die *Zusammenhänge von F_0 mit der spektralen Neigung* untersucht. Die Ergebnisse der Korrelationsanalyse sind in Tabelle 7.13 dargestellt. Bei *normalem Stimmaufwand* besteht ein signifikanter Zusammenhang für die Gesamtverteilung

LK	normaler SA	erhöhter SA
Gesamt	10825	11800
Obstr	648	652
Son	207	188
Vok	9970	10960

Tabelle 7.12: Stichprobengrößen für die Korrelationsanalysen von F_0 zu jeweils einem spektralen Parameter bei den normalem und erhöhtem Stimmaufwand

LK	SA	p-Wert	Korrelations-koeffizient
Gesamt	N	$7,67E{-}163$	$-0,257$
	L	$1,64E{-}47$	$-0,133$
Obstr	N	$1,41E{-}04$	$0,149$
	L	$0,679$	$-0,0162$
Son	N	$0,455$	$-0,0522$
	L	$0,318$	$0,0732$
Vok	N	$6,5E{-}160$	$-0,265$
	L	$8,4E{-}20$	$-0,0868$

Tabelle 7.13: Korrelationsanalyse der Lautklassen für F_0 und die spektrale Neigung bei normalem Stimmaufwand und bei lauter Sprache

sowie für die Obstruenten und Vokale. Für die Sonoranten kann die H_0-Hypothese, dass keine Korrelation besteht, nicht zurückgewiesen werden. Es besteht also kein signifikanter Zusammenhang für die Klassen der Sonoranten von F_0 und der spektralen Neigung. Die Beträge der Korrelationskoeffizienten der signifikanten Korrelationen sind insgesamt wesentlich niedriger, verglichen mit den Analysen für normale und laute Sprache aus Abschnitt 7.3. Weiterhin sind negative Korrelationen für die Gesamtverteilung und die Verteilung der Vokale zu finden. Die Korrelation der Obstruenten hingegen ist positiv und geringer als die der Vokale und der Gesamtverteilung. Auch wenn für die meisten Verteilungen signifikante, also echte Zusammenhänge feststellbar sind, sind diese nicht sehr groß.

Für *erhöhten Stimmaufwand* ist der beobachtete Zusammenhang noch geringer (siehe Tabelle 7.13). Nur für die Gesamtverteilung und die Verteilung der Vokale sind signifikante Zusammenhänge vorhanden. Der Betrag der Korrelationskoeffizienten ist gering. Besonders für die Vokale ist die Korrelation sehr klein. Obwohl ein signifikanter Zusammenhang besteht, ist der Zusammenhang zu klein, um für die Auswertung relevant zu sein. Ein signifikanter Zusammenhang mit sehr geringem Korrelationskoeffizienten ist dadurch erklärbar, dass der Signifikanztest und die Korrelationsanalyse voneinander unabhängige Aussagen machen. Der Korrelationskoeffizient sagt nichts über die Signifikanz aus, und umgekehrt. So kann ein signifikanter Zusammenhang bestehen, die Größe des Zusammenhangs jedoch sehr gering sein. Für sehr große Stichproben sind die Korrelationen meistens signifikant. Die Stichprobe der Vokale für erhöhten Stimmaufwand ist mit 10960 Paaren sehr groß (vergleiche Tabelle 7.12). Auf Grund der niedrigen Korrelationskoeffizienten wird auf ein Streudiagramm verzichtet.

Der *Vergleich normaler und lauter Sprache* zeigt, dass sich der geringe Zusammenhang für normale Sprache bei erhöhtem Stimmaufwand noch weiter verringert.

Als Nächstes folgt die Untersuchung des *Zusammenhangs zwischen F_0 und dem gewichteten spektralen Schwerpunkt* (siehe Tabelle 7.14). Die Betrachtung der Korrelationen für *normale Sprache* verdeutlicht, dass nur die Verteilung der Vokale für F_0 und den spektralen Schwerpunkt eine signifikante Korrelation zeigt. Bei den anderen Verteilungen kann die H_0-Hypothese nicht zurückgewiesen werden, sodass die festgestellten Korrelationen möglicherweise zufällig sind. Der Korrelationskoeffizient der Vokale ist negativ und vergleichbar zu dem Wert der Analyse von F_0 und der spektralen Neigung. Die Korrelation ist, wie bei der spektralen Neigung, gering.

LK	SA	p-Wert	Korrelations-koeffizient
Gesamt	N	$0,183$	$-0,0128$
	L	$7,87E-74$	$0,166$
Obstr	N	$0,326$	$-0,0386$
	L	$3,42E-09$	$0,229$
Son	N	$0,571$	$0,0395$
	L	$0,0896$	$0,124$
Vok	N	$6,02E-03$	$-0,0275$
	L	$2,87E-35$	$0,118$

Tabelle 7.14: Korrelationsanalyse der Lautklassen für F_0 und den spektralen Schwerpunkt bei normalem Stimmaufwand und bei lauter Sprache

Für *laute Sprache* sind die Korrelationen, außer für die Verteilungen der Sonoranten, signifikant. Dies ist interessant, da auch bei den Analysen normaler und lauter Sprache in Abschnitt 7.3 für die Sonoranten zwar signifikante Ergebnisse erzielt wurden; die p-Werte waren allerdings wesentlich größer als die der anderen Verteilungen. Bei beiden Korrelationsanalysen (normal versus laut; F_0 versus spektrale Parameter) ist die geringere Stichprobengröße der Verteilung der Sonoranten ausschlaggebend für die Unterschiede des p-Wertes verglichen mit den anderen Verteilungen (siehe Tabellen 7.3 und 7.12). Für die Analyse in diesem Abschnitt (F_0 versus spektrale Parameter) ist die Stichprobengröße der Obstruenten reduziert, da nur stimmhafte Obstruenten mit einbezogen wurden. Aus diesem Grund sind die Werte der Obstruenten häufig nicht signifikant, wie beispielsweise bei der spektralen Neigung (siehe Tabelle 7.13).

Beim *Vergleich normaler und lauter Sprache* zeigt sich ein anderes Verhalten als bei der spektralen Neigung. Die Korrelationen der Gesamtverteilung sowie der Obstruenten und Sonoranten sind nicht-signifikant für normale Sprache, während für die Gesamtverteilung und die Obstruenten lauter Sprache eine geringe echte Korrelation festgestellt werden kann. Die Vokale zeigen für normale und laute Sprache eine signifikante Korrelation. Der Vergleich ihrer Koeffizienten indiziert eine höhere Korrelation für laute Sprache. Dies steht ebenfalls im Gegensatz zu den Beobachtungen bei der spektralen Neigung. Insgesamt ist die Korrelation für sämtliche Verteilungen, auch wenn sie echt ist, zu gering, um weiter analysiert zu werden.

Die Ergebnisse der *Korrelationsanalyse von F_0 mit dem Energieverhältnis* ist in Tabelle 7.15 dargestellt. Für die Obstruenten *normaler Sprache* kann die H_0-Hypothese bei einem Signifikanzniveau von 0,05 zurückgewiesen werden. Die anderen Verteilungen normaler Sprache weisen keine signifikante Korrelation auf. Der Korrelationskoeffizient der Obstruenten ist negativ und sein Betrag so gering, dass für normale Sprache keine relevante Korrelation zwischen F_0 und dem Energieverhältnis festgestellt werden kann.

Bei *lauter Sprache* gibt es signifikante Korrelationen für die Gesamtverteilung und die Vokale. Die Korrelationskoeffizienten sind allerdings sehr gering, sodass auch für laute Sprache keine echte Korrelation zwischen F_0 und dem Energieverhältnis festgestellt werden kann.

LK	SA	p-Wert	Korrelations-koeffizient
Gesamt	N	$0,98$	$-2,37E-04$
	L	$1,9E-07$	$0,0479$
Obstr	N	$0,0447$	$-0,0789$
	L	$0,652$	$-0,0177$
Son	N	$0,561$	$-0,0406$
	L	$0,995$	$-4,44E-04$
Vok	N	$0,865$	$1,70E-03$
	L	$0,0142$	$0,0234$

Tabelle 7.15: Korrelationsanalyse der Lautklassen für F_0 und das Energieverhältnis bei normalem Stimmaufwand und bei lauter Sprache

LK	SA	p-Wert	Korrelations-koeffizient
Gesamt	N	$0,183$	$-0,0128$
	L	$7,94E-74$	$0,166$
Obstr	N	$0,327$	$-0,0386$
	L	$3,43E-09$	$0,229$
Son	N	$0,571$	$0,0395$
	L	$0,0896$	$0,124$
Vok	N	$6,01E-03$	$-0,0275$
	L	$2,88E-35$	$0,118$

Tabelle 7.16: Korrelationsanalyse der Lautklassen erste spektrale Moment bei normalem Stimmaufwand und bei lauter Sprache

Die Auswertung zeigt, dass kein *Unterschied zwischen normaler und lauter Sprache* bezüglich des Zusammenhangs von F_0 und dem Energieverhältnis besteht, da für beide Stimmaufwandsgrade keine Zusammenhänge vorhanden sind. Weiterhin verdeutlicht die Analyse, dass sich das Energieverhältnis von den anderen spektralen Merkmalen abhebt, da diese zumindest für eine Verteilung eine Korrelation größer 0,1 aufweisen. Dies stimmt überein mit den Ergebnissen aus Kapitel 6. Hier stellte sich das Energieverhältnis als am wenigsten geeignet zur Klassifikation des Stimmaufwands heraus und hob sich dadurch von den anderen Merkmalen ab.

Nachfolgend werden die *spektralen Momente* analysiert. Die Ergebnisse der *Korrelationsanalyse von F_0 mit dem ersten spektralen Moment* ist in Tabelle 7.16 zu finden. Für *normale Sprache* zeigt erneut nur eine Verteilung, die der Vokale, eine signifikante Korrelation. Der Korrelationskoeffizient ist negativ und sein Betrag so gering, dass nicht von einer Korrelation zwischen F_0 und dem ersten spektralen Moment für normale Sprache ausgegangen werden kann. Dies ist vergleichbar mit den Ergebnissen für das Energieverhältnis.

Die Sprache *erhöhten Stimmaufwands* zeigt signifikante Korrelationen für drei Verteilungen; die Gesamtverteilung, die Verteilung der Obstruenten und die der Vokale. Den höchsten Korrelationskoeffizienten der drei Verteilungen weisen die Obstruenten auf. Für die drei Verteilungen ist der Koeffizient größer 0,1. Er ist jedoch

LK	SA	p-Wert	Korrelations-koeffizient
Gesamt	N	$0,983$	$-1,99E-04$
	L	$1,09E-21$	$0,088$
Obstr	N	$0,0548$	$-0,0755$
	L	$1,69E-03$	$0,123$
Son	N	$0,759$	$-0,0215$
	L	$0,71$	$-0,0273$
Vok	N	$0,724$	$-3,54E-03$
	L	$9,53E-07$	$0,0468$

Tabelle 7.17: Korrelationsanalyse der Lautklassen für F_0 und das zweite spektrale Moment bei normalem Stimmaufwand und bei lauter Sprache

insgesamt für sämtliche Verteilungen klein. Die Zusammenhänge sind dementsprechend sehr gering.

Der *Vergleich der Stimmaufwandsgrade* zeigt, dass für laute Sprache eine echte Korrelation für drei Verteilungen vorhanden ist, während normale Sprache keine nennenswerte Korrelation aufweist. Es besteht folglich eine etwas größere Korrelation für laute Sprache als für normale. Die Korrelationen sind allerdings auch für laute Sprache gering. Die Ergebnisse sind vergleichbar mit denen des spektralen Schwerpunkts. Sogar die Korrelationskoeffizienten nehmen sowohl für normale als auch für laute Sprache ähnliche Werte an.

Die Ergebnisse der *Korrelationsanalyse von F_0 mit dem zweiten spektralen Moment* sind in Tabelle 7.17 dargestellt. Bei einem Signifikanzniveau von 0,05 ist keine der Korrelationsanalysen *normaler Sprache* signifikant.

Für *laute Sprache* sind die Ergebnisse der Gesamtverteilung, der Obstruenten und der Vokale signifikant. Für die Obstruenten wird ein Korrelationskoeffzient größer 0,1 erzielt. Für die anderen zwei Verteilungen ist der Koeffizient geringer. Insgesamt sind die Koeffizienten sämtlicher Verteilungen so gering, dass eine weitere Analyse der Ergebnisse nicht zielführend ist.

Beim *Vergleich normaler und lauter Sprache* ist ein minimaler Unterschied festzustellen. Für die Obstruenten lauter Sprache gibt es einen geringen echten Zusammenhang, welcher bei normaler Sprache nicht nachweisbar ist. Darüber hinaus sind die Verteilungen für beide Stimmaufwandsgrade entweder nicht signifikant oder sehr gering.

Die Untersuchungsergebnisse der *Korrelation zwischen F_0 und dem dritten spektralen Moment* ist in Tabelle 7.18 zu finden. Für *normale Sprache* können bei einem Signifikanzniveau von 0,05 signifikante Korrelationen für die Gesamtverteilung und die Vokale festgestellt werden. Die Korrelationskoeffizienten sind jedoch sehr gering ($< 0,1$), sodass insgesamt kein nennenswerter, echter Zusammenhang zwischen F_0 und dem dritten Moment normaler Sprache vorhanden ist.

Die Korrelationsanalysen *lauter Sprache* ergeben echte Korrelationen für die Gesamtverteilung, die Obstruenten und die Vokale. Für die Gesamtverteilung und die Obstruenten liegt der Betrag der Koeffizienten jeweils über 0,1, er ist allerdings für beide Verteilungen dennoch niedrig. Für die Vokale liegt der Betrag unter 0,1. Die Korrelation der Vokale ist folglich sehr gering. Da die Korrelationen der Gesamtverteilung und der Obstruenten ebenfalls insgesamt nur gering sind, werden sie nicht weiter analysiert.

LK	SA	p-Wert	Korrelations-koeffizient
Gesamt	N	$0,0331$	$0,0205$
	L	$2,15E-50$	$-0,137$
Obstr	N	$0,197$	$0,0508$
	L	$2,14E-05$	$-0,166$
Son	N	$0,0658$	$-0,128$
	L	$0,0531$	$-0,141$
Vok	N	$1,24E-03$	$0,0323$
	L	$7,65E-21$	$-0,0893$

Tabelle 7.18: Korrelationsanalyse der Lautklassen für F_0 und das dritte spektrale Moment bei normalem Stimmaufwand und bei lauter Sprache

LK	SA	p-Wert	Korrelations-koeffizient
Gesamt	N	$0,617$	$4,8E-03$
	L	$4,13E-39$	$-0,12$
Obstr	N	$0,143$	$0,0576$
	L	$1,79E-04$	$-0,146$
Son	N	$0,085$	$-0,12$
	L	$0,121$	$-0,113$
Vok	N	$0,157$	$0,0142$
	L	$1,15E-13$	$-0,0708$

Tabelle 7.19: Korrelationsanalyse der Lautklassen für F_0 und das vierte spektrale Moment bei normalem Stimmaufwand und bei lauter Sprache

Bei dem *Vergleich normaler und lauter Sprache* werden ähnliche Tendenzen für beide Stimmaufwandsgrade festgestellt. Für beide Stimmaufwandsgrade sind die Korrelationen gering, wobei die Korrelationen der normalen Sprache insgesamt, verglichen mit lauter Sprache, noch geringer sind. Eine Gegenüberstellung mit der spektralen Neigung zeigt keine Übereinstimmung. Wie in der statistischen Analyse aus Kapitel 6 unterscheiden sich die Werte für das dritte spektrale Moment und die spektrale Neigung voneinander, sodass möglicherweise komplementäre Information in beiden Merkmalen enthalten ist.

Als Letztes erfolgt die *Korrelationsanalyse zwischen F_0 und dem vierten spektralen Moment* (siehe Tabelle 7.19). Für *normale Sprache* sind keine signifikanten Korrelationen vorhanden.

Bei *lauter Sprache* sind für die Gesamtverteilung, die Obstruenten und die Vokale signifikante Korrelationen zu beobachten. Wie bereits bei den anderen Merkmalen beobachtet, sind auch hier die Korrelationen klein. Für die Gesamtverteilung und die Obstruenten ist jeweils eine negative Korrelation vorhanden, deren Betrag über 0,1 liegt. Der Betrag des Koeffizienten der Vokale liegt unter 0,1. Somit sind die Zusammenhänge auch für F_0 und das vierte spektrale Moment bei lauter Sprache gering.

Auch für das vierte spektrale Moment sind die *Gesamttendenzen normaler und lauter Sprache* gleich. Für beide Stimmaufwandsgrade sind keine oder nur geringe

Zusammenhänge sichtbar. Allerdings sind für laute Sprache echte Korrelationen zu finden. Für normale Sprache ist dies nicht der Fall.

Abschließend lässt sich festhalten, dass **keines der spektralen Merkmalen eine mittlere oder starke Korrelation zur F_0 aufweist**. Dies gilt für normalen und erhöhten Stimmaufwand. Die Idee, ein Merkmal für die Sprechererkennung zu entwickeln, welches sich aus dem Verhältnis von F_0 zum Spektrum berechnet, ist dementsprechend zu verwerfen. Es ist nicht zu erwarten, aus diesem Verhältnis robuste Merkmale für die Sprechererkennung entwickeln zu können.

7.5 Zusammenfassung

In diesem Kapitel wurden die F_0-Veränderungen bei erhöhtem Stimmaufwand im Kontext der Veränderungen des Spektrums untersucht. Hierfür wurde zunächst eine Analyse der F_0-Veränderungen durchgeführt (siehe Abschnitt 7.1). Es zeigte sich, dass F_0 bei steigendem Stimmaufwand stark verändert wird, besonders für Sonoranten und Vokale.

Die Korrelationsanalyse normaler und lauter Sprache aus Abschnitt 7.3 untersuchte sämtliche bisher betrachteten Parameter. Die Analyse ergab Zusammenhänge zwischen lauter und normaler Sprache, welche für die Gesamtverteilung und die Vokale am stärksten ausgeprägt waren. Die spektralen Parameter wiesen größere Zusammenhänge auf als F_0. Die relativ hohen Korrelationen für Vokale führen zu der These, dass für diese Lautklasse möglicherweise eine Vorhersage der spektralen Veränderungen durchführbar ist. Besonders hohe Korrelationen wurden für die Vokale des spektralen Schwerpunkts und die Vokale des ersten, dritten und vierten spektralen Moments beobachtet. Ähnliche Korrelationen zeigen sich auch für andere Parameter. Für den spektralen Schwerpunkt und das erste Moment besteht außerdem ein großer Unterschied zwischen normaler und lauter Sprache. Dementsprechend ist hier davon auszugehen, dass die Korrelation nicht auf Grund ähnlicher Gegebenheiten in normaler und lauter Sprache besteht, sondern ein tatsächliches Potential zur Prädiktion vorliegt. Die Analyse der Streudiagramme der Gesamtverteilungen zeigte, dass die teilweise sehr großen Korrelationen nicht zur Prädiktion mit einer einfachen linearen Funktion ausreichen. Es ist zu prüfen, ob eine Vorhersage mit Hilfe statistischer Modelle eher zielführend ist. Sowohl bei statistischen Modellen als auch bei Funktionen zur Prädiktion besteht aber das Problem, dass Werte ausgegeben werden, die normiert sind und keine sprecherspezifischen Informationen enthalten. Im Rahmen der Sprechererkennung ist eine solche Prädiktion folglich nicht Erfolg versprechend. Es kann lediglich untersucht werden, ob anhand der normalen Sprachdaten eines Sprechers, unter Einbezug eines allgemeingültigen Modells zur Veränderung der Sprache bei erhöhtem Stimmaufwand, eine sprecherspezifische Prädiktion vorgenommen werden kann.

Die Korrelationsanalysen aus Abschnitt 7.4, welche Zusammenhänge zwischen F_0 und den spektralen Parametern untersuchten, fanden keine größeren Zusammenhänge. Damit ist es anscheinend nicht möglich, basierend auf den Zusammenhängen zwischen F_0 und dem Spektrum robuste Merkmale für die Sprechererkennung zu entwickeln.

Kapitel 8

Vergleich verschiedener Merkmale für die Sprechererkennung

Das folgende Kapitel gibt einen Überblick über das im Rahmen dieser Arbeit umgesetzte Sprechererkennungssystem (Abschnitt 8.1), das Auswertungskonzept (Abschnitt 8.2) und die verschiedenen Merkmale, die zur automatischen Sprechererkennung bei unterschiedlichem Stimmaufwand in Trainings- und Testdaten verwendet werden (Abschnitte 8.3 bis 8.5). Ein Schwerpunkt liegt auf dem Vergleich verschiedener Standardmerkmale (Abschnitt 8.3). Dieser Vergleich ist notwendig, um die optimale Einstellung für das Basissystem und damit die Standardmerkmale zu bestimmen, die durch eine Veränderung des Stimmaufwands am wenigsten beeinflusst werden. Im Weiteren werden dann unterschiedliche Systeme mit F_0-basierten Merkmalen im gleichen Szenario separat und in Kombination mit dem Basissystem getestet (Abschnitt 8.4). Es wurden F_0-basierte Merkmale ausgewählt, da hier ein starker Abfall der Leistung erwartet wird. Zudem soll geprüft werden, inwiefern F_0-basierte Merkmale bei wechselndem Stimmaufwand verwendet werden können. Abschließend werden weitere spektrale Merkmale zur Verbesserung der Gesamtleistung der automatischen Sprechererkennung vorgestellt (Abschnitt 8.5).

8.1 Realisierung des Sprechererkennungssystems

Die Konzeption des Sprechererkennungssystems orientiert sich an den Ausführungen aus Kapitel 3. Es wurde als *Basissystem* ein GMM-UBM-basiertes Sprecherverifikationssystem umgesetzt. Das System umfasst einen Trainings- und einen Testmodus. Im Trainingsmodus wird zunächst ein Hintergrundmodell (UBM) trainiert. Das UBM repräsentiert die Sprache der betrachteten Sprecherpopulation. Es wird in Form eines Gauß'schen Mischverteilungsmodells (GMM) mit 1024 Mischungskomponenten umgesetzt. Das Modelltraining wird mit dem auf dem Forward-Backward-Algorithmus basierenden Tool HERest des *Toolkits HTK* (Hidden Markov Model Toolkit) realisiert (Young et al., 2006). Nach dem Training des Hintergrundmodells werden die Sprechermodelle erstellt. Diese werden, ausgehend vom gut trainierten UBM, mit den Daten des einzelnen Sprechers adaptiert. Im Testmodus der Sprecherverifikation werden die Merkmale der Testsignale gegen jeweils ein ausgewähltes Sprechermodell und das UBM getestet. Dieser Vorgang kann, je Testsignal, für einzelne ausgewählte Sprechermodelle durchgeführt werden

oder für alle trainierten Sprechermodelle. In der vorliegenden Arbeit wird jedes Testsignal gegen alle Sprechermodelle getestet. Aus den Ausgaben des betrachteten Sprechermodells und des Hintergrundmodells wird für das gegebene Sprechermodell die Log-Likelihood-Ratio (LLR) berechnet (siehe Abschnitt 3.2.1). Dieser Wert ist die Ausgabe, beziehungsweise die Punktzahl, des Sprecherverifikationssystems. Die Punktzahl kann, mittels eines vorab bestimmten Grenzwertes, zur Rückweisung oder Annahme der H_0-Hypothese genutzt werden. Die Log-Likelihood-Ratio kann auch ohne vorab bestimmten Grenzwert genutzt werden, um zu entscheiden, wie wahrscheinlich es ist, dass der betrachtete Sprecher im fraglichen Sprachsignal spricht. Außerdem wird sie zur Erstellung einer DET-Kurve (Detection Error Tradeoff) genutzt, welche im nächsten Abschnitt näher erläutert wird.

Dieses Basissystem wird im Folgenden als Framework genutzt, in das unterschiedliche *Standardmerkmale*, die *logarithmierte* F_0 *(log*F_0*)* und das *COG-Verhältnis* als Merkmale eingefügt werden. Ferner werden F_0*-Statistiken*, wie sie in der forensischen Sprechererkennung gebräuchlich sind, und eine *Kombination spektraler Merkmale* zur Verifikation der Sprecher getestet. Die F_0-Statistiken und die spektralen Parameter werden nicht in das GMM-UBM System eingebettet, da hier nur ein Vektor pro Sprachsignal vorliegt, anstatt einem Vektor pro Frame. Dementsprechend wird im Training eine Statistik pro Sprecher trainiert. Im Testmodus wird dann die Distanz zwischen Sprechermodell und Testsignalstatistik sowie die Distanz zwischen dem Testsignal und einer Hintergrundsprecherstatistik berechnet. Das Verhältnis dieser Distanzen wird dann als Ausgabe des Systems verwendet (Details zu diesen Systemen sind in den Abschnitten 3.2.2.2, 8.4 und 8.5 zu finden).

8.2 Metrik zur Bewertung der Systemleistung

Zur Bewertung der Ergebnisse der unterschiedlichen Systemkonfigurationen muss ein Auswertungskonzept bereitgestellt werden, welches im Folgenden beschrieben wird. Hierfür werden zunächst die möglichen Fehler erläutert, die ein Sprecherverifikationssystem produzieren kann. Anschließend werden unterschiedliche Darstellungsarten der Systemleistung vorgestellt.

Bei der Verifikation wird geprüft, ob ein Sprachsignal von einen Zielsprecher (auch Referenzsprecher genannt) gesprochen wurde oder nicht. Ist dieses Sprachsignal tatsächlich von dem Zielsprecher artikuliert worden, so handelt es sich um einem Zielversuch. Bei einem Nicht-Zielversuch ist das Sprachsignal von einem anderen Sprecher gesprochen worden. Für einen Zielversuch kann die Entscheidung des Systems entweder als korrekte Zuweisung zu dem Zielsprecher ausfallen oder das Sprachsignal kann fälschlicherweise zurückgewiesen werden. Eine solche Rückweisung wird *falsche Rückweisung* oder *fehlender Alarm*[9] genannt (miss detection). Für einen Nicht-Zielversuch bestehen ebenfalls zwei Möglichkeiten. Das Nicht-Zielsignal kann korrekterweise zurückgewiesen werden oder fälschlicherweise als Zielsprecher akzeptiert werden. Diese Fehlentscheidung wird *fehlende Rückweisung* oder *falscher Alarm* genannt (false alarm).

[9]Mit Alarm ist die Zuordnung eines Sprachsignals zu einem Sprechermodell durch das System gemeint; unabhängig von der Richtigkeit der Zuordnung. Ein Alarm liegt folglich immer dann vor, wenn keine Rückweisung erfolgt. Ein fehlender Alarm ist demnach eine Rückweisung, die fälschlicherweise erfolgt, da es sich um den Zielsprecher handelt, für den ein Alarm ausgelöst werden müsste.

Für die Visualisierung dieser Fehler gibt es unterschiedliche Darstellungsmöglichkeiten. Eine ältere Standarddarstellungsart ist die *ROC-Kurve (Receiver Operating Characteristic)*. Bei der ROC-Kurve werden die falschen Alarme in Abhängigkeit von den korrekten Erkennungen aufgezeichnet. Die falschen Alarme werden auf der x-Achse und die korrekten Erkennungen auf der y-Achse dargestellt. Je weiter die Kurve eines Systems in der Ecke links oben liegt, desto besser ist seine Leistung. Die Achsenbeschriftung der ROC-Kurve erfolgt linear.

Die neuere Variante der ROC-Kurve ist die *DET-Kurve (Detection Error Tradeoff)*. Bei der DET-Kurve werden nicht die korrekten Erkennungen dargestellt, sondern die beiden möglichen Fehler. Die Achsenskalierung ist nicht linear, sondern entspricht der Standardnormalverteilung (Martin, Doddington, Kamm, Ordowski & Przybocki, 1997). Dies wird gemacht, um den interessanten Bereich der niedrigen Fehlerraten zu fokussieren. Die nichtlinearen Achsen führen dazu, dass die Leistungskurven von Systemen mit einem ausgewogenen Verhältnis zwischen fehlendem und falschen Alarm annähernd linear sind. Je besser die Leistung des Systems ist, desto weiter links unten befindet sich die Kurve im Graphen. Ein bedeutungsvoller Punkt der DET-Kurve ist die *Gleichfehlerrate (Equal Error Rate, EER)*. Dies ist der Punkt der Kurve, an dem beide Fehler gleich groß sind. Dieser Punkt wird als Vergleichswert in dieser Arbeit genutzt. Weiterhin können für Systeme, die eine Entscheidung treffen, anstatt nur eine Punktzahl auszugeben, die Kosten berechnet werden, die von dem System für das gegebene Korpus verursacht wurden. Die Kostenberechnung stellt eine Möglichkeit dar, die Systemleistung auf das Szenario angepasst zu beurteilen. Die Kosten setzen sich zusammen aus den Kosten für fehlende und falsche Alarme, welche abhängig von der jeweiligen Gewichtung vom Nutzer festgelegt werden; den Fehlerraten für falschen und fehlenden Alarm; sowie der a priori Wahrscheinlichkeit für das Auftreten eines Zielsprechersignals. Weiterhin können für jede Systemausgabe die minimalen Kosten bei optimaler Einstellung des Systems berechnet werden. Eine detaillierte Beschreibung der Berechnung der Kosten ist in dem NIST Evaluationskonzept für die Sprechererkennung zu finden ("The NIST Year 2010 Speaker Recognition Evaluation Plan", 2010).

Liegen die ausgegebenen Punktzahlen des untersuchten Sprechererkennungssystems in Form von Log-Likelihood-Ratios vor, so kann die *APE-Kurve (Applied Probability of Error)* zur Visualisierung genutzt werden. Die APE-Kurve eignet sich besonders zur Darstellung der Kalibrierung eines Systems, während die DET-Kurve zur Visualisierung der diskriminativen Fähigkeiten eines Systems optimal ist (van Leeuwen & Brümmer, 2007). Da die Kalibrierung in dieser Arbeit nicht behandelt wird, wird die APE-Kurve nicht zur Analyse der Ergebnisse verwendet. Zudem liegen nicht alle Ausgaben als Log-Likelihood-Ratio vor.

Für die nachfolgenden Tests wird die **DET-Kurve als Darstellungsart gewählt**, da diese Form der Darstellung für unterschiedliche Ausgabeformen der Punktzahl geeignet ist. Weiterhin werden DET-Kurven auch in den Standardevaluationen des *National Institute of Standards and Technology* eingesetzt ("The NIST Year 2010 Speaker Recognition Evaluation Plan", 2010). Als Vergleichswert zwischen zwei Systemen wird die EER verwendet. Kosten werden nicht berechnet, da das hier erstellte Sprecherverifikationssystem keine Entscheidung ausgibt, sondern lediglich eine Punktzahl.

	MFCC	LPC	PLP
Vgl(TF)	2%	14%	2%
Basis	10%	29%	7%
Basis(TF)	6,04%	29%	6%
UBM+	11%	29%	9%
UBM+(TF)	6,43%	27%	7%

Tabelle 8.1: Vergleich der EER der unterschiedlichen Merkmale und Einstellungen (Vgl steht für den Vergleichstest: normale gegen normale Sprache; Basis steht für die oben beschriebenen Standardeinstellungen des Merkmals; UBM+ steht für das mit dem OLLO-Korpus erweiterte UBM; TF steht für Telefonfilter.)

8.3 Standardmerkmale

Das in Abschnitt 8.1 beschriebene Basissystem wird im Folgenden mit unterschiedlichen Arten von Merkmalen getestet, um herauszufinden, welches Merkmal am wenigsten durch die Erhöhung des Stimmaufwands beeinflusst wird. Die in Betracht gezogenen Merkmale sind die in Abschnitt 3.2.2.1 beschriebenen *MFCC-*, *LPC-* und *PLP-Merkmale*. Sämtliche Merkmale werden mit HTK (Young et al., 2006) extrahiert. Vorabtests mit einer anderen Korpuszusammenstellung wurden bereits in Harwardt (2010c)[10] vorgestellt.

Bei der Berechnung der MFCC-Merkmale werden entsprechend der Standardeinstellungen (Rosenberg, Bimbot & Parthasarathy, 2008) 15 Koeffizienten extrahiert sowie die zugehörigen Delta und Delta Delta Koeffizienten. Insgesamt ergibt dies einen Merkmalsvektor mit 45 Elementen. Für die DFT wird ein Hamming-Fenster von 20 ms gewählt sowie 26 Filterbankkanäle. Weiterhin wird eine cepstrale Mittelwertsubtraktion durchgeführt. Bei der LPC- und PLP-Analyse werden ebenfalls 15 Koeffizienten plus Delta und Delta Delta Koeffizienten berechnet.

Die Ergebnisse der Tests sind in Tabelle 8.1 dargestellt. Für die drei Merkmale wird zunächst ein Vergleichstest (Vgl) gemacht, bei dem Sprache normalen Stimmaufwands gegen Sprache normalen Stimmaufwands getestet wird. Diese Tests sind notwendig, um bestimmen zu können, wie stark die Erkennung durch den veränderten Stimmaufwand beeinflusst wird. Für die Vergleichstest wird ein Telefonfilter verwendet, da die MFCC-Merkmale kombiniert mit einem Telefonfilter bei den Folgetests die besten Ergebnisse lieferten (siehe Tabelle 8.1 — Basis(TF)). Die Audiodaten für diesen und die drei weiteren Tests stammen aus dem „Pool 2010“-Korpus (siehe Abschnitt 5.2.1). Für das Training des UBMs werden die männlichen Sprecher des Kiel-Korpus verwendet (siehe Abschnitt 5.2.3).

Die weiteren Tests prüfen normalen gegen erhöhten Stimmaufwand. Die Standardeinstellung ohne Telefonfilter wird Basistest (Basis) genannt. Basis(TF) ist eine Erweiterung des Basistests um einen Telefonfilter. Bei den Basistests und den Vergleichstest werden für das UBM nur Daten der männlichen Sprecher des Kiel-Korpus verwendet. Für die Tests UBM+ wird das UBM-Korpus um Daten männlicher Sprecher mit normalem Stimmaufwand aus dem OLLO-Korpus erweitert. Auch mit dem erweiterten UBM wird nach den Basistests (UBM+) eine Kombination mit einem Telefonfilter durchgeführt (UBM+(TF)).

[10] Ein Abdruck befindet sich im Anhang A.3.

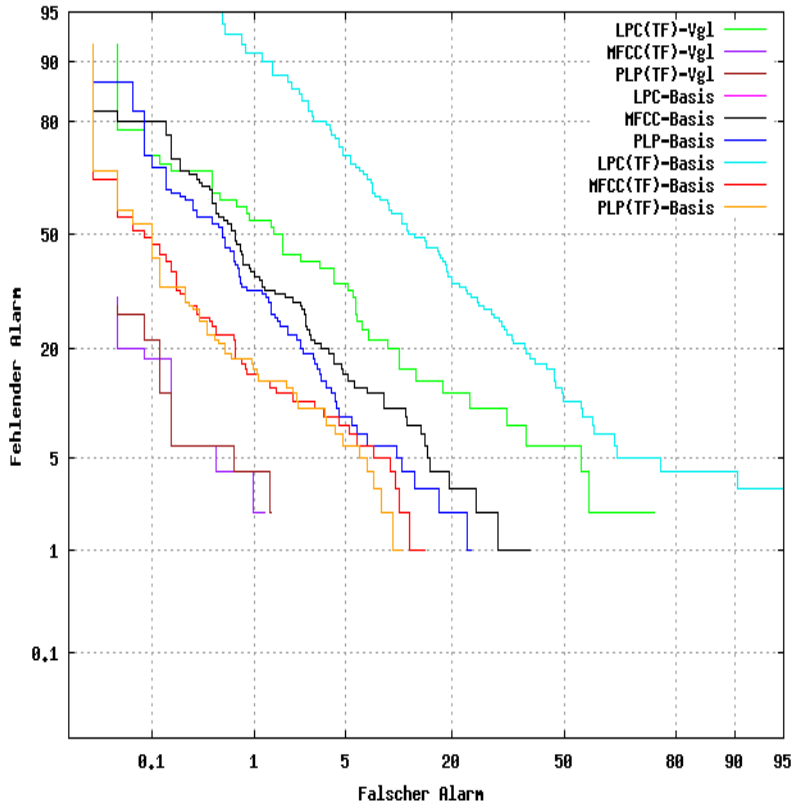

Abbildung 8.1: DET-Kurven zum Vergleich unterschiedlicher Merkmale mit den Basis UBM-Daten

Insgesamt wird sichtbar, dass die MFCC- und PLP-Merkmale ähnlich gute Ergebnisse erzielen. Die LPC-Merkmale zeigen wesentlich schlechtere Leistungen in sämtlichen Tests. Um die Ergebnisse besser interpretieren zu können, werden die Ergebnisse der Basistests mit und ohne Telefonfilter in Abbildung 8.1 als DET-Kurven präsentiert. Die Ergebnisse mit dem erweiterten UBM werden in Abbildung 8.2 dargestellt. Die EER zeigt sich in den Abbildungen als der Schnittpunkt der Kurve mit der Diagonale vom Nullpunkt zur oberen rechten Ecke der Graphik. Für die Vergleichstests wird, wie erwartet, für alle drei Merkmale die jeweils niedrigste EER erreicht. Da die EER für MFCC- und PLP-Merkmale bei nur 2% liegt, wird die EER bei erhöhtem Stimmaufwand im Testdurchlauf teilweise bis zu 5 Mal höher. Die EER der LPC-Merkmale hingegen, welche für die Vergleichstests bereits sehr hoch ist, wird lediglich verdoppelt. Die Erweiterung des UBMs führt nur für die LPC-Merkmale in Kombination mit dem Telefonfilter zu einer Verbesserung. Für die MFCC- und PLP-Merkmale werden die Ergebnisse sogar schlechter. Dies ist dadurch erklärbar, dass die zugefügten Daten aus dem OLLO-Korpus keine Spontansprache, sondern Logatome enthalten, welche andere Charakteristika aufweisen (beispielsweise Länge der Sequenz oder Betonung) als Spontansprache. Weiterhin zeigt sich, dass die Tests mit Telefonfilter zu besseren Ergebnissen führen als die ohne, sodass für die Vergleichstests die Einstellung Basis(TF) gewählt wurde.

Beim Vergleich der Ergebnisse des Basiskorpus zeigt Tabelle 8.1, dass die EER für MFCC- und PLP-Merkmale mindestens verdreifacht wurde. Für die LPC-Merkmale

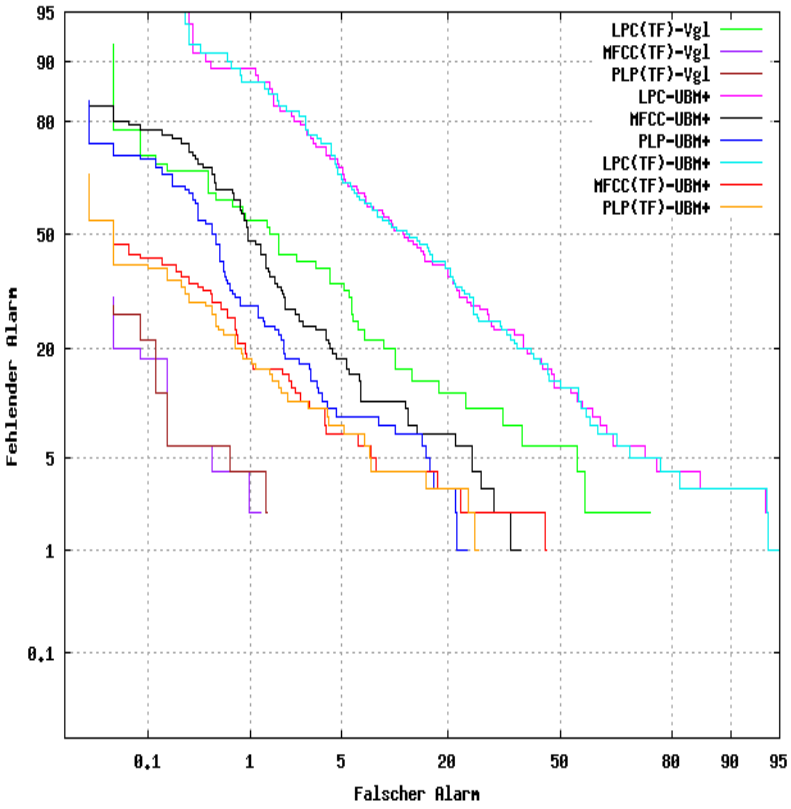

Abbildung 8.2: DET-Kurven zum Vergleich unterschiedlicher Merkmale mit dem erweiterten UBM

kann eine Verdopplung beobachtet werden. Die Ergebnisse der Tests mit und ohne Telefonfilter sind für die LPC-Merkmale gleich. Für MFCC- und PLP-Merkmale kann hingegen ohne Telefonfilter eine weitere Verschlechterung festgestellt werden. Diese Verschlechterung scheint für PLP-Merkmale zunächst nicht so groß zu sein wie für MFCC-Merkmale. Anhand der Kurven in Abbildung 8.1 wird deutlich, dass die Kurven der MFCC- und PLP-Merkmale ohne Telefonfilter eng beieinander liegen. Lediglich um den Bereich der EER schneiden die PLP-Merkmale deutlich besser ab. Die Kurve des Basistests der LPC-Merkmale (LPC-Basis) ist in Abbildung 8.1 nicht ersichtlich, da sie mit der Kurve des Basistests mit Telefonfilter (LPC(TF)-Basis) übereinstimmt.

Die Tests mit dem erweiterten UBM liefern insgesamt nur für die LPC-Merkmale unter Nutzung des Telefonfilters (LPC(TF)-UBM+) eine Verbesserung (siehe Abbildung 8.2). Die anderen Testergebnisse werden durch die zusätzlichen UBM-Daten verschlechtert. Die Ergebnisse der MFCC-Merkmale werden weniger stark verschlechtert als die der PLP-Merkmale. Insgesamt liegen die EER-Werte beider Merkmale nah beieinander, sodass auch die zugehörigen Kurven in Abbildung 8.2 sehr ähnlich sind.

Zusammenfassend kann festgehalten werden, dass je nach Merkmal und Einstellung eine **bis zu 5 Mal schlechtere EER bei nicht-übereinstimmendem Stimmaufwand** erreicht wird. Auch wenn insbesondere die MFCC- und PLP-Merkmale beeinflusst

werden, liegt die Gesamtleistung beider Merkmale sowohl bei übereinstimmendem als auch bei nicht-übereinstimmendem Stimmaufwand über der Leistung der LPC-Merkmale. Da die **MFCC-Merkmale** in der Sprechererkennung sehr häufig Anwendung finden und für ihre Robustheit in der Spracherkennung bekannt sind, werden sie, anstatt der PLP-Merkmale, **für das Basissystem verwendet**.

8.4 F_0-basierte Merkmale

Im vorherigen Abschnitt wurden verschiedene Standardmerkmale für die automatische Sprechererkennung im Kontext unterschiedlicher Stimmaufwandsgrade in Trainings- und Testmaterial untersucht. Nachfolgend wird die Untersuchung ausgeweitet auf unterschiedliche F_0-basierte Merkmale. Die F_0-basierten Merkmale werden genauer betrachtet, da die Grundfrequenz besonders in der forensischen Sprechererkennung bedeutend ist und bei erhöhtem Stimmaufwand nicht als zuverlässiger Parameter verwendet werden kann. Die Verschlechterung der Systemleistung bei F_0-basierten Merkmalen ist der Untersuchungsschwerpunkt dieses Abschnitts. Vorabtests mit einer anderen Korpuszusammenstellung wurden bereits in Harwardt (2010a)[11] vorgestellt.

Für die Untersuchung der F_0-basierten Merkmale werden je Merkmal vier verschiedene Tests durchgeführt (siehe Tabelle 8.2). Zunächst wird die Funktionsfähigkeit der Merkmale in einem Vergleichstest (Vgl) mit gleichbleibendem Stimmaufwand in Trainings- und Testdaten geprüft. Dafür wird je Sprecher eine Audiodatei normalen Stimmaufwands für das Training und eine weitere Aufnahme normalen Stimmaufwands für den Test verwendet. Die Audiodaten für diesen und die drei weiteren Tests stammen aus dem „Pool 2010"-Korpus (siehe Abschnitt 5.2.1). Für das Training des UBMs werden die männlichen Sprecher des Kiel-Korpus verwendet (siehe Abschnitt 5.2.3).

Die erste Untersuchung mit nicht-übereinstimmendem Stimmaufwand in Trainings- und Testdaten (Basis) verwendet dieselben Audiodaten für das Training wie der Vergleichstest. Die Testdaten umfassen Audiodateien erhöhten Stimmaufwands. In diesem Test liegen je Sprecher zwei Audiodateien zum Test vor.

Ein weiterer Test (UBM+) nutzt zusätzliche Audiodaten für das UBM. Dies bedeutet, dass zu den UBM-Daten des Basistests die Aufnahmen der männlichen deutschen Sprecher normalen Stimmaufwands des OLLO-Korpus (siehe Abschnitt 5.2.2) hinzugefügt werden.

Der letzte Test (MFCC(TF)+Basis) nutzt die Audiodaten des Basistests. Hier werden die Ergebnisse der F_0-basierten Merkmale mit denen des MFCC-Basissystems mit Telefonfilter fusioniert. Hierfür wird eine lineare Fusion mit Hilfe des FoCal Toolkits (Brummer & du Preez, 2006) durchgeführt. Zum Training der Fusionierung wurden die Daten der Vergleichstests herangezogen. Diese Daten sind nicht optimal, da hier eine Übereinstimmung im Stimmaufwand vorliegt, welche im späteren Einsatzszenario nicht gegeben ist. Aus Mangel besser geeigneter Audiodaten wurden dennoch die Ergebnisse der Vergleichtests zum Training der Fusion der Basistests angewendet. Eine Zusammenführung der Merkmale auf Merkmalsebene ist in diesem Fall nicht möglich, da nicht für alle Merkmale dieselbe Anzahl an Messwerten vorhanden ist. Für ein Merkmal (F_0-Statistik — siehe unten) ist pro Sprachsignal

[11] Ein Abdruck befindet sich im Anhang A.3.

	F_0-Statistik	logF_0	logF_0/E
Vgl	66%	16,86%	24%
Basis	56,1%	46,02%	37%
UBM+	56,27%	50%	37%
MFCC(TF)+Basis	6%	6,31%	6%

Tabelle 8.2: Vergleich der EER der unterschiedlichen F_0-basierten Merkmale (Vgl steht für den Vergleichstest: normale gegen normale Sprache; Basis steht für die oben beschriebenen Standardeinstellungen des Merkmals; UBM+ steht für das mit dem OLLO-Korpus erweiterte UBM; MFCC(TF) steht für das MFCC-Basissystem aus Abschnitt 8.3 mit Telefonfilter.)

nur ein Messwert vorhanden, während die anderen Merkmale einen Messwert pro Frame aufweisen.

In den Experimenten soll geprüft werden, wie stark die Verschlechterung der Erkennungsergebnisse automatischer F_0-basierter Systeme ist. Es werden drei unterschiedliche F_0-basierte Merkmale getestet.

Zunächst wird die *F_0-Statistik* mit dem Mittelwert, der Standardabweichung, der Schiefe und der Kurtosis angewendet (siehe Abschnitt 3.2.2.2). Da keine Information pro Frame erfasst wird, sondern eine Statistik pro Sprachsignal erstellt wird, kann ein GMM-UBM-basiertes System, wie in Abschnitt 3.2.3 beschrieben, nicht genutzt werden. Stattdessen wird pro Sprachsignal die zugehörige F_0-Statistik berechnet, um anschließend die Distanz zwischen der Trainingsstatistik und der Teststatistik sowie zwischen der UBM-Statistik und der Teststatistik zu berechnen. Diese Differenzen werden dann zueinander ins Verhältnis gesetzt (Distanz zum Trainingssignal/ Distanz zum UBM). Zur Kalkulation der Differenz zweier Statistiken wird die Kullback-Leibler Distanz verwendet, da diese sich in den Experimenten von Kinnunen und González Hautamäki (2005) bewährt hat.

Das zweite F_0-basierte System (logF_0) orientiert sich an den Ausführungen von Reynolds et al. (2002). Die Grundfrequenz wird für jeden Frame gemessen. Anschließend wird ein Medianfilter angewendet (Rabiner & Schafer, 1978). Die gewählte Fensterbreite umfasst 5 Werte. Das Medianfilter entfernt Ausreißer, indem es für jeden Wert den Median von diesem und den vier umliegenden Werten als neuen Wert festlegt. Nach der Filterung der F_0-Werte und der Entfernung stimmloser Segmente werden die verbleibenden F_0-Werte logarithmiert und ihre Delta Koeffizienten berechnet (Euler, 2006). Statt der 1024 Mischungskomponenten für das MFCC-Basissystem (siehe Abschnitt 8.1) werden hier nur 64 Mischungskomponenten verwendet, da der Merkmalsvektor kleiner ist.

Für *das dritte System (logF_0/E)* wurden zusätzlich zu den logarithmierten F_0-Werten die logarithmierten Energiewerte und ihre Delta Koeffizienten einbezogen. Dies entspricht der Standardlösung von Reynolds et al. (2002).

Die Ergebnisse der Tests mit den drei F_0-basierten Systemen sind in Tabelle 8.2 und den Abbildungen 8.3 und 8.4 dargestellt. Aus dem Vergleich der EERs in Tabelle 8.2 geht hervor, dass die F_0-Statistik nicht zur Nutzung in der automatischen Sprechererkennung geeignet ist. Obwohl die Ergebnisse von Kinnunen und González Hautamäki (2005) vielversprechend waren, zeigt das System auf dem „Pool 2010"-Korpus sowohl für übereinstimmenden als auch für nicht-übereinstimmenden Stimmaufwand in Trainings- und Testdaten keine guten Erkennungsraten. Überra-

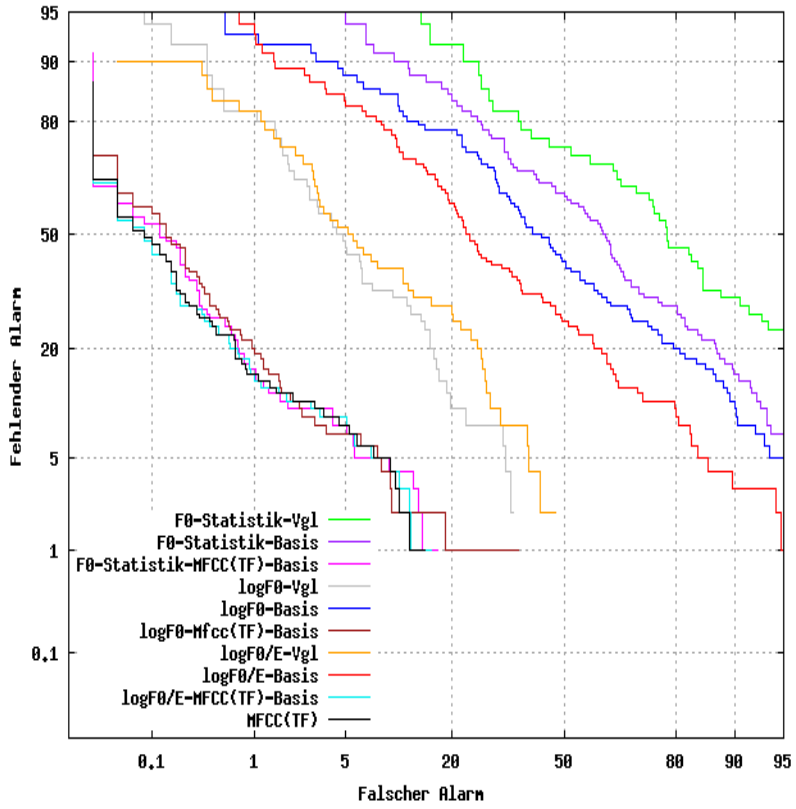

Abbildung 8.3: DET-Kurven zum Vergleich verschiedener Merkmale der unterschiedlichen F_0-basierten Systeme für das Basistestset

schenderweise ist die EER für übereinstimmenden Stimmaufwand sogar noch höher als für nicht-übereinstimmenden. Dies ist bei den anderen zwei Systemen nicht der Fall. Für übereinstimmenden Stimmaufwand zeigen die beiden Systeme mit logarithmierter F_0 gute Erkennungsraten, wobei das System ohne die logarithmierte Energie bessere Ergebnisse ausgibt. Bei den Tests mit nicht-übereinstimmendem Stimmaufwand ist das Gegenteil zu beobachten. Das System, welches die Information über die Energie nutzt, liefert bessere EERs als das System logF_0. Auch dieses Ergebnis ist interessant, da die Erwartung war, dass durch den veränderten Stimmaufwand die Energie derart stark beeinflusst wird, dass die Ergebnisse schlechter werden als bei einer Klassifikation, welche die Energie nicht einbezieht. Die Ergebnisse zeigen hingegen, dass die Energie zwar bei übereinstimmendem Stimmaufwand zu einer Verschlechterung führt; im Kontext nicht-übereinstimmenden Stimmaufwands kompensiert jedoch die veränderte logarithmierte Energie die F_0-Modifikationen. Daraus lässt sich schlussfolgern, dass die Energie im Kontext automatischer Sprechererkennung weniger stark durch Stimmaufwandsänderungen beeinflusst wird als F_0. Wie stark F_0 beeinflusst wird, zeigt sich an den Ergebnissen des Systems logF_0. Die EER wird beinahe verdreifacht. Das System logF_0/E verschlechtert sich dagegen lediglich um 13 Prozentpunkte. Auffällig ist weiterhin, dass, wie in Abschnitt 8.3, keine Verbesserung durch zusätzliche UBM-Daten erzielt werden kann. Die Kombination mit dem MFCC-Basissystem aus Abschnitt 8.3 liefert zwar eine Verbesserung, die Ergebnisse sind jedoch nicht so gut wie die des MFCC-Basissystems allein.

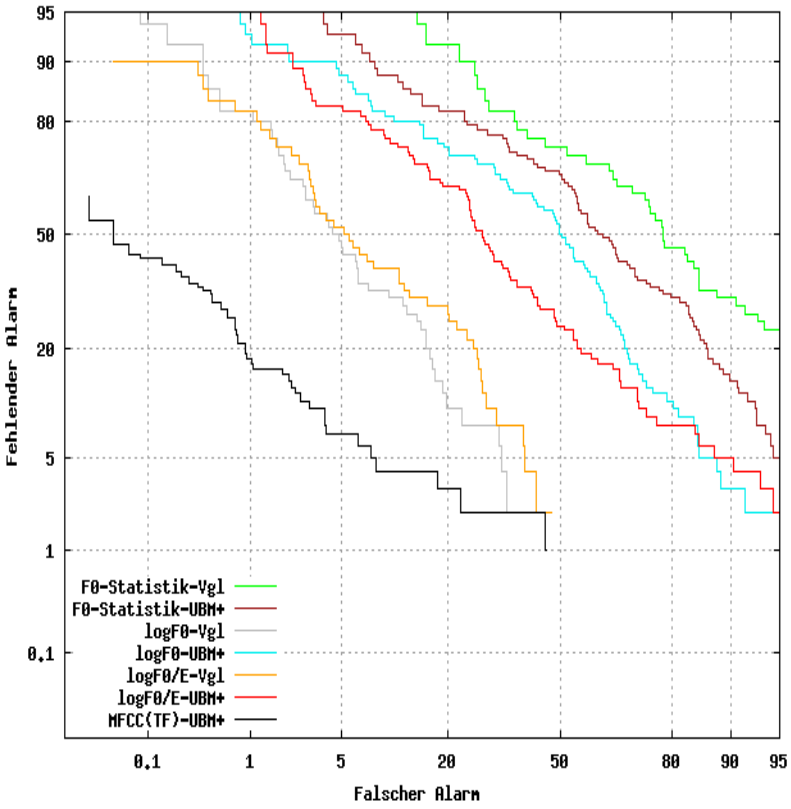

Abbildung 8.4: DET-Kurven zum Vergleich verschiedener Merkmale der unterschiedlichen F_0-basierten Systeme für die Tests mit erweitertem UBM

Zur genaueren Analyse der Ergebnisse über die EER hinaus können die DET-Kurven in den Abbildungen 8.3 und 8.4 genutzt werden. Die oben beschriebenen Ergebnisse werden durch die DET-Kurven bestätigt. Der Basistests (siehe Abbildung 8.3) liefert für die MFCC-Merkmale und für sämtliche F_0-Merkmale, kombiniert mit den MFCC-Merkmalen, ähnlich gute Ergebnisse. Die zugehörigen DET-Kurven sind links unten in der Abbildung zu finden. Die nächst besten Ergebnisse repräsentieren die Kurven der zwei Systeme mit logarithmierter F_0, wobei hier besonders die Verschlechterung durch die Nutzung der Energie deutlich wird. Die nachfolgenden Kurven sind die des Systems logF_0/E bei nicht-übereinstimmendem Stimmaufwand, gefolgt von denen des Systems logF_0. Oben rechts in der Darstellung sind die schlechtesten Ergebnisse, nämlich die des Systems F_0-Statsitik, abgebildet. Für die Tests mit erweitertem UBM (siehe Abbildung 8.4) zeigt sich die gleiche Leistungsrangfolge. Da die Ergebnisse dieser Tests insgesamt schlechter sind als die der Basistests, wurde keine Fusionierung durchgeführt.

Insgesamt zeigt sich durch die Tests, dass F_0-basierte Systeme teilweise (logF_0 und logF_0/E) im Kontext übereinstimmenden Stimmaufwands von Nutzen sind. Der veränderte Stimmaufwand ruft jedoch genau bei diesen Systemen eine große Verschlechterung hervor. Das System logF_0/E liefert im gegebenen Szenario die besten Ergebnisse. Die Kombination mit den MFCC-Merkmalen führt zu keiner weiteren Verbesserung im Vergleich zu dem MFCC-Basissystem. Dies bestätigt

die Ausgangsthese, dass **F_0-basierte Merkmale nur bedingt in diesem Szenario einsetzbar sind**.

8.5 Spezielle Merkmale bei erhöhtem Stimmaufwand

Dieser Abschnitt stellt zwei Merkmale vor, die speziell für veränderten Stimmaufwand entwickelt beziehungsweise entdeckt wurden. Zunächst wird das *COG-Verhältnis (center of gravity)* erläutert, welches auf dem spektralen Schwerpunkt basiert (siehe Abschnitt 2.4.3). Dann wird in Anlehnung an die statistische Analyse aus Kapitel 6 eine *Kombination verschiedener spektraler Parameter* für die automatische Sprechererkennung dargestellt. Es wird geprüft, ob diese Merkmale separat oder in Kombination mit den MFCC-Merkmalen eine bessere Leistung erzielen als das Basissystem allein.

Die ersten Tests der speziellen Merkmale sind entsprechend der vier Tests aus Abschnitt 8.4 konzipiert (Vgl, Basis, UBM+, MFCC(TF)+Basis). Zusätzlich zu diesen vier Tests wird ein fünfter Test (MFCC(TF)+UBM+) durchgeführt, bei dem die Ergebnisse der Tests UBM+ mit denen des MFCC-Basissystems fusioniert werden. Dieser fünfte Test ist eingeführt worden, da das erweiterte UBM für die in diesem Kapitel beschriebenen Merkmale teilweise zu einer Verbesserung führt. Für die F_0-basierten Merkmale ist dies nicht der Fall. Die Fusionierung der Ergebnisse wird mit dem FoCal Toolkit durchgeführt (Brummer & du Preez, 2006), da eine Kombination auf Merkmalsebene, wie in Abschnitt 8.4 erläutert, für die Kombination der spektralen Parameter nicht möglich ist.

Um das *COG-Verhältnis* erklären zu können, werden kurz die Eigenschaften des spektralen Schwerpunkts (siehe Abschnitt 2.4.3) in lauter Sprache dargestellt. Der spektrale Schwerpunkt ist ein Parameter, der signifikant durch wechselnden Stimmaufwand beeinflusst wird (siehe Abschnitt 6.2.3). Verschiedene Untersuchungen der spektralen Energie bei erhöhtem Stimmaufwand haben eine Migration von niedrigen zu mittleren und zu hohen Frequenzen hin gezeigt. Diese Energiemigration ist ursächlich für die Veränderung des spektralen Schwerpunkts. Liénard und Di Benedetto (1999) berechnen den spektralen Schwerpunkt für die verschiedenen Formanten separat. In diesem Fall steigt die Energie für F2 und F3 und damit auch der spektrale Schwerpunkt, jeweils bei erhöhtem Stimmaufwand.

Bei der Konzeption des COG-Verhältnisses wurde das Ziel verfolgt, ein Merkmal zu finden, welches robust gegenüber der veränderten Energiekonzentration des spektralen Schwerpunkts ist. Dafür wird das Verhältnis zwischen dem Schwerpunkt hoher und mittlerer Frequenzen berechnet, anstatt den Schwerpunkt über alle Frequenzen oder verschiedene spektrale Schwerpunkte für die einzelnen Formanten zu kalkulieren. Da in der Sprechererkennung meist bandbegrenzte Audiodaten verwendet werden, werden der hohe und mittlere Frequenzbereich so definiert, dass der hohe Frequenzbereich unterhalb 3400 Hz liegt. Außerdem soll, in Anlehnung an Liénard und Di Benedetto (1999), der mittlere Frequenzbereich mögliche F2-Werte enthalten und der hohe Bereich den spektralen Schwerpunkt um den dritten Formant repräsentieren. Der mittlere Frequenzbereich reicht dementsprechend von 800 bis 2200 Hz und der hohe von 2200 bis 3000 Hz. Die Berechnung eines Verhältnisses als Merkmal für die Sprechererkennung knüpft weiterhin an den Vorschlag Junqua (1993) an, der solche Verhältnismerkmale als Möglichkeit zur Verbesserung der

	COG-Verhältnis mit VAD	COG-Verhältnis ohne VAD	Momente Kombination
Vgl	38%	36%	60,61%
Basis	36%	37%	56%
UBM+	26%	26%	57%
MFCC(TF)+Basis	6%	5,37%	7%
MFCC(TF)+UBM+	6%	7%	7%

Tabelle 8.3: Vergleich der EER für unterschiedliche spektrale Merkmale (Vgl steht für den Vergleichstest: normale gegen normale Sprache; Basis steht für die oben beschriebenen Standardeinstellungen des Merkmals; UBM+ steht für das mit dem OLLO-Korpus erweiterte UBM; MFCC(TF) steht für das MFCC-Basissystem aus Abschnitt 8.3 mit Telefonfilter.)

Spracherkennung bei verändertem Stimmaufwand angibt (siehe Abschnitt 4.2). Für die Nutzung in der Sprechererkennung wird das COG-Verhältnis der mittleren und hohen Frequenzen sowie die erste und zweite Ableitung über fünf Frames berechnet (siehe Tabelle 8.3). In den nachfolgend beschriebenen Tests wird zusätzlich zu der Nutzung des COG-Verhältnisses ohne *Sprache-Pause-Detektion (voice activity detection — VAD)* (COG-Verhältnis ohne VAD) eine Version des Systems inklusive Sprache-Pause-Detektion (COG-Verhältnis mit VAD) realisiert. Beide Realisierungen sind in ein GMM-UBM Framework, wie in Abschnitt 8.1 beschrieben, eingebettet. Allerdings werden 128 statt 1024 Mischungskomponenten für das GMM verwendet, da sich diese Anzahl in ersten Tests als optimal herausgestellt hat (Harwardt, 2010b)[12].

Bei der *Kombination verschiedener spektraler Parameter* werden die Erkenntnisse aus Kapitel 6 genutzt. In Abschnitt 6.2.5.1 ergab die sprecherspezifische Analyse des ersten spektralen Moments, dass die mittlere Abweichung und der Maximalwert für einzelne Sprecher bei einer Erhöhung des Stimmaufwands relativ konstant bleiben. Weiterhin scheint eine Intersprechervariabilität für diese Parameter gegeben zu sein. Ob diese Intersprechervariabilität ausreichend groß ist, um eine zuverlässige Differenzierung zwischen unterschiedlichen Sprechern vornehmen zu können, wird durch die Nutzung der Parameter als Merkmale in einem Sprechererkennungssystem überprüft. Zusätzlich zu diesen zwei Parametern wird der Minimalwert des vierten spektralen Moments verwendet, da auch dieser Wert kaum durch eine Erhöhung des Stimmaufwands verändert wird aber für die unterschiedlichen Sprecher variiert. Diese drei Parameter werden als Merkmalsvektor für die Sprechererkennung genutzt (siehe Tabelle 8.3 — Momente Kombination). Da diese Merkmale nicht wie das COG-Verhältnis über einzelne kurze Frames, sondern über das gesamte Sprachsignal berechnet werden, kann kein GMM-UBM Framework zur Anwendung kommen. Stattdessen wird die Distanz zwischen Trainings-, Test- und Hintergrundmodelldaten berechnet, wie bei dem System, das F_0-Statistiken nutzt (siehe Abschnitt 8.4).

Die Ergebnisse der unterschiedlichen spektralen Merkmale sind in Tabelle 8.3 und den Abbildungen 8.5 sowie 8.6 dargestellt. Die Leistungen der zwei Versionen des COG-Verhältnisses sind sehr ähnlich. Die EER der Vergleichstests und der Basistests ist hoch, wobei interessanterweise die Basistests etwas besser abschnei-

[12]Ein Abdruck befindet sich im Anhang A.3.

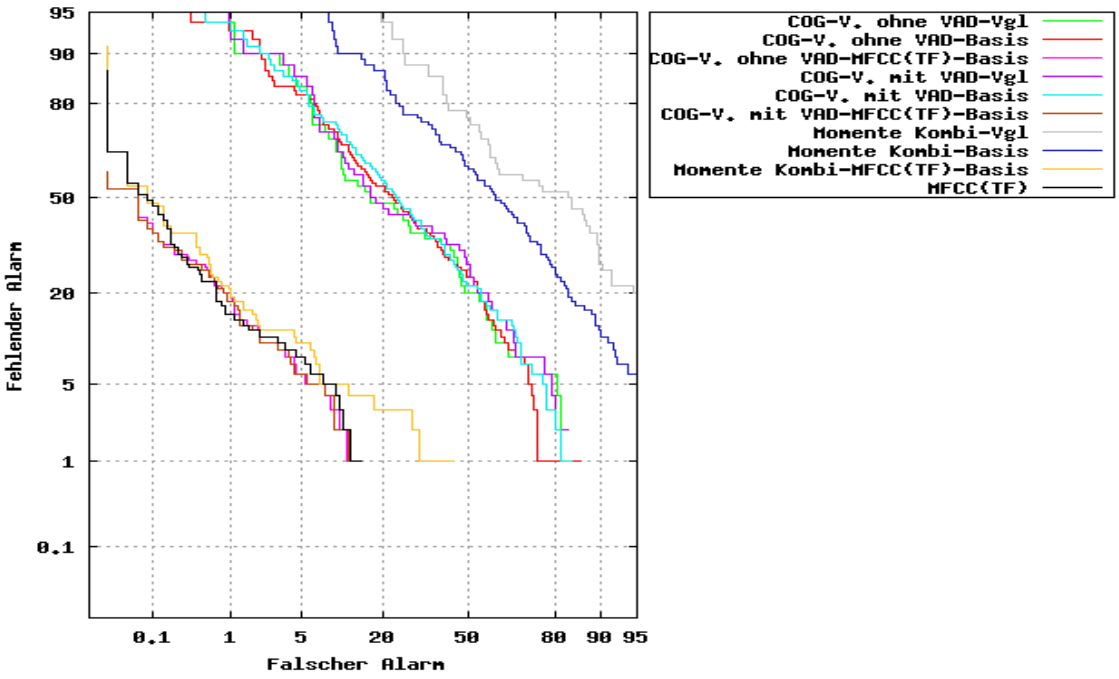

Abbildung 8.5: DET-Kurven zum Vergleich verschiedener Merkmale der unterschiedlichen Systeme spektraler Merkmale für das Basiskorpus

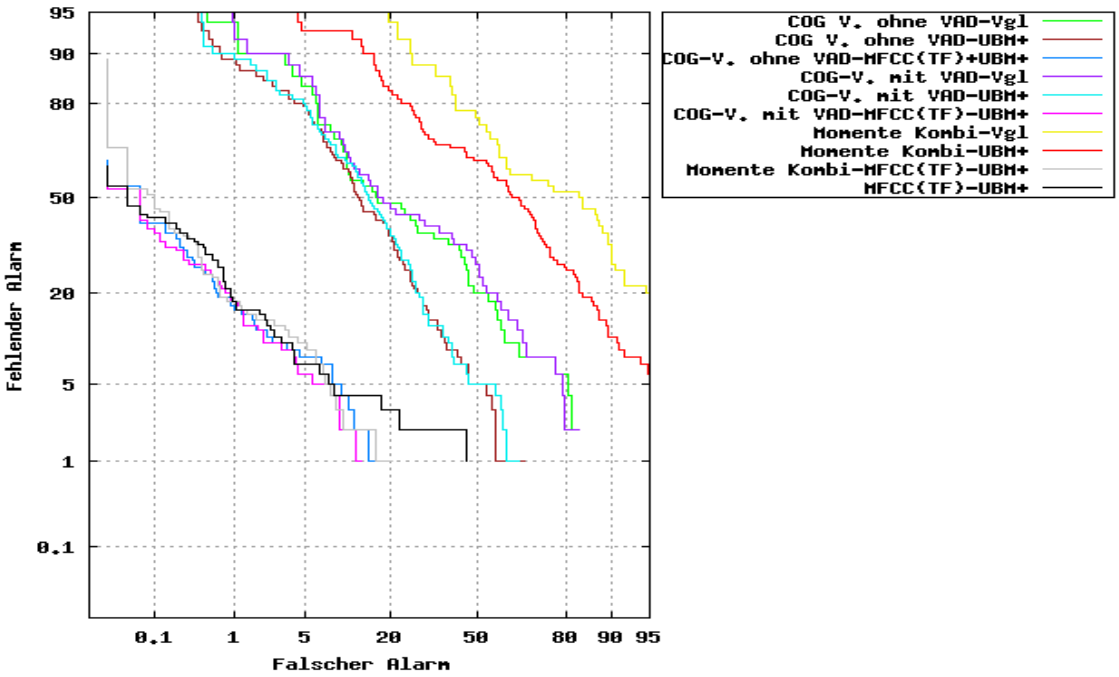

Abbildung 8.6: DET-Kurven zum Vergleich verschiedener Merkmale der unterschiedlichen Systeme spektraler Merkmale für das Korpus mit erweitertem UBM

den. Das Ergebnis des Basistests lässt sich durch eine Erweiterung der UBM-Daten (UBM+) stark verbessern. Die Kombination mit den MFCC-Merkmalen führt zu einer weiteren Verbesserung. Allerdings liegt die EER nur für das COG-Verhältnis ohne Sprache-Pause-Detektion unter der EER von 6% des MFCC-Basissystems. Das COG-Verhältnis mit Sprache-Pause-Detektion enthält folglich keine komplementäre Information zu den MFCC-Merkmalen; das COG-Verhältnis ohne Sprache-Pause-Detektion hingegen schon. Die Kombination einzelner spektraler Momente führt zu einer ähnlichen Leistung wie die F_0-Statistik. Die EERs sind für sämtliche Tests hoch. Die Kombination mit den MFCC-Merkmalen ist zwar wesentlich besser, die EER ist aber höher als nur für das MFCC-Basissystem allein. Die Abbildungen 8.5 und 8.6 präsentieren ähnliche Leistungen für die Tests mit den Basiseinstellungen im Vergleich zu denen mit erweitertem UBM. Für beide Testszenarien sind die guten Ergebnisse der MFCC-Merkmale und der Kombinationen mit diesen in den Darstellungen unten links zu finden. Es sind keine großen Unterschiede zwischen den einzelnen DET-Kurven erkennbar. Als Nächstes, in der Mitte der Abbildungen liegend, sind die weiteren Ergebnisse des COG-Verhältnisses mit und ohne Sprache-Pause-Detektion sichtbar. Hier zeigen die zwei Kurven der Tests UBM+ (COG-V. mit VAD-UBM+, COG-V. ohne VAD-UBM+) in Abbildung 8.6 etwas bessere Resultate als die zugehörigen Vergleichstests. Abschließend kann oben rechts die Leistung der Kombination der Momente als Merkmal für die Sprechererkennung abgelesen werden. Diese Merkmale liefern für beide Testszenarien im Vergleich zu den anderen Merkmalen die schlechtesten Ergebnisse.

Abschließend kann festgehalten werden, dass **keines der Merkmale für sich genommen so gute Ergebnisse wie das MFCC-Basissystem erzielt**. Das **COG-Verhältnis ohne Sprache-Pause-Detektion in Kombination mit den MFCC-Merkmalen** erzeugt hingegen **bessere Ergebnisse**. Die anderen Merkmale liefern keine komplementäre Information zu den MFCC-Merkmalen.

8.6 Zusammenfassung

In den vorausgegangenen Abschnitten wurden unterschiedliche Merkmale im Kontext automatischer Sprechererkennung bei nicht-übereinstimmendem Stimmaufwand in Trainings- und Testdaten verglichen. Es zeigte sich, dass die Merkmale, die über das gesamte Sprachsignal berechnet werden, statt über einzelne Frames, keine gute Leistung erzielen. Für die anderen Merkmale gilt, dass durch die Erhöhung des Stimmaufwands eine teilweise sehr starke Verschlechterung der Leistung erfolgt. Diese Verschlechterung kann durch die richtige Merkmalskombination abgeschwächt, jedoch nicht komplett aufgefangen werden. Um dies zu verdeutlichen, werden die besten Merkmalskombinationen im Folgenden vergleichend dargestellt.

Als Basissystem für die Tests wurde das GMM-UBM Framework mit den MFCC-Merkmalen inklusive Telefonfilter festgelegt (siehe Abschnitt 8.3). Dieses Basissystem sowie die besten Systemrealisierungen der Untersuchung werden in den DET-Kurven in Abbildung 8.7 gemeinsam präsentiert. Bei den besten Systemrealisierungen handelt es sich um zwei Kombinationen mit den MFCC-Merkmalen. Die erste Kombination nutzt den Logarithmus von F_0 und der Energie ($\log F_0/E$), während die zweite Kombination das COG-Verhältnis ohne Sprache-Pause-Detektion

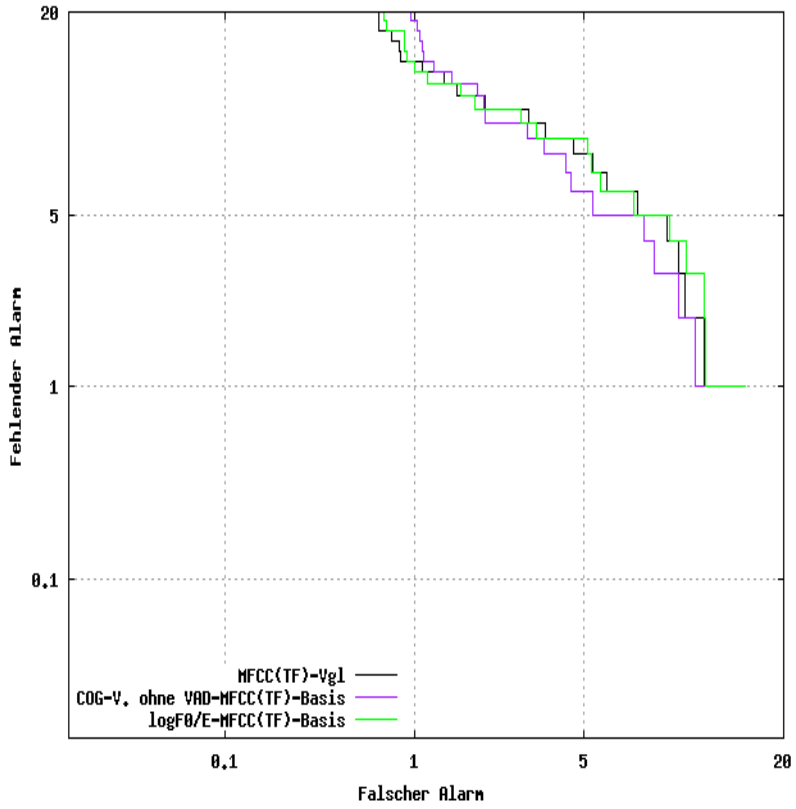

Abbildung 8.7: DET-Kurven zum Vergleich der besten Merkmale

nutzt. Allein schneiden diese Merkmale schlechter ab als die MFCC-Merkmale. Die Kombination liefert dagegen vergleichbar gute Ergebnisse. Mit den F_0-basierten Merkmalen kann keine Verbesserung der EER erzielt werden, jedoch ist die Leistung an einigen Operationspunkten der DET-Kurve etwas besser. **Das COG-Verhältnis, welches speziell für dieses Szenario entwickelt wurde, liefert in Kombination mit den MFCC-Merkmalen bessere Ergebnisse als das Basissystem.** Dies ist nicht nur an der EER zu erkennen. Auch an den meisten anderen Operationspunkten der DET-Kurve ist die Leistung der Kombination besser. **Folglich liefert das COG-Verhältnis komplementäre Information zu den MFCC-Merkmalen.**

Kapitel 9

Adaptionsverfahren

In Kapitel 8 wurden unterschiedliche Merkmale zur Nutzung in der automatischen Sprechererkennung verglichen. Das gegebene Szenario sieht normalen Stimmaufwand im Trainingsmaterial und erhöhten Stimmaufwand in den Testdaten vor. Um die Leistung der Sprecherverifikation für dieses Szenario weiter zu verbessern, werden in diesem Kapitel verschiedene Adaptionsverfahren untersucht. Es wird unterschieden zwischen Verfahren, die die Testdaten zu normaler Sprache hin transformieren (Abschnitt 9.1), und zwischen Verfahren, welche das Hintergrundmodell und die Sprechermodelle an erhöhten Stimmaufwand anpassen (Abschnitt 9.2).

Zur Überprüfung der Tauglichkeit beider Arten der Adaption werden für die nachfolgenden Tests unterschiedliche Merkmale aus Kapitel 8 gezielt ausgewählt. Die Adaption wird mit MFCC-Merkmalen inklusive Telefonfilter durchgeführt, da diese im Basissystem verwendet werden (siehe Abschnitt 8.3). Die in Abschnitt 8.4 dargestellten Merkmale $\log F_0$ und $\log F_0/E$ werden ebenfalls für die Adaption verwendet. Die F_0-Statistik ist nicht zur Adaption geeignet, da nur ein Merkmalsvektor pro Sprachsignal erstellt wird und für die Adaption nur Merkmale auf Frame-Basis Berücksichtigung finden. Aus diesem Grund ist auch die Kombination der Momente (siehe Abschnitt 8.5) nicht für die Adaption geeignet. Stattdessen wird das COG-Verhältnis ohne VAD (siehe Abschnitt 8.5) mit den unterschiedlichen Adaptionsverfahren getestet.

9.1 Adaption der Testdaten

Das in diesem Abschnitt beschriebene Adaptionsverfahren zur Kompensation des Lombardeffekts ist von Goldenberg, Cohen und Shallom (2006) entwickelt worden. Goldenberg et al. berechnen eine Transformationsmatrix, welche die Sprachsignale erhöhten Stimmaufwands in solche normalen Stimmaufwands transformieren soll. Die kompensierten Merkmale ergeben sich aus:

$$Merkmal_k(i) = \frac{Merkmal_L - \mu_T(i)}{\sigma_T(i)} \tag{9.1}$$

mit i=1,...,p, wobei p die Größe des Merkmalsvektors angibt. $Merkmal_k$ steht hierbei für die kompensierten Merkmale, während $Merkmal_L$ die Lombardsprache beschreibt. Die Werte für die Transformation μ_T und σ_T werden wie folgt berechnet:

$$\mu_T(i) = \mu_k(i) - \mu_L(i) \tag{9.2}$$

	MFCC(TF)	$\log F_0$	$\log F_0/E$	COG-Verhältnis ohne VAD
Basis (ohne Adaption)	6,04%	46,02%	37%	37%
Goldenberg-Basis	6%	52%	51%	37,18%
UBM+ (ohne Adaption)	6,43%	50%	37%	26%
Goldenberg-UBM+	6%	46%	47%	25,04%

Tabelle 9.1: Vergleich der EER für die Adaption unterschiedlicher Merkmale und Einstellungen (Basis steht für die oben beschriebenen Standardeinstellungen des Merkmals; UBM+ steht für das mit dem OLLO-Korpus erweiterte UBM; TF steht für Telefonfilter.)

$$\sigma_T(i) = \frac{\sigma_k(i)}{\sigma_L(i)} \tag{9.3}$$

$\mu_k(i)$ und $\sigma_k(i)$ geben den Erwartungswert und die Standardabweichung der Adaptionsdaten normalen Stimmaufwands an, während $\mu_L(i)$ und $\sigma_L(i)$ die statistischen Eigenschaften der Adaptionsdaten erhöhten Stimmaufwands repräsentieren.

Die Adaption nach Goldenberg et al. (2006) wird auf den Testdaten durchgeführt. Dies bedeutet, dass die Merkmale der Testdaten zu normaler Sprache hin normalisiert werden. Die Ergebnisse dieser Transformation für die unterschiedlichen Merkmale sind in Tabelle 9.1 dargestellt. Die Adaption wurde sowohl auf dem Basiskorpus als auch auf dem Korpus mit erweitertem UBM durchgeführt. Zur detaillierten Ergebnisanalyse wurde jeweils ein DET-Graph für die Basistests (siehe Abbildung 9.1) und ein DET-Graph für die Tests mit erweitertem UBM (siehe Abbildung 9.2) erstellt. Die DET-Kurven, wie auch Tabelle 9.1, präsentieren die Ergebnisse der Tests mit Adaption vergleichend zu den Ergebnissen ohne Adaption aus Kapitel 8.

Bei der Analyse der EER für die MFCC-Merkmale (siehe Tabelle 9.1) zeigt sich eine kleine Verbesserung durch die Adaption der Testdaten für den Basistest und den Test mit erweitertem UBM. Für das System $\log F_0/E$ werden die Ergebnisse für beide Tests schlechter. Die Leistung des Systems $\log F_0$ sinkt beim Basistest ebenfalls, während sie bei dem Test mit dem erweiterten UBM leicht verbessert wird. Für das COG-Verhältnis fallen die Ergebnisse andersherum aus; es ist eine kleine Verschlechterung bei den Basistests zu beobachten und eine etwas größere Verbesserung bei den Tests mit erweitertem UBM.

Insgesamt bleibt die Rangordnung der Systemleistung erhalten. Die MFCC-Merkmale liefern die besten Ergebnisse, gefolgt von dem COG-Verhältnis. Die F_0-basierten Merkmale führen zu der schlechtesten Systemleistung.

Die separate Betrachtung der Ergebnisse der Basistests in dem DET-Graph 9.1 bestätigt die oben beschriebenen Ergebnisse. Die MFCC-Merkmale werden durch die Adaption leicht verbessert, jedoch ist der Unterschied zwischen den Kurven minimal. In der Region um die EER herum wechseln sich die Kurven der beiden Systemrealisierungen mit MFCC-Merkmalen hinsichtlich der Rangfolge der Leistung ab. Treten für die falschen Alarme Werte kleiner eins auf, so ist nahezu durchgängig die Systemrealisierung mit Adaption (grüne Kurve) besser als die ohne Adaption (schwarze Kurve). Da die MFCC-Merkmale in diesen Tests die beste Leistung erbringen, sind die Kurven links unten, also in der Nähe des Urspungs positioniert. Die Kurven der als Nächstes dargestellten $\log F_0$-Merkmale liegen im Gegensatz dazu rechts oben und zählen mit zu den schlechtesten Systemen. Hier zeigt sich,

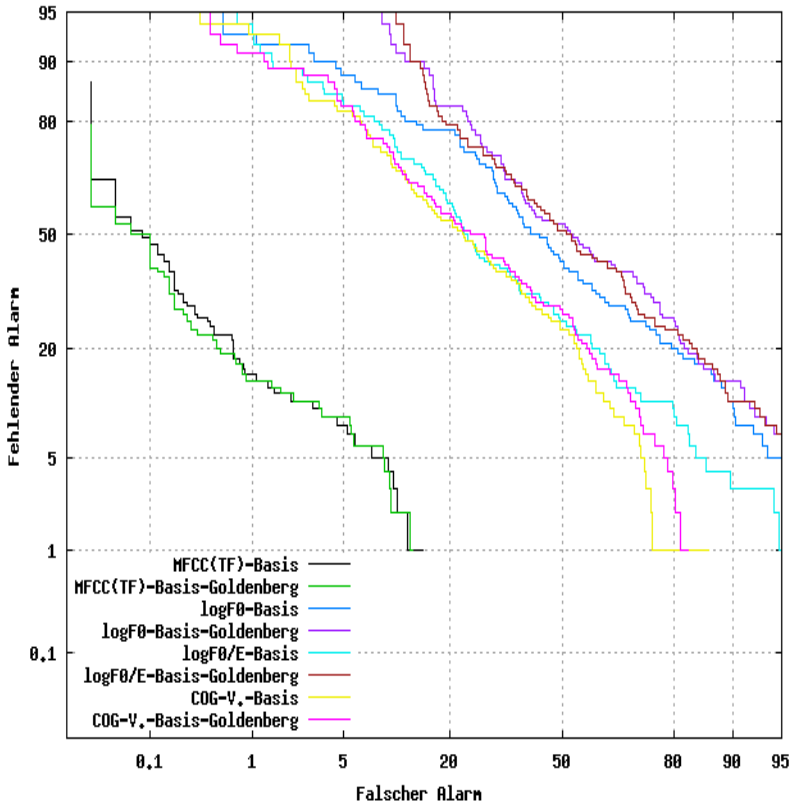

Abbildung 9.1: DET-Kurven zum Vergleich der Adaption nach Goldenberg für unterschiedliche Merkmale mit der Systemleistung der gleichen Merkmale ohne Adaption für die Basistests

wie bei den logF_0/E-Merkmalen, dass die Leistung durch die Adaption sinkt. Die Leistung der Systemrealisierung mit dem COG-Verhältnis als Merkmal bleibt relativ konstant. Es ist hingegen für die Systemrealisierung mit Adaption nicht nur die EER etwas schlechter (siehe Tabelle 9.1), sondern auch die Leistung insgesamt.

Die Kurven der Tests mit erweitertem UBM (siehe Abbildung 9.2) bestätigen die oben beschriebenen Ergebnisse aus Tabelle 9.1. Interessante Zusatzinformation kann besonders für die Kurven der Tests mit dem COG-Verhältnis gewonnen werden. Für eine Systemleistung mit hohen Raten für falsche Alarme und niedriger Anzahl fehlender Alarme ist die Realisierung ohne Adaption besser. Bei wenigen falschen Alarmen und mehr fehlenden Alarmen ist die adaptierte Version beinahe durchweg überlegen.

Nachdem teilweise Verbesserungen der Einzelsysteme gegenüber den nicht-adaptierten Systemen erzielt werden konnten, soll nun überprüft werden, ob eine Fusion einzelner adaptierter Systemvariationen miteinander zu einer weiteren Verbesserung der Gesamtsystemleistung führt. Zur linearen Fusion wurde, wie in Abschnitt 8.4, das FoCal Toolkit (Brummer & du Preez, 2006) verwendet. Die Ergebnisse der Fusionen sind in Tabelle 9.2 abgebildet. Die ersten Systemkombinationen enthalten die Punktzahlen der MFCC-Merkmale mit jeweils einem anderen Merkmal. Auch hier wurden bei sämtlichen Kombinationen sowohl die Basistests als auch die Tests mit erweitertem UBM berücksichtigt. Die MFCC-Merkmale wurden

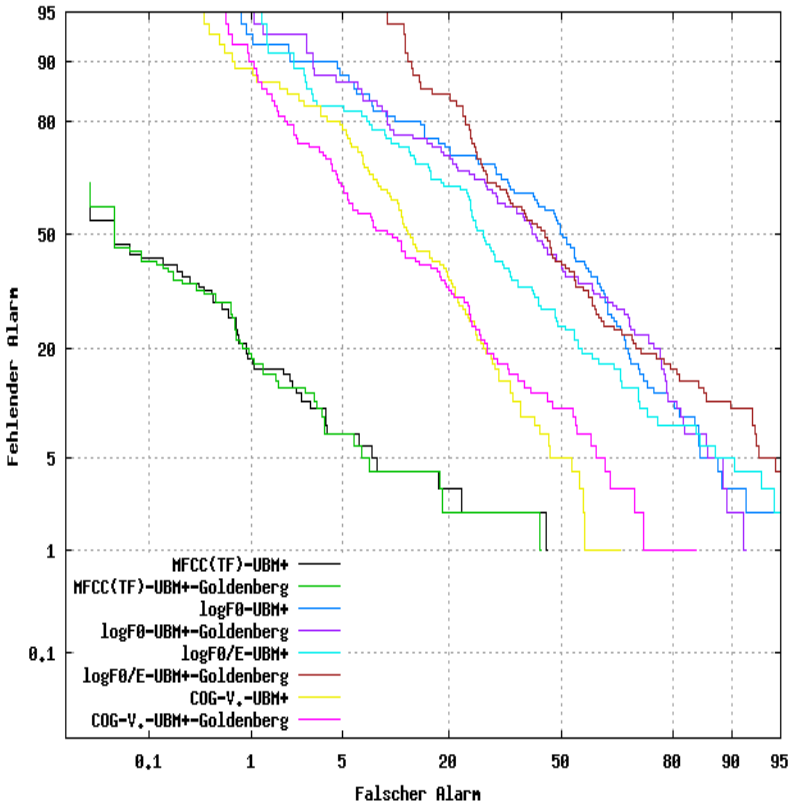

Abbildung 9.2: DET-Kurven zum Vergleich der Adaption nach Goldenberg für unterschiedliche Merkmale mit der Systemleistung der gleichen Merkmale ohne Adaption für das Korpus mit erweitertem UBM

mit jedem anderen Merkmal kombiniert, da sie die besten Ergebnisse liefern und deswegen eine gute Ausgangsbasis für eine Fusion bieten. Als Nächstes wurden die Ergebnisse des COG-Verhältnisses mit denen von jeweils einem F_0-basierten Merkmal kombiniert. Die F_0-basierten Merkmale wurden nicht miteinander kombiniert, da keine komplementäre Information in den Merkmalen vorhanden ist. Zuletzt wurden immer drei Merkmale miteinander kombiniert, wobei, im Wechsel, immer ein F_0-basiertes Merkmal ausgelassen wurde.

Insgesamt führen die Systemkombinationen, die die MFCC-Merkmale enthalten, zu den besten Ergebnissen. Die Kombinationen mit den logF_0-Merkmalen schneiden tendenziell schlechter ab als Kombinationen mit anderen Merkmalen, besonders bei den Tests UBM+. Dies kann bedingt sein durch das Training der Fusion. Möglicherweise müssen für den Trainingsprozess des logF_0-Systems mehr beziehungsweise besser angepasste Daten verwendet werden. Die niedrigste EER sämtlicher Fusionen liegt bei 6,02% und ist damit nicht niedriger als die der adaptierten Systemversion der MFCC-Merkmale (6%). Eine Fusion der adaptierten Systemvariationen führt dementsprechend nicht zu einer Verbesserung. Aus diesem Grund wurden zu den Fusionen keine DET-Kurven erstellt.

Abschließend kann festgehalten werden, dass die **Adaption nach Goldenberg für einige Merkmale, unter bestimmten Testbedingungen, zu Verbesserungen führt**.

Fusionierte Systeme	EER
MFCC(TF)+logF_0	11%
MFCC(TF)+logF_0/E	7%
MFCC(TF)+COG-Verhältnis ohne VAD	6,02%
MFCC(TF)+logF_0 (UBM+)	33%
MFCC(TF)+logF_0/E (UBM+)	7%
MFCC(TF)+COG-Verhältnis ohne VAD (UBM+)	7%
COG-Verhältnis+logF_0	50%
COG-Verhältnis+logF_0/E	42%
COG-Verhältnis+logF_0 (UBM+)	45,16%
COG-Verhältnis+logF_0/E (UBM+)	38%
COG-Verhältnis-MFCC(TF)+logF_0	11%
COG-Verhältnis-MFCC(TF)+logF_0/E	6,1%
COG-Verhältnis-MFCC(TF)+logF_0 (UBM+)	31%
COG-Verhältnis-MFCC(TF)+logF_0/E (UBM+)	9%

Tabelle 9.2: Fusionierung einzelner Ergebnisse der adaptierten Systemvariationen (Die mit UBM+ markierten Tests nutzen das mit dem OLLO-Korpus erweiterte UBM, alle anderen Tests sind Basistests.)

Doch wird die niedrigste EER des MFCC-Basissystems ohne Adaption nicht, beziehungsweise nur geringfügig unterschritten.

9.2 Adaption der Modelle

Nachdem im vorherigen Abschnitt die Transformation der Testdaten beschrieben wurde, stellt dieser Abschnitt zwei Formen der Adaption auf Modellebene vor. Bei beiden Formen wird der Unterschied zwischen dem UBM und den Adaptionsdaten berechnet. Davon ausgehend wird entweder eine Transformationsmatrix erstellt oder ein angepasstes UBM berechnet. In beiden Fällen hat die Adaption nicht nur Einfluss auf das UBM, sondern auch auf die Sprechermodelle, da das UBM als Ausgangspunkt für diese genutzt beziehungsweise die Transformationsmatrix für sämtliche Modelle verwendet wird.

Das erste Adaptionsverfahren ist die *MAP-Adaption* (siehe Abschnitt 3.2.3). Sie wird nach der Erstellung des UBMs durchgeführt. Dadurch wird das UBM, bevor es auf die einzelnen Referenzsprecher adaptiert wird, an laute Sprache angepasst. Für die MAP-Adaption werden, im Gegensatz zur Adaption der Testdaten in Abschnitt 9.1, als Adaptionsdaten nur die Daten erhöhten Stimmaufwands des „Pool 2010“-Korpus (siehe Abschnitt 5.2.1) verwendet. Genau wie bei der Adaption der Referenzsprecher werden nur die Erwartungswerte adaptiert.

Das zweite Adaptionsverfahren ist die *MLLR-Adaption (maximum likelihood linear regression)*. Bei der MLLR-Adaption handelt es sich um eine lineare Transformation. Die Berechnung der Transformationsmatrix erfolgt so, dass die resultierende Verteilung mit den adaptierten Parametern die Adaptionsdaten besser abbildet. Hierfür wird die Likelihood für das adaptierte Modell, gegeben die Adaptionsdaten, maximiert. Es wird also nach den optimalen Parametern für das zu berechnende adaptierte Modell gesucht. Die Maximierung erfolgt mit Hilfe des EM-Algorithmus

	MFCC(TF)	$\log F_0$	$\log F_0/E$	COG-Verhältnis ohne VAD
Basis (ohne Adaption)	6,04%	46,02%	37%	37%
MAP-Basis	4,29%	43,33%	37,04%	30%
MLLR-Basis	3,45%	48%	39,24%	34%
UBM+ (ohne Adaption)	6,43%	50%	37%	26%
MAP-UBM+	5%	49,29%	37%	20%
MLLR-UBM+	7,37%	36,04%	42%	25%

Tabelle 9.3: Vergleich der EER für die Adaption unterschiedlicher Merkmale und Einstellungen (Basis steht für die oben beschriebenen Standardeinstellungen des Merkmals; UBM+ steht für das mit dem OLLO-Korpus erweiterte UBM; TF steht für Telefonfilter.)

(siehe Abschnitt 3.2.3). Eine ausführliche Beschreibung der MLLR-Adaption ist in Legetter und Woodland (1995) zu finden.

Die Ergebnisse der Adaption der Modelle ist in Tabelle 9.3 und den Abbildungen 9.3 bis 9.6 dargestellt. Bei der Betrachtung der EER der Basistests für die MAP-Adaption (siehe Tabelle 9.3) wird eine Verbesserung für die meisten Merkmale beobachtet. Lediglich die $\log F_0/E$-Merkmale zeigen eine minimale Verschlechterung. Die Ergebnisse der UBM+-Tests zeigen die gleichen Tendenzen. Auch hier sind sämtliche Leistungen gestiegen, mit Ausnahme der Leistung der Merkmale $\log F_0/E$, welche unverändert ist. Für die MLLR-Transformation der MFCC-Merkmale auf dem Basistestset wird die insgesamt beste Systemleistung sämtlicher Experimente dieser Arbeit bei nicht-übereinstimmenden Stimmaufwand beobachtet. Mit 3,45% EER konnte die EER des MFCC-Basissystems beinahe halbiert werden. Weiterhin ist die Verschlechterung zu den in Abschnitt 8.3 beschriebenen Vergleichstest, in denen normale gegen normale Sprache getestet wurde, stark verringert worden. Die EER des Vergleichstests lag bei 2% (siehe Tabelle 8.1). Der negative Einfluss des erhöhten Stimmaufwands auf die automatische Sprechererkennung konnte dementsprechend durch die MLLR-Adaption mit zusätzlichen Daten auf den MFCC-Merkmalen stark eingeschränkt werden. Eine Verbesserung durch die MLLR-Adaption konnte auch für das COG-Verhältnis erzielt werden. Die Ergebnisse der F_0-basierten Merkmale wurden hingegen verschlechtert. Bei den Tests mit erweitertem UBM wurde im Gegensatz zu den Basistests die Leistung der MFCC-Merkmale verschlechtert; ebenso wie die Leistung der $\log F_0/E$-Merkmale. Das schlechte Ergebnis der $\log F_0$-Merkmale ohne Adaption konnte durch die MLLR-Adaption stark verbessert werden. Auch das Ergebnis des COG-Verhältnisses wird leicht verbessert. Für die MLLR-Adaption ist folglich ein weniger eindeutiges Muster wie für die MAP-Adaption ersichtlich.

Um die Ergebnisse besser interpretieren zu können, wurden DET-Kurven der einzelnen Tests und Adaptionsarten erstellt. Die Ergebnisse der MAP-Adaption auf dem Basistestset sind in Abbildung 9.3 dargestellt. Die MFCC-Merkmale zeigen für sämtliche Operationspunkte der Kurve eine Verbesserung oder zumindest gleichbleibende Leistung für die adaptierte Systemversion. Dies gilt auch für die meisten Operationspunkte der Kurve des COG-Verhältnis. Bei den F_0-basierten Merkmalen ist ein Wechsel der Leistungsrangfolge für adaptierte und nicht-adaptierte Systeme zu beobachten. Insgesamt ist das $\log F_0$-System, welches adaptiert wurde, tendenziell besser als die nicht-adaptierte Version. Für das $\log F_0/E$-System verhält es sich andersherum. Dies stimmt mit den Beobachtungen der EER aus Tabelle 9.3 überein.

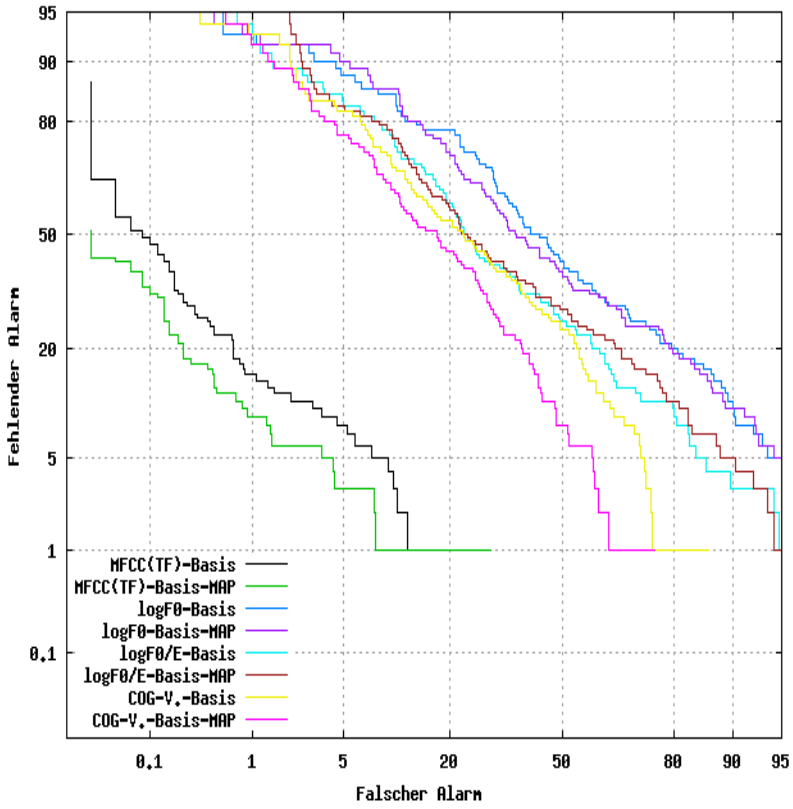

Abbildung 9.3: DET-Kurven zum Vergleich der MAP-Adaption für unterschiedliche Merkmale mit der Systemleistung der gleichen Merkmale ohne Adaption für die Basistests

Bei der MLLR-Adaption sind die Unterschiede zwischen der adaptierten und der nicht-adaptierten Versionen kleiner als bei der MAP-Adaption (siehe Abbildung 9.4). Es zeigen sich jedoch grundsätzlich ähnliche Muster. Für die MFCC-basierten Systeme sowie für die Systeme auf Basis des COG-Verhältnisses sind die adaptierten Systeme jeweils besser als das nicht-adaptierte Pendant. Die Leistung des Merkmals $\log F_0/E$ wird durch die MLLR-Adaption verschlechtert. Für das Merkmal $\log F_0$ ist, ähnlich wie bei der MAP-Adaption, ein ständiger Wechsel in der Rangfolge der Leistung zu beobachten. Auch hier schneidet tendenziell an den meisten Operationspunkten die adaptierte Systemrealisierung besser ab. Dies steht im Gegensatz zu unseren Erkenntnissen aus der Analyse der EER aus Tabelle 9.3. Hier ist eine bessere Leistung der nicht-adaptierten Version ersichtlich. Dies gilt jedoch in der DET-Kurve nur für den Bereich um die EER herum und nicht für die anderen umliegenden Operationspunkte.

Bei der Analyse der Tests mit erweitertem UBM für die MAP-Adaption (siehe Abbildung 9.5) zeigt sich erneut für die MFCC-Merkmale und das COG-Verhältnis, dass die Adaption eine große Verbesserung der Systemleistung mit sich bringt. Für die F_0-basierten Merkmale besteht kein großer Unterschied in der Systemleistung der adaptierten und nicht-adaptierten Systemrealisierungen. Es ist hingegen interessant, dass für die Kurven des Merkmals $\log F_0$ bei niedrigerer Anzahl fehlender Alarme die adaptierte Version um einiges besser abschneidet, während bei niedrigerer

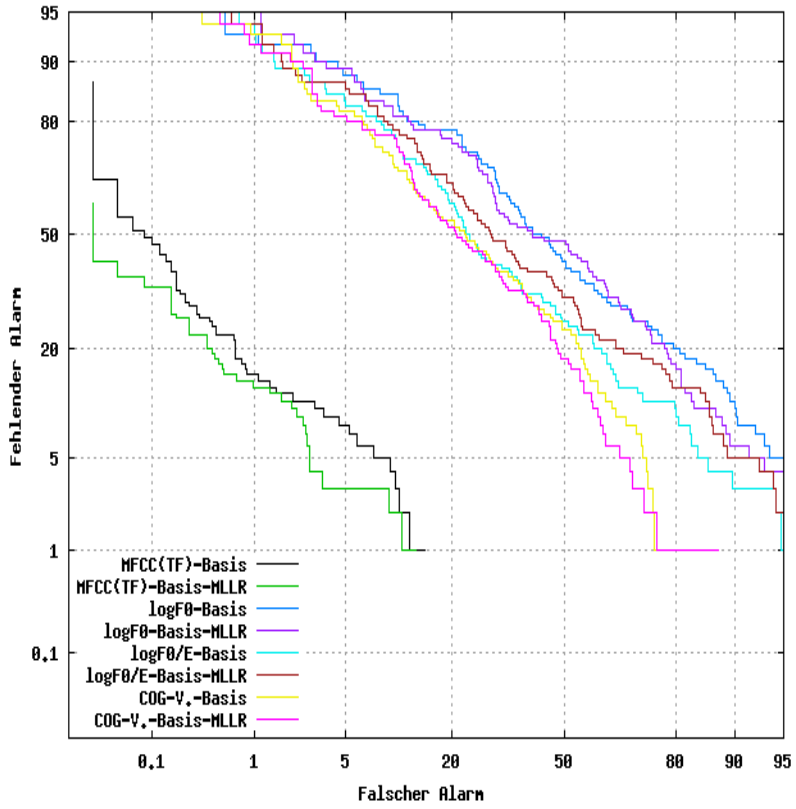

Abbildung 9.4: DET-Kurven zum Vergleich der MLLR-Adaption für unterschiedliche Merkmale mit der Systemleistung der gleichen Merkmale ohne Adaption für die Basistests

Anzahl falscher Alarme die nicht-adaptierte Version eindeutig bessere Ergebnisse erzielt.

Die Experimente mit der MLLR-Adaption auf dem Datenset mit erweitertem UBM zeigen ein anderes Muster als die bisher beschriebenen Tests (siehe Abbildung 9.6). Die MLLR-Adaption führt auf diesem Datenset für die MFCC-Merkmale zu einer Verschlechterung über die meisten Operationspunkte des DET-Graphen. Das COG-Verhältnis zeigt zwar, wie bei dem Vergleich der EER-Werte (siehe Tabelle 9.3), eine Verbesserung durch die Adaption; diese ist jedoch minimal. Für das Merkmal $\log F_0$ wird über die gesamte Kurve eine starke Verbesserung erzielt, während die Leistung der $\log F_0$/E-Merkmale sinkt. Die DET-Kurven der MLLR-Adaption bestätigen damit eindeutig, dass für die MLLR-Adaption kein eindeutiger Trend ersichtlich ist.

Nachdem die einzelnen Merkmale im Kontext der Adaption der Modelle untersucht wurden, werden nun unterschiedliche Systemfusionen vorgestellt. Die Auswahl der fusionierten Systeme erfolgt wie im vorigen Abschnitt (9.1). Zur linearen Fusion wurde, wie in Abschnitt 8.4, das FoCal Toolkit (Brummer & du Preez, 2006) verwendet. Die Ergebnisse der Fusionen sind in Tabelle 9.4 dargestellt. Die Ergebnisse der MAP-Adaption zeigen, dass für vier Merkmalskombinationen eine EER kleiner als die kleinste EER der nicht-fusionierten Systeme aus Tabelle 9.3 erzielt wird (4,29%). Diese vier Merkmalskombinationen enthalten stets die MFCC-Merkmale und nutzen nicht die erweiterten UBM-Daten. Zwei der Kom-

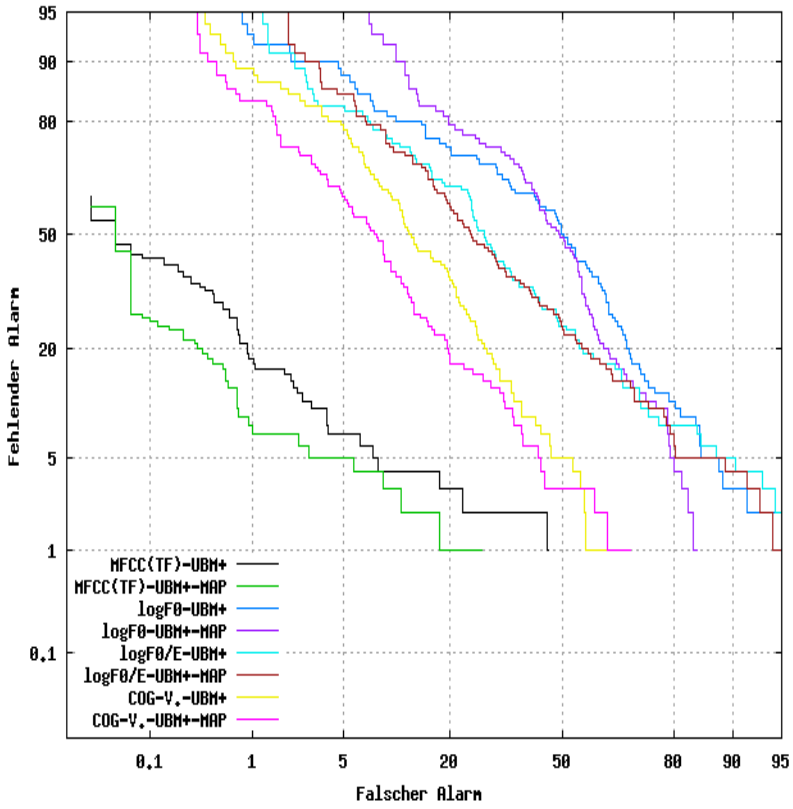

Abbildung 9.5: DET-Kurven zum Vergleich der MAP-Adaption für unterschiedliche Merkmale mit der Systemleistung der gleichen Merkmale ohne Adaption für das Korpus mit erweitertem UBM

binationen sind Zweierkombinationen: eine Kombination der MFCC-Merkmale mit den logF_0-Merkmalen und eine mit dem COG-Verhältnis. Die anderen zwei Systemkombinationen enthalten jeweils das COG-Verhältnis, die MFCC-Merkmale und ein F_0-basiertes Merkmal. Weiterhin ist auffällig, dass die Merkmale logF_0 bei erweitertem UBM immer zu schlechten Ergebnissen der Systemkombination führen. Dies ist auch bei der MLLR-Adaption in ähnlicher Form zu sehen und stimmt mit den Beobachtungen bei der Adaption der Testdaten aus Abschnitt 9.1 überein.

Bei der MLLR-Adaption schneidet die Fusion der MFCC-Merkmale mit dem COG-Verhältnis am besten ab. Das Ergebnis liegt jedoch oberhalb der 3,45% EER des MLLR-adaptierten MFCC-Systems ohne Fusion (siehe Tabelle 9.3). Insgesamt ist diese Merkmalskombination besser als die beste Systemkombination der MAP-Adaption.

Abschließend kann festgehalten werden, dass **die Adaption der Systeme auf Modellebene in den meisten Szenarien zu einer Verbesserung führt**. Besonders für die MAP-Adaption wird zuverlässig eine Verbesserung oder gleichbleibende Leistung beobachtet. Die **MLLR-Adaption erzielt das beste Ergebnis** sämtlicher in dieser Arbeit beschriebenen Tests. Allerdings führt die MLLR-Adaption für einige Szenarien auch zu Verschlechterungen, sodass der Einsatz dieser Art der Adaption nicht immer geeignet ist. Die Systemfusion führt für die MAP-Adaption teilweise

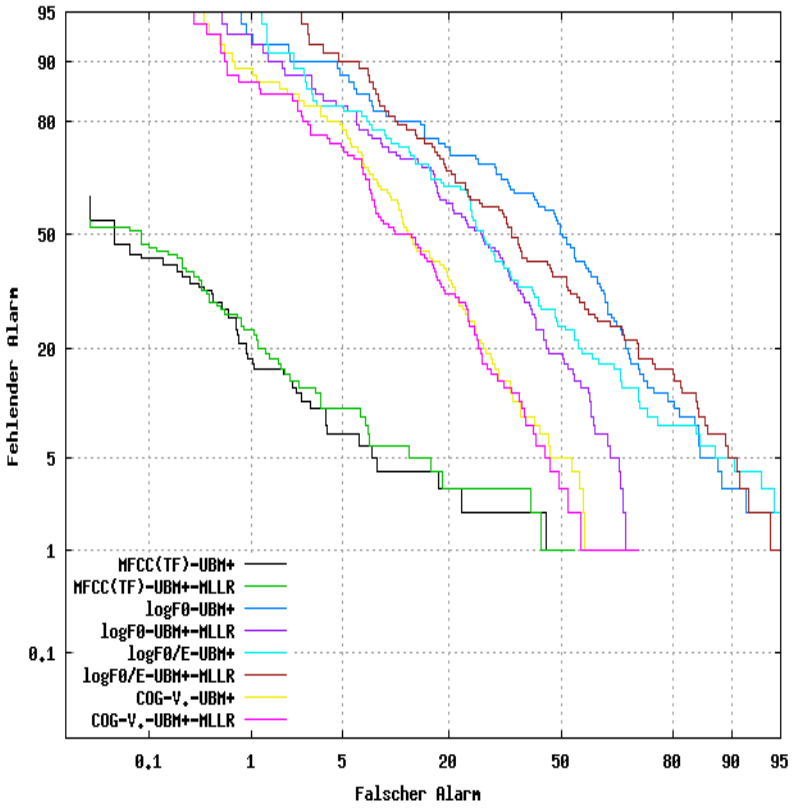

Abbildung 9.6: DET-Kurven zum Vergleich der MLLR-Adaption für unterschiedliche Merkmale mit der Systemleistung der gleichen Merkmale ohne Adaption für das Korpus mit erweitertem UBM

zu weiteren Verbesserungen. Die Leistung liegt jedoch weiterhin unter der der MLLR-Adaption für die MFCC-Merkmale. Für die MLLR-Adaption kann keine weitere Verbesserung durch eine Systemfusion beobachtet werden.

Fusionierte Systeme	EER MAP	EER MLLR
MFCC(TF)+logF_0	4%	9%
MFCC(TF)+logF_0/E	5%	4%
MFCC(TF)+COG-Verhältnis ohne VAD	4,02%	3,47%
MFCC(TF)+logF_0 (UBM+)	42%	16%
MFCC(TF)+logF_0/E (UBM+)	5%	9%
MFCC(TF)+COG-Verhältnis ohne VAD (UBM+)	5%	9%
COG-Verhältnis+logF_0	38%	44%
COG-Verhältnis+logF_0E	30%	34,18%
COG-Verhältnis+logF_0 (UBM+)	49,14%	36%
COG-Verhältnis+logF_0E (UBM+)	28%	39%
COG-Verhältnis-MFCC(TF)+logF_0	4,16%	7,02%
COG-Verhältnis-MFCC(TF)+logF_0/E	4,04%	5,02%
COG-Verhältnis-MFCC(TF)+logF_0 (UBM+)	41,27%	15%
COG-Verhältnis-MFCC(TF)+logF_0/E (UBM+)	5%	10%

Tabelle 9.4: Fusionierung einzelner Ergebnisse adaptierter Systeme (Die mit UBM+ markierten Tests nutzen das mit dem OLLO-Korpus erweiterte UBM, alle anderen Tests sind Basistests.)

Kapitel 10

Schlussfolgerungen und Ausblick

Das Ziel dieser Arbeit war die Analyse der Auswirkung erhöhten Stimmaufwands auf Sprache und sprachverarbeitende Systeme. Als Beispielszenario wurde die automatische Sprechererkennung bei ungleichem Stimmaufwand in Trainings- und Testdaten ausgewählt. Konkret wurden folgende Aufgabenstellungen formuliert:

- Untersuchungen akustischer Veränderungen bei erhöhtem Stimmaufwand
  - * Statistische Analyse spektraler Veränderungen bei Erhöhung des Stimmaufwands
  - * Evaluation spektraler Parameter zur automatischen Klassifikation des Stimmaufwands
  - * Zusammenhänge zwischen den spektralen Parametern und F_0
- Realisierung eines Sprecherverifikationssystems für nicht-übereinstimmenden Stimmaufwand
  - * Erstellen eines Frameworks zur automatischen Sprechererkennung sowie Evaluation des Sprechererkennungssystems hinsichtlich des Einflusses erhöhten Stimmaufwands
  - * Entwicklung von Strategien zur Verbesserung der Erkennungsraten des Sprechererkennungssytems für das gegebene Szenario

Die nachfolgenden Abschnitte fassen die Vorgehensweise und die erzielten Resultate zusammen. Abschließend werden mögliche Fortführungen dieser Arbeit dargestellt.

10.1 Untersuchungen akustischer Veränderungen bei erhöhtem Stimmaufwand

Statistische Analyse spektraler Veränderungen bei Erhöhung des Stimmaufwands Die statistische Analyse der akustischen Merkmale umfasste Untersuchungen verschiedener *spektraler Merkmale* (siehe Kapitel 6). Einbezogen in die Untersuchung wurden die Merkmale *spektrale Neigung, gewichteter spektraler Schwerpunkt, Energieverhältnis* sowie die *spektralen Momente*. Die Analyse der *Gesamtdaten* über alle Sprecher und Laute zeigt, dass für sämtliche spektralen Merkmale signifikante Unterschiede zwischen normaler und lauter Sprache bestehen. Dies bestätigt

die Erkenntnisse anderer Studien, die eine Veränderung der spektralen Energie bei erhöhtem Stimmaufwand beschreiben. Durch die Untersuchung der Momente der Verteilungen der spektralen Parameter und die Untersuchung der Z-Werte der Signifikanztests in der vorliegenden Arbeit lassen sich bestimmte Merkmale als stärker oder weniger stark beeinflusst festlegen. Ein solcher Vergleich erfolgte bisher in keiner anderen Arbeit.

Die *Analyse der Verteilungen* zeigt, dass die spektrale Neigung sowie das dritte und vierte Moment anscheinend stärker durch veränderten Stimmaufwand beeinflusst sind, während sich das Energieverhältnis als weniger sensitiv erweist. Die *Z-Werte* zeigen besonders große Modifikationen für den spektralen Schwerpunkt und das erste spektrale Moment. Das Energieverhältnis hingegen, ebenso wie das zweite spektrale Moment sind auch hier weniger verändert.

Der Vergleich der *Lautklassen* zeigt für sämtliche spektralen Merkmale, dass für alle Lautklassen ein signifikanter Unterschied zwischen normaler und lauter Sprache besteht. Für Obstruenten ist dieser jedoch geringer im Vergleich zu den anderen Lautklassen. Der größte Unterschied ist bei den Vokalen zu beobachten.

Die *sprecherspezifische Analyse* der spektralen Parameter ergibt unklare Muster über die unterschiedlichen Sprecher. Dies lässt den Rückschluss zu, dass die spektralen Veränderungen stark sprecherabhängig sind. Die Parameter, die bei erhöhtem Stimmaufwand nur eine geringfügige Veränderung erfahren und gleichzeitig über alle Sprecher wenig variieren, sind der Maximalwert und die mittlere Abweichung des ersten spektralen Moments sowie der Minimalwert des vierten spektralen Moments.

Evaluation spektraler Parameter zur automatischen Klassifikation des Stimmaufwands Nachdem signifikante Unterschiede für die spektralen Parameter festgestellt wurden, sind die verschiedenen spektralen Parameter im Rahmen eines Stimmaufwandsklassifikators getestet worden (siehe Abschnitt 6.3). Der Klassifikator wurde mit jeweils einem GMM von 64 Mischungskomponenten pro Stimmaufwand realisiert und mit den spektralen Merkmalen, MFCC-Merkmalen sowie unterschiedlichen Merkmalskombinationen evaluiert. Ziel der Untersuchung war die erfolgreiche Klassifikation des Stimmaufwands sowie das Finden des hierfür besten Parameters. Ein Stimmaufwandsklassifikator kann einem sprachverarbeitenden System vorgeschaltet werden, um automatisch ein Modell mit ähnlichem Stimmaufwandsgrad auszuwählen und so die Erkennungsrate zu verbessern. In dieser Arbeit wurde das System genutzt, um die Ergebnisse der statistischen Analyse zu verifizieren. Eine bessere Klassifikation bedeutet in diesem Kontext, dass größere Unterschiede zwischen normalem und erhöhtem Stimmaufwand bestehen. Die Evaluation der spektralen Merkmale als Parameter für die Stimmaufwandsklassifikation bestätigt die Analyseergebnisse der Z-Werte der Gesamtverteilung. Die besten Erkennungsraten werden mit dem spektralen Schwerpunkt und dem ersten spektralen Moment erzielt. Diese Parameter zeigen bei der Analyse der Z-Werte besonders große Veränderungen für erhöhten Stimmaufwand. Die schlechtesten Klassifikationsergebnisse erzielt das Energieverhältnis, gefolgt vom zweiten spektralen Moment. Dies entspricht ebenfalls den Resultaten der Z-Wertanalyse. Auch die oben beschriebenen Tendenzen der Analyse der Verteilungen können bestätigt werden, da die spektrale Neigung sowie das dritte und vierte Moment besser zur Klassifikation des Stimmaufwands geeignet sind als das Energieverhältnis.

Die vorliegende Arbeit verbindet die statistische Analyse mit der automatischen Klassifikation des Stimmaufwands und zeigt auf diese Weise geeignete Merkmale zur Stimmaufwandsklassifikation auf.

Zusammenhänge zwischen den spektralen Parametern und F_0 Nach der Analyse der spektralen Veränderungen wurden die *Modifikationen von F_0* bei erhöhtem Stimmaufwand untersucht (siehe Kapitel 7). Die statistische Analyse zeigt signifikante Unterschiede zwischen normaler und lauter Sprache für die *Gesamtdaten*. Die Analyse der *Lautklassen* für F_0 zeigt, wie bei den spektralen Merkmalen, den größten Unterschied für Vokale und den kleinsten für Obstruenten. Im Gegensatz zu den spektralen Merkmalen sind bei der *sprecherspezifischen Analyse* hingegen keine Muster mit großen Unterschieden für die verschiedenen Sprecher sichtbar. Die Tendenzen sind ähnlich denen der Gesamtverteilung für die 19 untersuchten Sprecher. Die allgemeinen Tendenzen sind folglich sprecherübergreifend ähnlich; ob die Veränderungen allgemeingültig anstatt sprecherspezifisch sind, muss jedoch in detaillierteren Analysen erforscht werden, da andere Studien (Jessen et al., 2005) sprecherspezifische Unterschiede für den Einfluss erhöhten Stimmaufwands auf F_0 nachgewiesen haben.

Als Nächstes wurde für jeden spektralen Parameter sowie F_0 getestet, ob *normaler und erhöhter Stimmaufwand korreliert* sind. Die Korrelationsanalysen zeigen vor allem bei der Gesamtverteilung und den Vokalen hohe Korrelationskoeffizienten. Die Obstruenten und Sonoranten weisen ebenfalls signifikante Zusammenhänge auf. Diese sind jedoch nicht so groß wie die der Vokale und der Gesamtverteilung. Die spektralen Merkmale zeigen größere Zusammenhänge als F_0. Der stärkste Zusammenhang zwischen normaler und lauter Sprache besteht für die Vokale des dritten und vierten spektralen Moments.

Um festzustellen, ob es einen *Zusammenhang zwischen den spektralen Veränderungen und F_0* gibt, wurden weitere Korrelationsanalysen durchgeführt. Die Tests ergeben keine nennenswerten Zusammenhänge zwischen den spektralen Parametern und F_0. Eine vergleichende Untersuchung dieser Art ist bisher nicht in der Literatur zu finden. Die Entwicklung eines Merkmals für die Sprechererkennung, welches sich auf den Zusammenhang zwischen F_0 und einem der spektralen Parameter bezieht, scheint auf Grund der Analyseergebnisse nicht sinnvoll zu sein.

10.2 Realisierung eines Sprecherverifikationssystems für nicht-übereinstimmenden Stimmaufwand

Erstellen eines Frameworks zur automatischen Sprechererkennung sowie Evaluation des Sprechererkennungssystems hinsichtlich des Einflusses erhöhten Stimmaufwands Nachdem die akustischen Unterschiede zwischen normalem und erhöhtem Stimmaufwand untersucht wurden, folgte eine Analyse der Auswirkungen erhöhten Stimmaufwands auf die automatische Sprechererkennung (siehe Kapitel 8). Hierfür wurde ein *Framework zur automatischen Sprecherverifikation* erstellt, welches in Form eines GMM-UBM-Systems realisiert wurde. Anschließend wurden Tests mit normalem Stimmaufwand in Trainings- und Testdaten durchgeführt sowie Tests mit normalem Stimmaufwand in den Trainingsdaten und erhöhtem

Stimmaufwand im Testkorpus. Die Tests zeigen starke Verschlechterungen der Erkennungsrate. Die EER ist bis zu 5-Mal größer.

Entwicklung von Strategien zur Verbesserung der Erkennungsraten des Sprechererkennungssytems für das gegebene Szenario Um zu prüfen, ob es sprecherspezifische Merkmale gibt, die weniger stark durch eine Veränderung des Stimmaufwands beeinflusst werden, wurde das Sprechererkennungssystem mit unterschiedlichen *Standardmerkmalen* und *F_0-basierten Merkmalen* evaluiert.

Die Analyse der *Standardmerkmale* zeigt, dass die MFCC-Merkmale inklusive Telefonfilter am besten als Standardmerkmale für das Basissystem geeignet sind.

Die *F_0-basierten* Merkmale wurden mit einbezogen, da F_0 ein essenzieller Parameter der forensischen Sprechererkennung ist. Es zeigt sich, dass die F_0-basierten Merkmale schlechtere Ergebnisse liefern als die Standardmerkmale. Dies war zu erwarten, da F_0-basierte Merkmale in der Regel nicht allein, sondern nur in Kombination mit anderen Merkmalen angewendet werden. Weiterhin ist bekannt, dass F_0 bei erhöhtem Stimmaufwand verändert wird, sodass eine Verschlechterung der Ergebnisse bei erhöhtem Stimmaufwand auftritt. Die Verschlechterung ist jedoch nicht so groß wie bei den Standardmerkmalen, da die Fehlerrate der F_0-basierten Merkmale insgesamt höher ist und sie deswegen im Verhältnis zur Ausgangsleistung des Systems weniger steigen kann. Die Kombination der F_0-basierten Merkmale mit dem Basissystem ergibt keine Verbesserung im Vergleich zum Basissystem allein.

Daraufhin wurden *selbst entwickelte* beziehungsweise *entdeckte Merkmale* vorgestellt. Die Merkmale wurden auf Grund der statistischen Analysen der vorherigen Kapitel und der Literatur entworfen. Das erste Merkmal ist das *COG-Verhältnis*. Der zweite Merkmalsvektor ist eine *Kombination unterschiedlicher spektraler Parameter* aus Kapitel 6. Die Kombination der spektralen Parameter mit dem Basissystem führt zu einer geringfügigen Verschlechterung. Das COG-Verhältnis hingegen verbessert die Leistung des Basissystem, wenn es mit Selbigem kombiniert wird.

Nach der Evaluation unterschiedlicher Merkmale wurden verschiedene *Adaptionsverfahren* auf den besten Merkmalen der vorherigen Tests geprüft (siehe Kapitel 9). Bei den ausgewählten Merkmalen handelt es sich um die MFCC-Merkmale inklusive Telefonfilter, logF_0, logF_0/E und das COG-Verhältnis ohne VAD. Bei der Adaption wurde unterschieden zwischen der *Adaption der Testdaten* und der *Adaption der Modelle*. Für die *Adaption der Testdaten* wurde die *Transformation nach Goldenberg et al. (2006)* realisiert. Durch diese Adaption werden teilweise Verbesserungen, teilweise jedoch auch Verschlechterungen gegenüber den nicht-adaptierten Systemen erzielt. Die Adaption der Testdaten nach Goldenberg et al. führt insgesamt für die unterschiedlichen Systeme und Systemkombinationen nicht zu einer nennenswerten Verbesserung gegenüber dem Basissystem.

Für die *Adaption der Modelle* wurden die *MAP-* und die *MLLR-Adaption* realisiert. Bei diesen Adaptionen werden nicht die Testdaten, sondern das UBM und die Sprechermodelle angepasst. Die MAP-Adaption zeigt durchgehend Verbesserungen oder gleichbleibende Leistung. Die MLLR-Adaption hingegen weist, abhängig vom Testszenario, sowohl Verbesserungen als auch Verschlechterungen auf. Das insgesamt beste Ergebnis wird mit der MLLR-Adaption auf MFCC-Merkmalen erzielt. Eine Kombination unterschiedlicher Merkmale mit den MFCC-Merkmalen führt zu keiner weiteren Verbesserung bei der MLLR-Adaption. Für die MAP-Adaption hingegen führen einige Systemkombination, je nach Testszenario, zu weiteren Ver-

Systemname	EER
Vgl	2%
MFCC(TF)	6,04%
MFCC(TF)+COG-Verhältnis ohne VAD	5,37%
MFCC(TF)+MAP	4,29%
MFCC(TF)+MLLR	**3,45**%
MFCC(TF)+logF_0+MAP	4%
MFCC(TF)+logF_0/E+MAP	5%
MFCC(TF)+COG-Verhältnis ohne VAD+MAP	4%
MFCC(TF)+logF_0+COG-Verhältnis ohne VAD+MAP	4,16%
MFCC(TF)+logF_0/E+COG-Verhältnis ohne VAD+MAP	4,04%
MFCC(TF)+logF_0/E+MLLR	4%
MFCC(TF)+COG-Verhältnis ohne VAD+MLLR	3,47%
MFCC(TF)+logF_0/E+COG-Verhältnis ohne VAD+MLLR	5,02%

Tabelle 10.1: Vergleich der besten Ergebnisse aller Tests auf dem Basiskorpus

besserungen. Insgesamt bleibt das Ergebnis des MFCC-Systems, adaptiert mit der MLLR-Adaption, das Beste der Gesamtarbeit.

Zur zusammenfassenden Darstellung listet Tabelle 10.1 die Resultate der besten Systemrealisierungen der vorliegenden Arbeit auf. Der Vergleichstest mit normalen gegen normalen Stimmaufwand erzielt eine gute EER von 2% (Vgl). Bei nicht-übereinstimmendem Stimmaufwand erzielt das Basissystem eine EER von 6,04% (MFCC(TF)). Diese beiden Ausgangstests sind grau hinterlegt. Die weiteren Ergebnisse sind die Resultate der Systeme, mit denen eine nennenswerte Verbesserung erlangt werden konnte. Es zeigt sich, dass die Verschlechterung durch erhöhten Stimmaufwand mit Hilfe der vorgestellten Methoden nicht komplett kompensiert werden kann. Allerdings kann die Verschlechterung der Systemleistung stark reduziert werden. Während die EER für das Basissystem verdreifacht wird, wird die EER für das beste System (MFCC(TF)+MLLR) nur um 75% erhöht. Die EER des Basissystems auf normaler und lauter Sprache kann durch die eingesetzten Verfahren um 25% gesenkt werden. Die hier vorgestellten Methoden können folglich erfolgreich zur Verbesserung von Sprechererkennungssystemen bei nicht-übereinstimmendem Stimmaufwand eingesetzt werden.

10.3 Ausblick

Die vorliegende Arbeit demonstriert die Veränderungen akustischer Parameter bei erhöhtem Stimmaufwand. Weiterhin werden Lösungen zur robusten Sprechererkennung bei nicht-übereinstimmendem Stimmaufwand in Trainings- und Testdaten vorgestellt. Die dargestellten Konzepte liefern Verbesserungen der automatischen Sprechererkennung im gegebenen Szenario. Darüber hinaus sind weitere Verbesserungsansätze denkbar. Diese ergeben sich aus den statistischen Analysen der akustischen Parameter.

Anhand der Ergebnisse dieser Arbeit lässt sich folgern, dass für die Entwicklung von robusten Sprechererkennungssystemen bei verändertem Stimmaufwand nur Obstruenten statt dem kompletten Sprachsignal gewählt werden könnten, da

diese weniger stark beeinflusst sind als Vokale und Sonoranten. Allerdings sind Obstruenten insgesamt nicht so gut geeignet für die Sprechererkennung wie Vokale, da stimmlose Laute häufig ausgeschlossen werden und damit bereits einige Obstruenten wegfallen. Dadurch kann der Anteil der Trainings- und Testdaten sehr gering werden. Sinnvoll wäre es daher, Merkmale zu verwenden, die gute Ergebnisse bei stimmhaften und stimmlosen Obstruenten erzielen. Hierzu müssen Evaluationen vorhandener Merkmale durchgeführt werden oder spezielle Merkmale für die Sprechererkennung auf Obstruenten entwickelt werden.

Denkbar wäre auch eine Sprechererkennung bezogen auf andere Lautklassen, wie beispielsweise die Nasale. Scheffer et al. (2011) zeigten bereits gute Ergebnisse für die Sprechererkennung bei nicht-übereinstimmendem Stimmaufwand mit einem System, welches ausschließlich Silben analysiert, die einen Nasal enthalten. Diese Untersuchung arbeitet mit cepstralen Merkmalen und könnte auf andere Merkmale ausgeweitet werden. Außerdem ist es denkbar weitere Merkmale zur textunabhängigen Sprechererkennung zu evaluieren, welche lautunabhängig gute Ergebnisse erzielen und damit, wie in der vorliegenden Arbeit, auf die Gesamtdaten angewendet werden können.

Zur Verbesserung sprachverarbeitender Systeme ist es ebenso denkbar, eine Prädiktion der Veränderung der akustischen Eigenschaften von Sprache vorzunehmen. Hierzu eignen sich die spektralen Merkmale, da sie eine hohe Korrelation zwischen normaler und lauter Sprache aufweisen. Insbesondere die Klasse der Vokale scheint hierfür geeignet zu sein. Es ist zu evaluieren, ob eine Prädiktion mit Hilfe einer Regression oder statistischer Modelle zielführend ist. Für die Sprechererkennung ergibt sich das Problem, dass die Prädiktion einer Normalisierung gleich kommt und sprecherspezifische Information verloren geht. Für die Spracherkennung kann ein solches Verfahren hingegen durchaus sinnvoll sein, da die Entfernung sprecherspezifischer Eigenschaften in der Spracherkennung von Vorteil ist. Gelingt es, in die Prädiktion neben allgemeingültigen Informationen über die Veränderung der Sprache bei erhöhtem Stimmaufwand auch sprecherspezifische Information einzubeziehen, so könnte eine solche Prädiktion auch für die Sprecherkennung nutzbar sein.

Eine weitere Fortführung der vorliegenden Arbeit könnte die erzielten Ergebnisse im Kontext gestresster oder emotionsbehafteter Sprache betrachten. Hier sind beispielsweise Untersuchungen zu unterschiedlichen Arten von Stress und Emotionen denkbar.

Abgesehen von den hier vorgestellten Ansätzen bestehen noch zahlreiche weitere Untersuchungsmöglichkeiten, um die Sprechererkennung bei nicht-übereinstimmendem Stimmaufwand zu verbessern. Die vorliegende Arbeit zeigt, dass eine Leistungssteigerung der Sprechererkennung im Kontext erhöhten Stimmaufwands möglich ist und Ansätze für weitere Forschungsaktivitäten existieren.

Abkürzungsverzeichnis

A3	Amplitude des dritten Formanten
A3*	Korrigierte Amplitude des dritten Formanten
ASR	Automatic speech recognition, automatische Spracherkennung
Basis	Basistest, welcher normalen Stimmaufwand in den Trainingsdaten nutzt und erhöhten im Testset
BIC	Bayesian information criterion, Bayes'sches Informationskriterium
COG	Center of gravity, gewichteter spektraler Schwerpunkt
DCT	Discrete cosine transform, Diskrete Kosinus Transformation
DFT	Diskrete Fourier Transformation
EER	Equal error rate
EM	Expectation maximation
EV	Energieverhältnis
F_0	Grundfrequenz
F1	Erster Formant
F2	Zweiter Formant
F3	Dritter Formant
F4	Vierter Formant
FFT	Fast Fourier transform, Schnelle Fourier Transformation
Gesamt	Gesamtdaten über alle Sprecher und Phoneme
GMM	Gauß'sche Mischverteilungsmodelle
H1	Amplitude der ersten Harmonischen
H1*	Korrigierte Amplitude der ersten Harmonischen
HMM	Hidden Markov Modell
HTK	Hidden Markov Toolkit
IDFT	Inverse DFT
LFCC	Linear frequency cepstrum coefficients)
LK	Lautklasse
LLR	Log-Likelihood-Ratio
logE	Logarithmierte Energie
$\log F_0$	Logarithmierte F_0
$\log F_0/E$	Logarithmierte F_0 und Energie
LP	Lineare Prädiktion
LPC	Linearen prädiktiven Codierung
LR	Likelihood-Ratio
MAP	Maximum a posteriori
Max	Maximalwert einer Verteilung
MFCC	Mel frequency cepstrum coefficients, Mel-Cepstrum-Koeffizienten
Min	Minimalwert einer Verteilung

Mittl.AW	Mittlere Abweichung einer Verteilung
ML	Maximum likelihood
MLLR	Maximum likelihood linear regression
Mom1	Erstes spektrales Moment
Mom2	Zweites spektrales Moment
Mom3	Drittes spektrales Moment
Mom4	Viertes spektrales Moment
Momente	Kombination der ersten vier spektralen Momente
Momente Kombi	Momente Kombination ausgewählter spektraler Momente
MW	Mittelwert
Obstr	Obstruenten
OLLO	Oldenburger Logathom Korpus
Phonem	Korpus nach Phonemen sortiert
PLP	Perzeptuelle lineare Prädiktion
RC	Rate of closure, glottale Verschlussrate
SA	Stimmaufwand
Sch	Schiefe
SN	Spektrale Neigung
Son	Sonoranten
Spk	Korpus nach Sprecher sortiert
St.AW	Standardabweichung
SUSAS	Speech under simulated and actual stress database, Korpus mit Sprache unter simuliertem und tatsächlichem Stress
SUSC-0, SUSC-1	Speech under stress databases, Korpora mit Sprache unter Stress
TF	Telefonfilter
UBM	Universal background model, Hintergrundmodell
UBM+	Test mit erweitertem UBM
VAD	Voice activity detection, Sprache-Pause-Detektion
Var	Varianz
Vgl.	Vergleichstest mit Daten normalen Stimmaufwands für das Training und den Test
Vok	Vokale
VTLN	Vokaltraktlängennormalisierung
Wöl	Wölbung
ZEPS	Abkürzung der Merkmalskombination: Zero Cross Rate, Energy, Pitch, Energy slope - Nulldurchgangsrate, Energie, Grundfrequenz, Steigung der Energie

Abbildungsverzeichnis

Tabellenverzeichnis

Anhang

A.1 Quantifizierung des Stimmaufwands

A.1.1 Spektrale Neigung

LK	SA	MW	Max	Min	Mittl.AW	St.AW	Var	Sch	Wöl
Gesamt	N	-40,915	73,044	-150,746	31,933	36,540	1335,182	-0,287	-1,046
	L	-34,184	59,090	-199,541	27,522	31,858	1014,951	-0,420	-0,830
Obstr	N	-2,635	73,044	-90,772	4,807	8,762	76,773	0,137	15,909
	L	-1,883	53,903	-199,541	3,707	8,610	74,131	-5,551	91,291
Son	N	-59,051	-17,639	-110,289	15,564	18,749	351,511	-0,048	-0,708
	L	-36,214	-3,952	-97,834	13,499	17,230	296,862	-0,954	0,704
Vok	N	-64,345	7,138	-150,746	21,478	25,904	670,991	-0,165	-0,588
	L	-50,272	59,090	-167,157	21,910	26,780	717,186	-0,017	-0,322

Tabelle A.1: Momente der Verteilung der SN insgesamt und nach Lautklassen ausgewertet, für normale und laute Sprache (Abweichungen über 10% nach oben sind rot und nach unten sind blau markiert.)

Phonem	SA	MW	Max	Min	Mittl.AW	St.AW	Var	Sch	Wöl
p	N	-3,378	0,584	-45,392	3,228	5,189	26,922	-3,541	16,749
	L	-5,150	-0,019	-199,541	6,264	14,322	205,109	-6,528	57,761
b	N	-9,351	-0,077	-36,407	6,381	7,944	63,108	-0,818	0,353
	L	-5,891	-0,039	-30,381	3,725	5,095	25,956	-1,802	4,461
t	N	-2,232	16,160	-90,772	2,427	4,742	22,487	-8,971	146,197
	L	-1,119	20,648	-59,749	1,513	3,351	11,232	-6,974	107,530
d	N	-11,951	0,079	-61,157	8,832	11,393	129,790	-1,328	2,310
	L	-6,877	0,057	-39,626	4,411	6,321	39,961	-1,940	6,046
k	N	-5,097	21,417	-61,226	5,207	7,537	56,810	-2,120	7,216
	L	-2,460	19,201	-40,797	3,250	5,314	28,234	-2,307	11,545
g	N	-6,498	1,391	-52,015	6,766	8,653	74,873	-1,747	3,661
	L	-4,536	0,065	-79,251	4,767	8,209	67,394	-4,635	29,597
f	N	-0,737	7,231	-35,086	1,418	2,374	5,636	-4,321	49,456
	L	-0,879	2,923	-72,226	1,479	4,454	19,842	-12,214	181,053
v	N	-17,517	47,069	-74,951	9,487	15,806	249,840	0,447	8,438
	L	-11,667	-5,762	-28,384	4,045	5,239	27,444	-1,197	1,286
s	N	1,166	26,976	-24,290	1,954	3,300	10,888	1,958	16,511
	L	0,445	11,927	-12,917	0,898	1,651	2,727	1,115	18,048
z	N	-12,621	7,143	-69,617	9,692	12,172	148,147	-0,845	1,214
	L	-6,936	3,364	-35,049	4,165	5,745	33,005	-1,340	3,644
ʃ	N	21,606	73,044	-10,742	13,887	17,084	291,864	0,766	-0,100
	L	19,066	53,903	-4,511	8,295	10,630	112,997	0,909	0,667
t͡s	N	-0,815	10,957	-9,172	1,675	2,504	6,268	0,566	3,588
	L	-0,498	5,518	-6,340	0,834	1,307	1,709	-0,779	5,213

Tabelle A.2: Momente der Verteilung der SN sämtlicher Obstruenten für normale und laute Sprache (Abweichungen über 10% nach oben sind rot und nach unten sind blau markiert.)

Phonem	SA	MW	Max	Min	Mittl.AW	St.AW	Var	Sch	Wöl
l	N	-59,920	-17,639	-104,769	15,395	18,786	352,901	0,142	-0,691
	L	-40,567	-3,952	-97,834	15,710	19,622	385,033	-0,598	-0,039
m	N	-56,253	-20,799	-91,700	16,316	19,241	370,205	-0,154	-1,105
	L	-34,059	-15,834	-81,909	11,765	15,265	233,011	-1,196	0,872
n	N	-61,461	-28,153	-110,289	14,375	17,916	320,985	-0,250	-0,298
	L	-32,498	-12,587	-83,855	11,327	14,515	210,677	-0,968	1,341

Tabelle A.3: Momente der Verteilung der SN sämtlicher Sonoranten für normale und laute Sprache (Abweichungen über 10% nach oben sind rot und nach unten sind blau markiert.)

Phonem	SA	MW	Max	Min	Mittl.AW	St.AW	Var	Sch	Wöl
iː	N	-44,491	7,138	-101,315	14,779	17,848	318,534	-0,163	-0,586
	L	-19,771	35,687	-75,705	13,326	16,908	285,875	-0,104	0,132
ɪ	N	-63,524	1,673	-131,361	17,950	22,128	489,644	0,121	-0,308
	L	-34,556	59,090	-116,127	19,116	24,027	577,275	-0,034	0,026
eː	N	-52,644	5,394	-99,095	17,355	20,698	428,408	0,133	-0,684
	L	-24,000	36,875	-86,946	15,196	19,191	368,300	-0,144	0,045
ɛ	N	-59,150	-3,419	-117,486	18,122	22,097	488,285	-0,016	-0,536
	L	-36,069	33,620	-104,942	16,513	20,726	429,561	-0,143	0,045
aː	N	-53,926	-8,280	-124,411	15,326	18,382	337,905	-0,036	-0,614
	L	-52,494	-13,048	-101,254	13,150	16,111	259,577	-0,180	-0,440
a	N	-65,092	-2,250	-137,421	15,173	19,192	368,350	0,321	0,285
	L	-52,909	-7,806	-112,369	14,355	17,826	317,769	-0,147	-0,275
ə	N	-44,575	-5,993	-109,535	14,976	18,807	353,717	-0,757	0,240
	L	-39,268	6,938	-107,024	12,263	16,338	266,936	-0,936	1,746
ʊ	N	-97,012	-9,847	-150,746	16,352	22,072	487,154	0,981	1,959
	L	-76,485	-14,607	-167,157	17,783	22,400	501,746	0,000	0,111
uː	N	-79,356	-19,657	-136,926	19,827	23,759	564,506	0,206	-0,693
	L	-68,131	-11,854	-122,534	18,033	21,669	469,532	0,254	-0,717
oː	N	-67,961	-12,652	-126,539	20,743	24,065	579,104	0,069	-0,994
	L	-63,642	-8,921	-122,172	15,819	19,380	375,591	0,106	-0,485
ɔ	N	-84,597	-8,018	-139,273	14,505	18,964	359,615	0,458	0,998
	L	-71,736	-15,673	-134,200	15,415	19,447	378,172	0,021	-0,035

Tabelle A.4: Momente der Verteilung der SN sämtlicher Vokale für normale und laute Sprache (Abweichungen über 10% nach oben sind rot und nach unten sind blau markiert.)

Spk	SA	MW	Max	Min	Mittl.AW	St.AW	Var	Sch	Wöl
S02	N	-41,623	61,202	-145,016	31,971	37,266	1388,754	-0,272	-0,933
	L	-40,548	49,062	-129,908	28,861	33,501	1122,290	-0,129	-0,961
S03	N	-37,481	21,069	-121,164	26,886	31,840	1013,772	-0,365	-0,884
	L	-35,134	27,499	-122,203	25,966	30,126	907,599	-0,293	-0,925
S05	N	-46,335	47,150	-122,517	27,432	32,974	1087,265	-0,017	-0,996
	L	-36,122	21,764	-118,215	26,466	32,040	1026,539	-0,513	-0,598
S06	N	-58,126	3,247	-138,415	29,651	36,033	1298,373	0,460	-0,955
	L	-56,750	25,076	-119,866	27,906	33,816	1143,505	0,560	-0,881
S10	N	-39,048	43,943	-141,375	33,711	38,164	1456,526	-0,453	-1,064
	L	-32,503	30,160	-147,043	27,454	32,564	1060,444	-0,704	-0,319
S13	N	-29,604	59,493	-133,204	24,349	29,363	862,172	-0,688	-0,096
	L	-23,640	45,624	-106,653	21,695	26,366	695,169	-0,826	0,064
S15	N	-52,328	36,564	-126,472	31,860	36,939	1364,510	0,273	-1,219
	L	-38,045	11,899	-167,157	25,185	29,629	877,896	-0,259	-0,714
S16	N	-31,809	53,666	-146,302	28,867	34,579	1195,673	-0,823	-0,320
	L	-30,753	36,496	-132,624	26,215	30,811	949,333	-0,662	-0,557
S17	N	-45,187	65,496	-139,273	31,589	36,281	1316,304	0,058	-1,047
	L	-30,386	19,588	-100,423	25,158	28,135	791,582	-0,360	-1,171
S20	N	-44,068	14,156	-135,269	33,032	37,975	1442,097	-0,301	-1,120
	L	-29,905	45,610	-103,477	25,230	29,073	845,218	-0,342	-0,982
S21	N	-41,079	49,379	-136,926	36,040	40,015	1601,166	-0,259	-1,250
	L	-31,707	30,450	-151,654	29,932	34,251	1173,121	-0,597	-0,880
S22	N	-36,623	52,991	-129,008	32,595	36,726	1348,836	-0,333	-1,103
	L	-31,712	41,643	-112,369	28,248	31,654	1001,958	-0,360	-1,245
S26	N	-33,004	55,315	-143,047	29,661	34,053	1159,627	-0,508	-0,891
	L	-31,832	59,090	-108,312	28,843	32,475	1054,607	-0,297	-1,134
S29	N	-39,451	56,773	-130,280	32,058	36,776	1352,505	-0,280	-1,123

	L	-30,584	34,021	-123,167	24,371	28,827	831,001	-0,445	-0,813
S30	N	-39,624	73,044	-136,849	35,363	39,740	1579,237	-0,240	-1,136
	L	-31,156	43,206	-110,083	27,123	31,314	980,591	-0,438	-0,965
S31	N	-46,651	65,562	-136,701	32,294	36,750	1350,556	-0,081	-1,123
	L	-41,366	53,903	-199,541	28,204	33,364	1113,150	-0,249	-0,411
S32	N	-40,227	26,976	-150,746	28,463	33,829	1144,425	-0,411	-0,745
	L	-37,011	22,987	-121,908	25,787	30,660	940,013	-0,304	-0,878
S36	N	-45,931	23,071	-147,582	33,243	38,092	1451,019	-0,372	-1,018
	L	-38,058	13,168	-137,860	27,224	32,432	1051,866	-0,492	-0,719
S39	N	-35,490	40,388	-122,912	28,164	32,609	1063,362	-0,379	-0,956
	L	-27,978	51,700	-124,460	27,442	31,865	1015,407	-0,728	-0,541

Tabelle A.5: Momente der Verteilung der SN nach Sprechern ausgewertet für normale und laute Sprache (Abweichungen über 10% nach oben sind rot und nach unten sind blau markiert.)

A.1.2 Gewichteter spektraler Schwerpunkt

LK	SA	MW	Max	Min	Mittl.AW	St.AW	Var	Sch	Wöl
Gesamt	N	1123,806	3123,932	111,591	442,475	546,435	298591,742	0,658	-0,175
	L	1260,368	3177,296	117,293	440,653	531,214	282188,301	0,270	-0,525
Obstr	N	1470,844	3123,932	111,591	533,912	624,187	389609,088	-0,172	-0,837
	L	1381,986	3177,296	117,293	603,485	683,311	466913,338	0,004	-1,137
Son	N	733,178	1658,386	314,700	190,700	251,232	63117,274	1,363	2,287
	L	1026,592	1897,307	458,865	221,058	271,149	73521,511	0,262	-0,029
Vok	N	916,068	2101,475	265,837	296,129	354,517	125682,327	0,310	-0,647
	L	1203,671	2566,283	270,430	356,871	426,695	182068,424	0,092	-0,814

Tabelle A.6: Momente der Verteilung des COG insgesamt und nach Lautklassen ausgewertet, für normale und laute Sprache (Abweichungen über 10% nach oben sind rot und nach unten sind blau markiert.)

Phonem	SA	MW	Max	Min	Mittl.AW	St.AW	Var	Sch	Wöl
p	N	907,347	2039,813	111,591	274,249	341,106	116353,270	0,141	-0,281
	L	736,352	1928,856	117,293	289,174	346,768	120248,044	0,639	-0,221
b	N	465,025	1638,512	224,469	239,719	306,029	93653,761	1,569	1,461
	L	406,349	1002,595	250,166	125,783	168,256	28309,921	1,676	2,327
t	N	1540,856	2434,844	133,991	368,783	472,921	223654,078	-0,802	0,053
	L	1491,925	2514,817	333,989	472,817	553,437	306292,283	-0,502	-0,989
d	N	515,177	2024,388	239,146	253,044	321,101	103105,857	1,713	3,308
	L	524,514	2219,544	237,175	258,598	340,103	115670,090	1,954	4,593
k	N	1383,022	2564,418	406,711	365,522	432,143	186747,742	0,228	-0,738
	L	1305,582	2946,345	296,412	421,344	505,800	255833,913	0,344	-0,669
g	N	793,636	2419,372	260,977	356,951	415,099	172307,379	0,578	-0,208
	L	760,501	2249,798	258,209	366,165	419,837	176263,440	0,678	-0,367
f	N	1914,842	2616,469	760,965	171,216	220,611	48669,097	-0,447	1,169
	L	1895,434	2412,101	545,073	194,951	261,208	68229,537	-1,407	3,555
v	N	800,604	2446,748	397,605	189,413	322,132	103768,746	3,301	14,298
	L	867,878	1214,471	591,865	156,198	191,255	36578,456	0,476	-1,000
s	N	2102,183	3024,343	737,877	206,366	273,042	74552,195	0,105	1,273
	L	2083,429	2773,502	1005,556	182,071	240,339	57762,661	-0,493	1,510
z	N	1120,586	2934,264	400,070	606,433	709,916	503981,210	1,034	-0,502
	L	970,165	2521,125	462,022	398,702	550,510	303061,616	1,694	1,481
ʃ	N	2536,446	3123,932	1724,766	254,676	310,729	96552,619	-0,262	-0,627
	L	2688,648	3177,296	1844,546	196,421	242,259	58689,280	-0,566	0,064
t͡s	N	1800,168	2675,254	863,427	264,047	336,794	113430,319	-0,010	-0,070
	L	1719,879	2631,590	601,121	325,790	405,768	164647,987	-0,459	-0,304

Tabelle A.7: Momente der Verteilung des COG sämtlicher Obstruenten für normale und laute Sprache (Abweichungen über 10% nach oben sind rot und nach unten sind blau markiert.)

Phonem	SA	MW	Max	Min	Mittl.AW	St.AW	Var	Sch	Wöl
l	N	924,848	1658,386	527,517	207,153	273,956	75052,122	0,989	0,460
	L	1235,736	1897,307	645,007	166,296	225,008	50628,392	0,199	1,067
m	N	628,657	981,100	359,947	123,449	152,817	23353,059	0,500	-0,526
	L	917,965	1313,984	458,865	181,195	212,458	45138,451	-0,150	-0,898
n	N	608,114	971,628	314,700	110,439	140,086	19624,091	0,580	-0,097
	L	854,836	1368,220	515,375	153,602	188,990	35717,365	0,213	-0,345

Tabelle A.8: Momente der Verteilung des COG sämtlicher Sonoranten für normale und laute Sprache (Abweichungen über 10% nach oben sind rot und nach unten sind blau markiert.)

Phonem	SA	MW	Max	Min	Mittl.AW	St.AW	Var	Sch	Wöl
iː	N	1137,042	2101,475	423,355	259,208	311,508	97037,220	0,266	-0,585
	L	1618,009	2344,440	274,040	277,087	335,374	112475,956	-0,447	-0,409
ɪ	N	1182,373	2025,739	487,190	218,431	265,731	70613,229	0,442	-0,358
	L	1556,296	2566,283	572,176	243,571	298,701	89222,076	0,077	-0,391
eː	N	1297,817	2067,035	626,906	243,917	293,347	86052,463	0,042	-0,724
	L	1701,007	2368,007	314,425	186,728	239,470	57345,936	-0,446	2,236
ɛ	N	1284,361	1952,246	664,215	187,871	231,331	53513,844	0,259	-0,303
	L	1585,781	2324,982	869,224	177,164	223,767	50071,447	0,089	-0,024
aː	N	1022,236	1560,070	581,507	138,643	169,327	28671,582	0,253	-0,341
	L	1260,922	1772,968	725,156	139,450	170,199	28967,799	0,077	-0,490
a	N	1117,715	1613,945	674,263	127,582	158,633	25164,460	0,305	-0,062
	L	1344,804	1865,501	922,990	131,886	163,946	26878,209	0,237	-0,251
ə	N	993,401	1908,012	401,871	204,457	249,264	62132,562	0,546	-0,034
	L	1335,422	2071,934	707,011	234,939	280,518	78690,438	-0,131	-0,811
ʊ	N	581,371	1302,364	325,457	105,987	138,751	19251,876	1,141	2,147
	L	786,847	1500,168	387,135	169,320	208,723	43565,088	0,626	-0,085
uː	N	462,413	999,821	265,837	73,294	97,085	9425,584	1,394	2,407
	L	635,254	1417,612	277,350	139,461	173,658	30156,940	0,951	0,530
oː	N	562,615	1139,596	334,974	102,374	131,037	17170,672	1,214	1,711
	L	787,292	1516,166	270,430	147,085	176,682	31216,466	0,376	-0,409
ɔ	N	803,278	1344,748	463,825	116,920	149,945	22483,532	0,764	0,388
	L	1001,089	1519,746	585,574	141,632	171,736	29493,266	0,073	-0,473

Tabelle A.9: Momente der Verteilung des COG sämtlicher Vokale für normale und laute Sprache (Abweichungen über 10% nach oben sind rot und nach unten sind blau markiert.)

Spk	SA	MW	Max	Min	Mittl.AW	St.AW	Var	Sch	Wöl
S02	N	1117,291	2962,156	170,941	486,656	594,035	352877,318	0,622	-0,472
	L	1174,673	2935,815	269,701	443,283	540,927	292601,900	0,455	-0,370
S03	N	1221,964	2854,161	244,554	409,523	500,068	250067,858	0,081	-0,617
	L	1308,788	2892,334	197,269	424,668	509,249	259334,059	0,009	-0,357
S05	N	1164,264	2683,792	337,645	414,311	504,414	254433,472	0,146	-0,613
	L	1296,884	2618,044	262,029	431,334	511,721	261858,398	-0,174	-0,975
S06	N	944,269	2511,921	313,574	388,205	491,765	241832,666	1,097	0,350
	L	1103,043	3018,992	397,465	363,017	460,090	211683,015	0,767	0,185
S10	N	1110,046	2738,894	224,469	520,317	627,689	393992,963	0,648	-0,632
	L	1255,353	2843,790	237,175	485,798	590,765	349003,510	0,291	-0,718
S13	N	1140,884	2722,615	239,146	464,372	558,192	311578,710	0,367	-0,731
	L	1323,804	2908,743	250,928	529,691	614,674	377824,471	-0,025	-1,024
S15	N	981,441	2871,919	230,716	415,233	551,795	304478,047	0,985	0,221
	L	1110,288	2652,761	164,904	401,035	483,026	233314,542	0,329	-0,682
S16	N	1140,236	2847,057	178,628	494,669	594,263	353148,484	0,486	-0,723
	L	1183,881	2782,656	160,583	471,834	575,510	331212,274	0,323	-0,744
S17	N	1136,185	3028,391	133,991	387,793	510,827	260944,383	0,959	0,522
	L	1355,000	3072,747	139,566	414,483	497,864	247868,100	0,058	-0,409
S20	N	1097,185	2593,506	208,661	409,250	519,791	270182,578	0,627	-0,268
	L	1274,564	2921,250	158,087	421,364	514,056	264253,812	0,289	-0,492
S21	N	1195,640	2785,006	180,231	463,653	569,056	323824,238	0,554	-0,528
	L	1383,186	2946,345	237,666	482,904	570,289	325229,232	0,107	-0,898
S22	N	1160,506	3123,932	426,851	396,259	516,096	266354,723	1,245	1,643
	L	1366,179	3177,296	221,109	403,494	487,094	237260,378	0,673	0,424
S26	N	1236,150	2985,413	338,811	448,365	545,880	297984,637	0,426	-0,544
	L	1372,417	2718,677	410,061	420,707	489,326	239440,406	0,209	-0,933
S29	N	1095,606	2897,823	111,591	451,929	571,975	327155,526	1,013	0,321

	L	1272,542	2934,592	164,009	435,014	544,487	296465,690	0,451	-0,238
S30	N	1196,904	3088,909	303,186	448,900	565,932	320278,781	0,925	0,574
	L	1248,189	3082,340	188,346	434,494	535,796	287077,159	0,654	0,272
S31	N	1041,092	2774,481	210,943	396,559	500,007	250006,523	0,861	0,190
	L	1114,691	2840,955	117,293	394,679	493,622	243662,394	0,586	-0,055
S32	N	1087,573	2555,485	189,916	398,913	492,225	242285,725	0,434	-0,395
	L	1227,460	2329,338	296,412	396,954	468,251	219259,111	-0,067	-0,842
S36	N	966,534	2657,608	258,647	390,739	490,411	240502,683	0,808	-0,039
	L	1191,767	2335,672	186,572	382,886	465,241	216449,338	0,133	-0,798
S39	N	1294,255	2768,176	242,105	407,949	499,340	249339,960	0,212	-0,530
	L	1347,676	2861,315	311,938	469,177	541,941	293699,631	-0,037	-1,004

Tabelle A.10: Momente der Verteilung des COG nach Sprechern ausgewertet für normale und laute Sprache (Abweichungen über 10% nach oben sind rot und nach unten sind blau markiert.)

A.1.3 Energieverhältnis

LK	SA	MW	Max	Min	Mittl.AW	St.AW	Var	Sch	Wöl
Gesamt	N	0,705	85,196	0,002	0,812	2,192	4,805	14,897	340,222
	L	0,794	90,480	0,002	0,916	2,150	4,622	13,421	339,076
Obstr	N	1,210	85,196	0,006	1,195	3,344	11,185	10,558	157,653
	L	1,243	90,480	0,007	1,271	3,238	10,484	10,934	190,585
Son	N	0,258	4,184	0,010	0,233	0,423	0,179	5,141	37,666
	L	0,448	4,969	0,011	0,421	0,718	0,515	3,623	15,497
Vok	N	0,401	9,818	0,002	0,507	0,823	0,678	3,772	20,014
	L	0,576	22,208	0,002	0,692	1,271	1,616	5,296	45,802

Tabelle A.11: Momente der Verteilung des EV insgesamt und nach Lautklassen ausgewertet, für normale und laute Sprache (Abweichungen über 10% nach oben sind rot und nach unten sind blau markiert.)

Phonem	SA	MW	Max	Min	Mittl.AW	St.AW	Var	Sch	Wöl
p	N	0,313	2,886	0,010	0,159	0,249	0,062	3,561	22,927
	L	0,256	2,957	0,010	0,159	0,258	0,066	3,925	25,655
b	N	0,128	1,113	0,013	0,120	0,169	0,029	2,713	10,227
	L	0,061	0,426	0,011	0,038	0,066	0,004	3,434	13,372
t	N	0,705	11,121	0,012	0,439	0,707	0,500	4,888	50,503
	L	0,699	6,477	0,018	0,417	0,604	0,365	2,773	14,587
d	N	0,171	2,162	0,012	0,164	0,286	0,082	4,582	26,450
	L	0,174	4,759	0,008	0,190	0,462	0,213	7,820	72,083
k	N	0,545	10,671	0,007	0,523	0,902	0,814	4,570	31,508
	L	0,643	19,397	0,007	0,675	1,264	1,598	6,174	62,815
g	N	0,297	6,833	0,006	0,262	0,579	0,335	8,360	88,592
	L	0,321	17,676	0,007	0,336	1,036	1,074	13,201	208,132
f	N	1,485	11,567	0,093	0,636	0,951	0,905	3,167	20,040
	L	1,551	8,413	0,141	0,687	0,967	0,934	2,081	7,329
v	N	0,447	3,377	0,029	0,386	0,705	0,497	3,289	10,428
	L	0,299	0,920	0,077	0,112	0,163	0,026	1,841	4,668
s	N	1,548	38,628	0,231	0,874	2,162	4,674	10,256	143,670
	L	1,906	11,865	0,406	0,921	1,288	1,659	2,346	9,900
z	N	0,643	4,339	0,051	0,579	0,796	0,633	2,093	4,323
	L	0,528	5,168	0,050	0,510	0,831	0,690	3,177	11,166
ʃ	N	13,971	85,196	1,423	8,606	12,389	153,480	2,307	7,022
	L	14,007	90,480	1,536	7,419	11,595	134,451	3,019	13,059
t͡s	N	1,130	5,537	0,120	0,625	0,894	0,798	2,227	6,351
	L	1,298	8,064	0,093	0,784	1,166	1,359	2,478	8,502

Tabelle A.12: Momente der Verteilung des EV sämtlicher Obstruenten für normale und laute Sprache (Abweichungen über 10% nach oben sind rot und nach unten sind blau markiert.)

Phonem	SA	MW	Max	Min	Mittl.AW	St.AW	Var	Sch	Wöl
l	N	0,412	4,184	0,018	0,336	0,580	0,336	4,135	22,073
	L	0,654	4,969	0,023	0,599	0,955	0,912	2,745	7,703
m	N	0,171	1,799	0,011	0,160	0,300	0,090	3,977	17,116
	L	0,345	3,545	0,011	0,323	0,533	0,284	3,718	17,919
n	N	0,161	1,038	0,010	0,121	0,187	0,035	2,840	9,201
	L	0,274	1,626	0,019	0,229	0,360	0,129	2,442	5,662

Tabelle A.13: Momente der Verteilung des EV sämtlicher Sonoranten für normale und laute Sprache (Abweichungen über 10% nach oben sind rot und nach unten sind blau markiert.)

Phonem	SA	MW	Max	Min	Mittl.AW	St.AW	Var	Sch	Wöl
iː	N	1,726	9,818	0,093	0,940	1,306	1,705	2,008	6,291
	L	2,774	22,208	0,015	1,682	2,428	5,896	2,690	12,646
ɪ	N	0,949	7,684	0,016	0,702	0,965	0,932	2,183	7,315
	L	0,980	10,635	0,015	0,792	1,266	1,603	3,482	16,642
eː	N	1,182	5,273	0,052	0,704	0,936	0,876	1,502	2,804
	L	1,204	5,693	0,022	0,919	1,157	1,338	1,345	1,238
ɛ	N	0,174	4,493	0,005	0,128	0,277	0,077	8,204	98,463
	L	0,247	4,831	0,006	0,179	0,327	0,107	5,904	56,082
aː	N	0,083	1,282	0,005	0,056	0,096	0,009	5,040	43,663
	L	0,244	1,968	0,009	0,193	0,285	0,081	2,566	7,820
a	N	0,134	2,519	0,004	0,102	0,176	0,031	5,206	48,198
	L	0,379	6,598	0,009	0,311	0,554	0,307	4,798	33,694
ə	N	0,092	1,242	0,004	0,061	0,099	0,010	4,350	33,802
	L	0,137	1,635	0,003	0,099	0,145	0,021	3,119	17,960
ʊ	N	0,063	0,769	0,003	0,045	0,069	0,005	3,859	25,908
	L	0,094	0,895	0,002	0,072	0,107	0,012	2,916	12,130
uː	N	0,055	0,529	0,004	0,035	0,051	0,003	3,032	15,673
	L	0,107	1,146	0,007	0,069	0,097	0,009	2,616	14,122
oː	N	0,044	0,434	0,002	0,030	0,043	0,002	2,917	14,192
	L	0,076	0,706	0,003	0,053	0,072	0,005	2,305	9,100
ɔ	N	0,035	0,310	0,002	0,025	0,036	0,001	2,360	7,980
	L	0,069	0,645	0,002	0,054	0,077	0,006	2,623	10,005

Tabelle A.14: Momente der Verteilung des EV sämtlicher Vokale für normale und laute Sprache (Abweichungen über 10% nach oben sind rot und nach unten sind blau markiert.)

Spk	SA	MW	Max	Min	Mittl.AW	St.AW	Var	Sch	Wöl
S02	N	0,742	35,752	0,003	0,844	2,120	4,493	10,143	132,007
	L	0,623	32,408	0,002	0,732	1,700	2,891	10,267	153,054
S03	N	0,868	19,377	0,012	0,949	1,624	2,637	4,942	37,146
	L	1,225	45,124	0,011	1,359	2,783	7,742	6,942	78,629
S05	N	0,644	22,859	0,006	0,780	1,501	2,253	7,234	81,628
	L	0,747	15,326	0,014	0,824	1,398	1,953	4,263	26,745
S06	N	0,257	3,436	0,004	0,273	0,394	0,155	2,789	11,140
	L	0,281	6,899	0,003	0,310	0,481	0,232	4,983	50,623
S10	N	0,970	29,879	0,006	1,093	2,185	4,776	6,564	59,183
	L	1,102	23,582	0,013	1,234	2,093	4,381	4,852	36,159
S13	N	0,626	20,098	0,007	0,693	1,299	1,687	6,389	66,582
	L	0,850	52,719	0,010	0,892	2,440	5,954	12,759	228,863
S15	N	0,602	11,112	0,002	0,701	1,030	1,060	3,776	24,333
	L	0,456	11,523	0,005	0,552	0,938	0,880	5,245	42,663
S16	N	0,642	25,881	0,006	0,714	1,869	3,494	9,824	114,765
	L	0,557	12,447	0,003	0,609	0,989	0,977	5,077	41,697
S17	N	0,568	52,737	0,005	0,693	2,560	6,553	13,985	233,602
	L	0,497	17,655	0,007	0,550	1,254	1,573	7,994	82,096
S20	N	0,524	16,377	0,002	0,605	1,089	1,186	6,209	62,994
	L	0,963	44,881	0,009	1,190	2,460	6,053	8,421	117,299
S21	N	0,898	19,340	0,003	0,928	1,724	2,972	5,715	44,255
	L	1,250	20,945	0,002	1,347	2,141	4,586	3,800	21,151
S22	N	1,034	85,196	0,004	1,351	4,890	23,909	11,015	144,049
	L	0,614	28,070	0,006	0,700	1,817	3,303	8,523	91,763
S26	N	0,993	29,656	0,004	1,111	2,423	5,870	7,793	78,563
	L	0,731	19,038	0,002	0,818	1,455	2,116	5,209	42,511
S29	N	0,832	58,780	0,004	1,047	3,471	12,049	12,849	186,114

	L	1,365	90,480	0,009	1,770	4,951	24,515	10,960	152,589
S30	N	0,837	29,656	0,002	1,013	2,332	5,439	7,645	73,259
	L	0,919	30,291	0,008	1,133	2,635	6,944	6,963	58,373
S31	N	0,365	10,274	0,003	0,405	0,741	0,548	6,373	62,194
	L	0,533	10,787	0,003	0,607	1,010	1,020	4,444	29,033
S32	N	0,716	38,628	0,008	0,790	1,989	3,955	12,154	195,313
	L	0,460	5,302	0,009	0,476	0,735	0,540	3,033	11,190
S36	N	0,415	10,671	0,004	0,422	0,660	0,435	6,046	73,253
	L	0,654	22,208	0,004	0,729	1,418	2,011	7,775	90,311
S39	N	0,857	35,267	0,009	0,846	2,005	4,021	10,278	141,915
	L	1,034	22,672	0,012	0,974	1,828	3,341	6,864	67,045

Tabelle A.15: Momente der Verteilung des EV nach Sprechern ausgewertet für normale und laute Sprache (Abweichungen über 10% nach oben sind rot und nach unten sind blau markiert.)

A.1.4 Spektrale Momente

LK	SA	MW	Max	Min	Mittl.AW	St.AW	Var	Sch	Wöl
Gesamt	N	1131,599	3131,744	117,145	442,482	546,443	298599,956	0,658	-0,175
	L	1268,147	3185,107	123,890	440,671	531,234	282209,986	0,270	-0,525
Obstr	N	1478,613	3131,744	117,145	533,945	624,222	389653,479	-0,172	-0,837
	L	1389,707	3185,107	123,890	603,543	683,370	466994,235	0,004	-1,137
Son	N	740,988	1666,198	322,512	190,700	251,232	63117,514	1,363	2,287
	L	1034,401	1905,119	466,673	221,058	271,149	73521,737	0,262	-0,029
Vok	N	923,876	2109,286	273,316	296,132	354,521	125684,868	0,310	-0,647
	L	1211,479	2574,095	278,242	356,874	426,698	182071,552	0,092	-0,814

Tabelle A.16: Momente der Verteilung des ersten Moments insgesamt und nach Lautklassen ausgewertet, für normale und laute Sprache (Abweichungen über 10% nach oben sind rot und nach unten sind blau markiert.)

Phonem	SA	MW	Max	Min	Mittl.AW	St.AW	Var	Sch	Wöl
p	N	914,994	2047,622	117,145	274,321	341,186	116407,553	0,141	-0,281
	L	743,921	1936,606	123,890	289,234	346,834	120294,017	0,639	-0,221
b	N	472,811	1646,305	232,280	239,690	305,999	93635,186	1,569	1,462
	L	414,115	1010,406	257,968	125,732	168,191	28288,332	1,676	2,330
t	N	1548,636	2442,606	138,391	368,803	472,951	223682,193	-0,802	0,054
	L	1499,663	2522,607	341,201	472,861	553,484	306344,001	-0,502	-0,989
d	N	522,970	2032,187	246,953	253,026	321,083	103094,464	1,713	3,309
	L	532,267	2227,281	244,984	258,528	340,038	115625,804	1,955	4,598
k	N	1390,812	2572,225	414,523	365,537	432,163	186764,949	0,228	-0,738
	L	1313,287	2954,156	304,090	421,411	505,880	255914,081	0,343	-0,669
g	N	801,413	2427,184	268,784	356,931	415,083	172293,708	0,578	-0,207
	L	768,253	2257,606	266,020	366,131	419,812	176242,286	0,679	-0,366
f	N	1922,633	2624,263	768,312	171,225	220,629	48677,346	-0,448	1,171
	L	1903,206	2419,834	552,845	194,966	261,232	68242,375	-1,407	3,555
v	N	808,411	2454,556	405,415	189,412	322,131	103768,608	3,301	14,298
	L	875,683	1222,280	599,673	156,200	191,256	36578,756	0,476	-1,000
s	N	2109,984	3032,152	745,683	206,369	273,046	74553,879	0,105	1,273
	L	2091,200	2781,314	1013,090	182,085	240,366	57775,751	-0,493	1,511
z	N	1128,393	2942,076	407,878	606,430	709,913	503976,874	1,034	-0,502
	L	977,975	2528,936	469,833	398,702	550,510	303061,419	1,694	1,481
ʃ	N	2544,257	3131,744	1732,578	254,677	310,730	96552,937	-0,262	-0,627
	L	2696,456	3185,107	1852,356	196,423	242,262	58690,706	-0,566	0,064
t͡s	N	1807,970	2683,062	871,231	264,047	336,797	113432,157	-0,010	-0,070
	L	1727,648	2639,376	608,923	325,811	405,787	164663,107	-0,459	-0,304

Tabelle A.17: Momente der Verteilung des ersten Moments sämtlicher Obstruenten für normale und laute Sprache (Abweichungen über 10% nach oben sind rot und nach unten sind blau markiert.)

Phonem	SA	MW	Max	Min	Mittl.AW	St.AW	Var	Sch	Wöl
l	N	932,659	1666,198	535,327	207,154	273,957	75052,255	0,989	0,460
	L	1243,547	1905,119	652,819	166,296	225,007	50628,333	0,199	1,067
m	N	636,467	988,911	367,757	123,450	152,817	23353,160	0,500	-0,526
	L	925,774	1321,792	466,673	181,195	212,458	45138,546	-0,150	-0,898
n	N	615,923	979,440	322,512	110,440	140,087	19624,254	0,580	-0,097
	L	862,645	1376,030	523,182	153,602	188,990	35717,360	0,213	-0,345

Tabelle A.18: Momente der Verteilung des ersten Moments sämtlicher Sonoranten für normale und laute Sprache (Abweichungen über 10% nach oben sind rot und nach unten sind blau markiert.)

Phonem	SA	MW	Max	Min	Mittl.AW	St.AW	Var	Sch	Wöl
iː	N	1144,851	2109,286	431,167	259,209	311,509	97037,986	0,266	-0,585
	L	1625,817	2352,251	281,849	277,089	335,377	112477,983	-0,447	-0,409
ɪ	N	1190,183	2033,551	495,000	218,432	265,732	70613,703	0,442	-0,358
	L	1564,106	2574,095	579,959	243,572	298,702	89222,901	0,077	-0,391
eː	N	1305,626	2074,848	634,717	243,919	293,349	86053,354	0,042	-0,724
	L	1708,815	2375,819	322,236	186,730	239,473	57347,504	-0,446	2,236
ɛ	N	1292,172	1960,058	671,890	187,872	231,332	53514,448	0,259	-0,303
	L	1593,591	2332,794	877,036	177,165	223,768	50072,198	0,089	-0,024
aː	N	1030,046	1567,883	589,317	138,645	169,329	28672,295	0,253	-0,341
	L	1268,732	1780,780	732,861	139,450	170,201	28968,215	0,077	-0,490
a	N	1125,527	1621,757	682,073	127,583	158,634	25164,692	0,305	-0,062
	L	1352,615	1873,313	930,802	131,887	163,947	26878,470	0,237	-0,251
ə	N	1001,207	1915,824	409,262	204,460	249,270	62135,690	0,546	-0,034
	L	1343,231	2079,747	714,695	234,942	280,522	78692,453	-0,131	-0,811
ʊ	N	589,179	1310,175	333,267	105,987	138,752	19252,054	1,141	2,147
	L	794,655	1507,980	394,944	169,320	208,724	43565,533	0,626	-0,085
uː	N	470,214	1007,633	273,316	73,298	97,092	9426,895	1,393	2,406
	L	643,053	1425,418	285,161	139,466	173,666	30159,765	0,951	0,530
oː	N	570,419	1147,408	342,783	102,377	131,042	17171,967	1,213	1,711
	L	795,097	1523,977	278,242	147,089	176,687	31218,263	0,376	-0,409
ɔ	N	811,088	1352,561	471,635	116,921	149,946	22483,699	0,764	0,388
	L	1008,899	1527,557	593,385	141,633	171,737	29493,607	0,073	-0,473

Tabelle A.19: Momente der Verteilung des ersten Moments sämtlicher Vokale für normale und laute Sprache (Abweichungen über 10% nach oben sind rot und nach unten sind blau markiert.)

Spk	SA	MW	Max	Min	Mittl.AW	St.AW	Var	Sch	Wöl
S02	N	1125,093	2969,966	178,333	486,656	594,035	352877,129	0,622	-0,472
	L	1182,471	2943,626	277,503	443,288	540,930	292605,327	0,455	-0,370
S03	N	1229,768	2861,973	252,360	409,526	500,071	250071,359	0,081	-0,617
	L	1316,581	2900,145	203,993	424,679	509,260	259345,964	0,008	-0,357
S05	N	1172,072	2691,604	345,454	414,313	504,415	254434,583	0,146	-0,613
	L	1304,685	2625,852	269,829	431,341	511,728	261865,431	-0,174	-0,975
S06	N	952,078	2519,730	321,383	388,204	491,764	241831,768	1,097	0,350
	L	1110,847	3026,802	405,271	363,014	460,085	211677,898	0,767	0,185
S10	N	1117,844	2746,706	232,280	520,323	627,693	393998,894	0,648	-0,632
	L	1263,148	2851,602	244,984	485,808	590,774	349014,218	0,291	-0,718
S13	N	1148,686	2730,428	246,953	464,376	558,195	311582,015	0,367	-0,731
	L	1331,601	2916,555	258,429	529,699	614,681	377832,338	-0,025	-1,024
S15	N	989,227	2879,731	237,997	415,234	551,799	304482,490	0,985	0,221
	L	1118,058	2660,569	172,678	401,050	483,038	233325,277	0,329	-0,683
S16	N	1148,021	2854,869	183,975	494,676	594,274	353161,903	0,486	-0,723
	L	1191,646	2790,468	168,321	471,851	575,531	331235,731	0,323	-0,744
S17	N	1143,973	3036,202	138,391	387,802	510,843	260960,428	0,959	0,522
	L	1362,784	3080,559	146,903	414,499	497,886	247890,393	0,058	-0,409
S20	N	1104,976	2601,317	216,311	409,261	519,800	270191,889	0,627	-0,268
	L	1282,346	2929,059	165,760	421,380	514,075	264272,901	0,289	-0,492
S21	N	1203,425	2792,818	186,370	463,661	569,064	323833,874	0,554	-0,528
	L	1390,958	2954,156	245,392	482,924	570,312	325255,546	0,107	-0,898
S22	N	1168,274	3131,744	434,661	396,272	516,109	266368,762	1,245	1,643
	L	1373,938	3185,107	227,834	403,523	487,136	237301,873	0,673	0,424
S26	N	1243,952	2993,225	346,618	448,370	545,882	297987,512	0,426	-0,544
	L	1380,177	2726,487	417,487	420,741	489,367	239480,038	0,209	-0,933
S29	N	1103,385	2905,635	117,145	451,946	571,994	327176,756	1,013	0,321

	L	1280,307	2942,400	170,137	435,040	544,523	296504,787	0,450	-0,238
S30	N	1204,678	3096,721	310,989	448,921	565,956	320306,650	0,925	0,574
	L	1255,929	3090,151	196,047	434,543	535,851	287136,042	0,653	0,271
S31	N	1048,875	2782,293	218,430	396,564	500,015	250014,947	0,861	0,190
	L	1122,467	2848,766	123,890	394,693	493,638	243678,645	0,586	-0,055
S32	N	1095,380	2563,297	197,725	398,916	492,228	242288,778	0,434	-0,395
	L	1235,244	2337,116	304,090	396,976	468,283	219289,205	-0,068	-0,842
S36	N	974,339	2665,356	266,456	390,740	490,411	240502,994	0,808	-0,039
	L	1199,562	2343,476	193,758	382,893	465,251	216458,703	0,133	-0,797
S39	N	1302,062	2775,988	248,270	407,952	499,345	249345,478	0,212	-0,530
	L	1355,446	2869,127	319,572	469,208	541,983	293745,367	-0,037	-1,004

Tabelle A.20: Momente der Verteilung des ersten Moments nach Sprechern ausgewertet für normale und laute Sprache (Abweichungen über 10% nach oben sind rot und nach unten sind blau markiert.)

LK	SA	MW	Max	Min	Mittl.AW	St.AW	Var	Sch	Wöl
Gesamt	N	821856,848	2159528,966	67838,506	360675,836	424059,791	179826706558,488	0,242	-0,800
	L	863812,073	2131286,077	81246,169	328373,031	390430,378	152435880184,314	0,257	-0,759
Obstr	N	943691,574	2159528,966	67838,506	261826,298	334165,117	111666325523,807	-0,400	0,011
	L	909285,781	2131286,077	81246,169	303041,179	366304,407	134178918726,009	-0,229	-0,589
Son	N	688175,732	1744519,665	145589,909	227982,711	299714,475	89828766617,038	1,216	1,609
	L	950578,330	1844418,632	302047,286	217186,543	275210,334	75740727950,406	0,211	0,139
Vok	N	748854,882	2061023,150	70588,305	388905,086	457347,966	209167162349,056	0,621	-0,674
	L	839627,497	1954128,743	108620,450	336348,832	401418,624	161136911700,062	0,478	-0,707

Tabelle A.21: Momente der Verteilung des zweiten Moments insgesamt und nach Lautklassen ausgewertet, für normale und laute Sprache (Abweichungen über 10% nach oben sind rot und nach unten sind blau markiert.)

Phonem	SA	MW	Max	Min	Mittl.AW	St.AW	Var	Sch	Wöl
p	N	895651,920	1500204,646	143037,719	204782,888	257971,991	66549548161,897	-0,696	0,014
	L	756591,826	1547654,752	136343,028	282960,700	324776,258	105479617437,704	0,119	-1,096
b	N	388176,667	1324774,329	81370,156	305997,605	372922,619	139071279675,288	1,239	-0,152
	L	297935,990	1127335,774	107948,901	152821,198	215329,293	46366704591,795	2,082	4,094
t	N	1169313,365	1831999,957	169629,429	184817,645	257973,956	66550562133,706	-1,041	2,082
	L	1157363,088	1856037,284	212521,662	230021,119	303745,737	92261472637,521	-0,801	0,419
d	N	474368,072	1714178,426	67838,506	332628,499	387008,133	149775295124,471	0,911	-0,467
	L	474510,944	1623012,897	94417,437	328674,248	398574,679	158861774909,683	1,115	-0,139
k	N	761393,047	1595316,393	184275,981	210430,491	253566,756	64296099649,467	0,164	-0,430
	L	735272,968	1574778,655	198802,425	196757,592	245746,947	60391562046,191	0,488	0,063
g	N	692362,662	1505088,129	88188,782	368832,926	407283,133	165879550734,346	0,048	-1,440
	L	665448,589	1687184,657	81246,169	418468,279	451997,696	204301917170,989	0,335	-1,490
f	N	1012461,450	1770823,025	608813,743	116223,420	154294,498	23806792174,489	0,955	2,219
	L	1063606,257	1579856,702	664027,797	126366,486	158749,467	25201393136,479	0,470	0,167
v	N	813973,385	1382831,518	202589,376	198638,761	250713,847	62857433234,803	0,232	-0,030
	L	894180,567	1377828,461	603752,155	160062,060	202135,543	40858777673,435	0,729	-0,474
s	N	1204048,331	2130248,401	571418,105	177714,607	225450,738	50828035319,151	0,695	0,733
	L	1145957,724	2131286,077	697729,005	165206,862	217494,349	47303791697,105	1,160	2,010
z	N	971038,189	2159528,966	386168,376	256773,456	325035,096	105647813523,136	0,627	0,420
	L	929047,072	1890052,703	433738,210	227558,190	295711,117	87445064543,339	0,987	0,658
ʃ	N	578534,599	1174254,075	221899,726	147916,479	180024,412	32408789003,223	0,452	-0,119
	L	612687,568	1534157,977	249814,945	158359,260	211040,406	44538053142,291	1,481	3,557
t͡s	N	1223578,197	1811639,290	852992,285	148880,870	186561,355	34805139322,259	0,679	0,239
	L	1228961,041	1923482,670	660582,825	160169,585	207553,045	43078266575,367	0,561	0,632

Tabelle A.22: Momente der Verteilung des zweiten Moments sämtlicher Obstruenten für normale und laute Sprache (Abweichungen über 10% nach oben sind rot und nach unten sind blau markiert.)

Phonem	SA	MW	Max	Min	Mittl.AW	St.AW	Var	Sch	Wöl
l	N	878435,163	1744519,665	370501,002	275911,645	342047,322	116996370816,636	0,762	-0,111
	L	1074599,797	1844418,632	441048,425	201564,247	266566,234	71057557031,947	0,340	0,379
m	N	580206,125	1198719,303	236042,595	166963,327	216425,101	46839824203,356	0,783	0,456
	L	892472,824	1464795,292	302047,286	215739,003	262355,715	68830520959,210	-0,058	-0,592
n	N	569435,472	1038533,112	145589,909	137795,549	179213,969	32117646832,715	0,495	0,378
	L	840359,339	1320032,333	362883,846	184365,345	233551,557	54546329590,627	0,091	-0,464

Tabelle A.23: Momente der Verteilung des zweiten Moments sämtlicher Sonoranten für normale und laute Sprache (Abweichungen über 10% nach oben sind rot und nach unten sind blau markiert.)

Phonem	SA	MW	Max	Min	Mittl.AW	St.AW	Var	Sch	Wöl
iː	N	1421771,865	2061023,150	385659,786	241550,827	300825,661	90496078079,720	-0,552	-0,043
	L	1509244,366	1954128,743	120737,628	148986,883	197746,491	39103674728,759	-1,106	3,231
ɪ	N	1270268,203	1951030,925	499928,919	217936,804	269157,207	72445602125,427	-0,132	-0,415
	L	1263866,361	1826909,671	616674,153	156513,899	199402,118	39761204724,060	-0,221	0,192
eː	N	1345192,470	1866563,897	724125,242	179509,997	227108,955	51578477492,285	-0,356	-0,233
	L	1319943,606	1689195,582	182305,998	137559,623	185698,115	34483789768,155	-1,343	3,857
ɛ	N	1019586,078	1806279,909	495351,147	152301,879	190975,853	36471776336,722	0,109	0,339
	L	987627,296	1491178,705	534501,386	115788,249	147968,233	21894598040,602	0,038	0,372
aː	N	573168,683	1341837,575	231418,795	147608,681	188369,329	35483004053,775	0,885	0,537
	L	622671,412	1146275,321	238607,284	147665,142	175891,367	30937772830,222	0,358	-0,578
a	N	603384,681	1334545,847	215601,178	149921,225	187004,499	34970682602,691	0,737	0,294
	L	617757,711	1234951,311	232569,735	132058,676	161174,260	25977142056,991	0,534	-0,121
ə	N	851895,024	1583654,022	282402,677	206214,478	249112,785	62057179578,486	-0,088	-0,669
	L	950602,015	1375097,096	404797,702	159158,367	193142,892	37304176756,085	-0,497	-0,507
ʊ	N	402937,895	1442313,550	70588,305	153111,749	201584,710	40636395182,033	1,298	2,346
	L	571654,148	1468510,428	158026,804	216790,632	260954,487	68097244270,229	0,747	-0,097
uː	N	292019,710	988790,890	84963,327	108092,252	141288,817	19962529919,814	1,412	2,415
	L	475021,172	1297019,547	108620,450	173430,839	219181,216	48040405319,789	0,952	0,460
oː	N	360226,004	1161140,152	76021,979	147500,690	181006,243	32763260048,825	0,987	0,596
	L	588328,690	1223386,480	123842,737	187208,424	223786,702	50080488065,520	0,231	-0,672
ɔ	N	503144,330	1244776,281	147666,172	178676,686	221113,336	48891107406,183	0,855	0,270
	L	618146,435	1205098,861	141712,173	187000,720	222844,515	49659677744,324	0,201	-0,787

Tabelle A.24: Momente der Verteilung des zweiten Moments sämtlicher Vokale für normale und laute Sprache (Abweichungen über 10% nach oben sind rot und nach unten sind blau markiert.)

Spk	SA	MW	Max	Min	Mittl.AW	St.AW	Var	Sch	Wöl
S02	N	741767,202	1929032,252	104213,686	367385,925	412374,111	170052407567,261	0,221	-1,179
	L	740247,588	1742100,618	96796,277	345511,182	389793,953	151939326030,864	0,311	-1,148
S03	N	920836,427	2061023,150	67838,506	353212,450	439169,924	192870222464,058	0,614	-0,347
	L	939176,187	1913389,967	124610,142	306633,184	378061,256	142930313114,263	0,404	-0,401
S05	N	932532,694	1952374,979	84963,327	384116,857	475134,120	225752432344,607	0,186	-0,775
	L	933932,313	1859941,294	143185,791	308692,241	392314,305	153910513810,961	0,152	-0,588
S06	N	655399,244	1825641,877	94477,950	389306,313	441445,053	194873734837,772	0,597	-0,863
	L	754514,813	1904698,142	153456,401	400606,072	443867,428	197018293555,206	0,459	-1,182
S10	N	761164,804	1793283,207	93185,629	345618,534	401450,108	161162188937,007	0,123	-0,969
	L	888300,053	1867234,456	81246,169	335673,572	412592,215	170232335666,405	0,271	-0,554
S13	N	809273,536	1899634,671	118324,512	347453,915	415121,822	172326127357,650	0,262	-0,668
	L	854871,179	1877464,208	107948,901	336912,660	409489,814	167681907455,681	0,256	-0,654
S15	N	698802,896	1773339,538	70588,305	399131,949	446921,716	199739020055,893	0,210	-1,335
	L	844453,932	1911765,589	108620,450	415763,348	469386,630	220323808193,520	0,287	-1,228
S16	N	773131,918	1917097,881	81370,156	382512,682	441802,597	195189534884,768	0,238	-0,945
	L	806649,973	2131286,077	94417,437	412563,731	470759,940	221614921270,665	0,322	-1,042
S17	N	814095,143	1606086,066	145667,449	315733,148	361593,607	130749936946,521	0,200	-1,179
	L	948094,014	1732622,923	122444,966	308536,291	359087,030	128943495386,017	0,042	-0,833
S20	N	780783,147	1986144,931	109420,712	361982,646	431412,521	186116763640,844	0,327	-0,664
	L	864867,712	1954128,743	176333,693	335997,696	395974,828	156796064082,545	0,548	-0,628
S21	N	848206,455	1866563,897	109704,061	379457,655	428875,050	183933808298,490	0,051	-1,193
	L	877100,829	1687184,657	151969,419	322983,972	370259,818	137092332892,065	0,088	-1,121
S22	N	913426,239	2130248,401	167780,188	353987,010	419349,067	175853639967,742	0,338	-0,789
	L	920375,197	1782975,738	249814,945	271360,228	321712,430	103498887421,588	0,034	-0,901
S26	N	914841,184	1936889,596	131345,841	336538,972	403222,329	162588246457,279	0,083	-0,746
	L	749877,938	1638462,860	167607,787	260931,605	313592,632	98340338776,264	0,591	-0,590
S29	N	774828,179	1673186,612	104171,266	302873,926	358587,748	128585173175,989	0,098	-0,776

	L	861080,482	1764337,615	175141,356	281995,006	341911,440	116903433109,424	0,247	-0,750
S30	N	847516,338	1869368,765	111156,007	357942,987	412342,999	170026748634,542	0,218	-0,999
	L	881440,240	1844418,632	187053,262	316128,987	367682,961	135190759850,443	0,256	-0,887
S31	N	770754,726	1992600,680	86755,003	366259,517	418680,921	175293713349,165	0,368	-0,874
	L	805874,131	2036255,915	131703,192	348512,184	405816,051	164686667198,056	0,486	-0,707
S32	N	935827,223	1929161,052	169500,097	325650,124	392334,864	153926645182,017	0,266	-0,670
	L	961101,924	1817068,646	206582,502	246725,195	317734,512	100955220283,364	0,286	-0,083
S36	N	776807,848	2159528,966	85223,594	401417,818	455797,539	207751396928,265	0,330	-1,060
	L	864738,798	1890981,460	95941,693	344784,981	401728,720	161385964250,474	0,356	-0,890
S39	N	958373,818	1928527,308	98027,136	311176,401	371055,050	137681850478,689	0,262	-0,794
	L	904149,577	1748298,845	132804,213	262599,611	322424,509	103957564095,631	0,302	-0,505

Tabelle A.25: Momente der Verteilung des zweiten Moments nach Sprechern ausgewertet für normale und laute Sprache (Abweichungen über 10% nach oben sind rot und nach unten sind blau markiert.)

LK	SA	MW	Max	Min	Mittl.AW	St.AW	Var	Sch	Wöl
Gesamt	N	1,482	7,234	-2,919	1,162	1,478	2,186	0,842	0,269
	L	1,182	6,996	-3,407	1,044	1,305	1,703	0,795	0,520
Obstr	N	0,740	7,234	-2,919	0,995	1,359	1,847	1,657	3,402
	L	0,944	6,996	-3,407	1,194	1,519	2,306	1,176	1,159
Son	N	1,875	4,637	0,052	0,618	0,783	0,614	0,402	0,429
	L	1,136	3,806	-0,406	0,450	0,598	0,358	1,121	2,728
Vok	N	1,934	6,612	-0,595	1,132	1,371	1,879	0,786	-0,251
	L	1,301	6,393	-1,792	0,953	1,176	1,383	0,584	-0,167

Tabelle A.26: Momente der Verteilung des dritten Moments insgesamt und nach Lautklassen ausgewertet, für normale und laute Sprache (Abweichungen über 10% nach oben sind rot und nach unten sind blau markiert.)

Phonem	SA	MW	Max	Min	Mittl.AW	St.AW	Var	Sch	Wöl
p	N	1,445	6,750	-0,601	0,719	0,974	0,950	1,536	3,098
	L	2,010	6,996	-0,334	0,990	1,186	1,407	0,619	0,000
b	N	4,143	7,234	-0,008	1,592	1,889	3,570	-0,634	-0,937
	L	4,127	6,219	0,995	1,071	1,290	1,665	-0,654	-0,524
t	N	0,480	5,591	-0,978	0,601	0,885	0,783	2,118	6,078
	L	0,591	4,516	-0,999	0,804	1,004	1,007	1,204	0,759
d	N	3,657	7,161	-0,500	1,649	1,885	3,553	-0,212	-1,274
	L	3,635	6,791	-0,928	1,531	1,819	3,310	-0,302	-0,959
k	N	0,758	3,946	-1,222	0,778	0,931	0,867	0,397	-0,429
	L	0,873	4,992	-2,072	0,828	1,022	1,045	0,350	0,034
g	N	2,317	6,764	-0,790	1,507	1,739	3,024	0,609	-0,865
	L	2,563	6,994	-1,151	1,604	1,824	3,327	0,392	-1,034
f	N	-0,048	1,321	-1,157	0,224	0,285	0,081	-0,008	0,772
	L	-0,064	1,841	-0,930	0,231	0,318	0,101	1,379	5,298
v	N	1,696	4,560	-0,791	0,533	0,803	0,645	0,406	3,861
	L	1,441	2,252	0,730	0,386	0,453	0,205	0,071	-1,300
s	N	-0,255	1,654	-1,885	0,276	0,380	0,145	-0,915	2,977
	L	-0,259	0,919	-1,400	0,222	0,290	0,084	-0,103	1,039
z	N	1,295	3,525	-1,661	0,967	1,161	1,348	-0,606	-0,786
	L	1,455	2,970	-0,759	0,663	0,883	0,779	-1,045	0,304
ʃ	N	-0,912	0,577	-2,919	0,547	0,682	0,466	-0,294	-0,423
	L	-1,177	-0,025	-3,407	0,426	0,527	0,277	-0,248	0,733
t͡s	N	0,065	1,486	-1,259	0,335	0,437	0,191	-0,102	0,471
	L	0,143	2,207	-1,100	0,423	0,554	0,307	0,784	0,702

Tabelle A.27: Momente der Verteilung des dritten Moments sämtlicher Obstruenten für normale und laute Sprache (Abweichungen über 10% nach oben sind rot und nach unten sind blau markiert.)

Phonem	SA	MW	Max	Min	Mittl.AW	St.AW	Var	Sch	Wöl
l	N	1,371	2,822	0,052	0,503	0,639	0,409	0,180	-0,327
	L	0,794	2,250	-0,406	0,300	0,421	0,177	0,603	2,090
m	N	2,173	4,035	0,977	0,563	0,713	0,508	0,578	-0,160
	L	1,317	3,806	0,374	0,453	0,635	0,403	1,667	3,993
n	N	2,175	4,637	1,030	0,520	0,702	0,493	0,861	1,361
	L	1,411	2,771	0,324	0,432	0,533	0,284	0,465	-0,269

Tabelle A.28: Momente der Verteilung des dritten Moments sämtlicher Sonoranten für normale und laute Sprache (Abweichungen über 10% nach oben sind rot und nach unten sind blau markiert.)

Phonem	SA	MW	Max	Min	Mittl.AW	St.AW	Var	Sch	Wöl
iː	N	0,891	3,599	-0,500	0,483	0,598	0,358	0,493	0,190
	L	0,102	5,925	-1,310	0,416	0,535	0,286	1,562	10,956
ɪ	N	0,796	2,986	-0,334	0,385	0,477	0,228	0,228	0,151
	L	0,219	1,862	-1,792	0,372	0,469	0,220	-0,004	0,301
eː	N	0,583	2,188	-0,595	0,400	0,491	0,241	0,416	-0,334
	L	-0,004	4,965	-1,099	0,260	0,402	0,162	4,227	51,818
ɛ	N	0,672	2,126	-0,480	0,316	0,397	0,157	0,179	0,185
	L	0,265	1,647	-1,102	0,285	0,361	0,131	-0,242	0,425
aː	N	1,448	2,873	0,319	0,300	0,375	0,140	0,431	0,050
	L	1,258	2,385	0,121	0,310	0,383	0,147	-0,105	-0,274
a	N	1,275	2,879	0,264	0,288	0,367	0,134	0,357	0,574
	L	1,115	2,414	0,039	0,303	0,375	0,140	-0,096	-0,136
ə	N	1,254	3,072	-0,259	0,437	0,541	0,292	0,223	-0,235
	L	0,687	2,275	-0,634	0,425	0,525	0,276	0,395	-0,275
ʊ	N	3,207	6,031	0,780	0,790	0,982	0,964	0,387	-0,184
	L	2,418	5,192	0,477	0,733	0,881	0,776	0,326	-0,533
uː	N	3,915	6,459	1,465	0,766	0,927	0,860	-0,106	-0,574
	L	2,835	6,393	0,621	0,718	0,869	0,755	0,258	-0,350
oː	N	3,428	6,612	1,073	0,788	0,964	0,930	0,201	-0,312
	L	2,407	5,906	0,413	0,644	0,810	0,655	0,778	0,360
ɔ	N	2,362	4,545	0,821	0,539	0,666	0,444	0,230	-0,289
	L	1,975	4,351	0,477	0,540	0,667	0,445	0,600	0,067

Tabelle A.29: Momente der Verteilung des dritten Moments sämtlicher Vokale für normale und laute Sprache (Abweichungen über 10% nach oben sind rot und nach unten sind blau markiert.)

Spk	SA	MW	Max	Min	Mittl.AW	St.AW	Var	Sch	Wöl
S02	N	1,719	6,641	-1,577	1,462	1,791	3,208	0,725	-0,452
	L	1,543	6,141	-1,680	1,310	1,552	2,408	0,565	-0,643
S03	N	1,218	7,161	-1,731	1,036	1,292	1,670	0,954	0,531
	L	1,000	5,024	-1,706	0,932	1,171	1,371	0,857	0,451
S05	N	1,500	6,242	-1,452	1,226	1,532	2,346	1,009	0,179
	L	1,127	5,441	-1,284	1,021	1,280	1,638	0,896	0,071
S06	N	1,979	5,941	-0,747	1,229	1,501	2,253	0,519	-0,584
	L	1,698	5,136	-2,047	1,129	1,346	1,813	0,377	-0,863
S10	N	1,676	7,234	-1,685	1,393	1,763	3,108	0,736	-0,257
	L	1,247	6,994	-1,400	1,165	1,488	2,213	1,022	0,709
S13	N	1,477	6,317	-1,147	1,312	1,621	2,626	0,981	-0,132
	L	1,134	6,393	-1,558	1,312	1,636	2,676	1,178	0,736
S15	N	2,035	6,612	-1,692	1,529	1,842	3,394	0,644	-0,713
	L	1,575	5,797	-1,037	1,137	1,381	1,906	0,662	-0,397
S16	N	1,570	6,764	-1,187	1,277	1,631	2,659	0,945	0,121
	L	1,476	6,791	-1,452	1,283	1,596	2,546	0,912	0,127
S17	N	1,266	5,213	-2,285	0,872	1,132	1,282	0,572	0,525
	L	0,981	6,311	-1,836	0,882	1,150	1,323	1,354	3,671
S20	N	1,650	6,132	-1,297	1,291	1,597	2,550	0,801	-0,350
	L	1,138	5,743	-1,803	0,963	1,169	1,367	0,374	-0,337
S21	N	1,315	5,833	-1,151	1,133	1,445	2,088	0,885	-0,054
	L	0,917	4,392	-2,072	1,080	1,260	1,588	0,539	-0,653
S22	N	1,244	4,466	-2,919	0,915	1,145	1,312	0,011	0,170
	L	0,887	4,077	-3,407	0,851	0,986	0,971	-0,151	-0,196
S26	N	1,156	5,609	-1,683	1,008	1,281	1,641	0,891	0,321
	L	0,954	3,736	-1,792	0,976	1,126	1,267	0,073	-1,079
S29	N	1,492	6,750	-1,899	1,097	1,418	2,012	0,623	0,291

	L	1,055	5,097	-2,242	0,988	1,193	1,424	0,374	-0,392
S30	N	1,246	5,401	-2,157	1,055	1,276	1,627	0,157	-0,305
	L	1,121	5,366	-2,208	0,937	1,140	1,299	0,183	-0,050
S31	N	1,658	6,287	-1,208	1,080	1,344	1,806	0,521	-0,404
	L	1,445	6,996	-1,190	1,033	1,273	1,619	0,619	0,026
S32	N	1,386	5,831	-1,542	1,000	1,243	1,544	0,698	0,207
	L	1,171	4,992	-0,754	0,839	1,045	1,093	0,630	0,040
S36	N	1,776	6,858	-1,264	1,134	1,429	2,041	0,683	0,033
	L	1,272	6,014	-0,645	0,935	1,177	1,385	0,835	0,789
S39	N	0,910	6,312	-1,684	0,864	1,086	1,180	0,944	1,005
	L	0,810	5,338	-1,766	0,940	1,151	1,324	0,756	0,230

Tabelle A.30: Momente der Verteilung des dritten Moments nach Sprechern ausgewertet für normale und laute Sprache (Abweichungen über 10% nach oben sind rot und nach unten sind blau markiert.)

LK	SA	MW	Max	Min	Mittl.AW	St.AW	Var	Sch	Wöl
Gesamt	N	4,163	66,247	-1,877	5,956	8,542	72,971	2,358	5,986
	L	2,429	63,623	-1,829	3,987	6,080	36,968	3,080	13,016
Obstr	N	1,738	66,247	-1,749	3,780	7,135	50,912	4,216	20,073
	L	2,651	63,623	-1,732	4,721	7,640	58,369	3,257	12,179
Son	N	3,276	23,587	-1,793	2,859	3,927	15,423	1,782	4,564
	L	0,400	15,340	-1,765	1,396	2,203	4,852	3,490	17,655
Vok	N	5,690	61,972	-1,877	6,843	9,049	81,880	1,784	3,004
	L	2,353	49,493	-1,829	3,673	5,171	26,741	2,253	6,527

Tabelle A.31: Momente der Verteilung des vierten Moments insgesamt und nach Lautklassen ausgewertet, für normale und laute Sprache (Abweichungen über 10% nach oben sind rot und nach unten sind blau markiert.)

Phonem	SA	MW	Max	Min	Mittl.AW	St.AW	Var	Sch	Wöl
p	N	2,167	49,259	-1,407	3,166	5,111	26,127	3,544	17,198
	L	4,992	53,076	-1,405	5,151	6,801	46,250	1,975	6,037
b	N	24,103	62,043	-1,188	14,594	17,120	293,080	-0,026	-1,072
	L	21,730	47,116	-0,333	10,158	11,831	139,963	-0,193	-1,042
t	N	-0,212	33,449	-1,715	1,564	3,302	10,903	5,586	38,570
	L	0,151	23,848	-1,718	1,998	3,061	9,371	3,143	12,589
d	N	18,837	66,247	-1,500	14,400	16,539	273,547	0,421	-1,022
	L	18,224	56,886	-1,558	13,089	15,438	238,344	0,495	-0,698
k	N	1,162	19,725	-1,583	2,083	2,823	7,972	2,034	5,324
	L	1,527	26,961	-1,652	2,234	3,234	10,460	2,711	10,915
g	N	8,901	57,070	-1,644	10,233	12,683	160,846	1,390	1,136
	L	10,740	63,623	-1,718	10,917	13,559	183,841	1,206	0,704
f	N	-0,943	0,907	-1,598	0,175	0,247	0,061	1,679	6,347
	L	-0,982	2,051	-1,459	0,200	0,304	0,092	3,743	28,134
v	N	2,655	24,177	-1,018	2,288	4,103	16,837	3,624	16,184
	L	1,196	4,389	-0,959	1,263	1,487	2,211	0,409	-1,088
s	N	-1,022	2,776	-1,690	0,283	0,500	0,250	3,682	18,255
	L	-0,996	1,193	-1,644	0,229	0,312	0,097	1,480	4,930
z	N	1,879	12,196	-1,749	2,255	2,762	7,629	0,863	0,345
	L	1,763	8,730	-1,491	1,772	2,164	4,683	0,521	-0,391
ʃ	N	1,273	10,979	-1,051	1,687	2,175	4,730	1,399	1,858
	L	1,545	14,008	-1,419	1,492	1,958	3,833	1,771	8,137
t͡s	N	-1,098	1,124	-1,724	0,221	0,333	0,111	2,269	9,033
	L	-0,979	4,078	-1,732	0,397	0,643	0,414	3,405	17,180

Tabelle A.32: Momente der Verteilung des vierten Moments sämtlicher Obstruenten für normale und laute Sprache (Abweichungen über 10% nach oben sind rot und nach unten sind blau markiert.)

Phonem	SA	MW	Max	Min	Mittl.AW	St.AW	Var	Sch	Wöl
l	N	1,208	8,513	-1,793	1,762	2,310	5,337	1,144	0,907
	L	-0,501	4,476	-1,765	0,671	1,027	1,055	2,389	7,472
m	N	4,631	17,909	-0,374	3,120	4,177	17,447	1,284	1,201
	L	0,963	15,340	-1,469	1,708	2,912	8,483	3,302	12,563
n	N	4,337	23,587	-0,478	2,836	4,200	17,643	2,161	6,649
	L	1,016	7,857	-1,630	1,499	1,979	3,918	1,297	1,532

Tabelle A.33: Momente der Verteilung des vierten Moments sämtlicher Sonoranten für normale und laute Sprache (Abweichungen über 10% nach oben sind rot und nach unten sind blau markiert.)

Phonem	SA	MW	Max	Min	Mittl.AW	St.AW	Var	Sch	Wöl
iː	N	-0,556	12,297	-1,877	1,032	1,435	2,059	2,421	9,776
	L	-1,365	41,781	-1,829	0,342	1,315	1,728	28,876	940,337
ɪ	N	-0,693	7,782	-1,809	0,732	0,991	0,981	2,144	9,177
	L	-1,206	2,679	-1,797	0,311	0,467	0,218	2,920	14,271
eː	N	-1,037	3,334	-1,754	0,580	0,786	0,618	1,864	3,938
	L	-1,339	27,935	-1,738	0,270	1,429	2,041	19,618	398,630
ɛ	N	-0,658	4,525	-1,742	0,573	0,772	0,596	1,815	5,438
	L	-1,008	1,885	-1,704	0,265	0,359	0,129	1,728	6,555
aː	N	2,318	11,107	-1,264	1,498	1,884	3,550	0,841	0,935
	L	1,403	9,446	-1,350	1,336	1,674	2,804	0,812	0,885
a	N	1,624	12,510	-1,359	1,324	1,691	2,860	1,044	2,573
	L	0,977	7,803	-1,381	1,139	1,396	1,949	0,643	0,114
ə	N	0,930	11,106	-1,688	1,511	1,973	3,892	1,389	2,255
	L	-0,406	5,082	-1,589	0,882	1,172	1,373	1,781	3,324
ʊ	N	12,390	53,228	-1,090	6,885	8,909	79,363	1,246	1,692
	L	6,336	31,586	-1,475	4,576	5,644	31,857	1,018	0,764
uː	N	18,756	54,138	0,849	8,048	9,770	95,450	0,503	-0,236
	L	9,063	49,493	-1,092	5,307	6,749	45,556	1,183	2,051
oː	N	14,466	61,972	-0,372	7,466	9,319	86,845	0,940	0,979
	L	6,017	42,366	-1,282	4,271	5,707	32,569	1,694	3,435
ɔ	N	6,484	24,001	-0,948	3,741	4,649	21,615	0,806	0,411
	L	3,942	24,706	-1,362	3,059	4,017	16,136	1,524	2,944

Tabelle A.34: Momente der Verteilung des vierten Moments sämtlicher Vokale für normale und laute Sprache (Abweichungen über 10% nach oben sind rot und nach unten sind blau markiert.)

Spk	SA	MW	Max	Min	Mittl.AW	St.AW	Var	Sch	Wöl
S02	N	6,294	53,424	-1,725	8,466	11,124	123,742	1,829	2,824
	L	4,615	46,900	-1,663	6,231	8,078	65,252	1,782	3,280
S03	N	2,249	66,247	-1,849	4,127	6,100	37,206	3,636	23,614
	L	1,298	30,780	-1,821	3,059	4,699	22,085	2,987	10,846
S05	N	4,047	51,346	-1,836	6,469	9,034	81,616	2,283	5,296
	L	1,972	35,554	-1,767	3,962	5,545	30,748	2,269	5,474
S06	N	7,109	47,121	-1,715	7,954	10,190	103,846	1,496	1,533
	L	4,665	30,466	-1,801	5,553	6,836	46,725	1,205	0,727
S10	N	6,009	62,043	-1,705	8,168	10,929	119,450	1,899	3,036
	L	3,169	63,623	-1,823	5,240	8,049	64,793	3,011	11,280
S13	N	4,481	47,136	-1,822	6,915	9,281	86,143	1,908	2,798
	L	3,336	49,493	-1,829	5,699	8,857	78,448	2,831	8,196
S15	N	8,534	61,972	-1,724	10,956	13,301	176,908	1,374	0,698
	L	4,052	39,025	-1,766	5,564	7,444	55,410	1,875	3,545
S16	N	5,407	57,070	-1,754	7,617	10,487	109,980	2,109	4,183
	L	4,745	56,886	-1,755	6,852	9,520	90,625	2,295	5,984
S17	N	2,428	30,117	-1,507	3,572	5,003	25,030	2,273	6,017
	L	1,478	45,416	-1,718	2,798	5,716	32,673	4,963	28,026
S20	N	5,422	47,540	-1,768	7,667	9,940	98,800	1,729	2,025
	L	1,870	35,982	-1,820	3,071	4,215	17,768	2,262	7,984
S21	N	3,488	44,138	-1,747	5,490	7,470	55,794	1,992	3,549
	L	1,615	23,207	-1,799	3,089	4,315	18,620	2,149	4,987
S22	N	2,201	25,304	-1,793	3,307	4,299	18,482	1,706	3,456
	L	0,820	19,324	-1,692	1,910	2,525	6,374	2,038	7,533
S26	N	2,253	38,860	-1,877	4,070	5,955	35,456	2,565	7,560
	L	1,541	18,059	-1,682	2,591	3,281	10,763	1,480	2,207
S29	N	3,919	49,259	-1,802	5,249	7,780	60,527	2,471	6,715

	L	1,654	30,924	-1,798	2,999	4,092	16,745	2,192	6,546
S30	N	2,633	35,171	-1,799	3,883	5,073	25,739	1,847	4,675
	L	1,683	31,339	-1,797	2,828	3,920	15,366	2,284	7,998
S31	N	4,505	51,997	-1,634	5,601	7,404	54,822	1,767	3,391
	L	3,284	53,076	-1,702	4,436	6,123	37,497	2,343	8,644
S32	N	2,548	36,041	-1,787	3,968	5,546	30,762	2,460	7,837
	L	1,390	26,961	-1,762	2,612	3,851	14,830	2,342	6,044
S36	N	5,383	59,277	-1,749	6,626	9,013	81,231	1,997	4,263
	L	2,388	43,708	-1,764	3,702	5,715	32,665	3,353	15,683
S39	N	1,015	51,500	-1,811	2,596	4,205	17,685	4,196	30,292
	L	0,919	34,815	-1,829	2,532	3,951	15,612	3,405	17,141

Tabelle A.35: Momente der Verteilung des vierten Moments nach Sprechern ausgewertet für normale und laute Sprache (Abweichungen über 10% nach oben sind rot und nach unten sind blau markiert.)

A.2 Zusammenhänge zwischen den spektralen Parametern und F_0

A.2.1 F_0

LK	SA	MW	Max	Min	Mittl.AW	St.AW	Var	Sch	Wöl
Gesamt	N	124,422	503,840	53,671	20,741	27,401	750,842	1,611	11,917
	L	176,865	493,090	53,845	38,537	48,980	2399,006	0,876	1,439
Obstr	N	117,433	418,508	55,285	17,230	25,144	632,236	2,691	25,098
	L	146,194	459,634	54,155	34,441	44,063	1941,506	1,146	3,158
Son	N	128,321	227,781	63,325	18,321	22,438	503,450	0,273	0,313
	L	205,071	345,554	121,135	28,248	36,513	1333,172	0,479	0,420
Vok	N	124,889	503,840	53,671	21,041	27,660	765,096	1,573	11,447
	L	178,639	493,090	53,845	38,243	48,612	2363,131	0,929	1,521

Tabelle A.36: Momente der Verteilung von F_0 insgesamt und nach Lautklassen ausgewertet, für normale und laute Sprache (Abweichungen über 10% nach oben sind rot und nach unten sind blau markiert.)

Phonem	SA	MW	Max	Min	Mittl.AW	St.AW	Var	Sch	Wöl
b	N	116,887	235,160	55,285	16,778	23,583	556,167	0,567	3,815
	L	153,303	320,907	56,189	35,705	44,737	2001,407	0,637	0,595
d	N	115,739	183,877	56,114	17,329	22,428	503,032	0,147	0,648
	L	144,354	274,750	56,457	31,163	37,799	1428,782	0,533	0,302
g	N	124,868	418,508	58,874	20,611	32,972	1087,173	3,332	24,950
	L	156,822	459,634	54,155	42,088	53,974	2913,204	0,921	2,472
v	N	120,070	214,002	87,195	15,835	20,255	410,251	1,036	3,181
	L	161,344	223,072	128,548	17,552	21,471	461,003	0,583	-0,210
z	N	108,833	199,388	80,530	9,736	14,080	198,249	1,754	7,941
	L	116,704	174,732	60,682	9,441	12,643	159,844	0,347	3,247

Tabelle A.37: Momente der Verteilung von F_0 für sämtliche Obstruenten normaler und lauter Sprache (Abweichungen über 10% nach oben sind rot und nach unten sind blau markiert.)

Phonem	SA	MW	Max	Min	Mittl.AW	St.AW	Var	Sch	Wöl
l	N	126,427	170,324	85,612	16,787	20,040	401,599	0,044	-0,784
	L	207,138	345,554	121,135	31,172	41,613	1731,614	0,578	0,344
m	N	128,542	177,797	89,730	17,961	21,360	456,264	-0,001	-0,926
	L	206,271	302,197	130,238	24,859	32,669	1067,248	0,174	0,394
n	N	130,229	227,781	63,325	20,278	25,945	673,120	0,462	0,893
	L	200,957	295,503	142,677	28,388	33,857	1146,283	0,425	-0,490

Tabelle A.38: Momente der Verteilung von F_0 für sämtliche Sonoranten normaler und lauter Sprache (Abweichungen über 10% nach oben sind rot und nach unten sind blau markiert.)

Phonem	SA	MW	Max	Min	Mittl.AW	St.AW	Var	Sch	Wöl
iː	N	124,575	234,184	66,593	20,301	25,554	652,998	0,932	0,729
	L	180,373	361,317	95,891	38,983	48,283	2331,264	0,955	0,433
ɪ	N	139,096	267,401	53,671	22,800	29,881	892,873	0,233	0,854
	L	200,295	430,351	53,845	40,770	50,985	2599,505	0,436	0,527
eː	N	125,182	185,175	88,525	15,351	19,342	374,107	0,524	0,095
	L	190,507	323,026	112,295	33,812	43,186	1865,032	0,851	0,145
ɛ	N	128,437	478,009	53,995	20,427	30,954	958,160	3,653	33,670
	L	180,451	321,941	57,603	34,803	43,727	1912,072	0,544	0,165
aː	N	110,807	187,776	63,755	15,376	19,044	362,670	0,878	0,653
	L	157,832	322,677	74,699	30,029	38,346	1470,388	1,085	0,940
a	N	121,681	237,478	55,247	18,218	23,558	554,980	0,269	1,225
	L	168,918	463,049	55,270	33,420	44,076	1942,730	0,805	3,021
ə	N	111,674	203,193	73,139	17,515	22,670	513,915	1,169	1,525
	L	154,228	295,968	78,754	31,507	40,703	1656,697	1,228	1,117
ʊ	N	141,027	420,607	57,128	25,515	35,593	1266,835	1,330	8,426
	L	204,971	493,090	55,345	45,734	60,972	3717,564	1,013	2,524
uː	N	126,360	207,111	79,165	20,031	24,845	617,280	0,840	0,264
	L	181,959	334,115	99,446	38,609	47,079	2216,448	0,818	0,040
oː	N	118,342	194,027	55,796	17,333	21,493	461,947	0,803	0,474
	L	168,641	324,085	92,492	34,685	43,321	1876,720	1,022	0,657
ɔ	N	128,718	503,840	56,381	20,030	28,440	808,846	2,566	30,943
	L	179,929	365,472	55,787	33,427	42,610	1815,600	0,437	0,395

Tabelle A.39: Momente der Verteilung von F_0 für sämtliche Vokale normaler und lauter Sprache (Abweichungen über 10% nach oben sind rot und nach unten sind blau markiert.)

Spk	SA	MW	Max	Min	Mittl.AW	St.AW	Var	Sch	Wöl
S02	N	124,209	214,765	56,114	16,727	21,680	470,016	0,638	1,203
	L	175,683	391,436	56,189	31,404	39,645	1571,688	0,270	0,969
S03	N	124,544	324,787	58,353	17,794	22,966	527,451	1,309	8,555
	L	158,677	237,413	69,876	23,140	28,155	792,720	-0,008	-0,396
S05	N	106,230	167,334	58,795	6,428	9,075	82,353	0,806	6,646
	L	131,889	203,236	60,682	13,843	17,748	314,996	0,420	0,821
S06	N	111,853	222,691	53,995	12,924	18,911	357,634	0,931	4,487
	L	149,507	333,177	78,334	17,804	22,366	500,218	0,712	6,375
S10	N	135,948	235,160	66,774	12,121	17,007	289,235	0,377	3,568
	L	155,019	244,362	75,779	17,451	22,238	494,524	0,407	0,939
S13	N	134,368	267,401	62,546	21,059	28,594	817,638	1,142	2,560
	L	178,158	320,501	90,397	40,384	47,125	2220,738	0,314	-0,856
S15	N	160,925	253,467	66,306	22,165	27,650	764,535	0,171	0,059
	L	239,871	330,878	59,525	42,744	49,171	2417,833	-0,461	-0,810
S16	N	133,383	420,607	55,285	25,039	31,639	1001,010	1,182	8,428
	L	165,574	434,586	57,709	39,715	47,381	2244,968	0,715	2,050
S17	N	112,054	307,285	59,109	8,421	12,608	158,957	5,270	79,920
	L	152,906	313,659	55,345	13,128	18,054	325,931	0,353	9,971
S20	N	119,436	418,508	63,501	13,975	19,916	396,634	4,815	68,493
	L	180,522	463,049	89,552	26,365	32,804	1076,129	0,967	8,971
S21	N	115,631	294,275	55,796	14,328	20,830	433,879	2,006	12,785
	L	181,143	316,594	59,386	28,262	36,336	1320,304	0,012	0,617
S22	N	132,295	246,147	55,932	15,992	22,860	522,593	0,395	3,152
	L	226,271	314,560	87,122	35,655	43,604	1901,305	-0,396	-0,370
S26	N	124,870	276,351	59,323	14,472	19,758	390,366	0,496	6,743
	L	255,565	493,090	66,796	33,225	46,087	2124,045	-0,581	3,113
S29	N	137,540	206,214	63,449	16,170	20,446	418,030	-0,734	1,358

	L	192,447	460,501	70,142	30,857	37,693	1420,796	0,717	4,943
S30	N	145,041	478,009	55,247	23,397	38,312	1467,846	2,568	20,813
	L	194,019	430,622	53,845	30,873	47,787	2283,627	0,354	4,905
S31	N	133,811	201,902	58,127	16,872	21,732	472,263	0,185	0,669
	L	153,036	420,519	64,003	22,499	29,590	875,597	1,031	9,240
S32	N	105,623	213,603	56,381	13,042	18,403	338,671	1,702	4,837
	L	170,385	461,030	57,603	38,201	49,200	2420,626	0,883	2,202
S36	N	99,918	503,840	53,671	12,094	22,081	487,590	9,266	159,225
	L	150,706	472,225	54,155	26,774	40,951	1676,969	3,065	21,718
S39	N	103,544	216,960	63,325	14,079	18,578	345,124	1,310	4,110
	L	156,374	430,351	65,096	28,681	37,756	1425,526	1,054	5,343

Tabelle A.40: Momente der Verteilung von F_0 nach Sprechern ausgewertet für normale und laute Sprache (Abweichungen über 10% nach oben sind rot und nach unten sind blau markiert.)

A.3 Veröffentlichungen

Vergleich von Merkmalsextraktionsverfahren für die automatische Sprecherverifikation bei Nichtübereinstimmung des Stimmaufwands in Trainings- und Testdaten

Corinna Harwardt

Fraunhofer FKIE, 53343 Wachtberg, Deutschland, Email: corinna.harwardt@fkie.fraunhofer.de

Einleitung

Die automatische Sprecherverifikation auf Audiodaten mit weitestgehenden Übereinstimmungen der Signaleigenschaften in Trainings- und Testmaterial liefert für viele Szenarien und Signalqualitäten bereits sehr gute Ergebnisse. Stimmen die Signaleigenschaften jedoch nicht überein, so sinkt die Erkennungsrate häufig rapide ab. Ein Fall der Nichtübereinstimmung ist die Erhöhung des Stimmaufwands in einem der Signale. Erhöht der Sprecher seinen Stimmaufwand, um beispielsweise Hintergrundgeräusche zu übertönen, so ändern sich die akustischen Eigenschaften des Sprachsignals stark. Bisher ist jedoch noch kein klares Muster zur Beschreibung dieser Veränderungen der verschiedenen akustischen Parameter gefunden worden, da die Veränderungen sprecherabhängig zu sein scheinen. Um dieses Problem speziell für die automatische Sprecherverifikation zu untersuchen, werden in dieser Arbeit bestehende Standardmerkmalsextraktionsverfahren auf ihre Leistung in einem solchen Szenario verglichen. Die Ergebnisse dieser Arbeit bilden demnach die Grundlage um bestimmte Merkmale für die weitere Nutzung in einem solchen Szenario auszuschließen und in weiteren Schritten das beste Merkmal auf seine Schwachpunkte für die Sprechererkennung bei verschiedenen Stimmaufwandsgraden zu untersuchen und gegebenenfalls zu verbessern.
Im Folgenden wird zunächst ein Überblick über den Einfluss von erhöhtem Stimmaufwand auf die akustischen Merkmale von Sprache gegeben. Dann wird das Sprechererkennungssystem vorgestellt mit dem die Tests durchgeführt wurden. Ein Schwerpunkt liegt hier auf der Vorstellung der unterschiedlichen Merkmale. Abschließend werden die Ergebnisse dargestellt und diskutiert.

Auswirkungen von erhöhtem Stimmaufwand auf das Sprachsignal

Bei der Produktion eines Sprachsignals verfolgt der Sprecher unterschiedliche Ziele. Möchte er lauter Sprechen - also den Stimmaufwand erhöhen - kann dies emotional oder durch die Kommunikationsgegebenheiten (Hintergrundgeräusche oder große Distanz) bedingt sein. In dieser Arbeit wird nur die Erhöhung des Stimmaufwands auf Grund von Hintergrundgeräuschen betrachtet.
Untersuchungen zum Stimmaufwand können perzeptiv, artikulatorisch und akustisch motiviert sein. In den Perzeptionsstudien wird untersucht ob der Sprecher sein Ziel - die Verbesserung der Verständlichkeit - trotz ungünstiger Kommunikationsgegebenheiten erreicht. Die artikulatorisch motivierten Studien analysieren wie der Sprecher sein Ziel erreicht. Während die akustischen Untersuchungen das Resultat begutachten. In der Sprechererkennung sind vor allem die akustisch motivierten Studien von Interesse. Bei der Untersuchung der Grundfrequenz zeigte sich in zahlreichen Studien (z.B. [1], [2]), dass die Grundfrequenz bei erhöhtem Stimmaufwand steigt. Die Stärke des Anstiegs lässt sich nicht klar definieren, sodass der Grad des Anstiegs möglicherweise sprecherabhängig ist [2]. Auch die Formanten und ihre Amplituden sind unter den verschiedensten Bedingungen analysiert worden. Hier ergibt sich jedoch kein klares Bild. Für F1 besteht ebenso wie für die Grundfrequenz, die Tendenz zum Anstieg (siehe z.B. [1], [3]). Für den zweiten und dritten Formanten lässt sich keine allgemeingültige Aussage machen. Es zeigt sich jedoch, dass die Veränderungen ausreichend stark sind, um die Leistung sprachverarbeitender Systeme negativ zu beeinflussen [4]. Welche der gängigen Standardmerkmale am wenigsten beeinflusst werden wird in dieser Untersuchung vorgestellt.

Das Sprechererkennungssystem

Um die verschiedenen Merkmalsextraktionsverfahren zu testen wurde ein GMM-UBM basiertes Sprecherverifikationssystem [5] verwendet. Man unterscheidet zwischen der Trainings- und der Testphase. Für beide Phasen muss vorab eine Vorverarbeitung durchgeführt werden, die eine energiebasierte Sprach-Pause-Detektion und die Merkmalsextraktion umfasst. Als Merkmalsextraktionsverfahren wurde variiert zwischen Mel-Cepstrum Koeffienten (MFCC), den Reflektionskoeffizienten der linearen Prädiktion (LPREFC) und der perzeptuellen linearen Prädiktion (PLP) als Mischung der beiden Verfahren, extrahiert mit HTK [6].
Bei der Berechnung der MFCCs wird zunächst eine Fourier Transformation durchgeführt. Anschließend wird eine Logarithmierung vorgenommen um später eine Trennung zwischen Anregung und Vokaltraktformung vornehmen zu können. Um die menschliche Wahrnehmung besser zu modellieren, werden die Frequenzen nach der Mel-Skala gewichtet. Die Frequenzen werden dann mittels Diskreter Cosinus Transformation in den cepstralen Bereich transformiert. Auf Grund der Trennung zwischen

Anregung und Vokaltraktformung enthalten die MFCCs Informationen über die Formantstruktur, nicht aber über die Grundfrequenz, sodass nur die Veränderung der Formanten durch erhöhten Stimmaufwand auf die MFCCs Einfluss nimmt, nicht aber die der Grundfrequenz.
Ein alternativer Ansatz ist die LPC Analyse. Sie folgt dem Grundsatz, dass sich die Parameter aus den vorherigen bestimmen lassen. Die Parameter Schätzung erfolgt rekursiv.
Die PLP Merkmale sind eine Kombination der beiden vorab genannten Verfahren. Eine detaillierte Beschreibung der Merkmale findet sich in [7].
Sind die Merkmale extrahiert können hiermit in der Trainingsphase ein Hintergrundmodell (UBM - Universal Background Model) und das Sprechermodell trainiert werden. Da für die einzelnen Sprecher häufig nicht viel Trainingsmaterial vorhanden ist, wird das UBM als Grundlage des Sprechermodells benutzt indem eine MAP (maximum a posteriori) Adaption des UBMs mit Sprecherdaten durchgeführt wird. Aus diesen Modellen wird dann ein Detektor zusammengestellt (siehe Abb. 1). Die Ausgabe des Sprecherverifikationssystems ergibt sich aus dem Verhältnis der logarithmischen Likelihood Werte (LLR - Log Likelihood Ratio) von UBM und Sprechermodell.

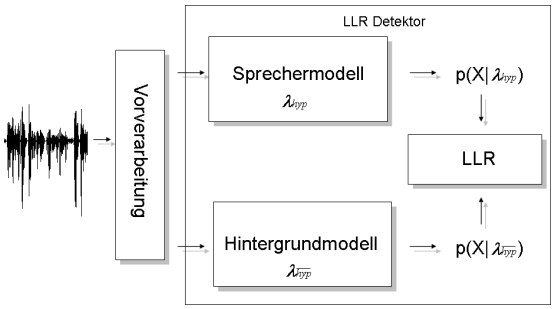

Abbildung 1: Detektor zur Sprecherverifikation

Ergebnisse

Zum Vergleich der Merkmalsextraktionsverfahren wurde das Pool 2010 Korpus verwendet [2]. Es wurden 105 Sprechermodelle auf spontansprachlichen Daten normalen Stimmaufwands mit durchschnittlich ca. 50 Sekunden Sprachanteil trainiert. Zum Testen wurden je Sprecher zwei Aufnahmen mit erhöhtem Stimmaufwand verwendet, die aus der gleichen Aufnahmesitzung stammen. Um den erhöhten Stimmaufwand hervorzurufen wurde den Sprechern weißes Rauschen per Kopfhörer zugeführt. Der durchschnittliche Sprachanteil liegt hier ebenfalls bei ca. 50 Sekunden.
Die Erkennungsergebnisse können in der DET (Detection Error Tradeoff) Kurve in Abbildung 2 abgelesen werden. Wählt man die Gleichfehlerrate (EER - Equal Error Rate) als Vergleichswert, so schneiden die MFCCs am besten ab (4,8%). Die LPC Analyse liefert mit Abstand die schlechtesten Ergebnisse (46,2%). Diese Merkmale scheinen für das gegebene Szenario nicht geeignet zu sein. Die PLP Merkmale hingegen schneiden ähnlich

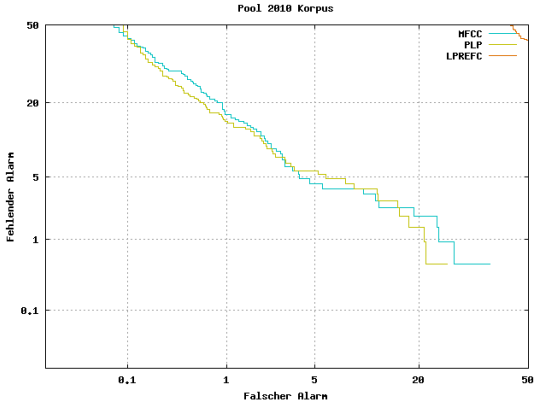

Abbildung 2: Ergebnisse verschiedener Merkmalsextraktionsverfahren auf dem Pool 2010 Korpus

gut wie die MFCCs ab (5,3%). Je nach Operationspunkt auf der DET Kurve liefern die PLPs sogar bessere Ergebnisse.

Schlussfolgerung

Die PLPs und die MFCCs scheinen gleichermaßen geeignet zu sein für Sprechererkennung mit erhöhtem Stimmaufwand. Die EER beider Verfahren ist relativ gering. Verglichen mit den Erkennungsraten für Sprachdaten gleichen Stimmaufwands (0,95% für MFCC) ist die Fehlerrate jedoch ungefähr um das Fünffache erhöht. Eine Verbesserung der Erkennungsrate durch zusätzliche oder an das Szenario angepasste Merkmale ist in nachfolgenden Studien zu untersuchen.

Literatur

[1] Schulman, R.: Articulatory Targeting and Perceptual Constancy of Loud Speech. Phonetic experimental research at the Institute of Linguistics (1985),University of Stockholm

[2] Jessen, M., Köster, O. und Gfroerer, S.: Influence of vocal effort on average and variability of fundamental frequency. International Journal of Speech, Language and the Law (2005)

[3] Liénard, J., Benedetto, M.: Effect of vocal effort on spectral properties of vowels. Journal of the Acoustical Society of America 106 (1999), 411–422

[4] Bořil H.: Robust Speech Recognition: Analysis and Equalization of Lombard Effect in Czech Corpora. Czech Technical University in Prague (2008)

[5] Reynolds, D., Quatieri T. und Dunn R.: Speaker Verification Using Adapted Gaussian Mixture Models. Digital Signal Processing 10 (2002), 19-41

[6] Young et al.: The HTK Book (for HTK Version 3.4). 2009, `http://htk.eng.cam.ac.uk/`

[7] Wendemuth A.: Grundlagen der stochastischen Sprachverarbeitung. 2004

Comparing Feature Extraction Methods for Speaker Verification with Vocal Effort Mismatch in Training and Test Data

Corinna Harwardt
Fraunhofer FKIE, Information Technology for Command & Control, Wachtberg, Germany
corinna.harwardt@fkie.fraunhofer.de

Automatic speaker verification techniques provide already very good performance in several scenarios. Especially for studio recorded speech the performance is excellent. But comparing speech samples from different recording situations or speech styles still leads to great performance degradations. One example for such a mismatch condition is different vocal effort in training and test data. This work presents investigations concerning this mismatch condition in respect to previous works (Jessen et al. 2005).

We carried out tests with a standard GMM-UBM System using MFCC features on the Pool 2010 corpus (Jessen et al. 2005). We found that the mismatch of vocal effort raises the equal error rate (EER) about 5 times (Harwardt 2010). This great performance loss is inacceptable and we therefore try to find features that are more robust. Hence, we compare the performance of different standard features for automatic speaker recognition systems in a GMM-UBM framework with each other as well as two F0 based systems. The two F0 based systems do not challenge with the standard features. Using F0 as a feature in speaker verification tasks normally produces higher error rates than standard features like MFCCs. After all the usage of F0 based features is recommendable because they can generate additional information which can not be captured with the standard features which often just capture low-level information about the vocal tract cavity. In forensics F0 features do have the additional advantage that they can be better explained to the court.

The standard features we compare are the Mel-Frequency Cepstrum Coefficients (MFCC), the reflection coefficients of the linear prediction (LPC) and the perceptual linear prediction (PLP). A detailed description of these features can be found in (Wendemuth 2004). Comparing these three types of features we find that the LPCs are not applicable to our scenario (EER: 46.2%). The PLP and MFCC features both lead to relatively good results (around 5% EER). Using just the EER as measurement of performance we find that the MFCCs outperform the other features. Therefore the MFCCs are kept as features in the baseline system. To improve the performance of this baseline system we investigate the performance of two F0 based system. One system uses F0 statistics as described in (Rose 2002). The distance between the statistics of two speech samples is calculated as weighted Euclidean distance. The second system is again a GMM-UBM system which uses logarithmic F0 values and the first derivative calculated over a five frame context. As in (Reynolds et al. 2002) we dismiss the first and last two frames of a voiced section to avoid discontinuities in the delta feature calculation. The results show that the F0 statistics are even worse than the LPC features (EER: 47%) whereas the log F0 system reaches 42% EER. In comparison to the MFCC features the degradation due to mismatch of vocal effort in percentage is lower (16.6%) but the overall performance of log F0 is much lower than the performance of MFCCs.

References

Jessen, M., O. Köster and S. Gfroerer. (2005). Influence of vocal effort on average and variability of fundamental frequency. *International Journal of Speech, Language and the Law*, 12, 174–213.

Harwardt, C. (2010). Vergleich von Merkmalsextraktionsverfahren für die automatische Sprecherverifikation bei Nichtübereinstimmung des Stimmaufwands in Trainings- und Testdaten, DAGA 2010.

Wendemuth A. (2004). Grundlagen der stochastischen Sprachverarbeitung.

Rose, P. (2002). Forensic Speaker Identification.

Reynolds, D. et al. (2002). SuperSID Project Final Report – Exploiting High-Level Information for High-Performance Speaker Recognition.

Proceedings of DiSS-LPSS Joint Workshop 2010,
25-26 September 2010, Tokyo, Japan, pp. 35-38

Investigating the COG Ratio as Feature for Speaker Verification on High-Effort Speech

Corinna Harwardt

Fraunhofer FKIE, Command and Control Information Systems, Germany
corinna.harwardt@fkie.fraunhofer.de

Abstract

Vocal effort mismatch in training and test data leads to immense degradations of speaker recognition systems. The changes on the acoustics of a speech signal induced by raised vocal effort are complex and despite several studies from various authors not completely known yet.

Instead of just gaining knowledge about these differences for automatic speaker recognition it is rather an essential to discover features that remain relatively stable in changing vocal effort conditions and contain speaker specific information. In this study we investigate the center of gravity (COG) ratio for high and mid frequency bands as feature for speaker recognition. We find that vocal effort mismatch leads to an equal error rate (EER) more than six times higher for a standard MFCC-based GMM-UBM system. For the COG ratio we observe a much smaller degradation of around 25%.

When adapting the UBM with additional high-effort speech data the EER of the COG ratio gets even better for the mismatch condition than for the matching task. Combining MFCC and the COG ratio leads to best results with an overall improvement of 16% compared to the standard MFCC-based system.

Index Terms: vocal effort, speaker recognition, center of gravity ratio

1. Introduction

Automatic speaker verification already yields good results for several tasks on spontaneous speech. However, a speaker produces many variations in spontaneous speech which can't be captured adequately with standard speaker recognition systems. Such variations might be for example disfluencies, emotions, influence of alcohol or drugs and others. The fact that often more than one of these variations occurs in spontaneous speech makes the investigation of spontaneous speech so challenging. In this paper we take into account just one variation: the change of vocal effort in spontaneous speech.

Vocal effort is the quantity a speaker raises his voice to adopt the loudness of his speech to the actual communication situation. A change of the communication situation, which induces an adjustment of vocal effort, might be for example a variation of the communication distance between the communication partners, hearing impairment, stress and other emotions or background noise. In this study we focus on high-effort speech induced by background noise, the so called Lombard speech [1]. We try to find a robust feature for a speaker verification scenario when the vocal effort does not match in training and test data. We consider normal-effort speech as training data and high-effort speech as test data. This scenario might be interesting for forensic case work, because the offenders' speech sample is often spoken with high vocal effort (e.g. when the offender makes the offence call from his mobile phone on a crowded place), whereas the suspect recordings are typically produced with normal vocal effort.

To be able to develop sufficient features for speaker verification with high-effort speech one should know which changes are induced by high vocal effort to the production and perception of speech. These differences are described in several studies, but due to the complexity and different presuppositions the results are not always consistent. Despite this we try to summarize the changes of the acoustics with focus on formants, pitch and spectral characteristics.

Changing vocal effort leads to several modifications in the acoustics of spontaneous speech. One major change is the higher F0 in high-effort speech [2]. Furthermore the formants are influenced by vocal effort, but compared to F0 these changes are not so clearly definable. The first formant increases in high-effort speech [3, 4]. For the second formant some authors don't notice any significant change [3], whereas others discover individual changes per phoneme [5]. Similarly some authors observe individual changes for F3 [6], other do not find significant changes [3, 4] and some report a shift of F3 to around 2600 Hz [7].

The distribution of energy in the speech spectrum has been part of different studies, too. Concerning the energy most authors go confirm with each other. They describe a tendency to shift the energy from the lower to the mid and higher frequencies in high-effort speech (e.g. [3]). This energy migration leads to changes of spectral features. Spectral tilt, which can be described as the slope of the spectral distribution, can be used to distinguish between vocal efforts [8]. Other studies focus on variations of the energy ratio, spectral balance or other spectral features. The COG which represents the weighted mean of the spectrum is further described in [9, 3]. The spectral center of gravity (COG) calculated over the whole spectrum is influenced by vocal effort too and might be additionally an indicator for stressed speech.

The modifications induced by change of vocal effort affect the traditional cepstral based features in speaker recognition [10]. By this study we want to address this problem and evaluate whether the COG ratio is suitable as feature for speaker verification with vocal effort mismatch.

We first describe the COG ratio as feature for speaker verification and our motivation to use the COG ratio. Then we describe the experiments including the corpora we used for our tests, the setup and the results. The results are divided into three subsections according to the different feature variations. Finally we draw some conclusions and give future perspectives.

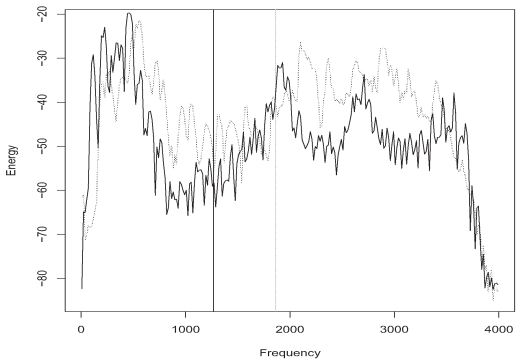

Figure 1: *Spectrum of the vowel [ə] pronounced with normal (black) and high (grey) vocal effort by the same speaker. The COG is marked as a vertical line for each spectrum.*

2. Center of Gravity Ratio

The spectral COG of a speech signal represents the weighted mean frequency of the spectrum. As in [9] we calculate the COG by

$$COG = \sum f_i * E_i / \sum E_i \quad (1)$$

with f_i representing the frequency and E_i standing for the spectral power as a function of the frequency. For energy migrations from low to high frequencies, as observed for high-effort speech, [3] the COG changes. Figure 1 illustrates the changes of COG in the speech spectrum of the vowel [ə].

As mentioned earlier we don't want to find features that change with vocal effort. Therefore we do not calculate the COG on the whole spectrum. As we know that the high and mid frequency components of the spectrum both get enriched in high-effort speech, we divide the ratio of the COG of high frequencies by the COG of mid frequencies. The choice of the frequency bands is motivated by the formant locations and by the fact that most speaker recognition applications must use band limited telephone speech. Hence the high frequency band contains frequencies from 2200 - 3000 Hz whereas the mid frequency band covers the frequency range from 800 to 2200 Hz. Figure 1 presents an example for the raise of the energy in the two frequency bands for the vowel [ə].

For the use of the COG ratio as feature in speaker verification systems we calculate the delta and delta delta features for each COG ratio over a five frame context. The resulting three dimensional vector is used as feature vector for the verification task.

3. Experiments

3.1. The Corpora

The data used in this study derives from the Pool 2010 corpus [2]. The Pool 2010 corpus contains audio data from 105 male native speakers of German. For each speaker four audio recordings are available. The different modes recorded cover read and spontaneous speech, each combined with the two modes normal speech and speech with increased vocal effort. Increase of vocal effort was induced by exposing 80 dB white noise to the speakers via headphones. For this study we used spontaneous speech with normal and high vocal effort transmitted via GSM. The data has been divided into a development set of 55 speakers and a test set of 50 further speakers. For training we used one recording per speaker containing about 50 seconds of normal-effort speech. For the tests with high-effort speech we had two recordings, each containing again around 50 seconds of speech. To get an impression of the systems performance loss due to vocal effort changes we needed another test set with normal-effort speech. This test set contained one recording per speaker, again with 50 seconds speech.

For the UBM (universal background model) we used additional data from the Kiel corpus [11]. The Kiel corpus consists of normal-effort speech. Both corpora contain German speech.

3.2. Experimental Setup

To test the COG ratio as feature for speaker verification we used a statistical framework based on gaussian mixture models (GMM) as proposed by [12]. First we trained an UBM that represents the speech and language from the speaker population under consideration, in this case German male speakers. To train this UBM we used the male speakers from the Kiel corpus. Next we adapted the speaker models from the well trained UBM with normal-effort speech from each of the 50 speakers from the Pool 2010 corpus. Lastly we ran the tests on the doubtful data. These doubtful data were either normal- or high-effort speech samples from the Pool 2010 corpus.

In this study we varied the features which are needed to train the models and to run the verification. The tests we ran are:

- Tests with standard MFCC features. The MFCC feature vector contained 15 MFCC and their first and second order derivate.
- Test with the COG Ratio. When using the COG ratio as feature we calculated a three dimensional feature vector with the COG ratio and the first and second order derivate.
- We extended our UBM training data by additional high-effort speech and adapted the existing UBM with this data. The adaptation data consisted of the development set from the Pool 2010 corpus.
- Additionally we fused the results from both systems to see whether we can improve the overall performance.

The results from these tests are described in the next sections.

3.3. Results

In the next subsections we will illustrate the experimental results with different features. First of all we describe the results with the standard MFCC features followed by the results with the newly proposed COG ratio and the combination of both features. The results of the adaptation are not listed separately. They are included in the description of the single features.

3.3.1. MFCC

In this subsection we describe MFCC as features for speaker verification with changing vocal effort conditions. We included the MFCC into a GMM-UBM system with 1024 mixture components. For the training procedure we used normal-effort audio data. As depicted in table 1 we reach an EER of 0.57% for normal-effort speech in training and test data. Usage of high-

Table 1: *EERs for the MFCC-based system.*

vocal effort	EER
normal-effort	0.57%
high-effort	4%
high-effort (adapted)	4%

effort speech as test samples raises the EER to 4%. Doing a MAP (maximum a posteriori) adaptation of the UBM with additional high-effort speech from the Pool 2010 corpus before training the speaker models does not improve the performance of the system (see table 1).

3.3.2. COG Ratio

Usage of the COG ratio is motivated by the fact that mid and high frequencies (as defined in section 2) in the speech spectrum get enriched if a speaker raises his vocal effort. Our underlying hypothesis for using this feature is that the ratio of the COG of these frequency regions seems to be relatively stable and additionally the COG ratio seems to have a great inter-speaker variability. If this thesis is correct the COG ratio would be a good feature for this speaker verification scenario. In table 2 we plotted the results of our tests. The standard GMM-UBM

Table 2: *EERs for the COG ratio based system. The number of mixture components is plotted in the rows and the kind of test data in the columns. For training we used always normal-effort audio data.*

	normal-effort	high-effort	high-effort (adapted)
64	24%	28%	23%
128	22.49%	28%	22%
256	24.65%	28%	23%
512	22.82%	29%	23%
1024	22%	29%	24.08%

systems operating with MFCC use 1024 or 2048 mixture components. Of course the feature vectors used in these systems are much larger than the features used for COG ratio, which have just three elements. Therefore we tried different numbers of mixture components and found a number of 128 mixtures to give best results. As shown in table 2 the system operating with 1024 components yields better results for normal-effort speech, but it is worse for high-effort speech. Additionally it has the disadvantage of a longer training procedure.

When comparing the results of normal and high-effort test samples we see a degradation of the performance for all the system variants. However, the degradation is not as great as for the MFCC-based system which has a more than six times higher EER for high-effort test speech. Hence we conclude that this feature is relatively robust against vocal effort mismatch. When we compare the EERs to those of the MFCC-based system we observe a much higher EER for the COG ratio. One reason for this might be the number of elements of feature vectors. To achieve better results with the COG ratio we should use them in combination with other features. One solution might be the combination of MFCC and COG ratio (see next subsection 3.3.3).

We observed the best results when adapting the UBM with additional high-effort speech data before speaker model training. This adaptation has the advantage that we don't need high-effort speech per speaker as proposed in [13]. We can rather use an extended UBM to get better results. For 128 mixtures the results on high-effort speech do even get better than the results on normal-effort test speech.

3.3.3. Fusion

To test whether the COG ratio can improve the performance of the standard MFCC-based system we fused the scores of both systems. For these merged scores additional audio data for training is needed, which should be as similar as possible to the data used later on. On the other hand the data used for other development purposes (e.g. the MAP adaptation of the UBM) should not be reused to train the fusion of scores. In the context of this work only the results of the high-effort test samples are regarded as important. Therefore we used the normal-effort speech samples as training data for merging the high-effort scores. We utilized the FoCal toolkit[1] for the fusion. The process is a linear fusion as described by [14].

The EERs of the single, as well as the fused features are presented in table 3. The MFCC system always operates with

Table 3: *Comparison of the EERs of the different systems on high-effort speech.*

Features	EER
MFCC	4%
COG Ratio	28%
COG Ratio (adapted)	22%
MFCC + COG ratio	3.35%
MFCC + COG ratio (adapted)	4%

1024 mixtures, whereas the COG ratio is used in conjunction with a 128 mixture model. As you can see in table 3 the fusion outperforms the system based on the COG ratio as well as the standard MFCC system. During the fusion process the scores get simultaneously calibrated by the FoCal toolkit. To ensure that the improvement doesn't originate in the previous calibration, we calibrated the scores of the MFCC-based system. Because of quantitative lack of training data (the same as for the fusion) it was not possible to perform a sufficient calibration.

Using scores from the adapted COG ratio system does not lead to an improvement over the MFCC alone. To provide a better performance overview, figure 2 plots the results as DET curves (detection error trade off). The systems operating on the COG ratio only are worse than those using the MFCC. The best performance is achieved by the system which makes use of both the MFCC and the COG ratio.

4. Conclusion

We evaluated the COG ratio as feature for speaker verification on high-effort speech. For this we used a GMM-UBM speaker verification framework with MFCC features as baseline.

Next we included the COG ratio and the first and second order derivate as features in this framework. We observed that compared to the MFCC features the COG ratio was relatively stable to changes in vocal effort.

We were able to further improve our results by adapting the UBM with additional high-effort speech data. By this MAP adaptation we gained better results in high-effort speech compared to the baseline test with normal-effort test samples. Hence

[1] http://www.dsp.sun.ac.za/ nbrummer/focal/

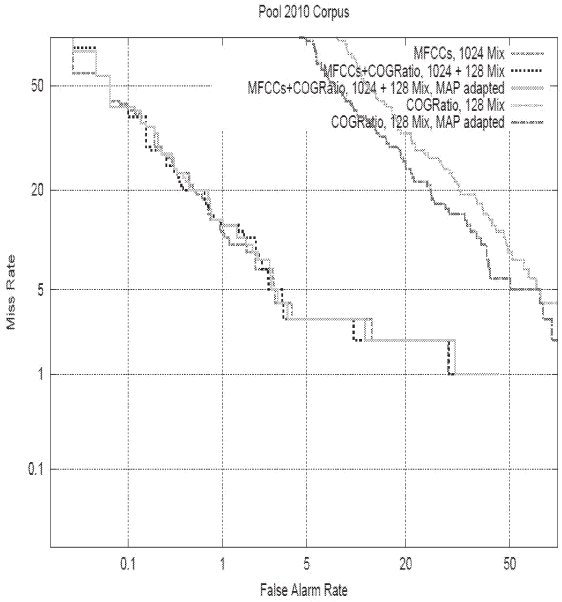

Figure 2: *DET curve for MFCC, COG Ratio, the fusion of both features and the adapted COG ratio.*

we conclude that the degradation due to vocal effort mismatch for COG ratio can be compensated by MAP adaptation of the UBM. The great advantage is that the user doesn't need additional high-effort speech for each enrolled speaker. He just needs additional data from some other speakers. For the MFCC the MAP adaptation of the UBM wasn't successful.

Comparing the EERs of MFCC and the COG ratio we observe that the performance of the COG ratio is much lower although it is stable to vocal effort changes. One reason might be that the MFCC vector consists of 45 components whereas the COG ratio vector has just three elements. In future work we will try to find further features which are stable to vocal effort changes and have a high inter- and low intra-speaker variability. These features could be used to extend the feature vector.

One other possibility is to fuse the scores generated with the COG ratio with other systems' results. We presented a linear fusion of MFCC and COG ratio in this paper. Linear fusion of both systems scores' yielded best results compared to the single systems. We could reach an improvement of around 16% compared to the standard MFCC-based system.

As future work tests with female speakers should be taken into account. Especially tests with both genders mixed in the training and test data are challenging with high-effort speech because male speakers are often identified as female person due to the raised F0. We need to check whether the COG ratio is robust against such gender-based false alarms.

5. References

[1] Lombard, Étienne. "Le signe de l'lvation de la voix", Ann. Mal. Oreil. Larynx 37, 1911, S. 101-119.

[2] Jessen, M. and Köster, O. and Gfroerer, S., "Influence of vocal effort on average and variability of fundamental frequency", International Journal of Speech, Language and the Law, 12:174–213, 2005.

[3] Liénard, J-S. and Di Benedetto, M-G., "Effect of vocal effort on spectral properties of vowels", J. Acoust. Soc. Am., 106(1): 411–422, 1999.

[4] Schulman, R., "Articulatory Targeting and Perceptual Constancy of Loud Speech", Phonetic experimental research at the Institute of Linguistics, University of Stockholm, 1985.

[5] Bond, Z. S. and Moore, T. J. and Gable, B., "Acoustic-phonetic characteristics of speech produced in noise and while wearing an oxygen mask", J. Acoustic Soc. Am., 85 (2): 907-912, 1989.

[6] Stanton, B. J. and Jamieson, L. H. and Allen, G. D., "Acoustic-Phonetic Analysis of Loud and Lombard Speech in Simulated Cockpit Conditions", ICASSP 88, New York City, 1988.

[7] Geumann, A., "Vocal intensity: acoustic and articulatory correlates", Conference on Motor Control, Nijmegen, 2001.

[8] Zhang, C. and Hansen, J.H.L., "Analysis and Classification of Speech Mode: Whispered through Shouted", Proc. Interspeech, pp.2289-2292, 2007.

[9] van Son, R. J. J. H. and Pols, L. C. W., "An acoustic description of consonant reduction", Speech Communiction, 28: 125–140, 1999.

[10] Shriberg, E. and Graciarena, M. and Bratt, H. and Kathol, A. and Kajarekar, S. and Jameel, H. and Richey, C. and Goodman, F., "Effects of Vocal Effort and Speaking Style on Text-Independent Speaker Verification", Proc. Interspeech, pp. 609612, Brisbane, Australia, 2008.

[11] Kohler, K. J. (editor), "Arbeitsberichte (AIPUK) Nr. 29", Institut für Phonetik und digitale Sprachverarbeitung, Universität Kiel, 1995.

[12] Reynolds, D. and Quatieri, T. and Dunn, R., "Speaker Verification Using Adapted Gaussian Mixture Models.", Digital Signal Processing, 10: 19–41, 2002.

[13] Hansen, J.H.L. and Varadarajan, V., "Analysis and Compensation of Lombard Speech Across Noise Type and Levels With Application to In-Set/Out-of-Set Speaker Recognition", IEEE Transactions on Audio, Speech and Language Processing, 17(2):366–378, 2009.

[14] Brummer, N. and du Preez, J., "Application-Independent Evaluation of Speaker Detection", Computer Speech & Language, 20 (2-3): 230–275, 2005.

Investigations on the speech spectrum of normal and loud speech

Corinna Harwardt
Fraunhofer FKIE, Wachtberg, Germany
corinna.harwardt@fkie.fraunhofer.de

Speaking up induces various modifications to acoustic characteristics of speech. In this paper we investigate the impact of raised vocal effort on the speech spectrum. In particular, we look at different spectral parameters and compare the changes we observe for each. The parameters we take into account are spectral tilt, spectral center of gravity, energy ratio and spectral moments. We carry out tests on the complete data set with all phonemes pooled into one distribution and tests with the data divided into the three phoneme classes vowels, sonorants and obstruents.

For our investigation we used the Oldenburger Logatome (OLLO) speech corpus (Wesker, 2005) which consists of data from 40 German speakers with four dialects (standard German, Bavarian, Eastfrisian, East Phalian). The available vowel phonemes are /a, aː, e, ɛ, i, ɪ. o, ɔ, u, ʊ, ə/. Furthermore /b, p, t, g, k, f, v, s, ʃ, z/ are the available obstruents and /l, m, n/.
To calculate spectral tilt (ST) we computed the regression line of the speech spectrum. The center of gravity (COG) is defined as in Son et al. (Son, 1999). The computation of the energy ratio was done as proposed by Wenndt et al. (Wenndt, 2002). Spectral moments (Mom1-Mom4) have been calculated according to Forrest et al. (Forrest, 1988). While examining the results, a special focus will be set on the comparison of spectral tilt and third moment as well as on center of gravity and first moment, because those parameters describe the same characteristics of the speech spectrum.

To compare spectral parameters of normal and loud speech we calculated the mean values of each parameter for both kinds of vocal effort (see Table 1). We marked those values with modifications of mean values greater than 10%. When we look at the complete distribution, which includes data from all phoneme classes, we see that all parameters, except the second spectral moment, are changed more than 10%. For sonorants and vowels we observe the same trend. The values of these two classes are all changed more than 10%. The obstruents are affected much lesser. For obstruents only spectral tilt as well as the third and forth spectral moment suffer from greater modifications.

To verify our observations we carried out Mann-Whitney tests (Hollander, 1999). Assuming $\alpha = 0.05$ ($z = 0 \pm 1.95996$) as significance level, we observe significant differences for all classes and parameters (see Table 2). Although we didn't find greater differences for the mean values of most spectral parameters of obstruents we monitor significant changes. Comparing the z-values of the different classes we see that the obstruents' values are much smaller than those of the other classes. Hence, we conclude that obstruents are less affected by changes of vocal effort than sonorants and vowels. Especially for vowels the z-values are great.
The parameters less affected are energy ratio and second spectral moment. Comparing the center of gravity and the first spectral moment we don't monitor great differences, whereas the comparison of spectral tilt and third moment shows great differences. A detailed analysis of results can be found in (Harwardt, 2011).

Table 1: Average mean values for different spectral parameters for normal (N) and loud (L) speech over the complete distribution as well as the vowel, sonorant and obstruent distributions. Changes of 10% or more are marked with an arrow downwards for decreasing values and upwards for increasing values.

		Complete Distribution		Vowels		Sonorants		Obstruents	
ST	N	-40.915	↑	-64.345	↑	-59.051	↑	-2.635	↑
	L	-34.184		-50.272		-36.214		-1.883	
COG	N	1123.806	↑	916.068	↑	733.178	↑	1470.844	
	L	1260.368		1203.671		1026.592		1381.986	
ER	N	0.705	↑	0.401	↑	0.258	↑	1.210	
	L	0.794		0.576		0.448		1.243	
Mom1	N	1131.599	↑	923.876	↑	740.988	↑	1478.613	
	L	1268.147		1211.479		1034.401		1389.707	
Mom2	N	821856.848		748854.9	↑	688175.7	↑	943691.6	
	L	863812.073		839627.5		950578.3		909285.8	
Mom3	N	1.482	↓	1.934	↓	1.875	↓	0.740	↑
	L	1.182		1.301		1.136		0.944	
Mom4	N	4.163	↓	5.690	↓	3.276	↓	1.738	↑
	L	2.429		2.353		0.400		2.651	

Table 2: Z-values obtained by the Mann-Whitney test for different spectral parameters over the complete audio data as well as the vowel, sonorant and obstruent distributions.

Parameter	Complete Distribution	Vowels	Sonorants	Obstruents
ST	15,9159	35,1518	10,9377	9,4189
COG	25,3868	48,2942	10,3844	-6,5763
ER	11,7371	25,007	3,7133	-2,7387
Mom1	25,3836	48,2937	10,3844	-6,5796
Mom2	9,8721	19,4614	9,1243	-4,2063
Mom3	-17,0732	-32,1066	-9,952	5,2644
Mom4	-16,7922	-27,3897	-9,8523	7,3907

References

Wesker, T., Meyer, B., Wagener, K., Anemüller, J., Mertins, A., Kollmeier, B. (2005): Oldenburger logatome speech corpus (OLLO) for speech recognition experiments with humans and machines. Interspeech.

Son, R. J. J. H. ; Pols, L. C. W.(1999): An acoustic description of consonant reduction. In: Speech Communication 28, p. 125–140.

Wenndt, S. J. ; Cupples, E. J. ; Floyd, R. M.(2002): A Study on the Classification of Whispered and Normally Phonated Speech. In: International Conference on Spoken Language Processing.

Forrest, K., Wiesmer, G., Milenkovic, P., Dougall, R. (1988): Statistical analysis of word-initial voiceless obstruents: Preliminary data. *J. Acoust. Soc. Am.* 84(1), p. 115-123.

Hollander, M., Wolfe, D. A. (1999), "Nonparametric Statistical Methods", John Wiley Sons, Inc..

Harwardt, C. (2011): "Comparing the Impact of Raised Vocal Effort on Various Spectral Parameters", Interspeech.

Comparing the Impact of Raised Vocal Effort on Various Spectral Parameters

Corinna Harwardt

Fraunhofer Institute for Communication, Information Processing and Ergonomics, Wachtberg, Germany
corinna.harwardt@fkie.fraunhofer.de

Abstract

Vocal effort changes induce various modifications to acoustic characteristics of speech. In this paper we investigate the impact of raised vocal effort on the speech spectrum. In particular, we look at different spectral parameters and compare the changes. The parameters we take into account are spectral tilt, spectral center of gravity, energy ratio and spectral moments. We carry out tests on the complete data set with all phonemes pooled into one distribution and tests with the data divided into three phoneme classes. Furthermore we run vocal effort classification tests to verify our results from statistical analysis. The results indicate significant changes for all parameters on the complete distribution as well as for vowels and sonorants. For obstruents we observe significant changes, too. But the modifications are much smaller than those for the other two phoneme classes. The parameters that are less affected by raised vocal effort are energy ratio and second spectral moment.

Index Terms: vocal effort, spectral tilt, spectral moments, spectral center of gravity, energy ratio

1. Introduction

The acoustic, articulatory and perceptional characteristics of speech are influenced by changes of vocal effort. Vocal effort can be defined as the quantity a speaker changes his voice when he adopts it to the demands of the communication situation. Such an adaptation might for example be necessary due to background noise or great distance to the communication partner. High vocal effort is associated with loud speech and low vocal effort corresponds to soft speech.

The influence of vocal effort on acoustic, articulatory and perceptional parameters has been investigated in several studies (e.g. [1, 2]). In this work we focus on acoustic changes, in particular on changes of the speech spectrum. In general, raising vocal effort leads to a shift of frequencies from lower to higher frequency ranges [3]. This results in modifications of several spectral features which rely on those frequency ranges. Standard features which characterize the spectrum are for example spectral tilt, spectral balance or spectral moments. These and other features are often described in literature (e.g. [3, 4]), but they are not always defined in the same way. Hence, the different studies are not comparable. Furthermore the preconditions of the investigations differ. For example some use Lombard speech[1] [1] whereas others use loud speech induced by instructing the speaker to raise his voice [2] or by letting him get over a distance to communicate with someone [3]. Other differences might for example be obtained by variation of recording settings, amount of data, gender of speaker, or usage of different parameters.

[1] Lombard speech is generated when a speaker adopts his vocal effort to background noise

To overcome this problem of non-uniform preconditions we decided to set up a comparison of *spectral tilt, spectral center of gravity, energy ratio* and *spectral moments*. One focus will be the comparison of first spectral moment and center of gravity as well as of third spectral moment and spectral tilt. The analysis is made on the complete distribution including all phonemes and, for deeper analysis, on the three phoneme classes vowels, sonorants and obstruents separately.

To describe our experiments we first introduce the methods and data we used (see section 2). Next we present the results in section 3, which are divided into the description of experiments on the complete data distribution and on the data divided into three phoneme classes. In section 4 we verify our results from section 3 with the help of an automatic vocal classification framework. At last we draw a conclusion and suggest future work.

2. Method

In this section we describe the preconditions for our experiments including the data base, the calculation of spectral parameters and the statistical analysis.

2.1. The Corpora

In our experiments we used two corpora. For the statistical analysis in section 3 we used the Oldenburger Logatome (OLLO) speech corpus which consists of data from 40 German speakers with four dialects. It provides standard German, Bavarian, East Frisian and Eastphalian. The speakers had to articulate logatomes. Logatomes are three-phoneme sequences with no semantic information. In the OLLO corpus logatomes are realized as different vowel-consonant-vowel or consonant-vowel-consonant combinations. The database includes speaker-dependent variabilities like age, gender and dialect. Furthermore it contains speaker-independent variabilities like speaking rate, speaking effort and speaking style [5].

For our investigation we used the normal- and high-effort speech samples from male speakers. The high-effort speech was invoked by instructing the speaker to raise the voice. For one of our experiments the data has been divided into three phoneme classes: vowels, sonorants and obstruents.

The available vowel phonemes are /a, aː, e, ɛ, i, ɪ, o, ɔ, u, ʊ, ə/. Plosives and fricatives are pooled together to one phoneme class, the obstruents. The obstruents in the OLLO corpus are /b, p, d, t, g, k, f, v, s, ʃ, z/. As sonorants /l, m, n/ were available.

To certify the results of the statistical analysis we accomplished a vocal effort classification which is done with the Pool 2010 corpus [6]. The Pool 2010 corpus contains spontaneous speech from 106 german male and one female speaker. The data from 57 speakers were taken as training data, whereas the others served as test data. For the test each speaker had 2 recordings, one with normal and one with high vocal effort. Overall we had 100 test trials for the vocal effort classification.

2.2. Spectral Parameters

The speech spectrum represents the energy of a signal at different frequencies. When a speaker raises his voice, the energy distribution in the speech signal changes. To investigate the changes on the spectrum induced by raised vocal effort, we have a closer look at different spectral parameters in normal and high effort speech.
For all spectral parameters we compute a 512-point fast Fourier transform.
As a first parameter we compute *spectral tilt (ST)*, which is defined in literature quite differently depending on the author [1, 7]. In this study we calculate the regression line over the speech spectrum and use the slope as measurement for spectral tilt.
The next spectral feature we look at is *spectral center of gravity (COG)*. The COG represents the weighted mean of the speech spectrum. To calculate the COG we use the numeric definition as proposed in [8].
Another feature we investigate is *energy ratio (ER)*, which was used by Wenndt et al. [9] and Zhang and Hansen [10] to distinguish between different degrees of vocal effort. The energy ratio results from the ratio of the energy of a high frequency band (2800-3000 Hz) to the energy of a low frequency band (450-600 Hz).
The last spectral parameters we take into account are *spectral moments* as defined by Forrest et al. [11]. We compute the first four moments, which are associated with mean, variance, skewness and kurtosis. One focus when comparing the features later on is the comparison of first moment and COG as wells as the comparison of spectral tilt and third moment. We will compare whether the different calculation methods lead to different results. Spectral moments will be abbreviated with *Mom1-Mom4* or with *Moments* for all four moments.

2.3. Statistical Analysis

To evaluate our experiments we compared the mean values of the distributions and made statistical significance tests. Due to the fact that most of the distributions are not normally distributed we picked a non-parametrical test, the Mann-Whitney test [12]. As significance level we defined a = 0.05 ($Z=0\pm1,95996$).

3. Comparison of Spectral Features

In this section we describe the experiments we run on differently sorted data sets. For each degree of vocal effort we first pooled together all data to one complete distribution. Then we divided the resulting two complete distributions each into three phoneme classes to further analyze the differences between normal and loud speech.

3.1. Complete Distribution

At first we look at the mean values of the complete distributions for all spectral parameters. They are displayed in Table 1.
We see that nearly all spectral parameters change more than 10% when a speaker raises vocal effort. Only the mean value of the second spectral moment is not changed that much. The means of third and forth moment decrease whereas the other mean values increase. Comparing the first spectral moment and the COG, which both represent the weighted mean of the spectrum, we see, as expected, the same trend. Both are increased more than 10%. Overall the values for both parameters do not differ much. Interestingly the third moment, which is associated with the spectral tilt, does not show the same trend as the spectral tilt. The third moment gets decreased and the tilt increased. The values for both parameters are completely different. Hence, we conclude that these parameters are not redundant and further tests need to be done to check whether additional information is included.

Table 1: *Average mean values for different spectral parameters over the complete distribution for normal (N) and loud (L) speech. Changes of 10% or more are marked with an arrow downwards for decreasing values and upwards for increasing values.*

Parameter		Complete Distribution	
ST	N	-40.915	↑
	L	-34.184	
COG	N	1123.806	↑
	L	1260.368	
ER	N	0.705	↑
	L	0.794	
Mom1	N	1131.599	↑
	L	1268.147	
Mom2	N	821856.848	
	L	863812.073	
Mom3	N	1.482	↓
	L	1.182	
Mom4	N	4.163	↓
	L	2.429	

To further analyze the difference between loud and normal speech we carried out Mann-Whitney tests for each of the spectral parameters. We listed the resulting Z-values in Table 2. The changes induced by raised vocal effort are significant for all spectral parameters including the second moment, which showed differences lower than 10% for the average mean value. According to this low difference one can see in Table 2 that the second moment shows lowest significance. The Z-values of the parameters with decreasing means are negative and those of the others positive.

Table 2: *Z-values obtained by the Mann-Withney for different spectral parameters over the complete audio data.*

Parameter	Complete Distribution
ST	15.9159
COG	25.3868
ER	11.7371
Mom1	25.3836
Mom2	9.8721
Mom3	-17.0732
Mom4	-16.7922

3.2. Analysis of Phoneme Classes

The results on the complete distribution indicate great spectral changes in high-effort speech. We want to go deeper into examination of the data in considering three phoneme classes instead of the complete distribution. The average mean values

of the spectral parameters for loud and normal speech over all phoneme classes are presented in Table 3. For vowels and sonorants we observe the same patterns. The third and forth spectral moment are decreased more than 10% whereas all other parameters get increased more than 10%. This confirms the results obtained for the complete distribution, except for the second spectral moment. The second spectral moment changes more than 10% for vowels and sonorants. As for the complete distribution, we monitor again same changes for COG and the first moment and different trends for spectral tilt and the third moment.

Table 3: *Average mean values for different spectral parameters over three phoneme classes for normal (N) and loud (L) speech. Changes of 10% or more are marked with an arrow downwards for decreasing values and upwards for increasing values.*

Parameter		Vowels		Sonorants		Obstruents	
ST	N	-64.345	↑	-59.051	↑	-2.635	↑
	L	-50.272		-36.214		-1.883	
COG	N	916.068	↑	733.178	↑	1470.844	
	L	1203.671		1026.592		1381.986	
ER	N	0.401	↑	0.258	↑	1.210	
	L	0.576		0.488		1.243	
Mom1	N	923.876	↑	740.988	↑	1478.613	
	L	1211.479		1034.401		1389.707	
Mom2	N	748854.9	↑	688175.7	↑	943691.6	
	L	839627.5		950578.3		909285.8	
Mom3	N	1.934	↓	1.875	↓	0.740	↑
	L	1.301		1.136		0.944	
Mom4	N	5.690	↓	3.276	↓	1.738	↑
	L	2.353		0.400		2.651	

The obstruents behave completely different. The third and forth moment which decrease for vowels, sonorants and the complete distribution get increased more than 10% for obstruents. The spectral tilt is increased, too. The other mean values do not show modifications greater than 10%. Especially the mean value of the energy ratio seems to be just little affected by raised vocal effort.

For obstruents the parameter pairs first spectral moment and COG as well as third moment and spectral tilt each show same modification. This is different to the findings of the other distributions.

Looking at the Mann-Whitney tests for different phoneme classes (see Table 4) we find again significant differences for all measured parameters. The Z-values of obstruents are less significant than those of sonorants and vowels. This supports our findings of changes in average mean values. Again, the energy ratio stands out, because the Z-values for obstruents and sonorants are closest to the threshold $Z = 0 \pm 1,95996$ and hence, less modified than the other spectral parameters.

Table 4: *Z-values obtained by the Mann-Whitney test for different spectral parameters over three phoneme classes.*

Parameter	Vowels	Sonorants	Obstruents
ST	35.1518	10.9377	9.4189
COG	48.2942	10.3844	-6.5763
ER	25.007	3.7133	-2.7387
Mom1	48.2937	10.3844	-6.5796
Mom2	19.4614	9.1243	-4.2063
Mom3	-32.1066	-9.952	5.2644
Mom4	-27.3897	-9.8532	7.3907

When looking at Table 4 we see that the COG is affected the most by changing vocal effort. To illustrate these modifications we plotted the distributions of COG for all three phoneme classes in Figure 1. The first plot illustrates the obstruents' COG values. Although the obstruents are less influenced by vocal effort for all spectral parameters investigated in this paper, we observe some changes of the distribution for loud speech. But, as expected, the modifications are not great. The overall shape of the two distributions stays the same, too.

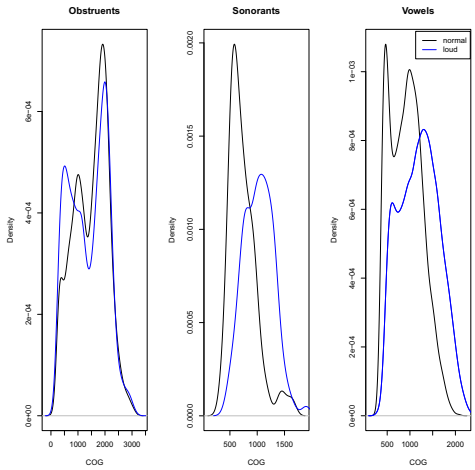

Figure 1: COG distributions for the phoneme classes

In contrast to the obstruents we expect great differences for vowels and sonorants. This is supported by the other two plots of sonorants and vowels. For vowels the shape of the curve does not change much in loud speech, but for sonorants we monitor a modification of the shape with different skewness for normal and loud speech.

After presenting the parameter with most changes we now demonstrate the parameter with fewest changes in Figure 2. As you can see, changes induceed by loud speech are very small for the energy ratio. Again obstruents are less affected than sonorants and vowels. Although the changes are significant for obstruents, we do not monitor great differences between the two distributions of the energy ratio. For all distributions of the energy ratio we observe similar curves with a leptocurtic shape and just one peak. For sonorants and vowels we see modifications when comparing normal and loud speech. But these modifications are much smaller than those of the COG (see Figure 1).

4. Automatic Classification of Vocal Effort with Spectral Parameters

As shown in the previous section spectral parameters seem to be able to distinguish between different degrees of vocal effort. We have seen that some parameters deliver better discrimination between normal and loud speech than others. In this section we will present the usage of the described spectral features in vocal effort classification. We will check whether the findings from

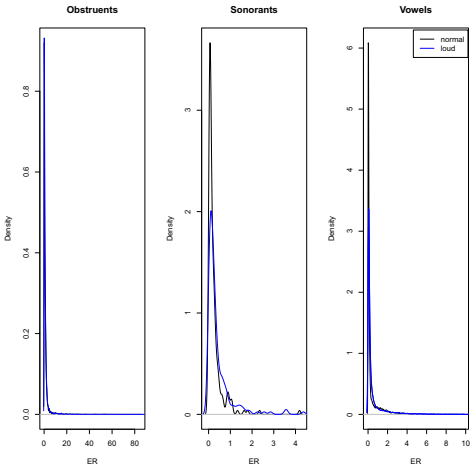

Figure 2: ER distributions for the phoneme classes

Table 5: *Results of vocal effort classification.*

Feature	Correct Detections
MFCC	94
ST	85
COG	90
ER	63
Mom1	90
Mom2	77
Mom3	82
Mom4	81
Moments	92
ST+COG+ER	95
Moments+COG+ST+ER	92

the previous section can be used in automatic classification of spontaneous speech.
The classifier uses gaussian mixture models with 64 components. As features we use the single spectral features, a combination of all spectral features and the standard Mel-Frequency Cepstrum Coefficients (MFCC) features as baseline system. The results can be seen in Table 5.
The first features in Table 5 are the standard MFCCs with 19 coeffients and without the zero'th-order coefficient, as proposed by [12]. We will compare the performance of the other features to each other and to the MFCCs. As you can see in the previous section (see section 3), the energy ratio seems to be worst for discrimination of vocal effort of all investigated features. This is supported by the results of the vocal effort classification system (see Table 5). Furthermore we can see from the Z-values in Table 4 that the second spectral moment is not approriate to distingiush between normal and loud speech, too. The results of the classification process certify this. The Z-values of the COG and the first spectral moment indicate, that these features are appropriate to classify different degrees of vocal effort. Hence, we monitor the highest number of correct detections when comparing them to the other single spectral features. The combination of all four spectral moments leads to a small improvement of the results of the first spectral moment. Combining all other features, except the spectral moments, we yield best results. These features provide even better performance than the standard MFCC features. The combination of all spectral features does not improve the result any further.

5. Conclusion

In this paper we presented a comparison of the influence of raised vocal effort on different spectral parameters. The results showed that the complete distributions of loud and normal speech differ significantly. The second spectral moment is influenced less than the other parameters. When dividing the data set into three phoneme classes we found that the vowels and sonorants behave more or less like the complete distribution, whereas the obstruents show completely different patterns. The obstruents are not changed as much as vowels and sonorants.
The comparison of different features demonstrated that the energy ratio and the second spectral moment are less affected by vocal effort changes than the other spectral parameters. This was certified by the automatic vocal effort classification.
As future work we propose to verify our results for female speakers. Additionally one could check whether a combination of MFCCs and spectral features for vocal effort classification leads to further improvements.

6. References

[1] Summers, W. van, Pisoni, D.B., Bernacki, R.H., Pedlow, R.I. and Stockes, M.A., "Effects of noise on speech production: Acoustic and perceptual analyses", J. Acoust. Soc. Am., 84(3):917–928, 1988.

[2] Schulman, R., "Articulatory dynamics of loud and normal speech", J. Acoust. Soc. Am., 85(1):295–312, 1989.

[3] Liénard, J.-S. and Di Benedetto, M.-G., "Effect of vocal effort on spectral properties of vowels", J. Acoust. Soc. Am., 106(1):411–422, 1999.

[4] Harwardt C., Gottsmann, F., Noubours, S., "On the Relationship between Vocal Effort and Spectral Moments", SiMPE, 2011.

[5] Wesker, T., Meyer, B., Wagener, K., Anemüller, J., Mertens, A. and Kollmeier, B., "Oldenburger logatome speech corpus (OLLO) for speech recognition experiments with humans and machines", Interspeech,1–4, 2005.

[6] Jessen, M., Köster, O. and Gfroerer, S., "Influence of vocal effort on average and variability of fundamental frequency", International Journal of Speech, Language and the Law, 12(2):174–213, 2005.

[7] Iseli, M., "Dependencies of Voice Source Measures on Age, Sex, Vowel Context, and Prosodic Features", University of California, Diss., 2007.

[8] Ron, R. J. J. H. and Pols, L. C. W., "An acoustic description of consonant reduction", Speech Communication, 28:125–140, 1999.

[9] Wenndt, S. J., Cupples, E. J. and Floyd R. M., "A Study on the Classification of Whispered and Normally Phonated Speech", International Conference on Spoken Language Processing, 2002.

[10] Zhang, C. and Hansen, J. H. L., "Effective Segementation based on Vocal Effort Change Point Detection", ITRW, 2008.

[11] Forrest, K., Wiesmer, P., Milenkovic, P. and Dougall, R., "Statistical analysis of word-initial voiceless obstruents: Preliminary data", J. Acoust. Soc. Am., 84(1):115–123, 1988.

[12] Hollander, M., Wolfe, D. A., "Nonparametric Statistical Methods", John Wiley Sons, Inc., 1999.

[12] Hansen, J. H. L. and Varadarajan, V., "Analysis and Compensation of Lombard Speech Across Noise Type and Levels With Application to In-Set/Out-of-Set Speaker Recognition", IEEE Transactions on Acoustics, Speech and Signal Processing, 17 (2): 366–378, 2009.

Investigating Robustness of Spectral Moments on Normal- and High-Effort Speech

Frederike Gottsmann, Corinna Harwardt

Fraunhofer-Institut für Kommunikation,
Informationsverarbeitung und Ergonomie FKIE, Germany
frederike.gottsmann@fkie.fraunhofer.de, corinna.harwardt@fkie.fraunhofer.de

Abstract

In this paper we are looking for a robust value of the spectral moments that does not change when a speaker varies his vocal effort from normal to loud speech. To do this we first calculate the first four spectral moments for normal and loud speech. Then we compare the results for each single phoneme. After this, we do a correlation analysis to check whether normal and loud speech are linked with each other linearly. The results of the investigations show that plosives and fricatives are robust to changes of vocal effort. Vowels and sonorants demonstrate significant differences in vocal effort.

Index Terms: vocal effort, spectral moments

1. Introduction

In general, the term vocal effort describes the quantity of voice a speaker uses when he adopts it to the communication situation. Changing vocal effort means that the speaker varies his voice from normal to soft/ whispered speech or from normal to loud/ shouted speech. Whispering is associated with low vocal effort, whereas loud speech corresponds to high vocal effort. Those changes might be induced for example by the presence of background noise, by varying the distance or by emotions.
Several studies have been made by different authors which investigated the impact of vocal effort changes on glottal, articulatory and acoustic parameters [6], [5]. Unfortunately the results over the different studies are not always comparable to each other due to different preconditions, e.g. gender of speaker or setting of recordings. In this investigation we just look at those studies which focus on spectral changes. Several of those studies found a shift of frequencies from lower to higher frequency bands [4]. Often investigated spectral parameters are spectral center of gravity [4], [2], spectral tilt [6], [2] or spectral emphasis [7]. The spectral moments as proposed by Forrest et al. [1] have not been part of an investigation of vocal effort changes, as far as we know. Hence, we decided to analyze the impact of raised vocal effort on spectral moments.
In this paper we describe the results of this analysis. At first we describe the methods we used in section 2. This incorporates presentation of the audio data we used, specification of the calculation of spectral moments and definition of the statistical tests needed for evaluation of results. Next we present the results in section 3. We show average spectral moments for each different phoneme and the associated t-tests. For deeper analysis of the relationship between normal and loud speech we do a correlation analysis. After presentation of results we draw a conclusion (see section 4).

2. Method

In the next sections we describe the preconditions for our investigations. First we present the corpus we used. Then we demonstrate our calculations and the method for statistical analysis.

2.1. The OLLO corpus

The Oldenburger Logatome (OLLO) speech corpus contains data of 40 German speakers which are spoken in four dialects: standard German (speaker from university population), Bavarian (speaker from Munich), East Frisian (speaker from Oldenburg) and Eastphalian (speaker from Magdeburg). The speakers articulate different vowel-consonant-vowel or consonant-vowel-consonant phoneme combinations without any semantic information. These phonemes are the vowels /a, aː, e, ɛ, i, ɪ, o, ɔ, u, ʊ, ə/, the plosives /b, p, d, t, g, k/ and the fricatives /f, v, s, ʃ, z/. As sonorants /l, m, n/ were available in the corpus. The database includes speaker characteristics like age, gender, dialect and speaker-dependent variabilities like speaking rate, speaking effort, speaking style [8]. For our investigations we used normal- and high-effort speech samples of 19 male speakers.

2.2. Spectral moments

The speech spectrum represents the signal's energy at different frequencies. Therefore the speech signal's energy distribution changes, when a speaker raises his voice. To investigate the changes on the spectrum induced by raised vocal effort, we have a closer look at the energy distribution of the spectrum. First we compute a 512-point fast Fourier transform. To analyze the distribution of the power spectrum, we compute the first four moments of the spectral distribution, which characterize the distribution of the power spectrum very well. The first moment (m1) is associated with the mean value of the spectral distribution and the unit is hertz (Hz). Whereas the second moment (m2) is linked with the variance around first moment. The third moment (m3) represents the skewness and shows the symmetry of the distribution. Positive skewness indicates a spectral tilt where we can find a higher concentration of energy in the lower frequency ranges. Negative skewness means a spectral tilt with a higher concentration of energy in the higher frequency ranges. Kurtosis is the fourth moment (m4) of the spectral distribution. Positive values show a compact spectrum, whereas, negative values indicate a flatter spectrum [3]. The units of skewness and kurtosis are dimensionless. The spectral moments were computed according to [1].

28 – 31 August 2011, Florence, Italy

2.3. Statistical Analysis

We accomplished a two-tailed t-test. We decided to use a critical value of $\alpha = 0.01$ which leads to considerable difference between normal and loud articulation. If the z-value in Table 2 is not located within the interval 2.58 to -2.58, we refuse the null hypothesis. In our case, the null hypothesis is true when spectral moments equal for normal and loud speech. Corresponding the alternative hypothesis says that both conditions values differ. In the case that we find significant differences, we assume a linear relation between the two vocal effort conditions. Therefore a Pearson product-moment correlation coefficient is calculated to check whether normal and loud speech correlate.

3. Results

For each single phoneme we calculated average values for the four moments for normal and loud articulation. We are looking for values that do not change between normal- and high-effort. A two-tailed t-test was used to recognize significant differences between these conditions. Then, we accomplished a correlation analysis to demonstrate an interrelation between normal- and high-effort.

3.1. Plosives

At first, we analyze the six phonemes /b, p, d, t, g/ and /k/. We start with the bilabial plosives /b/ as voiced and /p/ as voiceless phonemes. At the first, third and fourth moment the values for /b/ do not show any significant difference between normal and loud articulation (see Table 1). Hence, we have to maintain

Table 1: *Results of the t-tests for plosives for each spectral moment (m1-m4)*
s = significance, ns = no significance

	m1	m2	m3	m4
/b/	ns	s	ns	ns
/p/	s	s	s	s
/d/	s	s	ns	ns
/t/	s	s	ns	ns
/g/	s	ns	s	s
/k/	ns	ns	ns	ns

the null hypothesis. For the second moment there are significant differences between these two efforts. In our calculation we can see that the values for all four moments for normal articulation are higher than for loud (details see [**?**]). Normal values for the first moment are higher than loud values. Skewness is linked with a curve skewed to the right and kurtosis indicates a leptokurtic curve. The voiceless plosive /p/ has significant differences at all moments. For the first two moments of normal vocal effort, values are also higher than those for loud speech. However the results for the third and fourth moment illustrate another behavior. Skewness is positive for both efforts but normal values are lower. Likewise for kurtosis the curve is leptokurtic. In addition to these results, we perform a correlation analysis (see Table 2). As you can see both plosives /b/ and /p/ show positive values. The plosive /b/ includes the highest correlation between normal and loud articulation for the first moment. The lowest correlation between these two efforts has /p/ for the fourth moment. Next, we describe the alveolar plosives /d/ and /t/. The first two moments for the voiced /d/ show significant differences between normal and loud speech. Hence, we have

Table 2: *Correlation coefficient of plosives for each spectral moment (m1-m4)*

	m1	m2	m3	m4
/b/	0.7423	0.7112	0.6790	0.6158
/p/	0.6058	0.5536	0.4917	0.4103
/d/	0.7141	0.6347	0.6068	0.4388
/t/	0.7418	0.6450	0.7016	0.4769
/g/	0.8039	0.6056	0.7531	0.6069
/k/	0.8691	0.6034	0.8375	0.6430

to reject the null hypothesis. For these two moments the values for high-effort are higher than for normal-effort. On the other hand for the third and fourth moment, both conditions are not significantly different and hence, the null hypothesis has to be maintained. The normal values for the third moment are higher than those for loud speech. But both values are positive. Kurtosis shows a leptokurtic curve for both values but values for normal vocal effort are also higher. Similar results as for /d/ can be observed for the voiceless /t/ for the first, second and third moment. For the fourth moment we observe a platykurtic curve for both efforts. The correlation analysis illustrates a high correlation between normal and loud speech for the voiceless /t/ for the first moment. The lowest correlation between both efforts is evident for /d/ at the fourth moment.
As final plosives we look at the voiced velar /g/ and voiceless velar /k/. For /g/ the first, third and fourth moments illustrate significant differences between normal and loud articulation. Simply the second moment does not show any significant change. Loud values for the first two moments are higher than for normal articulation. On the other hand for the last two moments, normal values are higher. Skewness is positive, too, and kurtosis illustrates a leptokurtic curve for both. The voiceless /k/ does not present any significant difference for all four moments. Just for the first moment the loud values are higher than normal speech. For the other three moments, we observe higher values for normal-effort. A positive skewness and a leptokurtic curve are found for both, too. Here, the correlation analysis shows the highest correlation between normal and loud values for the voiceless /k/ for the first moment. This plosive indicates the lowest correlation between these two values for the second moment, too.
For all plosives, we can observe that /p/ shows always significant differences for all moments (see Table 2). The plosives /b, d, t, g/ and /k/ do not show any consistent differences for all moment. In the majority of cases the third and fourth moment do not illustrate any significant difference between normal and high vocal effort. We can also see that the first moment has the highest correlation for all phonemes between normal and loud articulation. When we look at individual phonemes we can perceive that /k/ has the highest correlation between normal and loud speech at the first moment (see Fig. 1). Hence, it could be that /k/ is most robust against changes in vocal effort for all four moments. However, we can say for the other plosives /b, d/ and /t/ that they are also robust against changes in vocal effort for the third and fourth moment.

3.2. Fricatives

In this section we investigate the fricatives /z, s, ʃ, v/ and /f/. First, the alveolar fricatives /z/ and /s/ are analyzed. The voiced fricative /z/ does not show any significant difference for

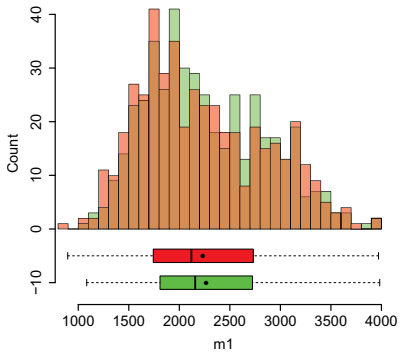

Figure 1: *Values for the first moment (m1) for loud (green) and normal (red) vocal effort for the plosive /k/*

the first, third and fourth moment (see Table 3). We can reject the null hypothesis only for the second moment. Average values for loud speech for the first and second moment are higher than average values for normal speech. For the third moment normal values are positive whereas loud values are negative. This illustrates a change in spectral tilt. Hence, the curve changes from a skew to the right to a skew to the left when a speaker changes his vocal effort from normal to loud speech. Kurtosis shows a platykurtic curve for normal as well as for loud articulation. The voiceless /s/ does not have any significant difference between normal and loud speech for the first and third moment. In contrast, the second and fourth moment offer significant differences between both efforts. The values for normal vocal effort for the first moment are slightly higher than the values for loud speech. We can observe higher values for loud articulation for the second moment. For the third moment, loud values are also higher but there is only a small difference to normal values. Both values are near null and negative skewness is linked with a curve skewed to the left. Values for the fourth moment are negative, too. This indicates a platykurtic curve for normal- and high-effort, in which normal values are higher. For these two fricatives the highest correlation between these two efforts can be found for the voiced /z/ of the second moment (see Table 4). The phoneme /s/ shows the lowest correlation between normal and loud articulation for the fourth moment.

Table 3: *Results of the t-tests for fricatices for each spectral moment (m1-m4)*
s = significance, ns = no significance

	m1	m2	m3	m4
/z/	ns	s	ns	ns
/s/	ns	s	ns	s
/ʃ/	s	ns	s	s
/v/	s	ns	s	s
/f/	ns	s	s	ns

Next we investigate the voiceless postalveolar fricative /ʃ/. The voiced variant /ʒ/ is not included in our corpus. The second moment does not illustrate a significant difference whereas for all other moments, we reject the null hypothesis. The first and fourth moment offer lower values for normal-effort than for

Table 4: *Correlation coefficient of fricative for each spectral moment (m1-m4)*

	m1	m2	m3	m4
/z/	0.8360	0.8601	0.7537	0.7205
/s/	0.7876	0.5959	0.7508	0.6043
/ʃ/	0.6826	0.8128	0.5601	0.5528
/v/	0.2287	0.6211	0.3210	0.2979
/f/	0.6314	0.8063	0.6139	0.5486

loud speech. Kurtosis displays a platykurtic curve for both values. For the second and third moment the values for normal speech are higher. We can also see a change in spectral tilt from normal to loud articulation, where the values for normal vocal effort are positive and the values for loud articulation are negative. The highest correlation between normal- and high-effort is for the second moment.
As labiodental fricatives we analyzed the voiced /v/ and voiceless /f/. The first, third and fourth moment for /v/ show significant differences between normal and loud articulation. For the second moment there are not any significant differences. We observe only for the first moment higher values for loud speech. Skewness is positive and linked with a curve skewed to the right. Kurtosis shows a leptokurtic curve. The voiceless /f/ has not any significant difference for the first and fourth moment. For the other two moments we reject the null hypothesis. The first and second moment illustrate lower values for normal effort. The third moment describes a positive skewness. Kurtosis is negative and therefore, we find a platykurtic curve for both efforts. The correlation analysis represents for each moment higher values for the voiceless /f/. The highest value is for the second moment. For the voiced fricative /v/ we observe the lowest value for the first moment.
In summary, we can see in Table 3 that the fricatives /z, s/ and /f/ do not show any significant difference for the first moment. The second, third and fourth moment have not always significant differences, too. In particular, it seems that for the first, third and fourth moment the fricative /z/ is robust against a change in vocal effort (see Fig. 2). There is also a high correlation between normal and loud articulation for these fricatives.

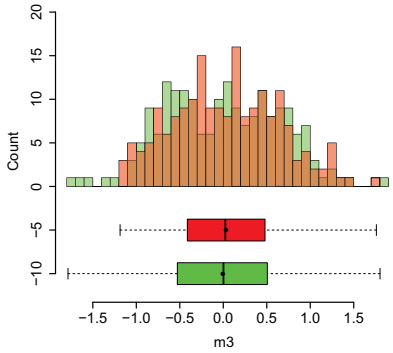

Figure 2: *Values for the third moment (m3) for loud (green) and normal (red) vocal effort for the fricative /z/*

3.3. Vowels

Vowels are the largest class of phonemes with the 11 single sounds /a, aː, e, ɛ, i, ɪ, o, ɔ, u, ʊ, ə/. At all moments all phonemes show significant differences between normal and loud articulation and hence, values for high vocal effort are always higher than for normal vocal effort. So, we can see, that they are not robust. There is a change in spectral energy when a speaker raises his vocal effort. Loud values for all phonemes for the first moment are higher than values for normal speech. As an example we can see in Fig. 3 the values for vowel /e/ for the first moment. This results we observe for the second moment, too. The third moment shows always a positive skewness and for all vowels the values for normal-effort are higher than for high-effort. Kurtosis indicates a leptokutic curve for all phonemes, where values for normal articulation are higher than for loud articulation. Correlation analysis demonstrates a positive correlation between normal and high-effort for all vowels, too. The highest correlation between these two efforts offers the central vowel /ə/ for the first, third and fourth moment. The vowel /ɛ/ possesses the highest correlation between normal and loud articulation for the second moment. In summary, we found that

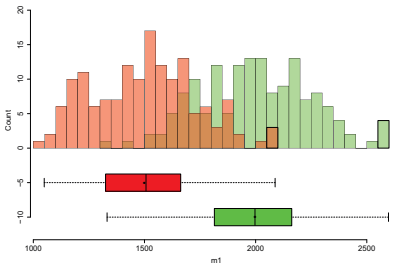

Figure 3: *Values for the first moment (m1) for loud (green) and normal (red) vocal effort for the vowel /e/*

vowels show no robust pattern when a speaker raises his vocal effort.

3.4. Sonorants

The sonorants /l, m/ and /n/ are the last phoneme class which is investigated. All moments show significant differences between normal and loud articulation. We observe for all sonorants lower values for normal speech than for loud speech for the first and second moment. In Fig. 4 we see an example for the first moment of the sonorants /l/. The values for the third and fourth moment for normal-effort are higher than the values for high-effort. For both moments, we observe positive values for normal and loud articulation. The correlation analysis for all sonorants is positive for all moments, too. These observations are comparable to those regarding the vowels. Hence we can assume that these values are not robust.

4. Conclusions

The aim of this investigation was to find out whether there is a robust value which is constant when a speaker changes his vocal effort from normal to loud speech. The results indicate that plosives and fricatives do not always show significant differences between normal and loud articulation. In particular, the plosive /k/ does not indicate any change for all moments. But also,

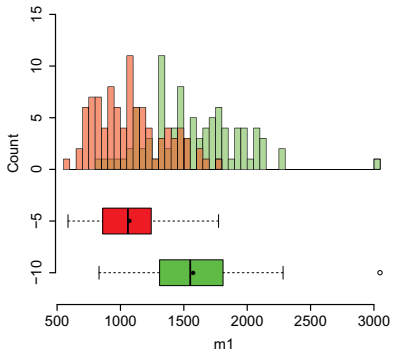

Figure 4: *Values for the first moment (m1) for loud (green) and normal (red) vocal effort for the sonorant /l/*

/b, d/ and /t/ have similar values for the third and fourth moment. The fricatives /z, s/ and /f/ are not significant for the first moment. This means for these plosives and fricatives, that they are consistent for each speaker when there is a change from normal to loud speech. Vowels and sonorants show significant differences and therefore the speech signal's energy distribution changes, when a speaker raises his vocal effort. Hence, we find no robust value for vowels and sonorants. In comparision to unvoiced phonemes we can see that voiced phonemes are less stable than unvoides ones. All phonemes and all moments show that the two efforts correlate with each other.

5. References

[1] Forrest, K., Wiesmer, P., Milenkovic, P. and Dougall, R, "Statistical analysis of word-initial voiceless obstruents: Preliminary data", J. Acoust. Soc. Am., 84(1):115–123, 1988.

[2] Junqua, J.C., "The Lombard reflex and its role on human listeners and automatic speech recognizers", J. Acoust. Soc. Am., 93(1):510–524, 1993

[3] Kardach, J., Wincowski, R., Metz, D.E., Schiavetti, N., Whitehead, R.L. and Hillenbrand, J. "Preservation of place and manner cues during simultaneous communication: a spectral moments perspective", Journal of Communication Disorders, 35 :533–542, 2002

[4] Liénard, J.-S. and Di Benedetto, M.-G., "Effect of vocal effort on spectral properties of vowels", J. Acoust. Soc. Am., 106(1):411–422, 1999.

[5] Schulman, R., "Articulatory dynamics of loud and normal speech", J. Acoust. Soc. Am., 85(1):295–312, 1989.

[6] Summers, W. van, Pisoni, D.B., Bernacki, R.H., Pedlow, R.I. and Stockes, M.A., "Effects of noise on speech production: Acoustic and perceptual analyses", J. Acoust. Soc. Am., 84(3):917–928, 1988.

[7] Traunmüller, H. and Eriksson, A., "Acoustic effects of variation in vocal effort by men, women and children", J. Acoust. Soc. Am., 107(6):3438–3451, 2000.

[8] Wesker, T., Meyer, B., Wagener, K., Anemüller, J., Mertens, A. and Kollmeier, B., "Oldenburger logatome speech corpus (OLLO) for speech recognition experiments with humans and machines", Interspeech,1–4, 2005.

[9] Gottsmann, F., "Realisierung einer Messmethode für den Stimmaufwand eines Sprechers", Magisterarbeit, Universität Bonn, 2010.

On the Relationship between Vocal Effort and Spectral Moments

Corinna Harwardt
Fraunhofer FKIE
Neuenahrer Str. 20
53343 Wachtberg, Germany
corinna.harwardt@fkie.fraunhofer.de

Frederike Gottsmann
Fraunhofer FKIE
Neuenahrer Str. 20
53343 Wachtberg, Germany
frederike.gottsmann@fkie.fraunhofer.de

Sandra Noubours
Fraunhofer FKIE
Neuenahrer Str. 20
53343 Wachtberg, Germany
sandra.noubours@fkie.fraunhofer.de

ABSTRACT

Speech processing applications in mobile environment strongly suffer from background noise and other disturbing factors as for example the submission channel. These factors mislead the speaker to raise his vocal effort. This so called Lombard effect leads to changes in the acoustics of a speech signal and hence influences the recognition performance. Many studies have shown great degradations of recognition performance for several speech processing applications on high-effort speech. Having an extended knowledge about the changes of acoustic measurements induced by changing vocal effort is the baseline for the development of robust features for successful speech processing tasks in mobile environment. Therefore we illustrate the effect of raised vocal effort on the speech spectrum, in particular on spectral moments.
To do this we first calculate the first four spectral moments for normal and loud speech. Then we compare the results over four different phoneme classes. To find out whether spectral changes depend on the speaker we add another investigation for each speaker separately. At last we do a correlation analysis to check whether normal and loud speech are linked linearly. The results show that spectral moments of vowels are significantly modified, whereas plosives aren't changed significantly. The speaker-specific analysis presents a relatively consistent pattern over all speakers for vowels and sonorants, but not for plosives and fricatives. The highest correlation can be found for fricatives and vowels.

Categories and Subject Descriptors

I.2.7 [**Natural Language Processing**]: Speech recognition and synthesis

General Terms

EXPERIMENTATION

Keywords

spectral moments, vocal effort, speaker-specific analysis

1. INTRODUCTION

In general, the term vocal effort describes the quantity of voice a speaker uses when he adopts it to the communication situation. Changing vocal effort means that the speaker varies his voice from normal to soft/ whispered speech or from normal to loud/ shouted speech. Whispering is associated with low vocal effort, whereas loud speech corresponds to high vocal effort. Those changes might be induced for example by the presence of background noise, by varying the distance or by emotions.

Several studies have been made by different authors which investigated the impact of vocal effort changes on glottal, articulatory and acoustic parameters [1], [2]. Unfortunately the results over different studies aren't always comparable to each other due to different preconditions like for example speaker sex or the setting of recordings. In this investigation we just look at those studies which focus on spectral changes. Several of those studies found a shift of frequencies from lower to higher frequency bands [3]. Often investigated spectral parameters are spectral center of gravity [3], [4], spectral tilt [1], [4] and spectral emphasis [5]. The spectral moments as proposed by Forrest et al. [6] haven't been part of an investigation on vocal effort changes, as far as we know. Hence, we decided to analyze the impact of raised vocal effort on spectral moments.

In this paper we present the results of this analysis. At first we describe the methods we used in Section 2. This incorporates presentation of the audio data we used, specification of the calculation of spectral moments and definition of the statistical tests needed for evaluation of results. Next we present the results in Section 3. We show average spectral moments for four different phoneme classes and the associated t-tests. To see whether the observed differences are speaker-specific we evaluate the data per speaker separately and analyze the speaker-specific character of changes of spectral moments in normal and loud speech. For deeper analysis of the relationship between normal and loud speech we do a correlation analysis. After presentation of results we draw a conclusion (see Sec. 4) and offer ideas for future work (see Sec. 5).

2. METHOD

In the next sections we describe the preconditions for our investigation. We first present the corpus we used. Then we illustrate our calculations and the method for statistical

analysis.

2.1 The OLLO Corpus

The Oldenburger Logatome (OLLO) speech corpus consists of data from 40 German speakers with four dialects: standard German (speakers from university population), Bavarian (speakers from Munich), East Frisian (speakers from Oldenburg) and Eastphalian (speakers from Magdeburg). The speakers had to articulate different vowel-consonant-vowel or consonant-vowel-consonant phoneme combinations with no semantic information. The database includes speaker-dependent variabilities like age, gender, dialect and speaker-independent variabilities like speaking rate, speaking effort, speaking style [7]. For our investigation we used the normal and high vocal effort speech samples from male speakers. The available vowel phonemes are /a, aː, e, ɛ, i, ɪ, o, ɔ, u, ʊ, ə/. For plosives there are /b, p, d, t, g, k/ and for fricatives /f, v, s, ʃ, z/. As sonorants /l, m, n/ were articulated.

2.2 Spectral Moments

The speech spectrum represents the signal's energy at different frequencies. Therefore the speech signal's energy distribution changes, when a speaker raises his voice. To investigate the changes on the spectrum induced by raised vocal effort, we have a closer look at the distribution of the spectrum. First we compute a 512-point fast Fourier transform. To analyze the distribution of the power spectrum, we compute the first four moments of the spectral distribution, which characterize the distribution very well. The first moment (m1) is associated with the mean value of the spectral distribution and the unit is hertz (Hz). Whereas the second moment (m2) is linked with the variance around first moment. The third moment (m3) represents the skewness and shows the symmetry of the distribution. Positive skewness indicates a spectral tilt where we can find a higher concentration of energy in the lower frequency ranges. Negative skewness means a spectral tilt where is a higher concentration of energy in the higher frequency ranges. Kurtosis is the forth moment (m4) of the spectral distribution. Positive values show a compact spectrum, whereas, negative values indicate a flatter spectrum [8]. Skewness and kurtosis are dimensionless. The spectral moments were computed according to [6].

2.3 Statistical Analysis

We accomplished a two-tailed t-test. We decided to use a critical value of $\alpha = 0.01$ because we want to measure a considerably difference between normal and loud articulation. If the z-value in Tab 2 is lying out of the interval 2.58 and -2.58, we refuse the null hypothesis. In our case the null hypothesis is that the spectral moments stay the same for normal and loud speech. Accordingly the alternative hypothesis says that the values of both conditions differ. In the case that we find significant differences, we assume a linear relation between the two vocal effort conditions. Therefore a Pearson product-moment correlation coefficient is calculated to check whether normal and loud speech are correlated.

3. RESULTS

We calculated average values of the phoneme classes for each spectral moment for normal and loud speech (see Tab. 1). The values of all speakers were taken into account. These values are supposed to show the difference between normal and loud speech. We used the two-tailed t-test to analyze the differences (see Tab. 2).

Table 1: Average values of four spectral moments (m1-m4) for normal and loud speech

		vowels	sonorants	fricatives	plosives
m1	normal	1145.444	870.406	3754.803	2113.703
	loud	1544.0	1292.805	3830.660	2142.560
m2	normal	1793822	1432247	4771904	3407331
	loud	2230923	2037239	5054837	3380926
m3	normal	2.5869	2.8772	0.1216	1.3440
	loud	1.8680	1.9443	0.0776	1.3556
m4	normal	9.2568	10.4181	-0.1885	2.5720
	loud	4.5178	4.5473	-0.3969	2.6699

Table 2: Z-values of spectral moments of phoneme classes

	vowels	sonorants	fricatives	plosives
m1	72.4501	26.4271	2.8366	0.1325
m2	41.4528	21.8342	8.0989	-1.4272
m3	-50.6823	-23.6744	-2.3613	0.6880
m4	-50.2668	-23.4056	-4.8513	1.1451

For the first moment of vowels we observe a significant difference between normal and loud articulation. That means that the null hypothesis is rejected by a confidence level of 99%. We discover a higher average value for loud speech compared to normal speech. Likewise for sonorants, the first moment shows a significant difference between normal and high-effort and the loud average values are higher. The first moment of fricatives is raised up significantly, too and values for loud speech are higher than values for normal speech. One can do another observation with regard to the significance of plosives. There is no significant difference between normal and loud speech. However, loud values are higher than normal (see Tab. 1).

The z-values of the second moment in Tab. 2 show significant differences for vowels, sonorants and fricatives. As for the first moment we found higher average values for loud than for normal speech at all three classes. The null hypothesis is rejected by a confidence level of 99%. Plosives show no significant difference and in contrast to the other phoneme classes loud speech values are lower than normal.

The Skewness is the third moment that we calculate. For vowels and sonorants we observe a significant difference between the two vocal effort conditions. Hence, the alternative hypothesis is accepted. Fricatives and plosives don't change significantly. This means that we maintain the null hypothesis. All average values of normal and loud speech for all phoneme classes show a positive skewness. A positive skewness is linked with a curve skewed to the right. This indicates a higher energy concentration in lower frequency range. Looking at both vocal effort classes we notice higher average values of normal speech for vowels, sonorants and fricatives. This signifies that loud speech distributions of those three classes are closer to the symmetric form than

Table 3: Number of speakers with significant (s) and without significant (ns) changes of spectral moments (m1-m4) from normal to high vocal effort, investigated on four different phoneme classes.

		vowels	sonorants	fricatives	plosives
m1	s	19	19	-	2
	ns	-	-	19	17
m2	s	18	16	13	4
	ns	1	3	6	15
m3	s	19	17	-	3
	ns	-	2	19	16
m4	s	19	19	2	4
	ns	-	-	17	15

normal effort distributions. Hence, compared to normal-effort speech, loud speech is shifted to higher frequencies. Plosives just show a minimal difference between normal and loud speech. However, the difference for plosives is too small to make a statement with regard to the frequency distribution between normal and loud speech.

The last moment kurtosis changes significant for vowels, sonorants and fricatives. Again, values for plosives are not significantly influenced by vocal effort variation. Average values for vowels, plosives and sonorants are all positive and show a leptokurtic curve for normal as well as for high vocal effort. As you can see in Tab. 1 kurtosis of vowels and sonorants for normal articulation is twice as big as for loud articulation. This means that we monitor a steep distribution around the mean for both conditions, but normal effort is steeper. Hence, kurtosis is more pronounced for normal values. Fricatives have a lower and wider peak around the mean; therefore it is a platykurtic curve.

The average values and significance tests illustrate differences in spectral moments for normal and loud articulation. To get more information about the speaker-specific characteristics of these changes we accomplish another study which investigates the differences in all four phoneme classes for each speaker separately (see Tab. 3).

When we look at the first moment of vowels, we observe that all values for loud articulation are significantly higher than those for normal speech for all 19 speakers. For the second moment we monitor again significantly higher values for loud speech for all but one speaker. The third and forth moment show significantly higher normal values for all speakers. For each speaker and both degrees of vocal effort we observe a curve skewed to the right. Even though for loud speech the skewness is lower than for normal speech. Looking at the forth moment, we note that all speakers have a leptokurtic curve.

For sonorants the values of the first and forth moment behave like the vowels across all speakers. The second and third moment behave similar, too. However, for the values of the second moment there are three exceptions instead of just one for vowels. For the third moment we observe two exceptions.

Table 4: Correlation coefficient of phoneme classes

	vowels	sonorants	fricatives	plosives
m1	0.8458	0.7025	0.9179	0.8416
m2	0.6785	0.5730	0.8386	0.6795
m3	0.8778	0.7083	0.9038	0.7688
m4	0.7888	0.5974	0.6923	0.6399

As already seen before, plosives and fricatives show different characteristics than vowels and sonorants. This is again the case for the speaker-specific evaluation. We found no consistent trend for the changes of spectral moments across all speakers in normal and loud speech. In most cases, we must verify the null hypothesis for all moments and thus, there are no significant differences between normal and high vocal effort.

Due to the results for all speakers we can assume a relationship between normal and loud speech, in particular for vowels and sonorants. If the values are low for normal speech, they are low for loud speech, too. This could indicate a correlation between these two conditions. Therefore, we accomplish a Pearson-correlation in order to establish a linear relation (see Tab. 4). The results for all phoneme classes show a value above 0.5 and, hence, there is a positive correlation between normal and loud articulation. Fricatives indicate the strongest linear correlation between normal and high-effort for the first three moments (Fig. 1 shows an example). However we assume that they aren't suitable for a prediction of vocal effort because the results from the t-test show no significant difference between normal and loud articulation. Vowels show a high correlation between these two efforts, too. In this case, the t-test shows a significant difference. Hence, we can assume that vowels are suitable for a prediction when a speaker changes his vocal effort from normal to loud speech. Vowels indicate a higher correlation for the first and third moment as plosives and sonorants. However, for the first moment the correlation coefficient for plosives are close to those of vowels. Finally, sonorants offer the lowest correlation between normal and loud articulation for these phoneme classes (Fig. 1 shows an example).

4. CONCLUSION

For each phoneme class we investigated the difference in spectral moments for normal and high-effort speech with focus on speaker specific changes. Next, we examined the correlation between normal and loud speech. The aim of this investigation was to find out whether there is a difference between spectral moments of normal and loud speech.

The results indicate that vowels show the most consistent pattern and greatest modifications when vocal effort changes from normal to loud speech (Fig. 2 shows an example). Especially the first moment offers high significance and correlation between normal and loud articulation.

Plosives are the phoneme class with fewest changes. They demonstrate no significant changes for spectral moments across all speakers. If we look at individual speakers, there is no consistent pattern, too.

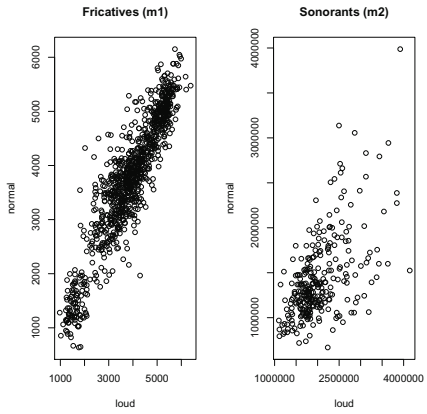

Figure 1: Correlation between loud and normal speech for fricatives (*left*) for the first moment (*m1*) and for sonorants (*right*) for the second moment (*m2*).

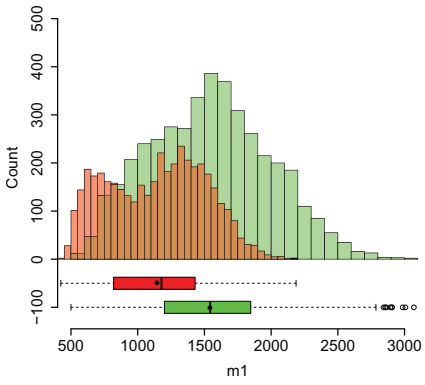

Figure 2: Values for the first moment (*m1*) for loud (*green*) and normal (*red*) vocal effort for vowels.

5. FUTURE WORK

Applying the knowledge from this study to a practical scenario, one could imagine using especially the first moment of vowels to build up a classifier to distinguish between different degrees of vocal effort. Additionally it might be used to develop a method to quantify vocal effort. However, one should keep in mind that using just the first moment for quantification or classification it isn't possible to compare data from different recording settings. The qualification of the first moment of vowels for these tasks needs to be approved in further studies.

The high correlation between normal and high vocal effort of vowels might be used to predict the changes induced to the speech spectrum by vocal effort changes. This needs to be checked in further work, too.

Plosives show no significant modification for the complete analysis and for most speakers in the speaker-specific analysis. This allows two possible interpretations. First the results may be caused by a bunch of changes induced by raised vocal effort, which do not follow a consistent trend and therefore don't lead to significant changes. The second interpretation is that plosives are relatively robust against vocal effort changes. The inhomogeneous results might arise from intraspeaker variabilities. These thesis used to be checked in another investigation, because finding a stable feature for changing vocal effort conditions would be appreciable for several speech processing applications.

6. REFERENCES

[1] Summers, W. van, Pisoni, D. B., Bernacki, R. H., Pedlow, R. I., Stockes, M. A.: Effects of noise on speech production: Acoustic and perceptual analyses. J. Acoust. Soc. Am. 84(3), 917–928 (1988)

[2] Schulman, R.: Articulatory dynamics of loud and normal speech. J. Acoust. Soc. Am. 85(1), 295–312 (1989)

[3] Liénard, J.-S., Di Benedetto, M.-G.: Effect of vocal effort on spectral properties of vowels. J. Acoust. Soc. Am. 106(1), 411–422 (1999)

[4] Junqua, J.-C.: The Lombard reflex and its role on human listeners and automatic speech recognizers. J. Acoust. Soc. Am. 93(1), 510–524 (1993)

[5] Traunmüller, H., Eriksson, A.: Acoustic effects of variation in vocal effort by men, women, and children. J. Acoust. Soc. Am. 107(6), 3438–3451 (2000)

[6] Forrest, K., Wiesmer, G., Milenkovic, P., Dougall, R.: Statistical analysis of word-initial voiceless obstruents: Preliminary data. J. Acoust. Soc. Am. 84(1), 115–123 (1988)

[7] Wesker, T., Meyer, B., Wagener, K., Anemüller, J., Mertins, A., Kollmeier, B.: Oldenburger logatome speech corpus (OLLO) for speech recognition experiments with humans and machines. Interspeech, 1–4 (2005)

[8] Kardach, J., Wincowski, R., Metz, D.E., Schiavetti, N., Whitehead, R.L. and Hillenbrand, J.: Preservation of place and manner cues during simultaneous communication: a spectral moments perspective, Journal of Communication Disorders 35, 533–542 (2002)

Literaturverzeichnis

Ahluwalia, R. (2008). *Entwicklung und Evaluation einer Sprachsteuerung zur Navigation einer Gruppe teilautonomer unbemannter Roboter*. Magisterarbeit, Universität Bonn, Philosophische Fakultät.

Andersson, A., Eriksson, A. & Traunmüller, H. (1996). *Cries and whispers: acoustic effects of variations of vocal effort* (Forschungsbericht). KTH.

Anglade, Y., Fohr, D. & Junqua, J.-C. (1992). Selectively Trained Neural Networks for the Discrimination of Normal and Lombard Speech. In *Proceedings of Second International Conference on Spoken Language Processing (ICSLP)* (S. 595-598).

Barfs, A. (2005). *Die Veränderung des Sprachsignals beim Schreien - eine akustische Analyse ausgewählter phonetischer Parameter*. Magisterarbeit, Universität Trier, Fachbereich II.

Becker, T. (2007). The Influence of intra-speaker variability in automatic speaker identification. In *Proceedings of the International Association of Forensic Phonetics and Acoustics 2007 Annual Conference (IAFPA).*

Becker, T. (2008). The influence of intra-speaker variability in automatic speaker verification using F0 features. In *Proceedings of the International Association of Forensic Phonetics and Acoustics 2008 Annual Conference (IAFPA)* (S. 24-25).

Becker, T. & Kreuzer, W. (2008). Automatische Sprecherverifikation basierend auf Stimmgrundfrequenz-Merkmalen mittels Hauptkomponentenanalyse. In *Konferenzband der 34. Jahrestagung für Akustik (DAGA)* (S. 257-258).

Bogert, B. P., Healy, M. J. R. & Tukey, J. W. (1963). The Quefrency Alanysis of Time Series for Echoes: Cepstrum, Pseudo-Autocovariance, Cross-Cepstrum and Saphe Cracking. In *Proceedings of the Symposium on Time Series Analysis* (S. 209-243).

Bond, Z. S., Moore, T. J. & Gable, B. (1989). Acoustic-phonetic characteristics of speech produced in noise and while wearing an oxygen mask. *Journal of the Acoustical Society of America, 85 (2)*, 907-912.

Bortz, J. (1993). *Statistik - Für Sozialwissenschaftler* (4. Aufl.). Berlin Heidelberg New York: Springer-Verlag.

Bořil, H. (2008). *Robust Speech Recognition: Analysis and Equalization of Lombard Effect in Czech Corpora*. Dissertation, Czech Technical University in Prague, Faculty of Electrical Engineering.

Bořil, H. & Hansen, J. H. L. (2009a). Reduced Complexity Equalization of Lombard Effect for Speech Recognition in Noisy Adverse Environments. In *Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech)* (S. 1243-1246).

Bořil, H. & Hansen, J. H. L. (2009b). Unsupervised Equalization of Lombard Effect for Speech Recognition in Noisy Adverse Environment. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*

(ICASSP) (S. 3937 - 3940).

Bořil, H. & Hansen, J. H. L. (2011). UT-Scope: Towards LVCSR Under Lombard Effect Induced by Varying Types and Levels of Noisy Background. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (S. 4472-4475).

Braun, A. (1992). Zur Bedeutung des Merkmals 'Mittlere Sprechstimmlage' in der Forensischen Sprechererkennung. In *Phonetik und Dialektologie - Joachim Göschel zum 60. Geburtstag* (S. 1-26). Marburg: Schriftenreihe der Universitätsbibliothek.

Brummer, N. & du Preez, J. (2006). Application-Independent Evaluation of Speaker Detection. *Computer Speech and Language*, *20*, 230-275. Verfügbar unter `http://www.dsp.sun.ac.za/~nbrummer/focal/index.htm`

Bußmann, H. (2002). *Lexikon der Sprachwissenschaft* (3. Aufl.). Stuttgart: Alfred Kröner Verlag.

Cabrera, D. & Gilfillan, D. (2002). Auditory Distance percption of Speech in the Presence of Noise. In *Proceedings of the International Conference on Auditory Display (ICAD)* (S. 1-9).

Dreher, J. J. & O'Neill, J. (1957). Effects of Ambient Noise on Speaker Intelligibility for Words and Phrases. *Journal of the Acoustical Society of America*, *29*, 1320-1323.

Duller, C. (2008). *Einführung in die nichtparametrische Statistik mit SAS und R*. Heidelberg: Physica-Verlag.

Edwards, A. W. F. (1972). *Likelihood*. Cambridge: Cambridge University Press.

Eriksson, A. & Traunmüller, H. (2002). Perception of vocal effort and distance from the speaker on the basis of vowel utterances. *Perception & Psychophysics*, *64 (1)*, 131-139.

Euler, S. (2006). *Grundkurs Spracherkennung – Vom Sprachsignal zum Dialog – Grundlagen und Anwendungen verstehen – Mit praktischen Übungen*. Wiesbaden: Vieweg Verlag.

Fan, X. & Hansen, J. H. L. (2008). Speaker Identification for Whispered Speech based on Frequency Warping and Score Competition. In *Proceedings of the 9th Annual Conference of the International Speech Communication Association (Interspeech)* (S. 1313-1316).

Fan, X. & Hansen, J. H. L. (2009). Speaker Identification with Whispered Speech based on Modified LFCC Parameters and Feature Mapping. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (S. 4553-4556).

Fant, G. (1960). *Acoustic Theory of Speech Production*. Mouton.

Forrest, K., Weismer, G., Milenkovic, P. & Dougall, R. N. (1988). Statistical analysis of word-initial voiceless obstruents: Preliminary data. *Journal of the Acoustical Society of America*, *84 (1)*, 115-123.

Garnier, M., Bailly, L., Dohen, M., Welby, P. & Loevenbruck, H. (2006). An Acoustic and Articulatory Study of Lombard Speech: Global Effects on the Utterance. In *Proceedings of the 9th International Conference on Spoken Language Processing (Interspeech - ICSLP)* (S. 1862-1865).

Garnier, M., Dohen, M., Loevenbruck, H., Welby, P. & Bailly, L. (2006). The Lombard Effect: a physiological reflex or a controlled intelligibility enhancement? In *Proceedings of the 7th International Seminar on Speech Production (ISSP)* (S. 255-262).

Garnier, M., Wolfe, J., Henrich, N. & Smith, J. (2008). Interrelationship between vocal effort and vocal tract acoustics: a pilot study. In *Proceedings of the 9th Annual Conference of the International Speech Communication Association (Interspeech)* (S. 2302-2305).

Gauvain, J.-L. & Lee, C.-H. (1991). Bayesian Learning of Gaussian Mixture Densities for Hidden Markov Models. In *Proceedings of the workshop on Speech and Natural Language (Human Language Technology Conference- HLT)* (S. 272–277). Verfügbar unter `http://www.aclweb.org/anthology-new/H/H91/H91-1053.pdf`

Gauvain, J.-L. & Lee, C.-H. (1994). Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains. *IEEE Transactions on Speech and Audio Processing*, *2*, 291–298.

Geumann, A. (2001). Vocal intensity: acoustic and articulatory correlates. In *5th International Conference on Speech Motor Control* (S. 2-5).

Goldenberg, R., Cohen, A. & Shallom, I. (2006). The Lombard Effect's Influence on Automatic Speaker Verification Systems and Methods for its Compensation. In *Proceedings of the International Conference on Information Technology: Research and Education (ITRE)* (S. 233 - 237).

Gottsmann, F. (2010). *Realisierung einer Messmethode für den Stimmaufwand eines Sprechers*. Magisterarbeit, Universität Bonn, Philosophische Fakultät.

Gottsmann, F. & Harwardt, C. (2011). Investigating Robustness of Spectral Moments on Normal- and High-Effort Speech. In *12th Annual Conference of the International Speech Communication Association (Interspeech)* (S. 2937-2940).

Gramming, P., Sundberg, J., Ternström, S., Leanderson, R. & Perkins, W. H. (1987). *Relationship between changes in voice pitch and loudness* (Bd. 28; Forschungsbericht Nr. 1). Dept. for Speech, Music and Hearing – KTH Computer Science and Communication.

Gross, J. (2009). *Package 'nortest'* (Forschungsbericht). CRAN. Verfügbar unter `http://cran.r-project.org/web/packages/nortest/nortest.pdf` (letzter Zugriff: 16.12.2010)

Górriz, J. M., Ramìrez, J., Lang, E. W. & Puntonet, C. G. (2008). Jointly Gaussian PDF-Based Likelihood Ratio Test for Voice Activity Detection. *IEEE Transactions on Audio, Speech and Language Processing*, *16* (8), 1565–1578.

Hansen, J. H. L. (1988). *Analysis and Compensation of Stressed and Noisy Speech With Application to Robust Automatic Recognition*. Dissertation, Georgia Institute of Technology, Faculty of the Division of Graduate Studies.

Hansen, J. H. L. & Varadarajan, V. (2009). Analysis and Compensation of Lombard Speech Across Noise Type and Levels With Application to In-Set/Out-of-Set Speaker Recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *17* (2), 366-378.

Hanson, H. M. (1997). Glottal characteristics of female speakers: Acoustic correlates. *The Journal of the Acoustical Society of America*, *101 (1)*, 466–481.

Hanson, H. M. & Chuang, E. S. (1999). Glottal characteristics of male speakers: Acoustic correlates and comparison with female data. *The Journal of the Acoustical Society of America*, *106 (2)*, 1064–1077.

Harwardt, C. (2009). Sprachverarbeitung militärisch relevanter Audiodaten. In M. W. und J. Grosche (Hrsg.), *Verteilte Führungsinformationssysteme* (S. 179-190). Berlin Heidelberg: Springer-Verlag.

Harwardt, C. (2010a). Comparing Feature Extraction Methods for Speaker Verification with Vocal Effort Mismatch in Training and Test Data. In *Proceedings of the International Association of Forensic Phonetics and Acoustics 2010 Annual Conference (IAFPA)* (S. 15).

Harwardt, C. (2010b). Investigating the COG Ratio as Feature for Speaker Verification on High-Effort Speech. In *Proceedings of the DiSS-LPSS Joint Workshop 2010 – The 5th Workshop on Disfluency in Spontaneous Speech – The 2nd International Symposium on Linguistic Patterns in Spontaneous Speech* (S. 35-38).

Harwardt, C. (2010c). Vergleich von Merkmalsextraktionsverfahren für die automatische Sprecherverifikation bei Nichtübereinstimmung des Stimmaufwands in Trainings- und Testdaten. In *Tagungsband der 36. Jahrestagung für Akustik (DAGA)* (S. 995-996).

Harwardt, C. (2011a). Comparing the Impact of Raised Vocal Effort on Various Spectral Parameters. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech)* (S. 2941-2944).

Harwardt, C. (2011b). Investigations on the speech spectrum of normal and loud speech. In *Proceedings of the International Association of Forensic Phonetics and Acoustics 2011 Annual Conference (IAFPA)* (S. 26-27).

Harwardt, C., Gottsmann, F. & Noubours, S. (2011). On the Relationship between Vocal Effort and Spectral Moments. In *Proceedings of the 6th Workshop on Speech in Mobile and Pervasive Environment (SiMPE)* (S. 1-4).

Heldner, M., Strangert, E. & Deschamps, T. (1999). A Focus Detector using Overall Intensity and High Frequency Emphasis. In *Proceedings of the 14th International Congress of Phonetic Sciences (ICPhS)* (S. 1491–1494).

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, *87*, 1738-1752.

Hess, W. J. (2008). Pitch and Voicing Determination of Speech with an Extension Toward Music Signals. In *Springer Handbook of Speech Processing* (S. 181-211). Berlin Heidelberg.

Hollander, M. & Wolfe, D. A. (1999). *Nonparametric Statistical Methods*. John Wiley & Sons, Inc.

Horthorn, T., Hornik, K., Wiel, M. A. van de & Zeileis, A. (2010). *Package 'coin'* (Forschungsbericht). CRAN. Verfügbar unter `http://cran.r-project.org/web/packages/coin/coin.pdf`

Iseli, M. (2007). *Dependencies of Voice Source Measures on Age, Sex, Vowel Context, and Prosodic Features*. Dissertation, University of California, Electrical Engineering Department.

Iseli, M., Shue, Y.-L. & Alwan, A. (2007). Age, sex, and vowel dependencies of acoustic measures related to the voice source. *Journal of the Acoustical Society of America*, *121*, 2283-2295.

Jessen, M., Köster, O. & Gfroerer, S. (2005). Influence of vocal effort on average and variability of fundamental frequency. *International Journal of Speech, Language and the Law*, *12(2)*, 174-213.

Junqua, J.-C. (1993). The Lombard reflex and its role on human listeners and automatic speech recognizers. *Journal of the Acoustical Society of America*, *93* (1), 510-524.

Junqua, J.-C. & Anglade, Y. (1990). Acoustic and Perceptual Studies of Lombard Speech: Application to Isolated-Words Automatic Speech Recognition.

In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (S. 841-844).

Kajarekar, S., Bratt, H., Shriberg, E. & Leon, R. de. (2006). A Study of Intentional Voice Modifications for Evading Automatic Speaker Recognition. In *Proceedings of the IEEE Odyssey - The Speaker and Language Recognition Workshop* (S. 1 - 6).

Kienast, M. & Sendlmeier, W. F. (2000). Acoustic Analysis of Spectral and Temporal Changes in Emotional Speech. In *In Proceedings of the ISCA Tutorial and Research Workshop on Speech and Emotion (SpeechEmotion)* (S. 92-97).

Kinnunen, T. & González Hautamäki, R. (2005). Long-Term F0 Modeling for Text-Independent Speaker Recognition. In *Proceedings of the 10th International Conference on Speech and Computer (SPECOM)* (S. 567-570).

Künzel, H. J. (1987). *Sprechererkennung - Grundzüge forensischer Sprachverarbeitung*. Heidelberg: Kriminalistik Verlag.

Kohler, K., Pätzold, M. & Simpson, A. (1995). *From scenario to segment - The controlled elicitation, transcription, segmentation and labelling of spontaneous speech* (Arbeitsberichte Nr. 29). Institut für Phonetik und digitale Sprachverarbeitung - Universität Kiel.

Labutin, P., Koval, S. & Raev, A. (2007). Speaker Identification based on the statistical analysis of F0. In *Proceedings of the International Association of Forensic Phonetics and Acoustics 2007 Annual Conference (IAFPA).*

Ladefoged, P. & McKinney, N. P. (1963). Loudness, Sound Pressure, and Subglottal Pressure in Speech. *The Journal of the Acoustical Society of America*, *35* (4), 454-460.

Legetter, C. J. & Woodland, P. C. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, *9*, 171-185.

Liénard, J.-S. & Di Benedetto, M.-G. (1999). Effect of vocal effort on spectral properties of vowels. *Journal of the Acoustical Society of America*, *106*, 411-422.

Lu, Y. (2010). *Production and Perceptual Analysis of Speech Produced in Noise*. Dissertation, The University of Sheffield, Department of Computer Science.

Lu, Y. & Cooke, M. (2009). The contributionof changes in F0 and spectral tilt to increased intelligibility of speech produced in noise. *Speech Communication*, *51*, 1253-1262.

Machelet, K. (2010). *Das Lesen von Sonagrammen - Kapitel I - Grundlagen: Sonagraph, Sonagramm, Koartikulation und Segmentierbarkeit.* Verfügbar unter `http://www.phonetik.uni-muenchen.de/studium/skripten/SGL/SGLKap1.html#Spektrum` (letzter Zugriff: 13.09.2010)

Makhoul, J. (1975). Linear Prediction: A Tutorial Review. In *Proceedings of the ieee* (Bd. 63, S. 561-580).

Martin, A., Doddington, G., Kamm, T., Ordowski, M. & Przybocki, M. (1997). The DET Curve in Assessment of Detection Task Performance. In *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech)* (S. 1895-1898).

Moore, T. J. & Bond, Z. S. (1987). Acoustic-Phonetic Changes in Speech due to Environmental Stressors: Implications for Speech Recognition in the Cockpit. In *Proceedings of the Fourth International Symposium on Aviation Psychology* (S. 77-83).

The NIST Year 2010 Speaker Recognition Evaluation Plan. (2010). Verfügbar unter `http://www.itl.nist.gov/iad/mig/tests/sre/2010/NIST_SRE10_evalplan.r6.pdf`

Noll, A. M. (1964). Short-Time Spectrum and "Cepstrum"Techniques for Vocal-Pitch Detection. *Journal of the Acoustical Society of America*, *36*, 296-302.

Pisoni, D. B., Bernacki, R. H., Nusbaum, H. C. & Yuchtman, M. (1985). Some Acoustic-Phonetic Correlates od Speech Produced in Noise. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (S. 1581-1584).

Pisoni, D. B. & Yuchtman, M. (1985). Acoustic-phonetic properties of vowels produced in noise. *Journal of the Acoustical Society of America, Supplement 1, 78*.

Pompino-Marschall, B. (2003). *Einführung in die Phonetik* (2. Aufl.). Berlin New York: Walter de Gruyter.

R Development Core Team. (2010). R: A language and environment for statistical computing [Software-Handbuch]. Vienna, Austria. Verfügbar unter `http://www.R-project.org` (ISBN 3-900051-07-0)

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

Rajasekaran, P. K., Doddington, G. R. & Picone, J. W. (1986). Recognition of Speech under Stress and Noise. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (S. 733-736).

Reynolds, D. (2002). An Overview of Automatic Speaker Recognition. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (S. 4072-4075).

Reynolds, D., Andrews, W., Campbell, J., Navrátil, J., Peskin, B., Adami, A. et al. (2002). *SuperSID Project Final Report – Exploiting High-Level Information for High-Performance Speaker Recognition* (Forschungsbericht). Department of Defense; National Science Foundation. (http://www.clsp.jhu.edu/ws2002/groups/supersid/, letzter Zugriff: 02.04.07)

Reynolds, D. & Campbell, W. M. (2008). Text-Independent Speaker Recognition. In Benesty, Sondhi & Huang (Hrsg.), *Springer Handbook of Speech Processing* (S. 763-781). Berlin Heidelberg: Springer-Verlag.

Reynolds, D., Quatieri, T. F. & Dunn, R. B. (2000). Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, *10*, 19-41.

Rose, P. (2002). *Forensic Speaker Identification*. London: Taylor & Francis.

Rosenberg, A. E., Bimbot, F. & Parthasarathy, S. (2008). Overview of Speaker Recognition. In Benesty, Sondhi & Huang (Hrsg.), *Springer Handbook of Speech Processing* (S. 725-741). Berlin Heidelberg: Springer-Verlag.

Rostolland, D. (1982a). Acoustic Features of Shouted Voice. *Acustica*, *50*, 118-125.

Rostolland, D. (1982b). Phonetic Structure of Shouted Voice. *Acustica*, *51*, 80-89.

Rostolland, D. & Parant, C. (1974). Analyse Physique de la Voix Criée. In *Proceedings of the 8th International Congress on Acoustics* (S. 240).

Sachs, L. (2002). *Angewandte Statistik - Anwendung statistischer Methoden - Zehnte, überarbeitete und aktualisierte Auflage* (10. Aufl.). Berlin Heidelberg New York: Springer-Verlag.

Scheffer, N., Ferrer, L., Graciarena, M., Kajarekar, S., Shriberg, E. & Stolcke, A. (2011). The SRI NIST 2010 Speaker Recognition Evaluation System. In *Proceedings of*

the International Conference on Acoustics, Speech and Signal Processing (ICASSP) (S. 5292-5295).

Schneider, K. & Möbius, B. (2007). Word stress correlates in spontaneous child-directed speech in German. In *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech)* (S. 1394-1397).

Schukat-Talamazzini, E. G. (1995). *Automatische Spracherkennung - Statistische Verfahren der Musteranalyse*. Braunschweig: Vieweg Verlag.

Schulman, R. (1985a). Articulatory Targeting and Perceptual Constancy of Loud Speech. In *Phonetic experimental research at the Institute of Linguistics, University of Stockholm (PERILUS)* (S. 86-91).

Schulman, R. (1985b). Dynamic and perceptual constraints of loud speech. *Journal of the Acoustical Society of America, Supplement 1, 78*.

Schulman, R. (1989). Articulatory dynamics of loud and normal speech. *Journal of the Acoustical Society of America, 85 (1)*, 295-312.

Shahin, I. (2006). Enhancing speaker identification performance unter the shouted talking condition using second-order circular hidden Markov models. *Speech Communication, 48*, 1047-1055.

Shriberg, E. (2007). Higher-Level Features in Speaker Recognition. In C. Müller (Hrsg.), *Speaker Classification I - Fundamentals, Features, and Methods* (S. 241-259). Berlin Heidelberg New York: Springer-Verlag.

Shriberg, E., Graciarena, M., Bratt, H., Kathol, A., Kajarekar, S., Jameel, H. et al. (2008). Effects of Vocal Effort and Speaking Style on Text-Independent Speaker Verification. In *Proceedings of the 9th Annual Conference of the International Speech Communication Association (Interspeech)* (S. 609-612).

Sjölander, K. & Beskow, J. (2000). WAVESURFER- AN OPEN SOURCE SPEECH TOOL. In B. Yuan, T. Huang & X. Tang (Hrsg.), *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP)* (S. 464-467).

Sluijter, A. M. C. & Heuven, V. J. van. (1996). Spectral balance as an acoustic correlate of linguistic stress. *Journal of the Acoustical Society of America, 100*, 2471-2485.

Sönmez, K., Shriberg, E., Heck, L. & Weintraub, M. (1998). Modeling Dynamic Prosodic Variation for Speaker Verification. In *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP)* (S. 3189-3192).

Stanton, B. J., Jamieson, L. H. & Allen, G. D. (1988). Acoustic-Phonetic Analysis of Loud and Lombard Speech in Simulated Cockpit Conditions. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (S. 331-334).

Stevens, K. N. (1989). On the quantal nature of speech. *Journal of Phonetics, 17*, 3-45.

Trancoso, I. (Hrsg.). (1996). *Technical Proceedings of the Workshop on Speech Under Stress Conditions* (Bd. AC/243(Panel 3)TP/5).

Traunmüller, H. (1997). Perception of speaker sex, age, and vocal effort. *Phonum, 4*, 183-186.

Traunmüller, H. & Eriksson, A. (2000). Acoustic effects of variation in vocal effort by men, women, and children. *Journal of the Acoustical Society of America, 107 (6)*, 3438–3451.

van Leeuwen, D. A. & Brümmer, N. (2007). An Introduction to Application-Independent Evaluation of Speaker Recognition Systems. In C. Müller (Hrsg.), *Speaker Classification I - Fundamental, Features, and Methods* (S. 330-353). Berlin Heidelberg New York: Springer-Verlag.

van Son, R. J. J. H. & Pols, L. C. W. (1999). An acoustic description of consonant reduction. *Speech Communication*, *28*, 125-140.

van Son, R. J. J. H. & van Santen, J. P. H. (2005). Duration and spectral balance of intervocalic consonants: A case for efficient communication. *Speech Communication*, *47*, 100-123.

van Summers, W., Pisoni, D. B., Bernacki, R. H., Pedlow, R. I. & Stockes, M. A. (1988). Effects of noise on speech production: Acoustic and perceptual analyses. *The Journal of the Acoustical Society of America*, *84 (3)*, 917–928.

Vary, P., Heute, U. & Hess, W. (1998). *Digitale Sprachsignalverarbeitung*. Stuttgart: Teubner.

Verlinde, V., Swail, Steeneken, Leeuwen van, Trancoso, South et al. (2000). *The Impact of Speech Under 'Stress' on Military Speech Technology* (Forschungsbericht). North Atlantic Treaty Organization (NATO).

Wenndt, S. J., Cupples, E. J. & Floyd, R. M. (2002). A Study on the Classification of Whispered and Normally Phonated Speech. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP – Interspeech)* (S. 649-652).

Wesker, T., Meyer, B., Wagener, K., Anemüller, J., Mertins, A. & Kollmeier, B. (2005). Oldenburg Logatome Speech Corpus (OLLO) for Speech Recognition Experiments with Humans and Machines. In *Proceedings of the 9th European Conference on Speech Communication and Technology (Interspeech – Eurospeech)* (S. 1273-1276).

Wissenschaftlicher Rat der Dudenredaktion (Hrsg.). (2001). *Duden - Deutsches Universalwörterbuch*. Mannheim: Dudenverlag.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X. A. et al. (2006). *The HTK Book (for HTK Version 3.4)*. Cambridge: Cambridge University Engineering Department.

Zhang, C. & Hansen, J. H. L. (2008a). Effective Segmentation based on Vocal Effort Change Point Detection. In *Proceedings of the ISCA Tutorial and Research Workshop on Speech Analysis and Processing for Knowledge Discovery.*

Zhang, C. & Hansen, J. H. L. (2008b). An Entropy based Feature for Whispered-Island Detection within Audio Streams. In *Proceedings of the 9th Annual Conference of the International Speech Communication Association (Interspeech)* (S. 2510-2513).

Zhou, B. & Hansen, J. H. L. (2005). Efficient Audio Stream Segmentation via the Combined T^2 Statistic and Bayesian Information Criterion. *IEEE Transactions on Speech and Audio Processing*, *13*, 467-474.

Zwicker, E. (1982). *Psychoakustik*. Berlin Heidelberg New York: Springer-Verlag.