

Eine akustisch–phonetische Untersuchung zur
Stimmverstellung

Schriftliche Hausarbeit zur Erlangung des Grades
eines Magister Artium (M.A.)
der Philosophischen Fakultät
der Christian–Albrechts–Universität
zu Kiel

vorgelegt von
Ramona Lorenzen
März 2004

Referent: Prof. Dr. Jonathan M. Harrington

Koreferent: PD Dr. Adrian P. Simpson

Tag der mündlichen Prüfung:

Zur Vervielfältigung genehmigt:

Kiel, den

Dekan: Prof. Dr. Albert Meier

Erklärung:

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig und ohne fremde Hilfe angefertigt und außer der angegebenen Literatur keine weiteren Hilfsmittel verwendet habe. Ferner versichere ich, dass diese Arbeit noch nicht zum Zwecke der Erlangung der Magisterwürde an anderer Stelle vorgelegen hat.

Inhaltsverzeichnis

1	Einleitung	1
2	Basis für eine Analyse zur Stimmverstellung	5
2.1	Sprechervariabilität und Sprechererkennung	5
2.1.1	Ursachen der Sprechervariabilität	7
2.1.2	Bereiche und Verfahren der Sprechererkennung	9
2.1.2.1	Aufgaben der Sprechererkennung	9
2.1.2.2	Bereiche der Sprechererkennung	10
2.1.3	Ermittlung sprecherspezifischer Merkmale	13
2.1.3.1	Auditiv–perzeptiver Ansatz	14
2.1.3.2	Akustisch–phonetischer Ansatz	16
2.1.3.3	Automatischer Ansatz	20
2.2	Definition und Auftreten von Stimmverstellung	25
2.2.1	Definitionen	25
2.2.2	Auftreten von Stimmverstellung	27
2.3	Studien zur Stimmverstellung	28
2.3.1	Grundlagenforschung zur Stimmverstellung	28
2.3.2	Studien im forensischen Kontext	34
3	Theorie des Untersuchungsgegenstands	39
3.1	Produktion der Frikative	39
3.2	Tendenzen im Spektrum	40
3.3	Variabilität bei Frikativen	43
4	Analyse zur Stimmverstellung	47
4.1	Sprecher und Sprachmaterial	47
4.1.1	Exkurs: Beschreibung der verstellten Stimmen	48
4.2	Hypothesen für die Analyse	52
4.3	Durchführung der Untersuchungen	54

4.3.1	Aufbereitung des Sprachmaterials	54
4.3.2	Datenanalyse	54
4.3.2.1	Spektralmomente und Wahrscheinlichkeitsverteilungen	55
4.3.2.2	Klassifikation	56
4.3.2.3	ANOVA	57
5	Ergebnisse	59
5.1	Beurteilung der Sprechervariabilität	59
5.2	Deskriptive Statistik der Spektralmomente	61
5.3	Klassifikationen von COG und Mom2	64
5.4	ANOVA von COG und Mom2	67
6	Diskussion	71
6.1	Ergebnis- und Methodendiskussion	71
6.2	Konsequenzen	76
7	Zusammenfassung	79
	Literatur	83
A	Abbildungen	89
A.1	Voruntersuchung: Frikativspektren	90
A.2	Wahrscheinlichkeitsverteilungen	92
B	R-Scripts	93
B.1	Berechnung der Spektralmomente	93
B.2	Klassifikation	94
B.3	ANOVA	95

Abbildungsverzeichnis

2.1	Kombinierter Ansatz für die forensische Sprechererkennung	11
2.2	Darstellung eines automatischen Sprechererkennungssystems	21
2.3	Darstellung des Likelihood-ratios (LR) anhand eines Beweises E .	22
3.1	Gemittelte Spektren der Frikative [f, h, s, ʃ] zum zeitlichen Mittel- punkt	41
A.1	Gemittelte Spektren über die Frikative [f, s, ʃ] für alle Stimmen von Sprecher B und D	90
A.2	Spektren der Frikative [f, s, ʃ] aller verstellten Stimmen von Spre- cher B und D	91
A.3	Normalkurven von COG und Mom2 der einzelnen Stimmen für die Frikative [f, s, ʃ]	92

Tabellenverzeichnis

2.1	Merkmale des auditiv–perzeptiven Ansatzes	14
2.2	Merkmale des akustisch–phonetischen Ansatzes	18
4.1	Aufteilung der verwendeten Frikative (n) pro Stimme und Sprecher	55
5.1	Mittelwerte und Standardabweichungen für COG und Mom2 von [f, s, ʃ] der einzelnen Stimmen und der Sprecher	62
5.2	Verwechslungsmatrizen der Parameter COG und Mom2 von [f, s, ʃ]	65
5.3	Korrekte Anteile (in %) der gesamten Klassifikationen von COG und Mom2 für [f, s, ʃ]	66
5.4	Ergebnisse des Proportionstests von COG und Mom2 bei [f, s, ʃ] .	66
5.5	Ergebnisse der ANOVA für das COG	67
5.6	Ergebnisse der ANOVA für Mom2	68

Kapitel 1

Einleitung

Viele Menschen sind in der Lage, bekannte Personen in Situationen ohne Sichtkontakt nur anhand ihrer Stimme innerhalb kürzester Zeit zu identifizieren, beispielsweise am Telefon oder in der Dunkelheit. Die Relevanz dieser menschlichen Fähigkeit steigt im forensischen Kontext, in dem die Notwendigkeit zur Identifikation auch unbekannter Stimmen gegeben ist. In solch einem Fall wirken aber viele Faktoren auf den Hörer ein, wie zum Beispiel Stress in der Tatsituation, der Zeitraum zwischen Tat und Aufforderung zur Identifikation, die Dauer des Gehörten usw. Die Einflüsse, denen der Hörer ausgesetzt ist, können dazu führen, dass eine Identifikation nicht mehr mit ausreichender Sicherheit möglich ist. In solch einer (hypothetischen) Situation ist ein forensischer Experte nicht mehr ausschließlich zur objektiven Urteilsabsicherung des linguistisch naiven Hörers erforderlich, sondern vor allem zur Durchführung einer forensischen Sprechererkennung. Die Vorgehensweise dieser basiert auf einer auditiven Beurteilung einer Sprach-Aufnahme, die von messphonetischen Verfahren unterstützt wird. Die Arbeit des Experten wird neben eventuell schlechter Qualität und kurzer Dauer des vorhandenen Sprachmaterials dadurch erschwert, dass menschliche Stimmen auch in sich selbst variieren können und sich nicht nur von den Stimmen anderer Sprecher unterscheiden.

Das Ziel der durchgeführten Analyse eines Experten liegt vor allem darin, Merkmale in dem vorliegenden Sprachmaterial zu finden, die ausschließlich den Sprecher kennzeichnen, möglichst wenig innerhalb des vorhandenen Materials variieren und dabei aber möglichst große Unterschiede zu vergleichbarem Material anderer Sprecher aufweisen.

Jedes Sprachmaterial in einem Kriminalfall, das wegen Stimmbeteiligung vom fo-

rensischen Experten¹ analysiert wird, weist die potentielle Gefahr von Stimmverstellung auf, die im forensischen Umfeld den Zweck erfüllt, individuelle Merkmale eines Sprechers zu verdecken und die Variabilität der eigenen Stimme zu erhöhen.

In der vorliegenden Arbeit wird es darum gehen, vor dem Hintergrund der forensischen Sprechererkennung inklusive einer Betrachtung der einzelnen Bereiche, die Problematik der Stimmverstellung herauszuarbeiten und eine Methode für eine eigene Analyse zur Stimmverstellung festzulegen.

Es ist das Ziel dieser Arbeit, herauszufinden, inwiefern Stimmverstellung einen Einfluss auf akustisch-phonetische Untersuchungsparameter im Sprachsignal ausübt und herauszufinden, ob mit diesen Parametern eine Sprecheridentifikation an verstellten Stimmen von zwei Sprechern (im Folgenden mit B und D abgekürzt), die in Stimmverstellung trainiert sind, möglich ist. Zusätzlich wird darauf geachtet, ob sich ein Unterschied in der Fähigkeit zur Stimmverstellung hinsichtlich der untersuchten akustischen Merkmale der beiden Sprecher belegen lässt, da sie beide in der Produktion verstellter Stimmen geübt sind. Die Intention der Stimmverstellung im forensischen Kontext liegt darin, individuelle Merkmale der Stimme und der Sprache soweit zu verdecken, dass ein Sprecher anhand dieser nicht mehr identifiziert werden kann. In der vorliegenden Arbeit handelt es sich jedoch nicht um einen realen forensischen Kontext, sondern um den speziellen Fall des Trainings in der Stimmverstellung. Die Folge davon ist ebenfalls, dass die Sprecher ihre eigenen sprecherspezifischen Merkmale soweit verdecken und nicht mehr erkannt werden kann, ob es sich um den vermuteten Sprecher handelt. Insofern ist eine Untersuchung im forensischen Kontext gerechtfertigt.

Die Hypothese der vorliegenden Arbeit ist, dass die verstellten Stimmen der beiden Sprecher B und D anhand der Energieverteilungen bei verschiedenen Frikativen trotz der Stimmverstellung korrekt identifiziert werden können. Die Basis für diese Hypothese stützt sich einerseits auf artikulatorische Gegebenheiten, andererseits auf bekannte Fakten über das spektrale Verhalten von Frikativen. Zur Untersuchung der Hypothese wird ein Klassifikationsverfahren aus der automatischen Sprach- und Sprechererkennung angewendet, das auf einem Wahrscheinlichkeitsmodell basiert.

Der Literaturteil dieser Arbeit besteht aus den Kapiteln 2 und 3. Kapitel 2 stellt die Ursachen und Auswirkungen der Sprechervariabilität, die Bereiche der Spre-

¹Es wird im Rahmen dieser Arbeit vorausgesetzt, dass der im forensischen Fall involvierte Experte phonetisches und sprachwissenschaftliches Wissen besitzt, das ihn qualifiziert, solche Untersuchungen durchzuführen.

chererkennung, mit denen sprecherspezifische Merkmale ermittelt werden können sowie theoretische Grundlagen und einige bisherige Studien zur Stimmverstellung vor. Kapitel 3 erläutert den theoretischen Hintergrund der Frikative, die aufgrund verschiedener Aspekte als Untersuchungsgegenstand ausgewählt wurden.

Der experimentelle Teil dieser Arbeit beginnt mit dem Kapitel 4, das die Durchführung einer eigenen Analyse zur Stimmverstellung darstellt. In Kapitel 5 werden die Ergebnisse dieser Analyse dargestellt. In Kapitel 6 werden die Ergebnisse der Analysen diskutiert.

Um die Struktur des Literaturteils darzustellen und um diese auch während des Lesens nachverfolgen zu können, folgt nun eine detaillierte Vorschau über den Literaturteil dieser Arbeit.

Unterschiede in gesprochener Sprache existieren zwischen verschiedenen Sprechern (interindividuell), aber auch innerhalb eines Sprechers (intraindividuell). In Abschnitt 2.1 wird diese so genannte Sprechervariabilität mit ihren Ursachen und Ausprägungen ausführlich dargestellt. Da sprecherspezifische Merkmale sowohl auditiv als auch messphonetisch ermittelt werden können, wird ebenfalls in diesem Abschnitt ein Überblick über verschiedene Ansätze der (forensischen) Sprechererkennung gegeben. In Abschnitt 2.2 wird das Phänomen Stimmverstellung vorgestellt. Zunächst werden verschiedene Definitionen betrachtet. Dann wird beschrieben, wann Stimmverstellung überhaupt auftritt und wieso Forschung in diesem Gebiet sinnvoll ist. Stimmverstellung stellt für verschiedene Bereiche der Sprechererkennung ein Problem dar. Anhand ausgewählter Studien in Abschnitt 2.3 wird diese Problematik näher erläutert.

Kapitel 3 reproduziert die theoretischen Grundlagen der Artikulationsart *Frikative*, die als Untersuchungsgegenstand ausgewählt wurde. Das Kapitel enthält eine kurze Darstellung der artikulatorischen Grundlagen in Abschnitt 3.1 und der akustischen Grundlagen der Frikativproduktion in Abschnitt 3.2. Im Anschluss daran wird in Abschnitt 3.3 eine Studie vorgestellt, in der Variabilität von Frikativen in artikulatorischer als auch akustischer Hinsicht untersucht wurde.

Kapitel 4 beschäftigt sich mit der Durchführung eines eigenen Experiments zur Analyse von Stimmverstellung. Die Sprecher und das Sprachmaterial für das Experiment werden in Abschnitt 4.1 vorgestellt. In Abschnitt 4.1.1 wird das verwendete Sprachmaterial auditiv oberflächlich in einem existierenden Rahmen beurteilt. In Abschnitt 4.2 wird die Hypothese auf Grundlage des Literaturteils und des vorliegenden Sprachmaterials entwickelt und dargestellt. Abschnitt 4.3 beschreibt dann die Durchführung der Analyse.

Kapitel 2

Basis für eine Analyse zur Stimmverstellung

In diesem Kapitel wird die theoretische Basis für eine Analyse zur Stimmverstellung festgelegt. Um die Gesamtproblematik der Stimmverstellung in der forensischen Sprechererkennung verstehen zu können, werden zunächst einige allgemeine Sachverhalte dargestellt.

2.1 Sprechervariabilität und Sprechererkennung

Bei der Verwendung von Sprache zwischen einem Sprecher und einem Hörer entsteht eine Kommunikation¹, die einen vielfältigen Zweck erfüllen kann und verschiedene linguistische Funktionen hat (Jakobson 1960). So kann ein Sprecher einen Hörer durch die *appellative Funktion* zu etwas auffordern oder ihn zu bestimmten Reaktionen bewegen und ihn durch die *referentielle Funktion* über einen Gesprächsgegenstand informieren. Er kann aber auch durch die *metasprachliche Funktion* über die verwendete Sprache selbst reden. Weitere sprachliche Funktionen, die möglich sind, werden als *expressive Funktion* und *phatische Funktion* bezeichnet. Letztere wird häufig in Sprache mit Kleinkindern oder Tieren verwendet; nach Jakobson (1960) dient sie auch dazu, das Kontaktmedium zu überprüfen und festzustellen, ob die Kommunikation noch erfolgreich ist. Die expressive Funktion dient der Übermittlung eigener Emotionen, Einstellungen und Meinungen zu einem Gesprächsgegenstand.

¹Kommunikation ist definiert als „[...] jede Form von wechselseitiger Übermittlung von Information durch Zeichen/ Symbole zwischen Lebewesen [...] oder zwischen Menschen und datenverarbeitenden Maschinen“ (Bußmann 1990, S. 392)

Die verschiedenen sprachlichen Funktionen von Jakobson (1960) stellen nur ein mögliches Kommunikationsmodell dar und es kann im Rahmen der vorliegenden Arbeit nicht näher darauf eingegangen werden, es ist jedoch wichtig, Folgendes zu berücksichtigen: Verschiedene Faktoren wirken auf eine Äußerung eines Sprechers im Rahmen von Kommunikation ein. So kann die Bedeutung eines Aussagesatzes mit referentieller Funktion durch Änderung der Intonation die Bedeutung eines Fragesatzes mit einer appellativen Funktion erhalten, die vom Hörer eine Reaktion erfordert. So erhält der Beispielsatz „Der Mann hat es getan.“ durch fallende Intonation die Qualität eines Aussagesatzes mit dem Zweck der Informationsübermittlung. Der gleiche Satz wird aber zu einer Frage mit dem Zweck des Informationserhaltes, wenn er mit steigender Intonation produziert wird, orthographisch in diesem Fall nur durch ein Fragezeichen gekennzeichnet: „Der Mann hat es getan?“. Zusätzlich ändert sich der Fokus mit Variation in der Akzentuierung (Kohler 1995). Die Sprechweise eines Sprechers kann also die linguistischen Funktionen ändern.

Unabhängig von linguistischen Funktionen der Sprache, die einen kommunikativen Zweck erfüllen, überträgt ein Sprecher auch immer zusätzlich para-linguistische Information, die Eigenschaften von ihm selbst preisgibt, beispielsweise sein Alter und Geschlecht, seinen Gesundheitszustand oder seine Stimmung:

„The acoustic speech signal therefore carries not only the linguistic message which the speaker wishes to convey, but also a host of para-linguistic information about the speaker’s identity, gender, emotions, state of health, and other characteristics which may be perceived by the listener.“ (Mokhtari 1998, S. 1)

Aus diesem Grund ist ein Hörer in der Lage, einen bekannten Menschen ohne Sichtkontakt nur anhand seiner Stimme als die betreffende Person zu erkennen. Solch eine Situation kann am Telefon auftreten, in der Dunkelheit oder in einer Menschenmenge. Sind Sprecher und Hörer einander unbekannt, können die para-linguistischen sprecherspezifischen Merkmale aber ebenfalls dazu führen, dass der Hörer den Sprecher wiedererkennt. Eine solche Situation liegt beispielsweise in einem Verbrechen vor, wenn ein Straftäter und ein Zeuge beteiligt sind, aber kein Sichtkontakt möglich war und andere Hinweise zur Aufklärung des Verbrechens nicht ausreichend sind. Der Zeuge kann, beziehungsweise muss dann aufgrund der stimmlichen und sprachlichen Informationen in der Lage sein, die wahrgenommenen Eigenschaften des Sprechers (wie zum Beispiel Alter, Geschlecht, Dialekt, Sprachfehler) an Sachverständige oder Polizisten weiterzugeben. Diese können

dann, bei einer genauen Beschreibung, daraus folgend ein Sprecherprofil erstellen und veranlassen eventuell eine auditive Gegenüberstellung (Broeders und Amelvoort 1999).

In einer Übersicht über verschiedene Bereiche der Sprechererkennung in Abschnitt 2.1.2 wird die gerade beschriebene *auditive Sprechererkennung* noch etwas ausgeführt.

Ein weiterer Aspekt, der in diesem Literaturüberblick berücksichtigt werden muss, ist die Variabilität in den phonetisch relevanten Bereichen Artikulation, Akustik und Perzeption sowie in der Sprache selbst (bezüglich der linguistischen Funktionen beispielsweise). Variation in diesen Bereichen, ob nun beabsichtigt oder unbeabsichtigt, ist ein großes Problem für die Forschung, beispielsweise bei der Erstellung eines einheitlichen Kommunikationsmodells, aber auch für Anwendung von Spracherkennungssystemen oder Hilfen für behinderte Menschen (Mokhtari 1998). Gerade auch für die Sprechererkennung stellt Variabilität ein Problem dar. Wie die Variabilität entsteht und welche Auswirkungen sie speziell auf die forensische Sprechererkennung hat, wird im nächsten Abschnitt dargestellt.

2.1.1 Ursachen der Sprechervariabilität

Sprechervariabilität kann sich in zwei Weisen ausprägen. Einerseits unterscheiden sich Äußerungen akustisch von einem Sprecher selbst bei einer Wiederholung eines Wortes, das heißt, wenn der linguistische Gehalt der Äußerung und die umgebenden Bedingungen gleich bleiben. Viele Muskeln und Prozesse müssen für die Sprachproduktion gesteuert werden und das ist in identischer Weise in zwei aufeinander folgenden Äußerungen fast unmöglich. Dies wird in der englischsprachigen Literatur als *intratalker-variability* bezeichnet (Tosi 1979). Die Variabilität von Äußerungen eines Sprechers kann zusätzlich noch durch Stimmverstellung oder Imitation erhöht werden, aber auch die Zeit, die zwischen zwei von ihm produzierten Äußerungen liegt, führt dazu, dass Unterschiede auftreten. Natürlich spielen auch emotionale und gesundheitliche Einflüsse eine Rolle und haben Auswirkungen auf Äußerungen eines Sprechers (Tosi 1979). Desweiteren ist Koartikulation ein Faktor, der zu intraindividuellem Variation führt, indem bestimmte Sprachlaute durch den Kontext beeinflusst werden.

Andererseits sind Äußerungen zwischen zwei Sprechern ebenfalls unterschiedlich, weil sich ihre Anatomie voneinander unterscheidet und/ oder weil sie ein unterschiedliches sprachliches Verhalten aufweisen. In der englischsprachigen Literatur wird die Variabilität zwischen verschiedenen Sprechern als *intertalker-variability*

(Tosi 1979) bezeichnet².

Im Folgenden wird die deutsche Bezeichnung *intraindividuell* für die Unterschiede, die bei einem Sprecher auftreten können, und die Bezeichnung *interindividuell* für die Unterschiede, die zwischen Sprechern existieren können, verwendet.

Diese beiden Formen der Sprechervariabilität können kombiniert den Grad der Sprecherspezifität eines untersuchten Merkmals angeben: Je größer die Variationen zwischen den Sprechern und je kleiner die Variationen innerhalb dieser sind, umso sprecherspezifischer ist das gewählte Merkmal (Wolf 1972). Diese Kombination ist Voraussetzung für eine erfolgreiche Identifikation und dient als Grundlage in forensischer, aber auch automatischer Sprechererkennung.

Vor allem interindividuelle Sprecherunterschiede haben nach Tosi (1979), Künzel (1987) und Mokhtari (1998) organische und erlernte Ursachen.

Organische Ursachen für interindividuelle Unterschiede sind das Alter, das Geschlecht, der Gesundheitszustand und die Größe des Sprechers. Aufgrund dieser Variablen sowie einem komplizierten Zusammenspiel von Muskeln, Knorpeln, Nerven und neuromuskulären Prozessen ergibt sich eine unterschiedliche Konfiguration des Artikulationsapparates bei der Sprachproduktion. Die Unterschiede zwischen den artikulatorischen Einstellungen des Ansatzrohres sind sprecherindividuell und werden als para-linguistische Information übertragen. Gerade die organischen Faktoren führen jedoch auch zu Variabilität in Äußerungen eines Sprechers, da es kaum möglich ist, in zwei aufeinanderfolgenden Äußerungen das Ansatzrohr auf identische Art zu formen, so dass eine exakt gleiche Äußerung entsteht (Tosi 1979; Mokhtari 1998).

Eine *erlernte* Ursache der Sprechervariabilität ist die Umgebung, in der ein Sprecher ständig lebt oder aufgewachsen ist und die sich auf sein sprachliches Verhalten auswirkt. Dies führt vor allem zur Ausprägung von Dialekt, Soziolekt und Idiolekt, die dazu dienen können, ihn von anderen Sprechern unterscheiden (Künzel 1987, 1989). Inwiefern organische und erlernte Faktoren zur Sprechererkennung beitragen können, wird in Abschnitt 2.1.3 näher beschrieben.

Zusätzliche Variation, die unabhängig von der Stimme und verwendeten Sprache des Sprechers ist, entsteht durch externe Faktoren, wie z.B. Störgeräusche in der Umgebung, „Echo“ und die Aufnahmebedingungen (Tosi 1979).

Die Gesamtheit der bestehenden intra- und interindividuellen Variabilität, wie sie

²Es existieren aber auch für die beiden Formen der Sprechervariabilität Bezeichnungen wie *intra- und interspeaker-variability* (Meuwly 2000b) oder *between-speaker-difference* und *within-speaker-variability* (Nolan 1983).

gerade beschrieben wurde, hat starken Einfluss auf die Leistung von Sprach- und Sprechererkennungssystemen und erschwert die Arbeit eines Experten im forensischen Umfeld.

2.1.2 Bereiche und Verfahren der Sprechererkennung

Sprecherabhängige Unterschiede hinsichtlich organisch produzierter Stimme und verwendeter Sprache sind gerade für die forensische Sprechererkennung von großem Nutzen. Wie schon in 2.1.1 beschrieben, entstehen inter- und intraindividuelle Sprecherunterschiede durch organische Faktoren, welche die im Artikulationsapparat produzierte Stimme beeinflussen, aber auch durch erlernte Faktoren, die sich auf die Verwendung der Sprache als Kommunikationsmittel auswirken. Ein Sprecher sollte also nur im Rahmen des gesamtsprachlichen Verhaltens beurteilt werden, wenn das Ziel ist, seine Identität in einem forensischen Kontext zu bestimmen.

Bevor nun die Ausprägungen einzelner sprecherspezifischer Merkmale in 2.1.3 beschrieben werden, folgt zunächst ein Überblick über die Aufgaben und Vorgehensweisen der forensischen Sprechererkennung.

2.1.2.1 Aufgaben der Sprechererkennung

Sprechererkennung erfüllt zwei Aufgaben: *Sprecheridentifikation* und *Sprecherverifikation* (Nolan 1983; Furui 1995; Broeders 2001). Es ist allerdings zu beachten, dass die genannten verwendeten Quellen Identifikation und Verifikation teilweise als Aufgaben sehen, die nur in der automatischen Sprechererkennung verwendet werden (Furui 1995), sie andererseits aber auch auf die gesamte Sprechererkennung beziehen, in der Experten involviert sind (Nolan 2001).

Die *Sprecheridentifikation* dient dazu, die Identität eines unbekanntem Sprechers in einer Menge von Sprechern anhand seiner Stimme zu ermitteln. Dieser Fall tritt auf, wenn nur das Sprachmaterial einer Straftat (*Tataufnahme*) zur Verfügung steht, aber noch kein Sprachmaterial eines Vergleichssprechers (*Vergleichsaufnahme*) vorhanden ist. Diese Situation wird nach Künzel (1987) als *Stimmenanalyse* bezeichnet. Das Ziel der Stimmenanalyse ist die Einschränkung des Personenkreises, aus dem der Straftäter stammen könnte.

Sprecheridentifikation kann im „open-set“ durchgeführt werden. Die Grundlage dafür bildet eine offene Population, beispielsweise die deutsch-sprechende Bevölkerung. Dabei besteht die Möglichkeit, dass der fragliche Sprecher nicht in der Po-

pulation vorhanden ist und somit nicht identifiziert werden kann. Sprecheridentifikation kann aber auch im „closed-set“ durchgeführt werden, so dass der fragliche Sprecher mit Sicherheit in der untersuchten Population vorhanden ist und darin nur noch zu identifiziert werden braucht. Ein Beispiel dafür bildet die Besatzung eines U-Bootes oder die Bevölkerung einer Insel. Die Sprecheridentifikation auf der Basis eines „closed-sets“ ist für die Forensik nicht ratsam (Champod und Meuwly 2000), da die Voraussetzungen eines solchen Verfahrens die Entscheidung negativ beeinflussen würden. Es kann im seltensten Fall von vornherein davon ausgegangen werden, dass ein Straftäter sich in einer festgelegten Gruppe von Menschen befindet, es sei denn es handelt sich um die sehr seltene und unrealistische Situation einer Straftat auf einem U-Boot mit einem unbekanntem Täter, der darauf gefunden werden muss, aber nicht durch andere Beweise außer seiner Stimme überführt werden kann (Meuwly 2004).

Im Gegensatz dazu handelt es sich bei der *Sprecherverifikation* um die Bestätigung der Identität eines Sprechers anhand festgelegter Untersuchungsparameter. Nach Gfroerer (2003) liegt oftmals in der forensischen Sprechererkennung die Sprecherverifikation vor, da eine bekannte Stimme (Verdächtiger) mit der Stimme der unbekanntem Person (Täter) verglichen und die Identität entweder bestätigt oder abgelehnt wird. Künzel (1987) bezeichnet dies als *Stimmenvergleich*.

Auch kommerzielle automatische Sprechererkennungssysteme erfüllen die Aufgabe der Sprecherverifikation, in denen die angegebene Identität durch „die Stimme als Schlüssel“ bestätigt oder verneint wird (Furui 1995).

2.1.2.2 Bereiche der Sprechererkennung

Sprechererkennung wird häufig in drei Bereiche unterteilt: Sprechererkennung durch „Laien“, (semiautomatische) Sprechererkennung durch Experten und automatische Sprechererkennung (Hollien 1990; Meuwly 2000a; Nolan 2001).

Die Sprechererkennung durch „Laien“ bezieht sich auf „linguistisch naive Personen“³, die Zeuge oder Opfer einer Straftat waren, dabei aber nur die Stimme des Täters gehört haben und ihn aufgrund dieser identifizieren. Die Identifikationsleistung der linguistisch naiven Person ist jedoch sehr subjektiv und von vielen Faktoren abhängig, wie z.B. der Dauer zwischen der Tat und der Gegenüberstellung mit der Stimme eines Verdächtigen, Stimmverstellung, der Länge des gehörten Sprachmaterials und anderen Einflüssen (Bull und Clifford 1984, Hollien, Majewski und Doherty 1982). Im günstigen Fall ist jedoch ein Experte anwesend, der die

³Bezeichnung nach Künzel (1990)

subjektiven und eventuell beeinflussten Urteile des Hörers mit messphonetischen Methoden absichert (Künzel 1987).

Da die vorliegende Arbeit sich jedoch auf Ansätze der Sprechererkennung durch Experten konzentriert, wird auf Vorgehensweisen und bestehende Probleme im Rahmen der naiven Sprechererkennung nicht weiter eingegangen.

Die forensische Sprechererkennung durch Experten basiert auf auditiv–perzeptiven, phonetisch–akustischen und automatischen Vorgehensweisen (Meuwly 2000a), die dem Zweck der Identifikation oder Verifikation dienen können. Diese Ansätze werden in Abbildung 2.1 schematisch dargestellt.

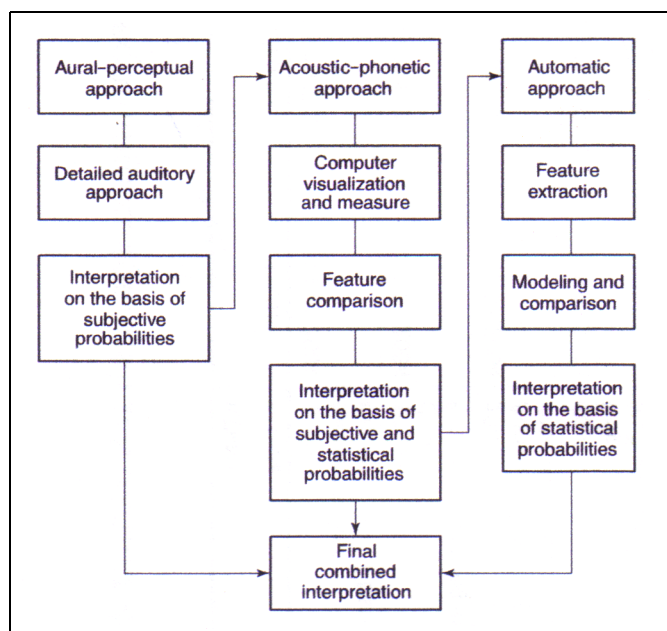


Abbildung 2.1: Kombiniertes Vorgehen für die forensische Sprechererkennung nach Meuwly (2000c)

Ein sehr wichtiger Bestandteil der forensischen Sprechererkennung ist die genaue auditive Beurteilung des vorliegenden Sprachmaterials (Künzel 1987; Meuwly 2000b), in Abbildung 2.1 als *auditiv–perzeptiver Ansatz* bezeichnet. Mit Hilfe bestimmter Anhaltspunkte, die sich in der Vergangenheit als günstig herausgestellt haben, werden sprachliche und stimmliche Eigenschaften des Sprechers vom phonetisch–forensischen Experten beurteilt. Sie werden in Abschnitt 2.1.3.1 ausführlicher dargestellt.

Die auditive Beurteilung der Sprechereigenschaften und eine daraus resultierende Interpretation des Sprachmaterials hängt stark von den Fähigkeiten und Erfah-

rungen des Gutachters ab und stellt somit trotz seiner phonetischen Kenntnis eine subjektive Beurteilung dar⁴. Dies betrifft vor allem auch das Erkennen der Stimmverstellung im Sprachmaterial, siehe Abschnitt 2.2.2. Dazu müssen allerdings umfangreiche Kenntnisse über das normalsprachliche Verhalten einer Person vorhanden sein (Künzel 1987). Ist das vorliegende Sprachmaterial des Täters und eines Verdächtigen derart unterschiedlich, dass eine Übereinstimmung der beiden Stimmen ausgeschlossen werden kann, so ist ein Abschluss der Analyse an dieser Stelle möglich (Abbildung 2.1). Im Fall einer auditiven Ähnlichkeit von Tat- und Vergleichsmaterial oder auch bei dem Verdacht von Stimmverstellung wird die Sprechererkennungsprozedur auf akustisch-phonetischer Ebene weitergeführt (Meuwly 2000b).

Der *akustisch-phonetische Ansatz* basiert auf der Visualisierung von Sprache mit Hilfe von Sonagrammen, die durch das Verfahren der Spektrographie aus dem vorhandenen Sprachmaterial erstellt werden. Besondere Eigenschaften des Sprechers, die einerseits durch auditive, aber auch visuelle Beobachtungen der Sonagramme auffallen, werden dann durch messphonetische Verfahren objektiviert (Künzel 1987). Mögliche sprecherspezifische Untersuchungsparameter im akustisch-phonetischen Bereich werden in Abschnitt 2.1.3.2 detaillierter beschrieben.

Da es sich in der forensischen Sprechererkennung um Identifikation oder Verifikation handelt, wird das Sprachmaterial möglicher Verdächtiger und das vorliegende Sprachmaterial des Straftäters beurteilt. Aufgrund dieser Auswertungen zwischen zwei oder mehreren Sprechern können statistische Verteilungen der untersuchten Merkmale erstellt werden, die dann Auskunft über die Ähnlichkeit der Sprecher geben, beispielsweise Histogramme (*Häufigkeitsverteilungen*) der Grundfrequenz. Unterscheiden sich die Histogramme erheblich (unter der Voraussetzung, dass Stimmverstellung oder andere stimmbeeinträchtigende Faktoren ausgeschlossen werden können), so ist es unwahrscheinlich, dass der Verdächtige auch der Täter war.

Auch der akustisch-phonetische Ansatz ist noch als subjektiv anzusehen, weil die Ergebnisse der messphonetischen Analysen vom Experten bewertet werden müssen. Allerdings ist eine höhere Objektivität durch Reproduzierbarkeit der Messungen möglich, im Gegensatz zu einer ausschließlich auditiven Analyse (Meuwly 2000a). Hat die akustisch-phonetische Analyse eine Übereinstimmung zwischen

⁴Die auditive Beurteilung durch einen Experten besitzt dennoch ein gewisses Maß an Objektivität durch Transkription mit der IPA, die von jedem phonetischen Experten produzierbar und nachvollziehbar ist bzw. sein sollte (Künzel 1987).

Tat- und Vergleichsaufnahme ergeben, ist wiederum ein Abschluss der Analyse des Sprachsignals möglich (Abbildung 2.1), oder sie kann mit Hilfe des automatischen Ansatzes noch weiter objektiviert werden.

Der *automatische Ansatz* basiert wiederum auf der Auswahl von geeigneten Untersuchungsparametern, gefolgt von einem automatischen Merkmalsvergleich, auf den ein menschlicher Betrachter, vor allem in kommerziellen Systemen, wenig Einfluss hat. Die Grundlage für diesen Ansatz bilden Modelle, die aus den erwarteten Wahrscheinlichkeiten des Auftretens eines Merkmals errechnet werden. Die genaue Vorgehensweise wird in Abschnitt 2.1.3.3 näher beschrieben.

2.1.3 Ermittlung sprecherspezifischer Merkmale

Unter der Berücksichtigung der Ursachen von intra- und interindividueller Variabilität, beschrieben in Abschnitt 2.1.1, sowie möglicher Ansätze (auditiv-perzeptiv, akustisch-phonetisch, automatisch) in der forensischen Sprechererkennung, dargestellt in Abschnitt 2.1.2, folgt nun eine Beschreibung der sprecherspezifischen Merkmale, die vom Experten mit Hilfe der verschiedenen Ansätze ermittelt und beurteilt werden können und zur Identifikation eines Sprechers beitragen. Diese Merkmale zeichnen sich dadurch aus, dass sie generell eine geringe intraindividuelle, aber eine hohe interindividuelle Variabilität aufweisen (Gfroerer 2003). Die „Identifikationsleistung“ der Merkmale nimmt jedoch stark ab, wenn wie sehr häufig im forensischen Kontext, das Telefon involviert ist und schlechte Übertragungsbedingungen vorherrschen (Masthoff 2004). Durch diese Faktoren werden die Frequenzbereiche beschnitten oder durch Störungen im Signal überlagert, so dass die Merkmale eventuell nicht mehr richtig bewertet werden können.

Eine Kombination aus auditiv-perzeptivem und akustisch-phonetischem Ansatz kann die Besonderheiten in der Sprache, Sprechweise und der Stimme eines Sprechers adäquat beschreiben. Der automatische Ansatz dient zur Objektivierung der beiden eher subjektiven Ansätze und basiert entweder auf den vorhergegangenen Analysen oder untersucht Parameter, die als sprecherspezifisch gelten, jedoch nicht auditiv oder akustisch im Sprachsignal zu bewerten sind (Meuwly 2000b).

2.1.3.1 Auditiv–perzeptiver Ansatz

Um eine umfangreiche und aussagekräftige Sprechercharakteristik zu erstellen, muss das gesamtsprachliche Verhalten betrachtet werden, da in einer Äußerung eines Sprechers nicht nur Eigenschaften der Artikulationsorgane codiert sind, sondern auch Eigenschaften, die Auskunft über die Herkunft, soziale Schicht sowie persönliche Bevorzugungen bestimmter sprachlicher Merkmale geben (Künzel 1987; Meuwly 2000b). Diese Merkmale werden auch als „high–level–speaker–characteristics“ (Van den Heuvel 1996) oder „high–level–information parameters“ (Battaner, Gil und Marrero 2003) bezeichnet.

Einer Beschreibung phonetischer und linguistischer Charakteristiken des Sprechers geht, soweit möglich, die Festlegung globaler Eigenschaften des Sprechers voraus. Dazu gehören Alter, Geschlecht, Sprechsituation (Monolog vs. Dialog, Lese– vs. Spontansprache) sowie emotionale und gesundheitliche Einflüsse, beispielsweise Stress, Depression, Erkältung (Gfroerer 2003).

Sprecher können durch auditiv–perzeptive Beurteilung hinsichtlich ihres Sprachverhaltens, ihrer organisch produzierten Stimme und ihrer Sprechweise näher betrachtet werden. Als vereinfachte Darstellung der dazugehörigen Merkmale dient Tabelle 2.1:

Sprache	Stimme	Sprechweise
Dialekt	Melodie	Sprechtempo
Soziolekt	Sprechstimmlage	Atemverhalten
Idiolekt	Stimmqualität	Pathologien

Tabelle 2.1: Merkmale des auditiv–perzeptiven Ansatzes nach Künzel (1987)

Im Bereich der Sprache (in der Bedeutung als sprachliches Verhalten) werden insbesondere Dialekt, Soziolekt, Eloquenz und Idiolekt beurteilt. Ein *Dialekt* kann dazu dienen, die Herkunft eines Sprechers auf einen regionalen Raum zu begrenzen und wird gerade in der Forensik als sehr wichtiges sprecherspezifisches Merkmal angesehen, sofern der Experte Wissen über die Verbreitung und die phonetische Ausprägung des gesprochenen Dialekts hat. Dies ist vor allem beim Verdacht von Stimmverstellung von äußerster Wichtigkeit (Künzel 1987). Auf die Problematik der Stimmverstellung wird jedoch in Abschnitt 2.2 gesondert eingegangen.

Der *Soziolekt* eines Sprechers gibt Auskunft über die gesellschaftliche Schicht, welcher er angehört. So zeichnen sich Professoren einer Universität sprachlich durch eine andere Wortwahl und Syntax aus als ein Fabrikarbeiter oder ein Handwer-

ker. Der Soziolekt hängt also sehr eng mit der *Eloquenz* (der sprachlichen Ausdrucksfähigkeit) zusammen, welche wiederum auf den Grad der Bildung zurückzuführen ist und sich zwischen verschiedenen Gesellschaftsschichten unterscheidet (Künzel 1987, 1989). Als *Idiolekt* werden sprachliche und phonetische Eigenschaften des Sprechers bezeichnet, die nur diesem Sprecher zuzuordnen sind und ihn charakterisieren. Dazu gehören Stereotypen (immer wiederkehrende Floskeln auf der Wortebene), Häitationen (Verzögerungen), Pausensetzung, besondere Eigenschaften in der Lautbildung, beispielsweise starke Nasalierung oder Behauchung, und Versprecher (Künzel 1989). Stark vereinfacht ist das Ziel des Sprechererkennungs-Experten, den Idiolekt eines Sprechers zu finden, inklusive dessen unverwechselbarer Merkmale in verschiedenen Bereichen der Sprache, Stimme und Sprechweise, und diesen von dem Idiolekt anderer Sprecher abgrenzen zu können.

Im Bereich der Stimme und der Sprechweise (siehe Tab. 2.1) gelten insbesondere „mittlere Sprechstimmlage“ und „Melodie“ der Stimme, die Stimmqualität, aber auch Pathologien, Sprechtempo, Atemverhalten und Rhythmus des Sprechers als sprecherspezifisch (Künzel 1987; Gfroerer 2003). Die *mittlere Sprechstimmlage* eines Sprechers ist der (bevorzugte) Tonhöhen-Bereich seiner Stimme beim Sprechen. Damit hängt auch die Abweichung der Tonhöhe nach oben bzw. nach unten zusammen, wodurch die wahrgenommene *Melodik* einer Stimme zustande kommt. Geringe Abweichungen führen zu einem monotonen Eindruck der Stimme. Ein wichtiges sprecherspezifisches Merkmal der *Stimmqualität* eines Sprechers ist die Glottalisierung, die entweder habituell (zum Beispiel immer am Ende einer Äußerung) oder durch Krankheit (Kehlkopfentzündung) bedingt sein kann.

Das *Sprechtempo* im Bereich der Sprechweise ist einerseits wichtig in Bezug auf Koartikulation bestimmter Sprachlaute oder Wortverbindungen (Künzel 1987)⁵. Schnell produzierte Sprache führt durch eine begrenzte Beweglichkeit der Artikulationsorgane zu verstärkter Assimilation, die sich aber an sprecherspezifischen Grenzen bemerkbar machen kann. Zudem gilt das Sprechtempo, beziehungsweise die mögliche Artikulationsgeschwindigkeit eines Sprechers als stabil innerhalb eines Sprechers (Künzel 1987).

Das *Atemverhalten* eines Sprechers ist im Bereich der Sprechweise ein ebenfalls wichtiges Merkmal. Sprecher können wiederum durch habituelles oder krankhaft bedingtes Atemverhalten auffallen. Werden Pausen an semantisch bzw. syntak-

⁵Es wird an dieser Stelle angenommen, dass der Begriff Koartikulation von Künzel (1987) *Assimilationen*, die durch zu schnelles Sprechen entstehen, miteinbezieht.

tisch ungünstigen Grenzen gesetzt, so kann dies einerseits auf eine Krankheit hindeuten, aber im Gegensatz dazu auch auf einen Bestandteil des Idiolekts des fraglichen Sprechers (Künzel 1987).

Die meisten bisher beschriebenen Merkmale sind situationsabhängig, d.h. Stress, Depression oder Angst können die Ausprägungen durch den Sprecher beeinflussen und führen zu Abweichungen der üblichen Stimmcharakteristiken, wobei diese dazu erst einmal bekannt sein müssen. Viele der beschriebenen Merkmale, vor allem im Bereich der Stimme (siehe Tab. 2.1) sind aber auch anfällig für Pathologien, die den Sprecher zusätzlich kennzeichnen. Ausgenommen von der Beeinflussung durch Pathologien ist lediglich der Dialekt (Künzel 1987).

2.1.3.2 Akustisch–phonetischer Ansatz

Basierend auf der auditiven Beurteilung des Sprachmaterials durch einen Phonetiker werden eventuell herausgefundene Besonderheiten durch akustische Messungen objektiviert.

Akustisch–phonetische Merkmale für die Sprechererkennung wurden in der Vergangenheit (und werden auch heute noch) häufig auf der Basis der Kurzzeit–Spektrographie ermittelt. Dieses Verfahren wurde als ein „hervorragendes Analyseverfahren für akustische Phänomene aller Art“ angesehen (Künzel 1987, S. 68). Die Spektrographie hat den Vorteil, Sprache durch *Sonagramme* sichtbar zu machen. Ein Sonagramm stellt das Sprachsignal in den Dimensionen Zeit, Frequenz und Intensität dar.

Der Vorteil, gesprochene Sprache visuell darstellen und Unterschiede zwischen Sprechern darin erkennen zu können, wurde von (Kersta 1962) in falscher Weise genutzt beziehungsweise falsch interpretiert. Er versuchte, Sonagramme als *Stimmabdruck* (*Voiceprint*) in Analogie zum Fingerabdruck zu verwenden, da Sonagramme akustische Charakteristiken wie Formanten abbilden und die Ausprägung dieser akustischen Eigenschaften sprecherspezifisch ist, beispielsweise die Lage der Formanten oder Energiekonzentrationen bei Frikativen. Auch wenn sich das Voiceprint–Verfahren als alleiniges Mittel zur Sprechererkennung als unwirksam erwiesen hat, so werden auch heute noch Sonagramme als visuell unterstützende Methode der auditiven Beurteilung verwendet, beispielsweise, um den auditiven Eindruck der Knarrstimme (Glottalisierung) zu bestätigen (Künzel 1987).

Verschiedene Lautkategorien, wie zum Beispiel Vokale oder Frikative, prägen sich im Sonagramm durch ihre Art und Weise in der Produktion unterschiedlich aus und sind dadurch bis zu einem gewissen Grad voneinander abtrennbar. Die Un-

terschiede entstehen einerseits durch die stimmhafte Quelle des Sprachschalls, der durch die Stimmlippen produziert wird, andererseits sind aber auch die Modifikationen des Ansatzrohres für unterschiedliche akustische und daraus folgend perzeptive Eigenschaften der Sprachlaute verantwortlich. Das modifizierte Ansatzrohr wirkt wie ein Filter auf das stimmhafte Signal der Quelle, wodurch die unterschiedlichen akustischen Eigenschaften der einzelnen Sprachlaute entstehen. Akustisch-phonetische Parameter können im Zeit- oder Frequenzbereich anhand der einzelnen abtrennbaren Laute betrachtet werden, die speziell durch Modifikationen des Ansatzrohres bei der Sprachproduktion entstehen (*segmentell*). Akustisch-phonetische Parameter können aber auch im Zeit- oder Frequenzbereich über mehrere Laute beziehungsweise die gesamte Äußerung hinweg betrachtet werden (*suprasegmentell*), wie beispielsweise die Grundfrequenz (Van den Heuvel 1996).

Generell sollten untersuchte Parameter nach Wolf (1972) einer akustisch-phonetischen Analyse gewisse Eigenschaften aufweisen, um geeignet für die Sprechererkennung zu sein:

Ein ausgewählter Untersuchungsparameter sollte zunächst häufig und natürlich in der jeweiligen Sprache auftreten und einfach zu messen sein. Zudem sollte sich der Parameter durch eine hohe interindividuelle, aber eine geringe intraindividuelle Variabilität auszeichnen. Nicht zuletzt sollte ein gewählter Parameter resistent gegen Übertragungsstörungen sein, aber auch gegen Stimmverstellung oder Änderungen im gesundheitlichen beziehungsweise emotionalen Zustand sein. Bis heute ist kein einzelner Parameter entdeckt worden, der alle von Wolf (1972) vorgeschlagenen Kriterien erfüllt und dadurch ein sehr guter Parameter für die Sprecheridentifikation wäre (Gfroerer 2003). Das globale Problem in der Sprechererkennung ist, dass zum Einen die Identität eines Sprechers im Sprachsignal nur durch mehrere, verschiedene Parameter und der Gefahr von Fehlern in der Identifikation belegt werden kann, zum anderen stellt auch der forensische Kontext selbst durch die mangelnde Bereitschaft der Sprecher, erkannt zu werden, eine Schwierigkeit dar. Dennoch bilden die Kriterien von (Wolf 1972) einen Rahmen, dessen Eigenschaften ein oder mehrere ausgewählte Parameter zumindest teilweise erfüllen sollten.

In der Literatur zur forensischen Sprechererkennung (Meuwly 2000a; Gfroerer 2003; Van den Heuvel 1996, u.a.) werden mehrere Parameter im Zeit- oder Frequenzbereich von segmentellen oder suprasegmentellen Kategorien genannt, die geeignet für die Sprecheridentifikation sind. Diese sind der besseren Übersicht

wegen in Tabelle 2.2 zusammengefasst, allerdings ohne auf die Einteilung in segmentelle oder suprasegmentelle Parameter einzugehen.

Frequenzbereich	Zeitbereich
Formanten	Dauern von Segmenten, der gesamten Äußerung
Energieverteilungen	Jitter
Grundfrequenz	Verhältnisse zwischen Segmentdauern
Langzeitspektrum	Artikulationsgeschwindigkeit
Melodik	

Tabelle 2.2: Merkmale des akustisch–phonetischen Ansatzes nach Künzel (1987), Hollien (1990), Meuwly (2000a), Gfroerer (2003)

Die Parameter aus Tabelle 2.2 werden auch als „*low level speaker characteristics*“ (Van den Heuvel 1996) beziehungsweise „*low-level-information parameters*“ (Battaner et al. 2003) bezeichnet. Teilweise handelt es sich bei den Darstellungen der Parameter um solche, die für Personen ohne ausgeprägten phonetischen Hintergrund geschrieben wurden (beispielsweise Künzel (1987)) oder um Darstellungen, die oberflächlicher Natur sind und vor allem akustisch–phonetische Parameter nur nennen, aber nicht weiter ausführen (Van den Heuvel 1996; Meuwly, Drygajlo und Alexander 2003).

Als wirksame und sprecherspezifische Parameter im *Frequenzbereich*, die anhand von Einzelsegmenten ermittelt werden können, gelten spektrale Energieverteilungen bei Frikativen und Verschlusslösungen von Plosiven sowie Formanten und deren Bandbreiten bei Vokalen (Battaner et al. 2003; Gfroerer 2003). Diese charakteristischen Merkmale werden vor allem durch die Konfiguration des Ansatzrohres beim Sprechvorgang bestimmt. Da jeder Sprecher sich durch eine individuelle Anatomie, aber auch durch ein individuelles Sprechverhalten auszeichnet, kann davon ausgegangen werden, dass die Position der Formanten oder die spektrale Ausprägung von Frikativen durch die Individualität des Sprechers beeinflusst wird. Für die Frikative wird dies in der vorliegenden Arbeit durch eine Studie bestätigt, die im Abschnitt 3.3 vorgestellt wird.

Auch in der forensischen Sprechererkennung ist die theoretische Eignung der Energieverteilungen von Frikativen als sprecherspezifisches Merkmal bekannt, es tritt jedoch ein Problem auf: Im forensischen Kontext ist oftmals das Telefon involviert. Da sich Sprecher durch Intensitätsgipfel in sprecherspezifischen Bereichen ausprägen, die über den Übertragungsbereich des Telefons hinausgehen, wird dieses Merkmal in der forensischen Praxis nicht angewendet (Masthoff 2004).

Es wird in einem eigenem Experiment jedoch in einem breiten Frequenzbereich

und unter Laborbedingungen versucht werden, die Energieverteilungen als Untersuchungsparameter in verstellten Stimmen zu verwenden und indirekt deren Sprecherspezifität zu überprüfen. Das Experiment wird in Abschnitt 4.3 beschrieben.

Ebenfalls im *Frequenzbereich*, jedoch nicht an einzelne Sprachlaute gebunden, gelten die Grundfrequenz, die Intonationskontur, die Intensität, aber auch die Stimmqualität sowie Frequenzverteilungen im Langzeitspektrum als sprecherspezifisch (Hollien 1990; Van den Heuvel 1996).

So ist vor allem die Grundfrequenz (F0) als Untersuchungsparameter in der forensischen Sprechererkennung zu nennen, die als die Tonhöhe der Stimme wahrgenommen wird (Künzel 1987). Der Grad der Abweichungen von einem sprechertypischen Bereich, in dem sich F0 bewegt, wird als Melodie der Stimme wahrgenommen. Sind wenig Abweichungen der Grundfrequenz nach oben oder unten vorhanden, so wird die Stimme als „monoton klingend“ bezeichnet.

Suprasegmentelle Parameter im *Zeitbereich* sind beispielsweise das Sprechtempo, aber auch die Dauern von Silben, Wörter und Toneinheiten zählen dazu (Van den Heuvel 1996), sowie eine mikroprosodische Frequenzvariation in der Phonation, „Jitter“ genannt, und der zeitliche Verlauf der Grundfrequenz (Künzel 1987; Meuwly 2000a; Gfroerer 2003).

Das Sprechtempo bzw. die Artikulationsgeschwindigkeit wurde von Künzel (1987) in einem Versuch an deutschen Sprechern des BKA Wiesbaden durch die so genannte „Netto-Silbenrate“ ermittelt. Abzüglich aller entstehenden Pausen wurden die Silben pro Sekunde berechnet. Durchschnittlich konnten bei deutschen Sprechern zwischen 4,4 und 6,0 Silben pro Sekunde realisiert werden.

Als segmentelle Parameter im *Zeitbereich* gelten vor allem die Dauern von einzelnen Segmenten sowie Dauerverhältnisse zwischen Konsonanten und Vokalen (Hollien 1990; Van den Heuvel 1996).

Die bisher beschriebene Vielfalt möglicher Untersuchungsparameter auf auditiv-perzeptiver oder akustisch-phonetischer Ebene zeigt bereits, dass es nicht einfach ist, richtige und wirksame Merkmale für die Sprechererkennung durch Experten zu finden. Sehr schwierig wird die Auswahl der Parameter bei Verwendung des Telefons und einer eventuell schlechten Übertragungsqualität, einem Verdacht der Stimmverstellung und sehr kurzer Dauer des Sprachmaterials (Gfroerer 2003; Künzel 1987).

2.1.3.3 Automatischer Ansatz

Der automatische Ansatz, der auf der menschlichen Stimme als Erkennungsmerkmal basiert, wird manchmal auch als *biometrische Technologie* bezeichnet (Meuwly 2003). Ein wichtiger Bestandteil des automatischen Ansatzes ist, dass die Interaktion mit dem Menschen vergleichsweise gering ist. Äußerungen eines Sprechers werden nicht mehr als gesamtsprachliches Verhalten angesehen, das sich phonetisch in irgendeiner sprecherspezifischen Weise ausprägt und das von einem Experten beurteilt wird.

„It is automatic in the sense that any subjective analysis or evaluation of the speech material is reduced to a minimum; it is global in the sense that it does not address specific acoustic speech parameters but treats the signal as a physical phenomenon, more specifically as a continuously varying complex vibration.“ (Broeders 2001, S. 60)

In der automatischen Sprechererkennung besteht die Prozedur aus einer Merkmalsauswahl und einem Merkmalsvergleich, wie im auditiv-perzeptiven und im akustisch-phonetischen Ansatz auch. Der Unterschied liegt darin, dass in den automatischen Erkennungssystemen Wahrscheinlichkeitsmodelle der ausgewählten Merkmale erstellt werden, durch die ein Sprecher adäquat beschrieben werden kann oder sich zumindest in diesem Merkmal soweit von einem anderen Sprecher abgrenzt, dass eine Entscheidung möglich ist.

Abbildung 2.2 auf Seite 21 stellt ein klassisches automatisches Sprechererkennungssystem dar, das auf akustischen Parametern des Sprachsignals basiert (Merlin, Bonastre und Fredouille 1999).

Den Ausgangspunkt eines automatisches Sprechererkennungssystems bildet eine Sprachaufnahme; bei der Sprechererkennung im forensischen Kontext ist dies eine Sprachaufnahme des Straftäters. Die erste Aufgabe besteht darin, einen oder mehrere Vergleichssprecher in einer Datenbank auszuwählen. Aus der Sprachaufnahme werden bestimmte Parameter extrahiert, die auf einer vorherigen Analyse mit auditiv-perzeptiven und akustisch-phonetischen Merkmalen basieren können, oder die unabhängig von solchen Analysen sind. Auch von den Sprachaufnahmen der Referenzpopulation mit Vergleichssprechern (Verdächtigen) werden Parameter extrahiert. Liegt die so genannte „open-set“-Situation vor, so ist die Stimme des Täters nicht mit Sicherheit in der Referenzpopulation enthalten. Dies ist meist der Fall in forensischen Sprechererkennungssystemen. Aus den Parametern der

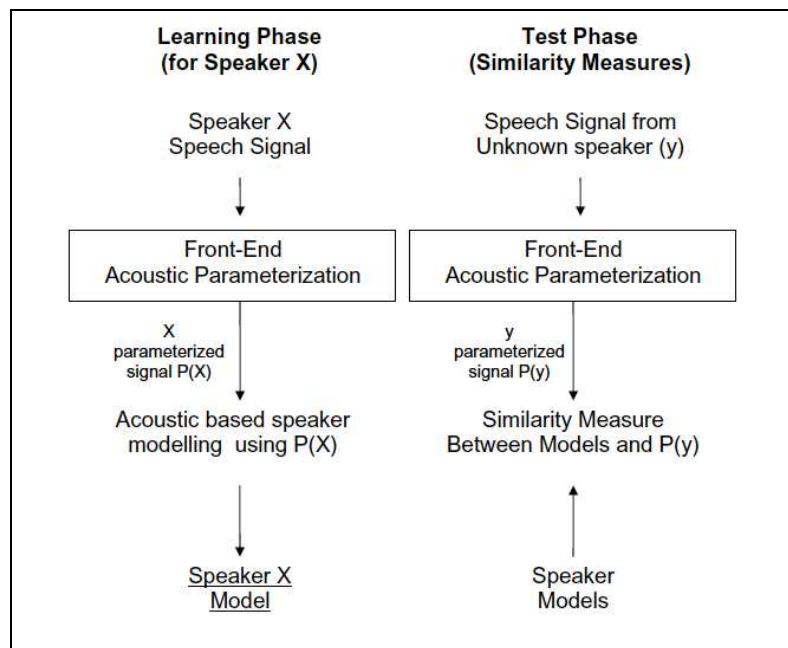


Abbildung 2.2: Darstellung eines automatischen Sprechererkennungssystems nach Merlin, Bonastre und Fredouille (1999)

Täterstimme sowie der Vergleichsstimmen werden Wahrscheinlichkeitsverteilungen erstellt, in Abbildung 2.2 als „speaker modelling“ bezeichnet. Dabei bildet die Wahrscheinlichkeitsverteilung des Verdächtigen das so genannte *Trainingsmodell*, in Abbildung 2.2 als „learning-phase“ bezeichnet, und die Wahrscheinlichkeitsverteilung des Täters das Testmaterial (Merlin et al. 1999; Broeders 2001). Darauf folgend werden dann die Ähnlichkeiten zwischen den Modellen der Parameter eines Sprechers berechnet. Je höher die Distanzen eines Parameters aus den Testdaten zu dem erstellten Trainingsmodell sind, umso höher ist die Wahrscheinlichkeit, dass Täter und Verdächtiger hinsichtlich des gewählten Parameters unterschiedlich sind (Künzel 1987; Furui 1995).

Die Berechnung der Wahrscheinlichkeitsmodelle ist abhängig von den Voraussetzungen und spielt eine große Rolle für die Stärke eines Beweises vor Gericht (Nolan 2001; Meuwly et al. 2003). Handelt es sich um eine „open-set“-Situation, so wird auf der Basis bedingter Wahrscheinlichkeiten (*conditional probabilities*) ein Verhältnis (*likelihood-ratio*) berechnet, das Auskunft darüber gibt, wie ähnlich sich die Wahrscheinlichkeiten für das Auftreten eines Merkmals bei dem Täter und bei dem/ den Verdächtigen sind. Dabei kann es durchaus der Fall sein, dass die zweite ermittelte Wahrscheinlichkeit des Merkmals nicht der bedingten Wahrscheinlichkeit des Täters entspricht. In einer „closed-set“ Situation kann

im Gegensatz dazu davon ausgegangen werden, dass der Täter, der sich durch eine gewisse Wahrscheinlichkeit (*a-posteriori*) eines Parameters repräsentiert, in der Population enthalten ist und die berechneten Wahrscheinlichkeiten entweder der Merkmalsverteilung des Verdächtigen oder des Täters zuzuordnen sind (Furui 1995; Meuwly et al. 2003).

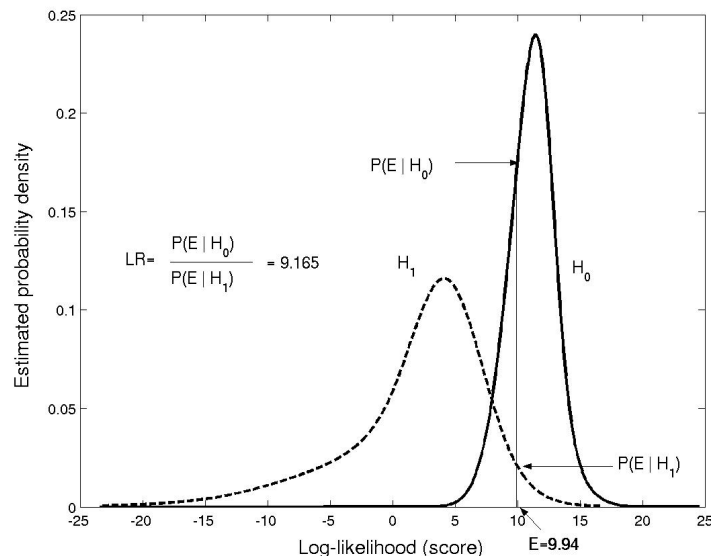


Abbildung 2.3: Darstellung des Likelihood-ratios (LR) anhand eines Beweises E nach Meuwly et al. (2003)

Der Ansatz des likelihood-ratios (LR) basiert auf der theoretischen Annahme, dass ein Beweis E eine gewisse Wahrscheinlichkeit besitzt, überhaupt aufzutreten. Dabei werden zwei gegensätzliche Hypothesen zugrundegelegt:

H_1 bezeichnet die Hypothese, dass es sich bei dem Verursacher von E (Täter) und dem Verdächtigen um ein und die selbe Person handelt, H_0 drückt die Gegenhypothese aus, dass der Beweis E von unterschiedlichen Personen stammt. Die so genannte Entscheidungsgrenze wird als ein Verhältnis LR aus diesen beiden Wahrscheinlichkeiten gebildet (siehe Abbildung 2.3). Dies wird auch als *Bayesian approach* bezeichnet (Nolan 2001). Notwendig für eine aussagekräftige Anwendung dieser Methode sind Referenzpopulationen, in denen die intraindividuellen Variationen bestimmter Parameter gespeichert sind. Zudem hängt die Stärke des Beweises, ausgedrückt durch das LR , eng mit der sprecherspezifischen Qualität der untersuchten Parameter und dessen Modellierung zusammen:

„The important point to be made here is that the estimate of the *LR* is only as good as the modelling techniques and databases used to derive it.“(Meuwly et al. 2003).

Wie im auditiv-perzeptiven und akustisch-phonetischen Ansatz, werden auch in der forensischen automatischen Sprechererkennung Entscheidungen nicht nur auf einem Parameter gegründet, sondern auf mehreren Parametern zusammen. Der Grund liegt in der Tatsache, dass bisher kein einzelner Parameter des Kurz- oder Langzeitspektrums bekannt ist, der eine korrekte Identifikation garantiert (Nolan 2001).

Seit Beginn der achtziger Jahre wurde von der Verwendung von direkt ermittelbaren Parametern des Kurzzeitspektrums, wie beispielsweise Formanten, Abstand genommen. Stattdessen wurde dazu übergegangen, Ableitungen aus dem Kurzzeitspektrum zu erstellen und für automatische Systeme zu verwenden (Furui 1995; Meuwly 2000a), zum Beispiel LPC-Koeffizienten, Kepstralkoeffizienten und weitere Ableitungen von diesen. Die abgeleiteten Kepstralkoeffizienten (*delta-Koeffizienten*) haben den Vorteil, dass sie von Modifikationen des Ansatzrohres unabhängig sind, da es sich um Ableitungen des Quellsignals beziehungsweise dessen spektraler Hüllkurve handelt. Daraus resultiert eine geringere Anfälligkeit gegenüber Langzeitvariationen innerhalb einer Äußerung (Furui 1995).

An dieser Stelle muss kurz auf zwei Probleme eingegangen werden, die sich für die forensisch-phonetischen Experten im Zusammenhang mit der automatischen Sprechererkennung und speziell durch die Verwendung des *LR* ergeben haben: Wie in diesem Abschnitt näher ausgeführt wurde, verwendet die automatische Sprechererkennung⁶ Untersuchungsparameter, die direkt aus der produzierten Stimme oder dem sprachlichen Verhalten beziehungsweise den akustischen Korrelaten nicht ersichtlich sind. Dies hat zur Folge, dass auch die Entscheidungsfindung im forensischen Fall durch menschlichen Sachverstand nicht nachvollziehbar ist, sondern stark von der Technik und der angewendeten Methode abhängt und sich in einer ausgegebenen Wahrscheinlichkeit repräsentiert. Der Richter muss sich demnach darauf verlassen können, dass die angegebene Wahrscheinlichkeit zur Identität korrekt ist und auch ausreichend sicher ist, sofern keine anderen Beweise vorhanden sind. Bezüglich der Wahrscheinlichkeiten, die durch den beschriebenen Ansatz mit dem *LR* ermittelt werden, stellt sich jedoch das angesprochene Problem, dass bis heute keine Datenbanken zur Verfügung stehen, in denen sowohl

⁶Als alleinige Untersuchungsmethode oder in Kombination mit dem auditiv-perzeptiven und dem akustisch-phonetischen Ansatz

das inter– als auch intraindividuelle Verhalten von Sprechern berücksichtigt ist und die vor allem groß genug sind, alle möglichen und tatsächlich vorkommenden Variationen in einer Population abzudecken:

„Worse, for very few of these [independent and mutually dependent parameters] do we have the population statistics which would be required for a quantitative application of the Bayesian approach.“ (Nolan 2001, S. 9)

Damit hängt zusammen, dass eine Wiederholung der Untersuchungen im forensischen Kontext nicht möglich ist. Da die Sprechererkennungsalgorithmen jedoch nicht in der Lage sind, „mögliche drastische Einflüsse der so genannten Kommunikations– oder Sprechsituation“ (Künzel 1987, S. 125) zu berücksichtigen, ist auch die Entscheidungsfindung des automatischen Systems nicht genügend flexibel. Der forensische Experte ist jedoch in der Lage, die Situation abzuschätzen und gegebenenfalls die Parameterauswahl, die zur Entscheidungsfindung führt, zu revidieren.

2.2 Definition und Auftreten von Stimmverstellung

Dieser Abschnitt definiert die Stimmverstellung für den forensischen Kontext und stellt zudem eine weitere Definition bereit, die für Grundlagenforschung im Allgemeinen und für die eigene Analyse, beschrieben in Abschnitt 4.3, verwendet werden kann.

2.2.1 Definitionen

Für eine detaillierte Betrachtung von Stimmverstellung ist der Kontext von entscheidender Bedeutung. Dabei müssen zwei Zielsetzungen der Erforschung zur Stimmverstellung berücksichtigt werden: Einerseits werden Untersuchungen speziell auf die forensische Fragestellungen ausgerichtet, andererseits existieren einige Studien zur Stimmverstellung, die unter Laborbedingungen produziert wurden und auf Grundlagenforschung ausgerichtet waren.

In der forensischen Sprechererkennung, wie sie in 2.1.2 dargestellt wurde, existiert eine Definition von Stimmverstellung, die im Rahmen einer auditiven Beurteilung gilt. Dort ist Stimmverstellung definiert als

”[...] die in der Absicht der Verbergung der eigenen Identität herbeigeführte Veränderung von Merkmalen der Stimme, Sprache und / oder Sprechweise eines Individuums.” (Künzel 1987, S. 103)

Bei jeder Sprechererkennungsaufgabe (siehe 2.1.2), und besonders wenn der Verdacht auf Stimmverstellung besteht, stellt die auditive Beurteilung den ersten und wichtigsten Schritt dar und erfordert die Betrachtung des gesamtsprachlichen Verhaltens des Sprechers (Künzel 1987). Basierend auf der auditiven Beurteilung werden Merkmale extrahiert, die auf Stimmverstellung hindeuten und die dann zur eventuellen Rekonstruktion der Originalstimme verwendet werden (Masthoff 2000). Welche auditiv-perzeptiven Merkmale des gesamtsprachlichen Verhaltens dafür geeignet sind und welche Auswirkungen Stimmverstellung auf akustisch-phonetische Parameter hat, wird in Abschnitt 2.3 anhand verschiedener Untersuchungen dargestellt.

Es stellt sich aber nicht nur das Problem, Stimmverstellung zu entdecken, sondern auch das Problem der eventuell vorhandenen Stimmimitation, d.h. ob eine Stimme, die als identifiziert angenommen wird, auch wirklich von dem Sprecher

stammt, und nicht von einem Imitator (Künzel 1987). Bei der Stimmimitation handelt es sich jedoch um eine besondere Art der Stimmverstellung:

„Die weitestgehende Art der Verbergung der eigenen Identität ist zweifellos die gleichzeitige Vortäuschung der Identität eines bestimmten anderen Individuums.“ (Künzel 1987, S. 103)

Stimmimitation ist daher in der Definition der Stimmverstellung von Künzel (1987) berücksichtigt und wird nicht gesondert betrachtet.

Hinsichtlich der Qualität und Beschaffenheit des Sprachmaterials einer eigenen Untersuchung in Abschnitt 4.1 gilt die Definition von Künzel (1987) in dieser Form nicht für die vorliegende Arbeit. Dafür gibt es einen Grund, der sich auf die Intention der Stimmverstellung bezieht:

Es wurden zwei Sprecher (B und D) verpflichtet. Beide Sprecher arbeiten als Radio-Moderatoren und sind trainiert in der Stimmverstellung. Sprecher B produziert eine Radio-Comedy, in der verschiedene Charaktere allein durch verschiedene Arten der Stimmverstellung kreiert werden. Das Ziel des Sprechers ist es also nicht, die eigene Identität zu verbergen, sondern durch Veränderung der eigenen Stimme neue „Personen“ zu kreieren.

Sprecher D bedient sich der Verstellungsart Imitation. Im Gegensatz zu Imitationen im forensischen Kontext ist der Zweck der Imitationen als eine Form der Stimmverstellung von Sprecher D ebenfalls nicht die Verbergung der eigenen Identität, sondern vorrangig die Unterhaltung bei öffentlichen Auftritten. Die englischsprachige Literatur stellt dafür den Begriff *Impersonation* zur Verfügung:

„Imitation can also be used for entertainment. [...] I call this type of imitation, when a speaker reproduces another speaker’s voice and speech characteristics, impersonation.“ (Zetterholm 1997, S. 269)

Die Intention der Stimmverstellung (Identitätsverdeckung vs. *Impersonation*) hat jedoch auf die weitere Untersuchung des Sprachmaterials keinen Einfluss, da die produzierten Stimmverstellungen zu einer Änderung des gesamtsprachlichen Verhaltens führen. Es muss aber berücksichtigt werden, wie auch im Abschnitt 2.3 gezeigt wird, dass Stimmverstellung direkt im forensischen Umfeld betrachtet werden kann, wie dies beispielsweise bei den Untersuchungen von de Figuereido und de Souza Britto (1996) oder Masthoff (2000) der Fall ist. Es sind aber auch Untersuchungen im Rahmen der Grundlagenforschung möglich, die dazu dienen, akustische und phonetische Grundlagen der Stimmverstellung zu ermitteln, beispielsweise die Untersuchung von Endres, Bambach und Flösser (1971).

Eine Definition, in welcher der Aspekt der Identitätsverbergung nicht zwangsläufig vorhanden sein muss, beschreibt Stimmverstellung als

„[...] any alteration, distortion, or deviation from the normal voice, irrespective of the cause.“ (Rodman 1998, S. 17).

Die Definition von Rodman (1998) ist weniger spezifisch als die Definition von Künzel (1987), sie ist aber wegen möglicher unterschiedlicher Intentionen der Stimmverstellung (Identitätsverbergung vs. *Impersonation*) sehr gut für die vorliegende Arbeit anwendbar.

Ein weiterer Grund rechtfertigt die Verwendung dieser Definition für das vorhandene Sprachmaterial: Die Definition wurde im Rahmen einer Empfehlung zur Durchführung von Studien zur Sprechererkennung mit dem Schwerpunkt der Stimmverstellung von Rodman (1998) erstellt. Durch diese Definition bleibt es dem Forscher offen, ob eine Grundlagen–Untersuchung unter Laborbedingungen durchgeführt wird oder es sich doch um eine Studie im forensischen Umfeld handelt, die sich an realen forensischen Bedingungen orientiert.

2.2.2 Auftreten von Stimmverstellung

Die Sprecheridentifikation durch Experten von Sprachmaterial einer Straftat beinhaltet immer die Möglichkeit der vorhandenen Stimmverstellung, vor allem, wenn der Täter der Straftat sich bewusst ist, dass seine Stimme aufgenommen wird. Dann würde ein Straftäter versuchen, durch Stimmverstellung die Feststellung seiner Identität zu erschweren, beispielsweise bei Erpressungen, in denen Kontakt häufig über das Telefon stattfindet und der Täter davon ausgehen kann, dass seine Stimme aufgezeichnet wird. Im Gegensatz dazu besteht für einen Straftäter kein Anlass, seine Stimme zu verstellen, wenn Telekommunikations–Überwachungsmaßnahmen verwendet werden, über die er nicht informiert ist und er sich daher nicht beobachtet fühlt (Gfroerer 2004).

Das Auftreten der Stimmverstellung ist stark von der Art des Verbrechens abhängig, bei Erpressungen und Entführungen ist die Frequenz wesentlich höher als bei Notrufmissbräuchen, sexueller Belästigung oder strafbaren Aktionen, die unter Alkohol– bzw. Drogeneinfluss geschehen (Masthoff 1996). Die Hauptinstitution in Deutschland für die Bearbeitung und Aufklärung von Kapitalverbrechen mit Stimmbeteiligung als Beweis, wie Erpressung oder Entführung, ist das Bundeskriminalamt. Im Zeitraum von 1989 bis 1994 hat es in 52% der dort registrierten

und bearbeiteten Verbrechen Stimmverstellung gegeben. Erpressungen bilden davon mit 69% Stimmverstellung den höchsten Anteil. Daher muss der involvierte Experte in der Lage sein, vorhandene Stimmverstellung zu erkennen und korrekt damit umzugehen (Masthoff 1996, 2000).

In anderen Einrichtungen, die Straftaten mit Stimmbeteiligung untersuchen, ist der Anteil der Stimmverstellung wesentlich geringer. So wiesen bearbeitete Fälle der Universität Trier nur 5% Stimmverstellung auf (Masthoff 1996). Jedoch tritt sie auf und muss auch bei kleineren Delikten richtig erkannt und bewertet werden.

2.3 Studien zur Stimmverstellung

Im vorherigen Abschnitt 2.1 wurde auf die Sprechervariabilität eingegangen, deren Ursachen und Ausprägungen und auch auf Bereiche und Vorgehensweisen der forensischen Sprechererkennung. In Abschnitt 2.1.3.2 wurden Kriterien genannt, die über die Qualität eines ausgewählten Merkmals des Sprechers entscheiden. Eines der Kriterien war Resistenz gegen Stimmverstellung.

Der aktuelle Abschnitt dient dazu, wichtige Studien zur Stimmverstellung zu reproduzieren, die das Problem der Stimmverstellung auf verschiedenen Ebenen verdeutlichen. Trotz der Relevanz für die forensische Sprechererkennung ist die Anzahl verfügbarer Studien vergleichsweise gering.

Wie in Abschnitt 2.2.1 angedeutet, kann die Zielsetzung der Studien auch über die Vorgehensweise entscheiden. Studien zur Sprechererkennung, die laut Definition von Künzel (1987) im forensischen Umfeld angesetzt sind, beziehen sich meist auf auditiv-perzeptive Merkmale, die aus Sprachmaterial eines Sprechers extrahiert werden können, beispielsweise der verwendete Soziolekt oder der sprecherspezifische Idiolekt. Einige akustisch-phonetische Merkmale werden nun speziell im Bezug auf die Stimmverstellung erweitert, beziehungsweise dahingehend betrachtet, ob sie auch im Falle von Stimmverstellung zur Identifikation eines Sprechers beitragen können oder ob Stimmverstellung einen Einfluss ausübt.

2.3.1 Grundlagenforschung zur Stimmverstellung

Nachdem Kersta (1962) seine Untersuchung zur Stimmidentifikation mit Hilfe von Sonagrammen als sogenannten *Voiceprint* bzw. *Stimmabdruck* veröffentlichte, wurden mehrere Untersuchungen im Bereich der Identifikationsmöglichkeiten von Sprechern durch Sonagramme, beziehungsweise im Rahmen automatischer Sprechererkennungssysteme, auch in Bezug auf das Problem der Stimmverstellung,

durchgeführt. Im Gegensatz zum invarianten Fingerabdruck haben sich die „Voiceprints“ durch die in Abschnitt 2.1.1 dargestellten Ursachen jedoch als zu variabel herausgestellt. Trotzdem ist das Sonagramm als Darstellung von Sprache noch heute ein sehr wichtiges Mittel zur Beurteilung von Stimmen, wenn auch nur in Kombination mit anderen Methoden (siehe Abschnitt 2.1.3.2).

Im Folgenden werden Studien von Endres, Bambach und Flösser (1971), McGlone, Hollien und Hollien (1977) sowie Hollien und Majewski (1977) näher dargestellt, da sie zu den wenigen Studien gehören, die sich mit der Auswirkung der Stimmverstellung auf Parameter im segmentellen Bereich auseinandergesetzt haben. Einschränkend muss allerdings erneut betont werden, dass die Studien im Rahmen der Voiceprint-Identifikationen erstellt wurden.

In der Studie von Endres et al. (1971) wurde der Einfluss von Alter, Stimmverstellung und Imitation auf die Formanten von verschiedenen Vokalen und dem Nasal /n/ sowie die Grundfrequenz berechnet und im Sonagramm betrachtet. Die Analyse zum Alter wird an dieser Stelle jedoch nicht weiter betrachtet. Das zweite Telexperiment untersuchte die Originalstimmen von fünf Sprechern und einer Sprecherin im Vergleich zu Stimmverstellungen in vier möglichen Merkmalen: Grundfrequenz, Sprechgeschwindigkeit, Akzent und Dialekt. Gemessen wurden die unteren vier Formanten von /a:, i:, n/. Der Ausgangstext wurde ein Mal mit der Originalstimme und drei Mal mit verstellter Stimme gelesen.

Das Ergebnis dieses Telexperiments war, dass die variierenden Arten der Stimmverstellung sich unterschiedlich stark auf die untersuchten Segmente auswirkten. Die Formantpositionen der Vokale verschoben sich, wobei inter-individuelle Unterschiede auftraten; bei einigen Versuchspersonen⁷ verschoben sich die Formanten nach oben, bei anderen VPn nach unten. Lediglich die Position des ersten Formanten war stabiler als die restlichen untersuchten Formanten im Vergleich zur Originalstimme (Endres et al. 1971, S. 1844). Manche VPn neigten dazu, ihre Stimme in der Art und Weise zu verstellen, wodurch sich die Intensität der Formanten derart abschwächte, dass diese im Sonagramm nicht mehr sichtbar waren. Unklar in der Darstellung der Ergebnisse blieb jedoch, wie sich die Stimmverstellung auf die Nasale auswirkte, das heißt, ob die Aussage zur Verschiebung der Formantpositionen die Nasale mit einbezog. Auch wurde nicht dargestellt, inwiefern Unterschiede hinsichtlich der Formanten oder der Grundfrequenz zwischen den einzelnen Verstaltungstypen auftraten. Grundsätzlich kritisch ist die Aussage

⁷im Folgenden VPn genannt

zur Stabilität von F1 anzusehen, da es unwahrscheinlich ist, dass sich die Vokalqualität, die durch F1 und F2 kodiert wird, nicht ändert. Dies betrifft vor allem die Verstellungen hinsichtlich des Akzents oder Dialekts.

Das dritte Telexperiment untersuchte die Imitation in mehreren Schritten. Zwei deutsche Imitatoren wurden aufgefordert, fünf Stimmen nachzuahmen. Die Originalstimmen der imitierten Personen lagen ebenfalls vor, um sie mit den Imitationen zu vergleichen. Es wurden wiederum die unteren vier Formanten der Vokale /a, e, i:/ verwendet, um den Einfluss der Imitation auf die Stimme zu analysieren. Das Ergebnis war, dass die Imitatoren teilweise in der Lage waren, die eigenen Formantpositionen derart zu variieren, dass sie sich an die Originalstimmen der imitierten Personen annäherten. Dabei kam es aber zu keiner zufriedenstellenden Annäherung der Formantpositionen, vor allem nicht im höheren Frequenzbereich. Auch in der Grundfrequenzanalyse ergaben sich Variationen (nach oben oder unten) zwischen den Imitationen und den imitierten Stimmen.

Anschließend führten Endres, Bambach und Flösser einen Hörtest mit den Originalstimmen der VPn, den imitierten Stimmen und der Originalstimme eines Imitators durch. Der Hörtest ergab, dass die zehn Hörer trotz Abweichungen der Formanten und der Grundfrequenz zwischen Imitationen und der Originalstimmen der VPn nicht in der Lage waren, die Originalstimme des Imitators zu erkennen und zudem überhaupt zu erkennen, dass die selbe Person auch Stimmen imitierte. Die Autoren wiesen aber darauf hin, dass

„[...] the sound of the voice and the mean pitch frequency alone does not play a predominant role in the identification of the imitated speaker by other persons.“ (Endres, Bambach und Flösser 1971, S. 1847)

Es bleibt in der Beschreibung der Methode zum Hörtest als auch in der Darstellung der Ergebnisse unklar, wie die Qualität des Sprachmaterials war, so dass ein Grund für das Hörtestergebnis vielleicht auch darin liegen könnte. Es scheint unwahrscheinlich, dass keiner von zehn Hörern eine Imitation als solche erkennen konnte. Die anscheinend kurze Dauer des präsentierten Sprachmaterials könnte dabei jedoch eine Rolle gespielt haben⁸.

Eine weitere Analyse zur Imitation zielte auf einen Vergleich zwischen Originalstimme eines Sprechers, einer Imitation von diesem Sprecher und der Originalstimme des Imitators ab. Wieder wurden die Formanten von /a, e, i:/ analysiert, in diesem Fall die unteren sechs. Vor allem bei /e/ waren jedoch kaum Unterschiede

⁸Möglicherweise handelt es sich bei den VPn um naive Hörer, deren Identifikationsleistung bei kurzer Dauer des Sprachmaterials stark reduziert ist (Bull und Clifford 1984).

in den ersten drei Formanten zu erkennen, weder zwischen den verglichenen Originalstimmen noch im Vergleich zur Imitation, was angesichts der geringen Zahl der VP ein Zufall gewesen sein könnte oder auch an der Methode der Formantberechnung gelegen haben könnte. Die Formanten F1-F3 von /e/ der imitierten Stimme wichen ebenfalls nur unwesentlich ab. Unterschiede traten erst bei den höheren Formanten F4-F6 auf, die für diese Analyse berechnet wurden. Dies war prinzipiell ebenfalls für /a/ der Fall, auch hier ähnelten sich die unteren Formanten der verglichenen Stimmen mehr als die höheren Formanten. Ergebnisse zu /i:/ wurden nicht dargestellt.

Diese Studie belegt, trotz Mängeln in der Methode und der ungenauen Darstellung der Ergebnisse (Tosi 1979), dass Stimmverstellung und Stimmimitation Einfluss auf Vokal-Formanten und die Grundfrequenz ausüben.

Ebenfalls auf den Vergleich akustischer Parameter von unverstellten und verstellten Stimmen zielte die Studie von McGlone, Hollien und Hollien (1977) ab.

23 männliche Sprecher wurden aufgefordert, einen Text in normaler und in frei verstellter Stimme zu lesen. Aus diesem Text wurde dann ein Satz („*I do not set the same store.*“) extrahiert, Sonagramme erstellt und die Grundfrequenz berechnet. Zusätzlich wurden die Formanten F1-F3 und deren Bandbreiten der Vokale /u, a, ε/ und der Diphthonge /ai, ei, œ/ ermittelt, die Dauern dieser Vokale und der stimmlosen Konsonanten /t, s/, der stimmhaften Konsonanten /d, m, n/ sowie die Gesamtdauer des Satzes. Diese Untersuchungsparameter wurden von drei unabhängigen Phonetikern ermittelt. Das Ziel der Untersuchung war es, die Differenzen der spektralen Parameter zwischen unverstellter und verstellter Stimme aufzuzeigen.

Die Studie ergab interindividuelle Unterschiede in der Grundfrequenz; einige der VPn erhöhten die Grundfrequenz bei der Verstellung, andere VPn senkten sie ab. Bezüglich der Formanten traten Abweichungen zwischen unverstellter und verstellter Stimme auf, die sich durch hohe intra-individuelle Variation, d.h. hohe Standardabweichung, auszeichneten. Es wurden allerdings nicht die einzelnen Sprecher, sondern alle Sprecher zusammen für jeden untersuchten Vokal berechnet, so dass die Betrachtung der einzelnen Sprecherdaten pro untersuchten Vokal nicht möglich ist. Die Bandbreiten der Vokal-Formanten wiesen ebenfalls große Sprechervariabilität auf. Auch die ermittelten Dauern wichen voneinander ab, wobei die größten Unterschiede hinsichtlich der untersuchten Gruppe (Konsonant, Vokal, Diphthong) zwischen unverstellten und verstellten Stimmen bei den Diphthongen auftraten.

Zusammengefasst bestanden also große inter- und wohl auch intraindividuelle Variationen zwischen unverstellter und verstellter Sprache in den gemessenen Parametern Grundfrequenz, Dauer und Formanten.

Im Gegensatz zu den bisher beschriebenen Studien wurde von Hollien und Majewski (1977) eine Studie durchgeführt, in der es nicht die Untersuchung akustischer Parameter aus dem Kurzzeitspektrum ging, sondern um das Langzeitspektrum⁹. Grundlage für die Untersuchungen war die Ermittlung von intra- und interindividuellen Unterschieden in zwei Telexperimenten. Am ersten Experiment nahmen 50 amerikanische und 50 polnische männliche Sprecher teil. Jeder Sprecher las einen Text, dadurch standen 2,5 Minuten aufgenommene Sprache pro Sprecher zur Verfügung. Es wurden über das gesamte Sprachmaterial 4 Langzeitspektren (sogenannte Vektoren) von maximal 32 Sekunden Länge mit 23 Frequenzbändern über 1/3 Oktave erstellt, um einen Frequenzbereich von 80 bis 12500 Hz abzudecken. Die ersten drei Vektoren wurden als Referenz des Sprechers ausgewählt, der vierte Vektor galt als Testmaterial. Die Analyse bestand in der Berechnung euklidischer Distanzen zwischen den Referenz- und den Testvektoren, getrennt für die amerikanischen und polnischen VPn. Durch die Distanzen des Testvektors auf die Referenzvektoren des selben Sprechers konnten die intraindividuellen Variationen ermittelt werden, durch die Distanzen zu den Referenzvektoren der anderen Sprecher wurden die interindividuellen Variationen berücksichtigt.

Am zweiten Experiment nahmen 25 amerikanische Sprecher teil und von ihnen wurde das gleiche Sprachmaterial wie im ersten Telexperiment aufgenommen, die Sprecher wurden jedoch aufgefordert, zusätzlich ihre Stimme zu verstellen. Sie wurden dabei einer Stresssituation (leichte Elektroschocks) ausgesetzt, so dass im Endeffekt drei separate Sprechbedingungen aufgenommen wurden. Inwiefern die VPn ihre Stimme verstellten, blieb ihnen frei überlassen, sie durften lediglich keinen fremden Dialekt oder Flüstern als Verstellungstyp verwenden. Ihnen wurde geraten, ihre Stimme nur im modalen Stimmregister zu verändern. Die Analyse bestand wie im ersten Telexperiment in der Berechnung euklidischer Distanzen, jedoch in einem Frequenzbereich von 80 bis 10000 Hz, so dass nur 22 Frequenzbänder zur Verfügung standen. Ein weiterer Unterschied zum ersten Telexperiment war, dass der erste Vektor von 32 Sekunden Länge als Testvektor diente und das gesamte Sprachmaterial von 128 Sekunden den Referenzvektor darstellte. Anhand der „normalen“ Stimme wurden die Identifikationsgrenzen für

⁹Untersuchungen des Langzeitspektrums schließen den sprachlichen Gehalt des Sprachsignals weitgehend aus (Hertrich 1986)

die euklidischen Distanzen festgelegt, um zu sehen, inwiefern korrekte Identifikationen mit dieser Methode möglich waren.

Zusätzlich wurde für beide Telexperimente noch der Frequenzbereich auf 315 bis 3150 Hz reduziert, um Bedingungen ähnlich denen des Telefons zu schaffen.

Die Experimente ergaben Folgendes: Im ersten Telexperiment, der Identifikation von Amerikanern und Polen, wurden Identifikationsraten von 96% für die polnischen Sprecher und 94 % für die amerikanische Sprecher erreicht. Sie sanken jedoch bei der Einschränkung des Frequenzbereichs auf 82% bzw. 70% ab.

Für das zweite Telexperiment in der Bedingung der Stimmverstellung war ein extremer Abfall der Identifikationsraten zu erkennen, bei Verwendung des gesamten Frequenzbereichs von 80 bis 10000 Hz waren nur noch 20% korrekte Identifikationen möglich, bei dem eingeschränkten Frequenzbereich von 315 bis 3150 Hz gab es jedoch 32% korrekte Identifikationen. Ein Grund für diese Abweichung im Vergleich zum anderen Telexperiment könnte sein, dass nicht nur die korrekten Identifikationen berücksichtigt wurden, sondern auch Falschidentifikationen, die bei Stimmverstellung im gesamten Frequenzbereich immerhin 12% betrogen, bei der Reduktion des untersuchten Frequenzbereichs aber auf 23% anstiegen, so dass trotz korrekter Identifikationen auch eine hohe Fehlerrate vorhanden war.

Die bisher beschriebenen Studien von Endres et al. (1971) und McGlone et al. (1977) untersuchten auf segmenteller Basis Parameter des Kurzzeitspektrums im Zeit- und Frequenzbereich. Beide Studien ergaben eine unsystematische Variation der untersuchten Parameter im Fall der Stimmverstellung, wobei die Frage nach der Vergleichbarkeit, aber auch die unkritische Akzeptanz der Ergebnisse problematisch ist¹⁰, da nach persönlicher Meinung Mängel in der Darstellung der Materialerstellung (Aufnahmebedingungen) und der Ergebnisse vorhanden waren, speziell bei der Untersuchung von Endres et al. (1971). Eine ausführlichere Bewertung der Studie wurde von Tosi (1979) und Nolan (1983) vorgenommen. Die Studie von Hollien und Majewski (1977) zeigte, dass auch im Langzeitspektrum, welches als unabhängig von sprachlichen Einzelkomponenten anzusehen ist, bei Stimmverstellung intra- und interindividuelle Variationen auftreten.

¹⁰Die Vergleichbarkeit von Ergebnissen verschiedener Studien ist aber bekanntermaßen nicht nur ein Problem, das die Studien zur Stimmverstellung betrifft.

2.3.2 Studien im forensischen Kontext

Wie in Abschnitt 2.2.2 erwähnt, ist Stimmverstellung immer potentiell vorhanden, wenn der Straftäter davon ausgehen kann, dass seine Stimme aufgenommen wird. Der erste Schritt einer Analyse, in der Stimmverstellung vorhanden sein könnte, besteht darin, herauszufinden, ob und in welchen Bereichen der Sprecher seine Stimme verstellt hat. Dies stellt natürlich ein großes Problem dar, weil der involvierte Experte einerseits wissen muss, beziehungsweise erkennen muss, dass Stimmverstellung vorhanden ist. Andererseits muss er in der Lage sein, aus den verstellten Merkmalen der Stimme die tatsächlichen Eigenschaften des Sprechers zu rekonstruieren (Gfroerer 2004). Im Folgenden werden zwei Studien vorgestellt, die sich auf die forensische Praxis beziehen und Ansätze darstellen, aufgrund real vorkommender Stimmverstellungsarten Rückschlüsse auf die Stimmverstellungen zu ziehen.

Eine Untersuchung, die sich speziell mit der Frage der Stimmverstellung im auditiv-perzeptiven Bereich beschäftigt hat, ist die von Masthoff (2000). Das Ziel der Studie war die Rekonstruierbarkeit modaler sprecherspezifischer Merkmale aus der verstellten Stimme.

Es gibt bestimmte Arten der Stimmverstellung, die von Sprechern bevorzugt werden. Diese werden für den Zweck der genauen Beschreibung der Stimmverstellung, die als Grundvoraussetzung einer forensisch-phonetischen Analyse dient, in vier Kategorien unterteilt, die den in Abschnitt 2.1.3 genannten Bereichen ähneln.

Die Kategorie *Respiration* betrifft Atemgeräusche, d.h. entweder die Atmung beim Sprechvorgang selbst oder Ein- bzw. Ausatemungsgeräusche zwischen Artikulationen. Die Kategorie *Phonation* berücksichtigt die Stimmlippenaktivität, z.B. eine Erhöhung oder Absenkung der Schwingungsfrequenz, Periodizität, Behauchung, Flüstern und Registerwechsel von Brust- zu Kopfstimme. In der Kategorie *Artikulation* sind Aktivitäten und Veränderungen des Ansatzrohres enthalten, wie z.B. auffällige Bildung von Einzellauten, aber auch Aufsetzen eines Dialekts oder fremdsprachiger Akzents und die Änderung der Ansatzrohrkonfiguration durch Anheben oder Absenken des Kehlkopfes. Aber auch Auswirkungen, die sich durch äußere Beeinflussung ergeben, werden in der Kategorie Artikulation berücksichtigt, so z.B. das Zuhalten der Nase, die Fixierung der Lippen oder das Wangenkneifen. Die Kategorie *Sprechweise* beinhaltet in Analogie zur Stimmbeschreibung von Künzel (1987) das Sprechtempo und den Redefluss.

Die Studie von Masthoff (2000) untersuchte mit Hilfe des Standardtextes „Nordwind und Sonne“ an 23 VPn (11 weiblichen und 12 männlichen) die normale

Stimmelage der Sprecher im Vergleich zur verstellten Stimme im Abstand von zehn Tagen. Die einzige Bedingung für die VPn war, dass die eigene Identität ohne technische Hilfsmittel so weit wie möglich verdeckt werden sollte.

Das Ergebnis der Studie von Masthoff (2000) lässt sich in wenigen Punkten zusammenfassen: Es konnten maximal drei Einzelparameter aus zwei Beschreibungskategorien auf einmal verstellt werden, öfter kamen aber Verstellungen in nur zwei Einzelparametern vor, am häufigsten waren Stimmverstellungen in den Kategorien Phonation und Artikulation anzutreffen. Als Einzelparameter wurde am häufigsten die Stimmelage verstellt, d.h. die VPn erhöhten ihre Grundfrequenz oder senkten sie ab. Stimmverstellung wurde mit zunehmender Dauer der Aufnahme inkonsistent. Zudem war die Stimmverstellung mit Ausnahme einer VP auditiv nicht wirkungsvoll, wobei nicht angegeben wird, ob das anhand des Expertenwissens des Autors beurteilt oder ob ein zusätzlicher Hörtest durchgeführt wurde. Des Weiteren konnte Masthoff die normale Stimmelage in Verstellungen dieses Parameters mit einer Genauigkeit von ein bis zwei Halbtonschritten rekonstruieren, die Fragestellung nach der Rekonstruierbarkeit der normalen Stimmelage eines Sprechers konnte zumindest bei Verstellungen der Stimmelage bestätigt werden und ist somit ein guter Anhaltspunkt, wenn es darum geht, Stimmverstellung in Sprachmaterial herauszufinden.

Eine weitere Studie von de Figueiredo und de Souza Britto (1996) wurde aufgrund real vorkommender Kriminalfälle mit Stimmverstellung in Brasilien durchgeführt. Straftäter verstellen ihre Stimme dort häufig in Entführungen, weil sie davon ausgehen können, dass ihre Stimme aufgenommen wird (siehe auch Abschnitt 2.2.2). Da die Anzahl der Entführungen in den letzten Jahren in Brasilien zugenommen hat, konnte auch eine steigende Zahl der Fälle mit Stimmverstellung beobachtet werden. Ein weit verbreitetes Mittel zur Stimmverstellung ist die Verwendung eines Stifts, der parallel zu den Lippen zwischen die Zähne genommen wird. Um den Einfluss dieser Art von Stimmverstellung zu untersuchen, ließen die Autoren drei männliche VPn einen Text mit 150 Wörtern in normaler Sprache und mit einem Stift zwischen den Zähnen lesen. Untersuchungsgegenstand waren die Formanten F1–F3 der Vokale /a, ε, e, i, ɔ, o, u/. Es wurden jeweils zehn Vokale pro Kategorie ausgewählt.

Der Vergleich der verstellten und der unverstellten Vokale ergab einen erheblichen Einfluss auf die Formanten, der jedoch tendenziell bei den VPn gleich war.

In der verstellten Stimme stieg F1 der hohen bzw. halb-hohen Vokale¹¹ /i, o, u/ stark an. Die mittleren Vokale /e, ε, ɔ/ wurden hinsichtlich F1 nur wenig beeinflusst. F1 des tiefen Vokals /a/ wurde durch den Stift zwischen den Zähnen abgesenkt. F1 wurde tendenziell stärker beeinflusst, je weiter hinten und je gerundeter der Vokal produziert wurde.

F2 wies prinzipiell ähnliche Ergebnisse wie F1 auf, erneut war ein Anstieg von F2 bei den gerundeten hinteren Vokalen /ɔ, o, u/ festzustellen, F2 der ungerundeten Vokale /i, e, a/ wurde abgesenkt. Eine Ausnahme bildete /ε/, bei diesem Vokal trat interindividuelle Variation auf, F2 wurde bei zwei VPn minimal erhöht, bei der dritten VP abgesenkt. Erneut war die Tendenz zur Erhöhung von F2 umso stärker, je weiter hinten und gerundeter der Vokal produziert wurde.

F3 der ungerundeten vorderen Vokale /i, e, ε, a/ wurde stark abgesenkt. Die gerundeten hinteren Vokale /o/ und /u/ wiesen einen Anstieg in F3 auf, allerdings wesentlich schwächer als bei F1 und F2. F3 des mittleren Vokals /ɔ/ dagegen wies kaum einen Unterschied zwischen verstellter und unverstellter Stimme auf, tendenziell (bei zwei VPn) war aber ein Absinken von F3 zu beobachten.

Insgesamt gesehen war in diesem Experiment eher die Tendenz zur Erhöhung der Formanten bei hinteren gerundeten Vokalen und eine Absenkung bei den vorderen ungerundeten Vokalen im Vergleich zur unverstellten Stimme vorhanden.

Durch diese Art der Stimmverstellung existieren Einschränkungen in der Artikulation in dreifacher Hinsicht: Der Stift zwischen den Zähnen sorgt dafür, dass die Kieferöffnung konstant bleibt, die Lippenrundung erschwert wird und auch die Beweglichkeit der Zunge stark eingeschränkt ist. Dies wirkt sich unterschiedlich stark auf die einzelnen Vokale aus:

„[...] the disguise [...] provokes alterations, whose magnitude will depend on the intrinsic quality of the segment of speech.“

(de Figueiredo und de Souza Britto 1996, S. 169)

Hintere gerundete Vokale sind stärker von der Verstellung betroffen, weil Lippenrundung nicht möglich ist, wodurch das Ansatzrohr verkürzt wird. In normaler Sprache werden die Formanten, vor allem F2, generell durch Lippenrundung abgesenkt. Ein Vergleich zwischen normaler und verstellter Stimme ergibt höhere Differenzen, wenn die Formanten eines Vokals in normaler Stimme abgesenkt werden, in der verstellten Stimme die Rundung aber fehlt, die Formanten sind also

¹¹es werden für die Vokalqualität die Bezeichnungen *hoch*, *halb-hoch*, *halb-tief*, *tief* von de Figueiredo und de Souza Britto (1996) übernommen

vergleichsweise höher. Je gerundeter der Vokal sein sollte, beispielsweise /u/, umso mehr werden die Formanten bei Stimmverstellung erhöht. In den Ergebnissen dieses Experiments zeigt sich dies vor allem durch einen Anstieg der ersten beiden Formanten bei /ɔ, o, u/, wobei der prozentuale Anstieg von F2 höher ist als von F1. Es ist jedoch festzustellen, dass der gerundete Vokal /u/ stärker in den ersten drei Formanten angehoben wird als die anderen Vokale.

Die Unbeweglichkeit des Kiefers durch den Stift wirkt sich vor allem auf den tiefen Vokal /a/ aus, deren Formanten bei allen VPn abgesenkt wurden. Im Vergleich zur normalen Stimme eines Sprechers ist der Kiefer bei der Produktion mit einem Stift zwischen den Zähnen geschlossener. Der Öffnungsgrad wird durch F1 kodiert, das heißt, wird ein Vokal geschlossener produziert, sinkt F1 ab; wird er mit einer größeren Kieferöffnung produziert, steigt F1. Da dies nicht möglich ist, wenn der Stift zwischen den Zähnen liegt, sinkt F1 gerade bei /a/ stark ab, aber auch die anderen Formanten werden dadurch beeinflusst.

Für den Anstieg von F1 bei /i/ ist zusätzlich die Zungenposition ausschlaggebend: Die Zunge wird für die Produktion zurückgezogen, dadurch wird F1 angehoben, F2 und F3 wird jedoch abgesenkt.

Die erzielten Ergebnisse zeichnen sich durch eine Reduktion des Vokalraums aus. Dies hat zur Folge, dass die peripheren Vokale /i/ und /u/ zentralisiert werden.

Anhand der ausführlich vorgestellten Studien ist erkennbar, dass die Stimmverstellung große Auswirkung auf die Identifikation in allen Bereichen ausübt. Ähnlich verhält es sich mit der Identifikationsleistung naiver Hörer, die bei Stimmverstellung drastisch absinkt (Bull und Clifford 1984). Die dargestellten Untersuchungen haben sich zudem nur auf einige Bereiche gesamtsprachlichen Verhaltens konzentriert, beziehungsweise auf Auswirkungen auf ausgewählte akustische Parameter. Sehr häufig wurden die Formanten untersucht, da deren Lage im Frequenzbereich als stark sprecherspezifisch gilt und sie (unter günstigen Umständen, d.h. störungsfreies Sprachmaterial) einfach zu ermitteln sind. Es besteht trotzdem noch keine Sicherheit darüber, inwiefern sich Stimmverstellung tatsächlich bei unterschiedlichen Individuen ausprägt, da jeder Täter unterschiedlich „talentiert“ zur Stimmverstellung ist und die Möglichkeiten, die Stimme zu verstellen, vielfältig sind. Das Hauptproblem besteht zudem darin, dass die akustischen Korrelate sprecherspezifischer Merkmale nicht zweifellos und mit nötiger Sicherheit bekannt sind (Masthoff 1985) und so auf alle möglichen Verdächtigen angewendet werden können. Es ist demnach sehr schwierig, die Stimmverstellung überhaupt zu erkennen (Künzel 1987), sofern die Originalstimme nicht bekannt ist.

Sowohl die beschriebenen Untersuchungen in Abschnitt 2.3.1 als auch die eigene Analyse sind demnach nur als Grundlagenforschung anzusehen, weil sie unter der Voraussetzung der bekannten Stimmverstellung durchgeführt wurden. Die Anwendbarkeit der Ergebnisse hängt stark von der jeweiligen benutzten Art der Stimmverstellung sowie der Situation in der forensischen Praxis ab.

Kapitel 3

Theorie des Untersuchungsgegenstands

In diesem Kapitel werden die artikulatorischen und akustischen Eigenschaften der Frikative erläutert. Zunächst wird die Frikativproduktion in Abschnitt 3.1 erläutert und darauf basierend werden die akustischen Auswirkungen in 3.2 dargestellt. Abschnitt 3.3 stellt eine Untersuchung zur Variabilität der Frikative vor. Die Abschnitte 3.1 und 3.2 beziehen sich vor allem auf die stimmlosen Frikative, weil sich die artikulatorischen Prozesse stimmhafter Frikative (abgesehen von der gleichzeitigen Erzeugung der Stimmhaftigkeit und einer geringeren Intensität) nicht wesentlich von der Produktion stimmloser Frikative unterscheiden (Fu, Rodman, Bitzer und Xu 1999). Der Abschnitt 3.2 stellt zudem nur die Frikative [f, s, ʃ] dar, da sie auch in der eigenen Untersuchung verwendet werden. Eine Beschreibung des spektralen Verhaltens anderer Frikative ist im Rahmen der vorliegenden Arbeit nicht von Nutzen.

3.1 Produktion der Frikative

Für die Produktion eines Frikativs ist eine Verengung zwischen zwei Artikulatoren erforderlich, durch die der pulmonale Luftstrom mit einer bestimmten Geschwindigkeit und mit einem bestimmten Druck gepresst werden muss, um eine Turbulenz zu erzeugen, die als Friktionsgeräusch wahrgenommen wird (Kohler 1995). Der entstehende auditive Eindruck eines Frikativs ist unter anderem abhängig von der Position der Engebildung, die an jeder Stelle im Ansatzrohr möglich ist. Zusätzlich sind aber auch Formanttransitionen umgebender Sprachlaute für die Wahrnehmung eines Frikativs verantwortlich (Wright, Frisch und Pisoni 1996).

Die Verengung unterteilt das Ansatzrohr in zwei Resonanzräume. Abhängig von der Position der Friktionsenge (dem Artikulationsort) variieren Form und Länge des vorderen und des hinteren Resonanzraums. Generell kann festgestellt werden, dass je kürzer der vordere Resonanzraum ist, umso höher ist die entstehende Resonanzfrequenz. Die daraus entstehenden Auswirkungen werden in Abschnitt 3.2 näher beschrieben. Vor allem der vordere Resonanzraum wirkt sich auf das Spektrum eines Frikativs aus, dies wiederum dient zur Unterscheidung verschiedener Artikulationsorte durch einen Hörer, insbesondere bei den *Sibilanten* [s, z, ʃ, ʒ] (Wright et al. 1996; Harrington und Cassidy 1999).

Aber nicht nur die Länge des Resonanzraumes ist ausschlaggebend für die spektrale und auditive Ausprägung eines Frikativs, auch die Form der Zunge spielt eine Rolle, da sie die Geometrie der Verengung beeinflusst. Die Sibilanten [s] und [ʃ] unterscheiden sich hinsichtlich der Zungenform durch eine gerillte oder flache Friktionsenge, in welcher der pulmonale Luftstrom in Turbulenz gerät. [s] prägt sich durch eine enge Rille in der Zunge aus, [ʃ] durch eine weitere Rille, die Zunge wird hier flacher gehalten. (Clark und Yallop 1995; Kohler 1995; Tabain 2001). Die dritte Frikativkategorie, [f], die untersucht wird, zeichnet sich durch einen sehr kurzen vorderen Resonanzraum aus. Dadurch, dass die Friktionsenge nur zwischen den Zähnen des Oberkiefers und der Unterlippe gebildet wird, ist keine zweite Friktionsquelle vorhanden, im Gegensatz zu den Frikativen. Das führt bei [f] zu einer geringeren Intensität des Spektrums sowie zu einer stärkeren Beeinflussung durch umgebende Sprachlaute.

Die Frikativproduktion ist sehr komplex und teilweise auch stark durch Koartikulation beeinflusst, und eine Änderung in der Position oder Geometrie der Verengung beziehungsweise der Resonanzräume wirkt sich auf die Wahrnehmung des Frikativs aus. Auch wenn eine artikulatorische Variabilität durchaus möglich ist, so muss sie vom Sprecher kontrolliert werden, sodass nicht unabsichtlich ein anderer Frikativ produziert wird.

3.2 Tendenzen im Spektrum

Unterschiedliche Artikulationsorte von Frikativen sind im Spektrum erkennbar, zudem bestehen intra- und interindividuelle Unterschiede. Es existieren aber allgemeine Tendenzen, wie Frikative verschiedener Artikulationsorte sich spektral ausprägen und anhand dessen erkannt werden können (Künzel 1987; Harrington und Cassidy 1999; Tabain 2001). Gerade die interindividuellen Unterschiede

und eine gewisse intraindividuelle Stabilität macht Frikative für die forensische Sprechererkennung interessant, allerdings nur unter der Voraussetzung, dass gute Übertragungsbedingungen vorliegen und das Telefon nicht involviert ist. Die folgenden spektralen Eigenschaften der Frikative erläutern dies näher.

Frikative zeichnen sich im Spektrum durch Aperiodizität beziehungsweise Quasi-periodizität aus. Stimmlose Frikative sind im Spektrum durch aperiodische Energie gekennzeichnet, sie entstehen durch die Verengung, in der die Luft verwirbelt wird. Stimmhafte Frikative sind quasiperiodisch, da die Friktion zusätzlich von den periodischen Schwingungen der Stimmlippen überlagert wird (Kohler 1995). Abhängig vom Artikulationsort ergeben sich Unterschiede in den Spektren der einzelnen Frikative, manche Frikative zeichnen sich durch Energiekonzentrationen in bestimmten Frequenzbereichen aus, andere weisen ein flaches Spektrum ohne wesentliche Energiegipfel (Amplitudenmaxima) auf. Das Vorhandensein der Energiegipfel (Resonanzen) ist abhängig von der Form der Resonanzräume und vor allem von deren Resonanzfrequenzen. Sie liegen in Bezug auf den vorderen Resonanzraum bei [s] im Bereich von 4 bis 7 kHz, während sie bei [ʃ] im Bereich von 2 bis 4 kHz liegen Harrington und Cassidy (1999, Tabain (2001). Es ist gut in Abbildung 3.1 erkennbar, dass die Amplitude des Spektrums in diesen Frequenzbereichen am größten ist.

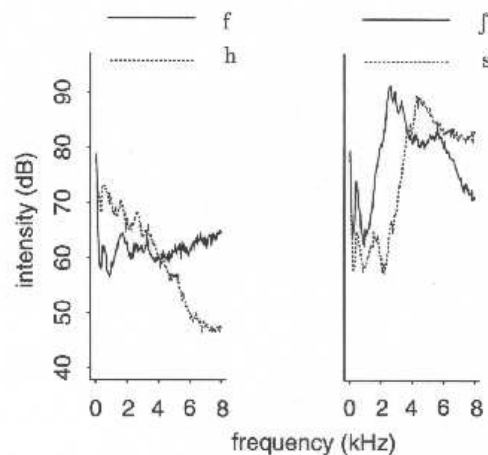


Abbildung 3.1: Gemittelte Spektren der Frikative [f, h, s, ʃ] zum zeitlichen Mittelpunkt nach Harrington und Cassidy (1999)

Anhand der typischen Frequenzbereiche für Frikativresonanzen lässt sich auch erklären, warum Frikative nicht in Sprachaufnahmen, die über das Telefon aufge-

nommen wurden, verwendet werden können: Telefonleitungen übertragen nur Frequenzen von 300–3400 Hz, sodass wichtige spektrale Eigenschaften, wie beispielsweise die Amplitudenmaxima, die für eine Untersuchung der Frikative nützlich sein könnten, nicht übertragen werden.

Spektren, deren Energie sich nicht in einem bestimmten Frequenzbereich konzentriert, werden *diffus* genannt. Beispiele dafür sind die Frikative [f] und [h] in Abbildung 3.1. Im Gegensatz dazu existieren aber auch Spektren, die in einem Frequenzbereich eine Energiekonzentration aufweisen, beispielsweise [s] und [ʃ] in Abbildung 3.1. Solche Spektren werden *kompakt* genannt.

Die Bezeichnungen *diffus* und *kompakt* gehen zurück auf das phonologische System von Jacobson, Fant und Halle (1952), mit dem sich durch verschiedene Merkmalspaare alle Phoneme aller Sprachen auf akustischer Grundlage beschreiben lassen sollten (Pelz 1996). Eines der Merkmalspaare galt der Unterscheidung zwischen diffusen und kompakten Spektren.

Wie diffus oder kompakt ein Spektrum ist, kann mit den so genannten Spektralmomenten quantifiziert werden. Sie dienen dazu, statistische Verteilungen der Frequenzen eines Spektrums zu beschreiben (Nittrouer 1995).

- Spektralmoment 1 (Energie-Hauptkonzentration in einem bestimmten Frequenzbereich [Hz])
- Spektralmoment 2 (Varianz des Spektrums [Hz])

Spektralmoment 1 ist das so genannte *spectral centre of gravity* (im Folgenden als COG bezeichnet) und gibt an, in welchem Frequenzbereich die Energiekonzentration liegt, da es einen gewichteten Mittelwert aus einem Amplitudenwert [dB] und der Frequenz [Hz] für jede Spektralkomponente bildet (Harrington und Cassidy 1999). Da [s] sich durch den kleineren vorderen Resonanzraum bei der Artikulation auszeichnet, liegt das COG oberhalb von dem COG von [ʃ]. [f] hat kein ausgeprägtes COG.

Spektralmoment 2 (im Folgenden als Mom2 bezeichnet) wird auch als *spectral variance* bezeichnet und gibt an, ob das Spektrum diffus oder kompakt ist. Die Berechnung der spektralen Varianz basiert auf dem COG. Je höher das zweite Spektralmoment ist, umso flacher ist das Spektrum. Weist das zweite Spektralmoment einen niedrigen Wert auf, so handelt es sich um ein Spektrum, das in einem Frequenzbereich eine Energiekonzentration aufweist. [f] zeichnet sich, verglichen mit [s] und [ʃ], durch ein diffuseres Spektrum aus, was sich durch einen höheren Wert in Mom2 ausdrückt (Harrington und Cassidy 1999).

Nicht berücksichtigt werden in der eigenen Analyse die Neigung des Spektrums (skewness, Spektralmoment 3) und die Wölbung (kurtosis, Spektralmoment 4), die ebenfalls als statistische Maße zur Beschreibung eines (Frikativ-) Spektrums verwendet werden können.

3.3 Variabilität bei Frikativen

Tabain (2001) führte eine Untersuchung zur Variabilität in der Produktion und der Akustik von Frikativen in Bezug auf zwei Theorien der Sprachproduktion durch.

Die Untersuchung von Tabain (2001) wurde an den Sibilanten [θ, ð, s, z, ʃ, ʒ] in verschiedenen Vokalkontexten in CV-Silben durchgeführt. Das Sprachmaterial wurde von vier Australisch-Englisch sprechenden Frauen produziert. Zusätzlich zu den Sprachaufnahmen wurden elektropalatographische (EPG-) Daten ermittelt, die durch 62 Elektroden auf einer dünnen Acrylplatte im Mundraum die Bewegung der Zunge beziehungsweise die Position der Verengung aufzeichneten. Das Ziel war die Beantwortung von zwei Fragestellungen: Erstens sollte herausgefunden werden, ob und wie stark koartikulatorische Einflüsse durch die Vokalkontexte in der Produktion und der spektralen Ausprägung der Frikative vorhanden waren und inwiefern es Unterschiede zwischen den verschiedenen Frikativen gab. Zweitens sollte überprüft werden, ob Variation in der Produktion von Frikativen sich direkt auf das Spektrum auswirkte. Besonders betrachtet wurde dabei der Unterschied der Sibilanten [s, z, ʃ, ʒ] untereinander, aber auch im Vergleich zu den Nicht-Sibilanten [θ] und [ð].

Zur Untersuchung der Fragestellungen wurde ein spektrales COG sowie ein EPG-COG ermittelt und miteinander verglichen.

Die Studie ergab zur Fragestellung der Koartikulation Folgendes:

Bezüglich der EPG-Daten wurden vor allem bei den dentalen Frikativen [θ] und [ð] große koartikulatorische Einflüsse festgestellt. Die Sibilanten [s, z, ʃ, ʒ] wurden konsistenter produziert, durch Koartikulation bedingte Variation war bei diesen Frikativen geringer. [ʃ] war am wenigsten betroffen.

Die koartikulatorische Wirkung auf die Frikative war vor allem bei gerundeten Vokalen am stärksten. Die dentalen Frikative waren davon (in einer unsystematischen Weise) am meisten betroffen, wobei die Einflüsse phonetisch nicht erklärt werden konnten. Bei [s] und [z] führte Lippenrundung zu einer geringeren aperiodi-

schen Energie im Frequenzbereich von 4 bis 6 kHz. Der Einfluss auf postalveolare Frikative durch die Vokale war vergleichsweise geringer, obwohl auch hier leichte Einwirkungen durch gerundete Vokale vorhanden waren. Die Wirkung der Lippenrundung kam durch eine Verlängerung des Ansatzrohres zustande, die sich auch auf den vorderen Resonanzraum auswirkte und dadurch die Resonanzfrequenzen absenkte.

Die Betrachtung der Frikativ-Spektren ohne Unterscheidung der einzelnen Vokale ergab Unterschiede zwischen den Artikulationsorten. Die alveolaren Frikative [s] und [z] wiesen einen Energieanstieg ab 4 kHz auf, wo die Energie sich darüber hinaus konzentrierte, war sprecherabhängig. Die postalveolaren Frikative [ʃ] und [ʒ] erreichten das Energiemaximum zwischen 3 und 4 kHz, wobei auch hier wieder interindividuelle Unterschiede vorhanden waren.

Die Fragestellung zur Übereinstimmung von artikulatorischen und spektralen Daten konnte bejaht werden, es war eine gute Korrelation zwischen spektralem und EPG COG vorhanden.

Begründet wurden die Ergebnisse von (Tabain 2001) damit, dass die dentalen, alveolaren und postalveolaren Frikative artikulatorisch sehr nah beieinander liegen, so dass eine ungenaue Artikulation sich sehr schnell zu einem perzeptorisch unterschiedlichen Frikativ wandeln würde und dies sich auch im Spektrum ausprägen würde. Dabei spielt nicht nur die Position der Engstelle eine Rolle, sondern auch deren geometrische Form (Länge, Höhe, Breite) sowie die gesamte Form der Zunge. Die Artikulation bei den vorderen Frikativen, speziell den Sibilanten, muss demnach sehr präzise sein, da die Artikulationsorte sehr nahe beieinander liegen (Tabain 2001):

„These three fricative types¹ are physically adjacent in articulatory terms. [...] the differences between the coronal fricatives are extremely subtle, involving not just differences in constriction location, but in constriction length, width and depth, as well as in tongue body shape overall and degree of contact. Variation in any one of these parameters may lead to the perception of another fricative sound. [...] fricative production requires a very precise articulatory configuration in order for the „correct“ acoustic output of noise coupled with frequency peaks to be produced.“ (Tabain 2001, S. 86)

¹Gemeint sind die Artikulationsorte der untersuchten Frikative [θ, ð, s, z, ʃ, ʒ]

Zusätzlich zu der Studie von Tabain (2001) existieren auch andere Untersuchungen, in denen die spektralen Eigenschaften von Frikativen analysiert wurden und die verschiedenen Artikulationsorte, vor allem bei den vorderen Frikativen, teilweise durch Spektralmomente unterschieden werden konnten (Shadle und Mair 1996; Fu et al. 1999; Jesus und Shadle 2000). Die Studie von Fu et al. (1999) bestätigte zusätzlich das Vorhandensein interindividueller Unterschiede in der Frikativproduktion, vor allem in einem Frequenzbereich bis 11 kHz, obwohl der Zweck dieser Untersuchung war, die verschiedenen Artikulationsorte automatisch zu differenzieren und die interindividuellen Variationen demnach für diese Fragestellung nicht von Vorteil waren.

Folgende Gründe führen nun dazu, Frikative als Untersuchungsgegenstand auszuwählen: Die Artikulation muss speziell bei den Sibilanten präzise sein, und die Physiologie eines Sprechers wird bei Stimmverstellung nicht geändert, wie beispielsweise seine Größe oder die Abmessungen seines Ansatzrohres. Stimmverstellung kann lediglich dazu führen, die intraindividuelle Variation zu erhöhen. Dies kann jedoch nur innerhalb der physiologischen Grenzen geschehen, die für die Frikativproduktion notwendig sind und unter der Voraussetzung, dass der Sprecher trotz des Ziels, nicht erkannt zu werden, Information an einen Hörer übertragen möchte, deshalb sollten die phonetischen Kategorien nicht geändert werden, da sich sonst die Perzeption ebenfalls ändern würde. Ein weiterer Grund, der für die Sibilanten als Untersuchungsgegenstand spricht, ist die nach Tabain (2001) belegte Korrelation von Artikulation und Spektrum. [f] wird durch als eine Referenzkategorie ausgewählt, die zeigen soll, ob sich auch andere Frikativkategorien außer den Sibilanten eignen, verschiedene Sprecher zu identifizieren. Es wird erwartet, dass eine höhere intraindividuelle Variation bei diesen Frikativen auftritt, und intraindividuell relativ konstant sind, wie in der Studie von Tabain (2001) vorgestellt wurde. Bestimmte Frikative, wie [s] oder [ʃ], sind in der forensischen Sprechererkennung als „Träger sprecherspezifischer Merkmale“ (Masthoff 2004) bekannt, sofern keine Telefonleitung verwendet wird.

Aufgrund dieser Fakten wird im Folgenden durch Spektralmomente untersucht werden, ob bestimmte Frikative dazu geeignet sind, zwei verschiedene Sprecher im Rahmen einer „open-set“ Situation zu identifizieren, wenn nur verstellte Stimmen von ihnen vorhanden sind.

Kapitel 4

Analyse zur Stimmverstellung

In diesem Kapitel werden die Sprecher und das Sprachmaterial vorgestellt (4.1). In einem Exkurs (Abschnitt 4.1.1) werden die stimmlichen und sprachlichen Eigenschaften der verstellten Stimmen nach auditiv-perzeptiven Kriterien beurteilt. Im Anschluss daran wird die Hypothese für die Experimente (Abschnitt 4.2) dargestellt. Darauf folgend werden in Abschnitt 4.3 die eigentlichen Analyseschritte, bestehend aus der Berechnung der Spektralmomente, dem Klassifikationsverfahren und den Varianzanalysen, beschrieben.

4.1 Sprecher und Sprachmaterial

Die Grundlage für die experimentellen Untersuchungen bilden zwei männliche Sprecher im Alter von ca. 30 Jahren (im folgenden als Sprecher B und D bezeichnet), die beide als Moderatoren bei einem lokalen Radiosender arbeiten.

Sprecher B ist Produzent und alleiniger Sprecher verschiedener Rollen einer Radio-Comedyproduktion. Die Rollen wurden durch Produktion verschiedener Sozio- und Idiolekte geschaffen. Auf die genauen Ausprägungen wird in 4.1.1 näher eingegangen.

Sprecher B wurde telefonisch kontaktiert, um herauszufinden, ob überhaupt Interesse bestand, Sprachaufnahmen zu produzieren. Daraufhin wurde ihm der Text „Nordwind und Sonne“ per Email übermittelt. Zusätzlich erhielt er die Anweisung, den Text in den am häufigsten von ihm produzierten Rollen möglichst frei zu lesen und dabei aber möglichst wenig vom vorgegebenen Text abzuweichen. Das fertige Sprachmaterial wurde dann auf CD persönlich übergeben.

Nach den Aufnahmen stellte Sprecher B den Kontakt zu Sprecher D her. Auch er wurde zunächst telefonisch kontaktiert, um Interesse und Möglichkeiten her-

auszufinden, und auch hier wurden der Text und die Anweisungen per Email übermittelt.

Die Anweisungen für Sprecher D unterschieden sich von denen für Sprecher B, weil nur bekannt war, dass Sprecher D Stimmen prominenter Deutscher (wie z.B. Udo Lindenberg, Marcel Reich-Ranicki) imitieren kann und damit auch in der Öffentlichkeit auftritt. Ihm wurde aufgetragen, seiner Meinung nach gut imitierte Stimmen zu verwenden, und den Text mit diesen Stimmen möglichst frei zu lesen. Das aufgenommene Sprachmaterial von Sprecher D wurde als Email zurückgeschickt, es bestand also kein persönlicher Kontakt zu diesem Sprecher.

Die Anweisungen hinsichtlich der Stärke oder Form der Verstellung an beide Sprecher waren bewusst vage gehalten, um die Sprecher nicht zu beeinflussen. Die Aufnahmen beider Sprecher wurden im Studio des Radiosenders gemacht. Auch hierzu wurden keine weiteren Anforderungen gestellt, außer dass die Qualität den Standard von Studioaufnahmen erfüllen sollte.

Die gesprochenen Texte beider Sprecher wurden in jeweils einer zusammenhängenden, ungeschnittenen Datei übermittelt. Beide Aufnahmen wurden im Studio des Senders als Audiodateien mit 44,1 kHz Abtastrate und 16 Bit Auflösung aufgenommen. Sprecher D komprimierte das Sprachmaterial ins MP3-Format, um es per Email verschicken zu können. Die Dateien beider Sprecher wurden nach Erhalt mit Cool Edit Pro2.0TM zur Weiterbearbeitung ins WAV-Format mit 16 kHz Abtastrate und 16 Bit Auflösung konvertiert und anschließend geschnitten, so dass die Texte der verstellten Stimmen beider Sprecher vorlagen.

In beiden Dateien der Sprecher B und D waren ursprünglich sieben verschiedene Stimmverstellungen enthalten.

- Sprecher B: Stimmen ALI, ANS, HEI, IRM, JUE, PAT, STO
- Sprecher D: Stimmen BEC, KER, LIN, LOR, UNB, UNK, REI

4.1.1 Exkurs: Beschreibung der verstellten Stimmen

Aus den vorliegenden Stimmen wurden jeweils drei Stimmen pro Sprecher ausgewählt. Der Hintergrund der Auswahl war, dass die Stimmverstellungen nicht nur Änderungen in der produzierten Stimme selbst, sondern auch im gesamt-sprachlichen Verhalten aufwiesen. Es war jedoch geplant, eine segmentell basierte Methode zu verwenden, die auf spektralen Untersuchungen im Frequenzbereich an verschiedenen Frikativen basierte. Es erschien günstig, die Menge der zu Verfügung

stehenden Frikative durch eine Vorauswahl der verwendeten Stimmverstellungen einheitlich zu halten.

Stimmverstellungen, die eine lexikalische Abweichung vom vorgegebenen Text „Nordwind und Sonne“, einen starken Dialekt oder auch starke Änderungen in der Stimmqualität aufwiesen, wurden aus der Analyse herausgenommen: Zwei Stimmverstellungen wiesen einen starken Dialekt beziehungsweise Akzent auf (STO und ALI), zwei Stimmverstellungen wichen stark auf einer lexikalischen Ebene vom vorgegebenen Text ab, indem sie Wörter verdrehten, wegließen oder die Satzstellung insgesamt änderten (PAT und ANS) und drei Stimmen änderten Eigenschaften ihrer Stimmqualität derart stark, dass auch sie von der Analyse ausgeschlossen wurden (LIN, REI, UNB2). Die übrigen Stimmen HEI, IRM, JUE von Sprecher B und die Stimmen KER, LOR, UNB von Sprecher D wiesen vor allem Variationen in stimmlichen Eigenschaften auf. Es liegt auf der Hand, dass durch eine Vorab-Auswahl von Sprachmaterial kein realer forensischer Hintergrund bestehen bleibt, da gerade in der forensischen Praxis die gesamtsprachlichen Merkmale wie Dialekt, Idiolekt und Soziolekt dazu dienen, die Herkunft des Sprechers einzugrenzen und zur Identifikation beizutragen (Künzel 1987, Abschnitt 2.1.2).

Um allerdings die Hypothese zu untersuchen, dass die Sprecher B und D in ihren Stimmverstellungen anhand gewisser segmenteller Parameter (Spektralmomente) mit einem automatischen Klassifikationsverfahren identifizierbar sind, und vor dem Hintergrund, dass die Stimmverstellung Bedingung für die Aufnahmen war und die Sprecher vor allem darin trainiert waren, wurden jegliche Faktoren ausgeschaltet, welche die segmentelle Vergleichbarkeit der Stimmen herabsetzten. Diese Bedingung traf auf die genannten Stimmen HEI, IRM, JUE von Sprecher B und die Stimmen KER, LOR, UNB von Sprecher D zu.

Dennoch war in diesen verstellten Stimmen immer noch phonetische Variation, bedingt durch Koartikulation, bezüglich der Stimmhaftigkeit bei der Realisierung von /s/ und /f/ vorhanden. Die Folge der Koartikulation war, dass wortinitiale phonologisch /z/ phonetisch entstimmte wurden. Deshalb wurde der zu untersuchende Frequenzbereich daran angepasst (siehe Abschnitt 4.3), so dass diese Variation keinen Einfluss hatte.

Um Uneindeutigkeiten in der Terminologie vorzubeugen, muss der Begriff „Stimme“ für den aktuellen Abschnitt sowie die gesamte Analyse der vorliegenden Arbeit näher erläutert werden. In Abschnitt 2.1.2 wurde festgestellt, dass im forensischen Kontext nur das gesamtsprachliche Verhalten eines Sprechers beurteilt werden darf. In den darauf folgenden Abschnitten wurde zur eindeutigeren Be-

schreibung auditiv–perzeptiver Sprechereigenschaften zwischen dem sprachlichen Verhalten als auch der produzierten Stimme selbst getrennt. Im Sinne der akustischen Analyse und im Rahmen des vorliegenden Sprachmaterials muss *Stimme* jedoch in einem anderen Zusammenhang betrachtet werden:

Eine Stimme von B oder D ist das vorliegende Resultat seiner Stimmverstellung inklusive dem gesamtsprachlichen Verhalten.

Im Folgenden werden die verwendeten Stimmverstellungen in einem groben Rahmen auditiv–perzeptiv ohne messphonetische Unterstützung beurteilt. Stimmimitation, oder genauer die *Impersonation* (Zetterholm 1997), wird an dieser Stelle und auch in der Analyse in Abschnitt 4.3 als eine Form der Stimmverstellung angesehen. Die Tatsache, dass die Sprecher B und D die Stimmverstellungen zu Unterhaltungszwecken produzierten, ist von untergeordneter Bedeutung, da Änderungen im gesamtsprachlichen Verhalten auftreten, und die Sprecher B und D anhand akustischer Parameter in den einzelnen Stimmen identifiziert werden sollen.

In der folgenden Beurteilung der Stimmen wird auf Änderungen in den Kategorien Respiration, Phonation, Artikulation und Sprechweise Bezug genommen. Dies entspricht der Kategorisierung von Masthoff (2000) in Abschnitt 2.3.2.

Sprecher B verstellte seine Stimme in mehrfacher Weise, um verschiedene Rollen zu kreieren, indem er verschiedene Soziolekte, Idiolekte und teilweise auch Akzente produzierte, um Personen des soziolektalen Umfeldes „Kaufhaus“ zu schaffen. Die in der Analyse verwendeten Stimmen von Sprecher B dienen der (ironischen) Charakterisierung eines Verkäufers (HEI), der eine sehr dominante und herrische Person (einen „prototypischen Macho“) darstellt, einer weiblichen Ansagerin (IRM) und eines männlichen Verkäufers (JUE) mit Bevorzugung des weiblichen Geschlechts („prototypischer Transvestit“).

HEI zeichnet sich durch eine laute, beziehungsweise geschriene Stimme mit niedriger Grundfrequenz aus, die relativ monoton ist. Es werden längere Pausen zur Akzentuierung gesetzt. Zusätzlich ist die Stimme satzfinal glottalisiert. Es scheint, als ob hier der so genannte *Frequency–Code* von Ohala (1983) realisiert wird. Der *Frequency–Code* besagt, dass Stimmen mit einer niedrigen Grundfrequenz etwas Großes und Starkes suggerieren. Dies stimmt überein mit den Vorstellungen, die hinsichtlich des dominanten Verhaltens des Sprechers beim Hören der Comedy–Serie entstehen. IRM prägt sich durch eine Stimme mit sehr hoher Grundfrequenz, wenig Pausen und sehr variabler Melodie aus. JUE ist gekennzeichnet durch eine mittelhohe Stimme, die nach perzeptivem Eindruck aber sehr sonor ist und zur

Realisation dieses Eindrucks sehr stark behaucht wird. Zudem ist die Stimme sehr melodisch, stark akzentuiert und nasaliert.

Sprecher D produzierte im Gegensatz dazu Stimmverstellungen, um verschiedene prominente Personen zu imitieren und änderte dadurch ebenfalls Idiolekt, Soziolekt und teilweise auch Dialekt, um das gesamtsprachliche Verhalten der imitierten Person zu übernehmen.

Die produzierten Verstellungen sind demnach nicht nur einem Soziolekt zuzuordnen, sondern abhängig vom gesamtsprachlichen Verhalten der imitierten Person. Auch bei den Stimmen dieses Sprechers wurde darauf geachtet, dass keine Stimmverstellung verwendet wurde, die die segmentelle Vergleichbarkeit des Textes stark herabsetzte. Aufgrund dieser Kriterien wurden die Stimmimitationen von Kermit (KER), Lorient (LOR) und einer namentlich unbekannt Person (UNB) ausgewählt.

Die Imitationen zeichnen sich durch folgende Eigenschaften aus:

Die Sprechweise des Textes von KER ist sehr schnell und ohne wesentliche wahrnehmbare Pausen. Die Stimmqualität ist stark pharyngalisiert, ansonsten wurden vom Sprecher aber keine weiteren wesentlichen Änderungen an der Stimmqualität vorgenommen, wodurch eine starke auditive Ähnlichkeit mit der bekannten Originalstimme bestehen blieb.

Die Stimme LOR zeichnet sich wiederum durch ein sehr langsames Sprechtempo aus, eine sehr starke Akzentuierung, die aber zu einem sehr variablen Rhythmus und einem sehr „abgehackten“ Eindruck führt. Es ist eine starke Glottalisierung vorhanden. Mehrere Vokale sind stark gelängt und es sind viele Pausen zwischen einzelnen Wörtern eingefügt. Die Stimme klingt gepresst, Teile des Artikulationsapparates werden demnach stark angespannt und es wird ein sehr hoher subglottaler Druck aufgebaut. Ein- und Ausatemgeräusche werden stark hörbar und wesentlich häufiger als bei den anderen Stimmen produziert, was sich vermutlich auf den erhöhten Luftverbrauch zurückführen lässt.

Die Stimme UNB wird wiederum in den Parametern Sprechweise und Artikulation variiert. Sie weist eine sehr angenehme Stimmqualität auf, die durch eine günstige Akzentuierung beim Lesen, vor allem aber durch eine stark behauchte Stimme erreicht wird. Das Sprechtempo ist etwas schneller als das der Stimme LOR und durch wesentlich weniger Pausensetzung geprägt, wodurch die einzelnen Sätze des Textes ineinander übergehen und sehr melodisch klingen.

Nach der beurteilten stimmlichen und gesamtsprachlichen Qualität der Aufnahmen kann Sprecher B seine individuellen Stimmmerkmale besser verstellen als

Sprecher D. Inwiefern dieser Eindruck durch Analysen an verschiedenen Frikativen bestätigt werden kann, wird zu zeigen sein.

Neben diesen Beobachtungen ist es wichtig zu erwähnen, dass Sprecher D bei seiner Aufnahme in Form von Spontansprache Information über ein vorhandenes Lispeln bei [s, ʃ] Auskunft gab, das auditiv jedoch nicht aufgefallen ist, weil es durch die Mikrofon-Positionierung, nach Aussagen des Sprechers, sehr gut kompensiert wurde.

4.2 Hypothesen für die Analyse

Das Ziel der durchgeführten Analyse ist, unter künstlich geschaffenen Voraussetzungen und unter Studiobedingungen herauszufinden, ob trotz Stimmverstellung eine korrekte Identifikation zweier Sprecher durch die Energieverteilungen in verschiedenen Frikativen möglich ist oder ob Stimmverstellung eine Identifikation durch die Spektralmomente COG und Mom2 verhindert.

Dem zugrunde liegt eine artikulatorische Annahme für unverstellte Stimmen: Nach Tabain (2001) müssen die Sibilanten [s] und [ʃ] aufgrund der nahe beieinander liegenden Artikulationsorte präzise produziert werden. Da die Ansatzrohre verschiedener Sprecher jedoch unterschiedlich geformt sind, wird sich die Artikulation bei den verschiedenen Sprechern auch akustisch unterschiedlich ausprägen, da es eine Korrelation zwischen Artikulation und Spektrum gibt. In der Studie von Tabain (2001) konnte eine gewisse interindividuelle Variabilität des COG und der artikulatorischen Produktion belegt werden. Zusätzlich konnte sie belegen, dass sich vor allem die alveolaren und postalveolaren Frikative akustisch und artikulatorisch intraindividuell relativ stabil verhielten.

Unter der Voraussetzung, dass die verschiedenen verstellten Stimmen der beiden Sprecher B und D prinzipiell als extreme intraindividuelle Variation (siehe Abschnitt 2.1.1) in Bereichen der Sprache, Stimme und Sprechweise angesehen werden könnten, die sich auch akustisch auswirkt, ergeben sich folgende Annahmen:

Eine korrekte Identifikation der Sprecher B und D ist durch die Untersuchungsparameter COG und Mom2 bei [f, s, ʃ] möglich, weil die intraindividuellen Unterschiede von Sprecher B beziehungsweise Sprecher D geringer sind als die interindividuellen Unterschiede zwischen beiden Sprechern. Auch im Fall von Stimmverstellung ist es möglich, die Sprecher B und D durch Berechnung von COG und Mom2 an den einzelnen verstellten Stimmen durch ein automatisches Klas-

sifikationsverfahren zu identifizieren. Die Sprecher müssen ihre „artikulatorischen Grenzen“ einhalten, um die vorderen Frikative korrekt zu produzieren. Für [f] wird genau das Gegenteil erwartet. Diese Kategorie dient jedoch vor allem als Referenz für die Sibilanten.

Unter Berücksichtigung der beschriebenen Literatur zur Stimmverstellung in Abschnitt 2.3 ist davon auszugehen, dass die Stimmverstellung des vorliegenden Sprachmaterials zu einer erheblichen intraindividuellen Variation führt, vor allem weil die Sprecher in der Stimmverstellung trainiert sind. Dies wird natürlich die Ergebnisse der Klassifikationen beeinflussen. Vor allem Sprecher B ist in der Stimmverstellung trainiert, da er sie über einen langen Zeitraum mit gleich bleibenden Charakteren für das Radio produzierte. Sprecher D hat die Stimmen ebenfalls häufig produziert, er hat aber durch die selteneren Auftritte mit den Imitationen weniger Übung darin als Sprecher B.

Es werden sich möglicherweise stärkere Unterschiede in der Ausprägung der Stimmen innerhalb von B ergeben, weil die produzierten Stimmen sich in ihrem Idiolekt von einander unterscheiden sollen und der Hörer der Radioproduktion nicht merken soll, dass ein Sprecher alle Rollen produziert. Im Gegensatz dazu nähern sich die Imitationen von Sprecher D an ein bestehendes gesamtsprachliches Verhalten an und werden nicht in einem zusammenhängenden Stück produziert, sodass es nicht auffallen würde, wenn die Imitationen noch stimmliche Eigenschaften des Imitators enthalten. Inwiefern tatsächlich Unterschiede in der Verstellungsleistung der Sprecher vorhanden sind, wird indirekt ebenfalls untersucht, indem die intraindividuellen Unterschiede der Sprecher B und D im Vergleich zueinander beobachtet werden.

Zusammengefasst ergeben sich folgende Hypothesen für die Analyse:

- Sprecher können durch COG und Mom2 identifiziert werden, weil die Spektren von Sibilanten sprecherspezifisch sind. Im Gegensatz dazu wird für [f] eine größere Variation erwartet, ihre Spektren sind nicht derart sprecherspezifisch.
- Die Spektralmomente COG und Mom2 in Sibilanten sind auch in verstellten Stimmen geeignete Parameter für eine Identifikation.
- Es besteht auch in dieser Analyse ein Zusammenhang zwischen der Fähigkeit, die Stimme zu verstellen und den intraindividuellen Unterschieden der Sprecher.

4.3 Durchführung der Untersuchungen

4.3.1 Aufbereitung des Sprachmaterials

Die Aufbereitung und Analyse des Sprachmaterials wurde mit der Software EMUTM (URL 1) und der Programmiersprache RTM (URL 2) durchgeführt.

Das EMU-System ist eine Software, mit der aus aufgenommenen Sprachdateien eine Sprachdatenbank aufgebaut werden kann. Eine Sprachdatenbank besteht aus Signal- und Labeldateien, die über eine sogenannte Templatedatei in die Sprachdatenbank integriert werden. In der Templatedatei stehen die Speicherorte der Signal- und Labeldateien, die Labeltypes und welcher Art die Signaldateien sind. Als Signaldateien lagen Sprachsignale im WAV-Format mit 16 kHz Abtastrate vor sowie DFT-Werte, die aus einer Fourier-Transformation gewonnen wurden.

Durch die Nyquistfrequenz ergab sich ein Untersuchungsbereich bis max. 8 kHz. Das angewendete Fenster in der DFT hatte eine Länge von 512 Punkten, so dass für jedes phonetische Segment der Sprachdatenbank 257 Frequenzkomponenten für die Berechnung der Untersuchungsparameter zur Verfügung standen.

Insgesamt waren für jeden Sprecher drei Aufnahmen vorhanden, jede Aufnahme stellte eine verstellte Stimme dar (siehe Tabelle 4.1). Das Sprachmaterial wurde segmentiert und orthographisch gelabelt und zusätzlich auf einer phonetischen Ebene mit dem SAMPA-Alphabet (URL 3) segmentiert und transkribiert. Die Kriterien zur phonetischen Zuordnung waren vor allem auditiver Art, die Segmentgrenzen wurden mit visueller Unterstützung eines Sonagramms in EMU festgelegt.

Für die akustische Analyse der Stimmverstellung wurden alle mit [f, s, ʃ] transkribierten Frikative zur weiteren Untersuchung verwendet. Darin enthalten waren auch reduzierte Frikative, deren Dauer in einigen Fällen weniger als 20 ms betrug. Eine Trennung der Frikative nach Vokalkontext oder Position im Wort wurde nicht vorgenommen.

4.3.2 Datenanalyse

Um die Analysen an den Frikativen [f, s, ʃ] durchzuführen, wurden folgende Daten aus der Datenbank entnommen:

- die DFT-Werte zum zeitlichen Mittelpunkt von jedem zu untersuchenden Frikativ
- die Sprecherlabels HEI, JUE, IRM, KER, LOR, UNB

- die Frikativlabels [f, s, ʃ]

Tabelle 4.1 stellt die Anzahl n der verwendeten Frikative für die Stimmen der Sprecher B und D dar.

Sprecher	Stimme	[f]	[s]	[ʃ]
B	Hei	9	16	6
	Irm	13	17	6
	Jue	10	14	6
Σ B		32	47	18
D	Ker	14	16	6
	Lor	11	12	6
	Unb	8	12	6
Σ D		33	40	18

Tabelle 4.1: Aufteilung der verwendeten Frikative (n) pro Stimme und Sprecher

Die Anzahl n der Frikative für die Sprecher B und D ist relativ einheitlich, unterscheidet sich jedoch bei den Frikativen [f] und [s] zwischen den Stimmen HEI, IRM, JUE innerhalb von B und KER, LOR, UNB innerhalb von D. Der Grund dafür liegt darin, dass stimmlose, aber auch entstimmte Frikative in die Untersuchung einbezogen wurden. Die entstimmten Frikative kamen vor allem wortinitial vor, jedoch nicht bei allen Sprechern. Hinzu kam, dass bei zwei Stimmen (LOR, UNB) von Sprecher D phonetisch [s̥] für phonologisch /s/ realisiert wurde. Das Kriterium für die Etikettierung der Frikative war auditiv begründet, so dass diese abweichend realisierten Frikative aus der Untersuchung ausgeschlossen wurden. Für eine subjektive Beurteilung der Sprecherunterschiede und auch der Variationen der einzelnen Stimmen bei den einzelnen Artikulationsorten wurden Spektren zum zeitlichen Mittelpunkt im Frequenzbereich von 0 bis 8 kHz erstellt. Sie sind in den Abbildungen A.1 und A.2 dargestellt. Die Basis für die Berechnung der Spektren bildeten die DFT-Werte, die aus einem Amplitudenwert pro Frequenzkomponente bestanden.

4.3.2.1 Spektralmomente und Wahrscheinlichkeitsverteilungen

Unter Berücksichtigung der erstellten Spektren wurden die Spektralmomente 1 und 2 der Frikative [f] und [s] in einem Frequenzbereich von 1-8 kHz und die Spektralmomente von [ʃ] in einem Bereich von 2-7 kHz berechnet. Dadurch ist ein direkter Vergleich von [f, s] zu [ʃ] jedoch nicht günstig. Es handelt sich bei

den Spektralmomenten um statistische Maße, die zur Beschreibung von Verteilungen und Spektren verwendet werden können. Wie auch schon in Abschnitt 3.2 dargestellt, beschreibt das erste Spektralmoment den gewichteten Mittelwert beziehungsweise die Energiekonzentration des jeweiligen Spektrums, das zweite Spektralmoment die Streuung des Spektrums. Beide Spektralmomente zusammen geben Information über die Verteilung eines Frikativspektrums und sind die zugrundeliegenden Untersuchungsparameter der Analyse. Das r-Skript zur Erstellung der Segmentlisten sowie der Berechnung der Spektralmomente befindet sich im Anhang B.1.

Zusätzlich wurden statistische Wahrscheinlichkeitsverteilungen¹ erstellt. Die Wahrscheinlichkeitsverteilung eines Parameters wird, sofern eine Normalverteilung vorliegt, durch den Mittelwert μ und die Standardabweichung σ des ausgewählten Parameters festgelegt. Der Mittelwert besitzt dabei die höchste Wahrscheinlichkeit, dass ein Untersuchungsparameter gerade diesen Wert erreicht. Dies spielt auch für die Klassifikation eine Rolle.

4.3.2.2 Klassifikation

Die Hypothese der Analyse ist, dass die Stimmen der Sprecher B und D durch die Spektralmomente korrekt identifizierbar sind, obwohl Stimmverstellung vorliegt. Die Identifikation zwischen den Sprechern B und D wurde durch eine Klassifikation auf der Basis eines Wahrscheinlichkeitsmodells durchgeführt. Die berechneten Spektralmomente COG und Mom2 von [f, s, ʃ] jeder der 6 Stimmen dienen als Grundlage für die Klassifikationen. Die Spektralmomente COG und Mom2 wurden in \mathbb{R}^{TM} für jeden Frikativ der Kategorien [f, s, ʃ] separat berechnet. Die Klassifikation wurde jedoch getrennt für die Spektralmomente jeder Kategorie durchgeführt und anschließend mit einem Proportionstest statistisch abgesichert.

Insgesamt wurden $2 * 3$ Klassifikationen durchgeführt. Das Skript für die Klassifikation befindet sich kommentiert im Anhang B.2. Die Klassifikation der Daten in der Matrix erfolgte in zwei Phasen (Trainingsphase und Klassifikationsphase). Es wurde eine „open-set“-Situation ausgewählt, um zumindest in dieser Hinsicht einen realen forensischen Hintergrund zu bewahren. Dazu wurde eine Hälfte der Daten von COG oder Mom2 jeder Stimme im Trainingsmodell verwendet, die andere Hälfte im Testmodell. Im Trainingsmodell wurden zwischen den Wahrscheinlichkeitsverteilungen die Entscheidungsgrenzen festgelegt, die zur Trennung der Sprecherkategorien B und D dienen. Beide Modelle basierten auf den in Ab-

¹*probability distributions*

schnitt 4.3.2.1 beschriebenen Wahrscheinlichkeitsverteilungen.

Da es bei den Analysen auf die Trennbarkeit der Sprecher B oder D ankam, wurden im Trainingsmodell die Entscheidungsgrenzen zwischen den Sprecherkategorien B und D durch den gewählten Parameter COG oder Mom2 festgelegt.

Nach der Erstellung der Trainingsmodelle wurde die Klassifikation durchgeführt. In der Klassifikation wurden die Testdaten den Kategorien aus dem Trainingsmodell zugewiesen, d.h. ein Wert, welcher der höchsten Wahrscheinlichkeit im Trainingsmodell entsprach, wurde dieser Kategorie zugeordnet. Dabei entschieden die im Trainingsmodell erstellten Entscheidungsgrenzen über die Zugehörigkeit zu Kategorie B oder D. Als Resultate der Klassifikationen entstanden Verwechslungsmatrizen, in denen die korrekten und falschen Identifikationen für jede Kategorie in absoluten oder in relativen Häufigkeiten (%) angezeigt werden.

Zur Absicherung der Klassifikation wurde ein Proportionstest (*2-sample test for equality of proportions with continuity correction*) durchgeführt. Dieser Test prüft, wie sich die durch die Klassifikation ergebenden richtigen Zuordnungen statistisch verhalten. In anderen Worten: Der Test überprüft, ob die korrekten Klassifikationen der Testdaten zu den Trainingsmodellen statistisch signifikant höher sind als die falschen Klassifikationen. So kann die Identifikationsleistung des Parameters bei der jeweiligen untersuchten phonetischen Kategorie herausgefunden werden. Der Proportionstest basiert auf der Grundlage der Anzahl n pro Frikativ [f, s, ʃ] (siehe Tabelle 4.1 auf Seite 55), den tatsächlich korrekten Identifikationen sowie den theoretisch möglichen Klassifikationen, die sich durch $n/2$ ergeben. Mit n , $n/2$ und den tatsächlich korrekten Klassifikationen wurde dann eine Gegenüberstellung aller Werte durchgeführt, die letztendlich zu einem signifikanten oder einem nicht signifikanten Ergebnis führte. Auch das Durchführungsskript des Proportionstests befindet sich im Anhang B.2.

4.3.2.3 ANOVA

Stimmverstellung führt zu einer Variation in verschiedenen Merkmalen, die als sprecherspezifisch gelten, beispielsweise die Grundfrequenz oder die Vokalformananten (siehe Abschnitt 2.3). Es wird vermutet, dass Stimmverstellung zusätzlich zu einer Änderung in der artikulatorischen Frikativproduktion führt und sich dies durch die Schaffung anderer Identitäten auf das Spektrum auswirkt (entweder durch die Kreation von „neuen“ Personen von Sprecher B oder die Impersonationen von Sprecher D, siehe Abschnitt 4.1.1). Da sich die einzelnen Stimmen auditiv stark unterschieden, wurden im Anschluss an die Klassifikation Varianz-

analysen (*ANalysisOfVariance*) durchgeführt, um herauszufinden, wie stark die intra- und interindividuellen Unterschiede durch die verstellten Stimmen bei den Sprechern B und D ausgeprägt sind. Die Hypothese für die ANOVA war, entsprechend der Darstellung in 4.2, dass die Sprecher sich in dem Parameter COG oder Mom2 unterscheiden. Für die phonetischen Kategorien [f, s, ʃ] getrennt wurden sowohl zweifaktorielle als auch einfaktorielle ANOVAs durchgeführt. Die abhängigen Variablen waren jeweils COG oder Mom2. Die unabhängigen Variablen bei der zweifaktoriellen ANOVA waren die Sprecher und die Stimmen. Da mit einer zweifaktoriellen ANOVA festgestellt werden konnte, dass Unterschiede im statistischen Sinne zwischen den Stimmen vorhanden waren, aber keine Aussagen darüber gemacht werden konnten, ob sie innerhalb von Sprecher B oder D auftraten, wurden zusätzlich einfaktorielle ANOVAs durchgeführt. Sie dienten der Analyse der Unterschiede zwischen den Stimmen innerhalb der Sprecher B oder D. Aus beiden Formen der ANOVA ergab sich folgende Auflistung der unabhängigen Variablen:

- Sprecher vs. Sprecher
- Stimmen innerhalb der Sprecher
- Stimmen innerhalb von Sprecher B
- Stimmen innerhalb von Sprecher D

Im Anhang B.3 befindet sich das Skript, in dem die Durchführungen der ANOVAs aufgeführt sind.

Kapitel 5

Ergebnisse

Dieses Kapitel dient der Darstellung der Ergebnisse der durchgeführten Analysen. In Abschnitt 5.1 werden die Ergebnisse der Voruntersuchungen beschrieben. In Abschnitt 5.2 werden die Ergebnisse der Sprecher und Stimmen mit Hilfe der deskriptiver Statistik, der ANOVA und den Klassifikationen dargestellt. In Abschnitt 5.3 werden die Klassifikationen von COG und Mom2 vorgestellt und in Abschnitt 5.4 werden die Ergebnisse der ANOVA analysiert.

5.1 Beurteilung der Sprechervariabilität in der Voruntersuchung

Abbildung A.1 im Anhang stellt die interindividuellen Unterschiede zwischen Sprecher B und D in den phonetischen Kategorien [f, s, ʃ] im Frequenzbereich von 0–8 kHz und im Amplitudenbereich von 0–70 dB dar. Die Spektren der Sprecher B und D sind über die jeweiligen Stimmen HEI, JUE, IRM beziehungsweise LOR, KER, UNB gemittelt (siehe auch Tabelle 4.1 auf Seite 55).

Auffällig in Abbildung A.1 ist, dass bei den untersuchten Frikativen die Intensität der Spektren des Sprechers B höher ist als die Intensität der Spektren des Sprechers D. Dies kann jedoch daran liegen, dass die Sprecher unterschiedlich weit vom Mikrophon entfernt waren oder unterschiedlich laut gesprochen haben. Auf diesen Unterschied wird deswegen nicht weiter eingegangen. Der so genannte *d.c. offset*, eine Frequenzkomponente bei 0 Hz, die durch die DFT zustande kommt, ist in allen Spektren beider Sprecher zu erkennen.

Das Spektrum von [f] für D verläuft im Bereich von ca. 1–8 kHz sehr flach, für das Spektrum von B ist geringfügig mehr Variation festzustellen. Dort gibt es eine

Abschwächung der Amplitude im Bereich von 4 kHz. Abgesehen von diesem „Tal“ steigt die Amplitude des Sprechers B bis 8 kHz kontinuierlich an.

Für [s] ist festzustellen, dass die Spektren der Sprecher sich in ihrem Verlauf sehr ähneln, abgesehen von dem Unterschied in der gesamten Energie, und die Amplitude relativ kontinuierlich von ca. 1,5–8 kHz ansteigt. Das bei [f] beobachtete „Tal“ von Sprecher B im Bereich von 4 kHz ist auch bei [s] festzustellen, aber in wesentlich schwächerer Ausprägung. Das Spektrum von Sprecher B weist zudem knapp darunter einen sprunghaften Anstieg der Intensität kurz unterhalb von 2 kHz auf.

Die Spektren von [ʃ] weisen insgesamt gesehen die größte Bewegung auf. Die beschriebene Abschwächung der Amplitude im Spektrum des Sprechers B ist erneut zu erkennen. Bis zu diesem Frequenzbereich von 4 kHz steigt die Intensität jedoch stark an und fällt ab ca. 5–8 kHz wieder kontinuierlich ab. Das Spektrum von D weist im Allgemeinen ebenso diese Tendenz auf. Es ist ebenfalls eine Abschwächung der Amplitude bei ca. 4 kHz zu erkennen, allerdings sind Anstieg der Intensität bis 4 kHz und die Abschwächung der Intensität ab 5 kHz wesentlich stärker ausgeprägt.

Alle Spektren des Sprechers D enthalten unterhalb von 0,5 kHz einen unterschiedlich stark ausgeprägten Gipfel, der vermutlich durch die Grundfrequenz zustande kommt, da in der Segmentliste auch entstimmte Frikative enthalten waren. Es ist möglich, dass die Extraktion der DFT-Werte zum zeitlichen Mittelpunkt des Segments noch im stimmhaften Bereich lag. Aus diesem Grund wird der Untersuchungsbereich auf minimal 1 kHz begrenzt.

Abbildung A.2 wurde erstellt, um die Unterschiede zwischen den einzelnen verstellten Stimmen zu verdeutlichen. Der Übersichtlichkeit halber wurden die Grafiken nicht pro Frikativ getrennt, sondern eine Abbildung für jede verstellte Stimme erstellt. Die Darstellungen auf der linken Seite der Abbildung zeigen die Spektren für die Stimmen von Sprecher B, während die Darstellungen auf der rechten Seite die Spektren für die Stimmen von Sprecher D zeigen. Analog zu Abb. A.1 wurden alle verfügbaren Frikative zu einem Spektrum pro phonetischer Kategorie [f, s, ʃ] zusammengefasst.

In dieser Abbildung fällt der sehr ähnliche Verlauf in Teilen des Spektrums von [f] und [s] auf. Dies bezieht sich vor allem auf die Stimmen IRM von Sprecher B und die Stimmen KER und LOR von Sprecher D. Lediglich für die Stimme HEI ist die Amplitude ab ca. 4 kHz auffällig unterschiedlich.

Die Spektren von [ʃ] zeichnen sich bei fast allen Stimmen durch eine hohe In-

tensität im Bereich 2–7 kHz aus. Eine Ausnahme davon bildet nur die Stimme UNB, dort ist die Intensität von [ʃ] sehr gering und die Amplitudenwerte aller drei Frikative überlappen sehr stark.

Bei den meisten Stimmen, ausgenommen JUE, fällt die Amplitude ab 6–7 kHz ab. Im Gegensatz dazu steigt die Amplitude der Spektren von [ʃ] für die Stimmen KER und LOR bis 4 kHz stark an, um dann relativ stark abzusinken.

Die Abbildungen A.1 und A.2 dienen dazu, die Frequenzbereiche für die Berechnung der Spektralmomente festzulegen. Aufgrund der genannten Besonderheiten wurden sie folgendermaßen eingegrenzt: Die Frikative [f] und [s] wurden im Bereich wegen des *d.c. offsets* erst ab 1 kHz untersucht, aber bis 8 kHz, da die Amplitude teilweise bei [f] und [s] bis zu dieser Frequenz noch anstieg. [ʃ] wurde im Bereich von 2–7 kHz untersucht, da sich die Energie in diesem Bereich konzentrierte und von 7–8 kHz wieder absank. Als vorläufiges und subjektives Fazit aus den Abbildungen A.1 und A.2 war festzustellen, dass Unterschiede zwischen den Spektren der verschiedenen Frikativen existierten, aber im Vergleich zwischen den Sprechern nicht sehr stark waren. Die Spektren in Abbildung A.2 wiesen etwas mehr Variabilität im spektralen Verhalten der Frikative auf.

5.2 Deskriptive Statistik der Spektralmomente

In diesem Abschnitt werden die Spektralmomente COG und Mom2 mit Hilfe der deskriptiven Statistik beschrieben, um die Beobachtungen aus Abschnitt 5.1 durch die Betrachtung der Mittelwerte und Standardabweichungen zu objektivieren.

Tabelle 5.1 auf Seite 62 zeigt die gerundeten Mittelwerte und Standardabweichungen für das COG und Mom2 der Kategorien [f, s, ʃ]. Die Werte wurden aus allen vorhandenen Frikativen pro Stimme berechnet, für die Sprecher B und D in Tabelle 5.1 schließen die Mittelwerte und Standardabweichungen alle Realisationen der zugehörigen Stimmen mit ein.

Generell betrachtet sind Unterschiede in den Mittelwerten und Standardabweichungen beider Spektralmomente abhängig vom Artikulationsort vorhanden. Wegen des abweichenden Frequenzbereichs von [ʃ] ist ein Vergleich von [f, s] mit diesem Frikativ jedoch ungünstig.

Mom2 von [ʃ] ist am geringsten. Bei [f, s] handelt es sich um sehr flache beziehungsweise diffuse Spektren, bei [ʃ] um ein eher kompaktes Spektrum. Der Vergleich von [f] und [s] im Parameter Mom2 zeigt Überschneidungen der Mittelwerte sowohl der einzelnen Stimmen als auch der Mittelwerte von Sprecher B und

D. Mom2 bei [f] von Sprecher B liegt bei 1651 Hz und von Sprecher D bei 1628 Hz. Mom2 bei [s] liegt für Sprecher B bei 1629 Hz und für Sprecher D bei 1623 Hz. Im Vergleich der beiden Frikative im Parameter COG weist jedoch [s] höhere Werte auf. Gerade für diese Kategorie ist solch ein hohes Mom2 jedoch sehr ungewöhnlich, da [s] sich eher durch ein kompaktes Spektrum mit einem niedrigen Mom2 auszeichnet (siehe Abschnitt 3.2).

	Stimme/ Sprecher	μ COG [Hz]	σ COG [Hz]	μ Mom2 [Hz]	σ Mom2 [Hz]
[f]	HEI	3612	98	1649	30
	IRM	3677	85	1670	29
	JUE	3646	95	1628	29
	B	3649	93	1651	33
	KER	3664	85	1627	22
	LOR	3655	105	1634	29
	UNB	3612	109	1621	23
	D	3649	97	1628	25
[s]	HEI	3973	111	1602	51
	IRM	3847	171	1661	35
	JUE	3868	117	1621	40
	B	3896	146	1629	49
	KER	3782	90	1616	43
	LOR	3801	124	1618	30
	UNB	3880	111	1638	22
	D	3817	113	1623	35
[ʃ]	HEI	3640	89	1162	18
	IRM	3651	65	1125	13
	JUE	3747	85	1119	17
	B	3679	90	1135	25
	KER	3720	27	1093	35
	LOR	3710	49	1120	20
	UNB	3748	28	1109	36
	D	3726	38	1107	31

Tabelle 5.1: Mittelwerte und Standardabweichungen für COG und Mom2 von [f, s, ʃ] der einzelnen Stimmen und der Sprecher

Zwischen den verschiedenen Stimmen innerhalb der Sprecher B und D sind ebenfalls Überschneidungen in den Mittelwerten festzustellen. Insbesondere die Mittelwerte und Standardabweichungen für die Sprecher B und D insgesamt zeigen, dass das COG von [f] sich zwischen den Sprechern nicht unterscheidet und auch die Betrachtung der Mittelwerte der einzelnen Stimmen zeigt keine starken Unterschiede.

Große Variation zwischen den Stimmen und Sprechern zeigt sich beim COG von [ʃ], dort weist die Stimme JUE von Sprecher mit 3747 Hz einen wesentlich

höheren Mittelwert auf als die beiden anderen Stimmen des Sprechers. Zudem überschneidet er sich mit Mittelwerten der Stimmen von Sprecher D. Im Parameter Mom2 weist die Stimme JUE jedoch den geringsten Mittelwert auf.

Auch bei [s] sind die Mittelwerte unterschiedlich, es ergeben sich Abweichungen zwischen den Sprechern B und D. Der Gesamt-Mittelwert für Sprecher B liegt über dem von Sprecher D. Dies trifft auch für Mom2 zu.

Bezogen auf die Mittelwerte von Mom2 sind die Unterschiede zwischen B und D erneut bei [ʃ] am größten, wobei angemerkt werden muss, dass eine Differenz von 28 Hz zwischen Mittelwerten vernachlässigbar ist, die Differenzen bei den Frikativen [f] (23 Hz) und [s] (6 Hz) sind geringer.

Vor allem bei Mom2 fällt auf, dass die Mittelwerte von Sprecher B generell höher sind als von Sprecher D.

Auffällig ist, dass die Standardabweichungen für das COG sehr klein sind, außer bei [s], dort treten im Verhältnis zu den anderen beiden Frikativen die größten Standardabweichungen auf, was sich auch in den Gesamtwerten für die Sprecher B (146 Hz) und D (113 Hz) widerspiegelt. Die Standardabweichungen für Mom2 sind ebenfalls relativ gering, aber wie bei [s] ist auch hier der Maximalwert bei [s] zu finden (49 Hz für Sprecher B). Lediglich bei COG von [f] und Mom2 von [ʃ] weist Sprecher B eine geringere Standardabweichung auf.

Als Fazit aus diesen deskriptiven Werten ergibt sich, dass in den Spektralmomenten der Sprecher B und D Unterschiede vorhanden sind, die teilweise jedoch sehr gering sind. Innerhalb der Sprecher B beziehungsweise D sind keine klaren Tendenzen erkennbar, wodurch zum Teil interindividuelle Überschneidungen von COG und Mom2 auftreten.

Die Beobachtungen anhand der Mittelwerte und Standardabweichungen werden durch die Darstellung als Wahrscheinlichkeitsverteilungen (Normalkurven) in Abbildung A.3 im Anhang verdeutlicht.

In dieser Abbildung sind die Normalkurven der Untersuchungsparameter COG und Mom2 für Sprecher B und D getrennt für die einzelnen Frikative [f, s, ʃ] dargestellt. Der Mittelwert der Normalverteilung wird durch die höchste Wahrscheinlichkeit auf der y-Achse angegeben, die Standardabweichung wird durch die Breite der Normalverteilung erkenntlich. Je geringer die Standardabweichung, umso höher ist die Wahrscheinlichkeit, dass ein bestimmter Wert erreicht wird und umso schmaler ist die Normalkurve.

Eine extreme Überschneidung der Kurven von B und D zeigt sich beim COG von [f]. Die geringe Abweichung der beiden Kurven voneinander kommt zustan-

de durch die minimal unterschiedliche Standardabweichung von 4 Hz bei beiden Sprechern, da die Mittelwerte identisch sind (siehe Tabelle 5.1).

Die anderen Verteilungen überschneiden sich mehr oder weniger stark und entsprechen somit den Beschreibungen, wie sie schon zu den Mittelwerten und Standardabweichungen in diesem Abschnitt gegeben wurde. An den Visualisierungen der deskriptiven Werte fällt jedoch sehr gut auf, dass die Standardabweichungen zwischen den Sprechern unterschiedlicher sind als die Abweichungen in den Mittelwerten. Klar dargestellt wird auch, dass die Standardabweichungen von B häufiger höher sind als die von D, was dadurch zu erkennen ist, dass die Kurven von B bis auf COG von [f] und Mom2 von [ʃ] flacher sind als die Kurven von D. Auf eine Darstellung der Verteilungen für die einzelnen Stimmen kann verzichtet werden, da sie durch starke Überschneidungen, wie sie auch schon bei den Mittelwerten und Standardabweichungen beschrieben wurden, unübersichtlich sind und keine neuen Erkenntnisse erbringen.

Fazit der deskriptiven Statistik:

Hinsichtlich der Konsistenz der Frikativ-Produktion scheint es Unterschiede zu geben. Die Frequenzbereiche, in denen sich die Spektralmomente (im Vergleich von Sprecher B zu D als auch deren verstellten Stimmen) ausprägen, sind an sich jedoch nicht stark unterschiedlich.

5.3 Klassifikationen von COG und Mom2

Die Klassifikationen dienen dazu, die Sprecher B und D durch die Parameter COG und Mom2 an drei verschiedenen Frikativkategorien [f, s, ʃ] zu identifizieren. Die Grundlage dafür bildeten Wahrscheinlichkeitsverteilungen, wie sie in Abbildung A.3 dargestellt sind.

Tabelle 5.2 fasst die Resultate der Klassifikationen von COG und Mom2 in insgesamt sechs Verwechslungsmatrixen getrennt für die Frikative [f, s, ʃ] zusammen. Die Sprecher B und D enthalten wiederum jeweils die drei verstellten Stimmen¹. Eine Verwechslungsmatrix stellt die korrekten und falschen Zuordnungen von Daten auf ein erstelltes Trainingsmodell in einer Klassifikation dar. Die Diagonale von links nach rechts in einer Matrix ergibt summiert die Anzahl der richtigen Klassifikationen für das COG oder Mom2 des jeweiligen Frikativs, die Werte au-

¹HEI, IRM, JUE sind von Sprecher B, KER, LOR, UNB sind von Sprecher D.

ßerhalb von dieser Diagonale in den Spalten und Reihen und zeigen die Anzahl der Falschklassifikationen.

Allgemein ausgedrückt entsprechen korrekte Klassifikationen dem Bereich innerhalb der Wahrscheinlichkeitsverteilungen, wie sie in in Abbildung A.3 im Anhang dargestellt sind. Falsche Klassifikationen entsprechen den Überschneidungen der Wahrscheinlichkeitsverteilungen zwischen den Sprechern B und D.

Parameter		[f]		[s]		[ʃ]	
		B	D	B	D	B	D
COG	B	9	7	13	10	7	2
	D	10	6	7	13	1	8
Mom2	B	7	9	15	8	8	1
	D	3	13	7	13	5	4

Tabelle 5.2: Verwechslungsmatrizen der Parameter COG und Mom2 von [f, s, ʃ]

Insgesamt gibt es für das COG von [f] 15 korrekte Zuweisungen gegenüber 17 Falschklassifikationen. Der Sprecher B wird durch COG neun Mal richtig (als B) und sieben Mal falsch als D klassifiziert. Sprecher D wird sechs Mal richtig erkannt und zehn Mal falsch als B klassifiziert. Mom2 von [f] führt zu 20 korrekten und 12 falschen Klassifikationen. Sprecher B wird dabei sieben Mal korrekt und neun Mal falsch klassifiziert, Sprecher D 13 Mal korrekt und drei Mal falsch.

Bei dem Frikativ [s] ergibt der Parameter COG insgesamt 26 korrekte und 17 falsche Identifikationen. Sprecher B und Sprecher D werden gleich häufig, nämlich 13 Mal, richtig klassifiziert. Die Anzahl der korrekten Klassifikationen für Mom2 liegt für beide Sprecher ebenfalls nahe beieinander. Es gibt 15 richtige Klassifikationen von Sprecher B und 13 richtige Klassifikationen von Sprecher D. Insgesamt werden beide Sprecher durch Mom2 15 Mal verwechselt.

Beim Frikativ [ʃ] werden die Sprecher B und D anhand des COG insgesamt 15 Mal korrekt klassifiziert. Sprecher B wird sieben Mal und Sprecher D acht Mal richtig identifiziert. Die Anzahl der Sprecherverwechslungen ist bei diesem Frikativ sehr gering, es gibt nur drei Falschklassifikationen. Die Verwendung des Mom2 bei [ʃ] führt zu 12 korrekten und sechs falschen Klassifikationen. Sprecher B wird anhand dieser Daten nur ein Mal falsch klassifiziert.

Die Ergebnisse der Klassifikation werden übersichtlicher, wenn die korrekten Zuordnungen der Parameter als prozentuale Anteile der gesamten Klassifikationen ausgedrückt werden.

	[f]	[s]	[ʃ]
COG	46,88	60,47	83,33
Mom2	62,50	65,12	66,67

Tabelle 5.3: Korrekte Anteile (in %) der gesamten Klassifikationen von COG und Mom2 für [f, s, ʃ]

Es zeigt sich in Tabelle 5.3 auf Seite 66, dass es (bis auf Parameter COG bei [f]) häufiger korrekte als falsche Klassifikationen gibt; es werden mindestens 60 % der Daten richtig klassifiziert. Das COG von [ʃ] zeichnet sich mit 83,33 % durch den höchsten Anteil der korrekten Identifikationen aus. Lediglich bei COG von [f] werden weniger als die Hälfte der entsprechenden Frikative korrekt klassifiziert.

Die Klassifikationen wurden mit einem Proportionstest statistisch überprüft. Die Ergebnisse sind in Tabelle 5.4 dargestellt. Sie zeigt die Resultate des Tests für die Untersuchungsparameter COG und Mom2 sowie die untersuchten phonetischen Kategorien [f, s, ʃ].

		χ^2	p-value	Signifikanz
COG	[f]	0	0.5	n.s.
	[s]	0.5761	0.2239	n.s.
	[ʃ]	3.125	0.03855	*
Mom2	[f]	0.5714	0.2248	n.s.
	[s]	1.4399	0.1151	n.s.
	[ʃ]	0.4571	0.2495	n.s.

Tabelle 5.4: Ergebnisse des Proportionstests von COG und Mom2 bei [f, s, ʃ]

Zunächst fällt auf, dass es nur ein statistisch signifikantes Ergebnis für das COG von [ʃ] gibt. Nur dieser Parameter bei diesem Frikativ ist demnach dazu geeignet, die Stimmen korrekt zu klassifizieren und auf einem statistisch signifikanten Niveau eine Identität der Sprecher auszuschließen. Bei allen anderen Frikativen ergeben sich durch die untersuchten Parameter nicht-signifikante Ergebnisse, das heißt, die Sprecher B und D waren in dem untersuchten Parameter nicht zu unterscheiden.

Fazit der Klassifikation:

Die Klassifikationen ergaben nur bei COG von [f] ein statistisch signifikantes Ergebnis, obwohl des öfteren mindestens 60 % der Daten von COG oder Mom2 richtig zugeordnet werden konnten.

5.4 ANOVA von COG und Mom2

Im Folgenden werden die Analysen für die beiden Untersuchungsparameter COG und Mom2 getrennt ausgewertet.

Tabelle 5.5 auf Seite 67 stellt die gerundeten Ergebnisse der ANOVA zwischen den Sprechern und den Stimmen für das COG und Mom2 für [f, s, j] dar. *Sprecher* steht für die aufgetretenen Unterschiede zwischen den Sprecher B und D, *Stimmen* bezeichnet einen mit Hilfe der ANOVA ermittelten Unterschied innerhalb der Stimmen eines Sprechers B oder D, während *intra B* und *intra D* diese Unterschiede zwischen den Stimmen, getrennt für den Sprecher B beziehungsweise Sprecher D, bezeichnen.

		F-ratio	Pr(>F)	Sign.
[f]	Sprecher	0.00	0.9872	n.s.()
	Stimmen	1.03	0.4019	n.s.()
	intra B	1.34	0.2787	n.s.()
	intra D	0.75	0.4800	n.s.()
[s]	Sprecher	8.73	0.0041	**
	Stimmen	3.49	0.0110	*
	intra B	3.88	0.0280	*
	intra D	3.05	0.0593	n.s.(.)
[j]	Sprecher	5.15	0.0305	*
	Stimmen	3.01	0.0335	*
	intra B	3.25	0.0672	n.s.(.)
	intra D	1.84	0.1933	n.s.()

Tabelle 5.5: Ergebnisse der ANOVA für das COG

Es zeigt sich, dass das COG von [f] sich weder zwischen den Sprechern B und D (*Sprecher*) noch zwischen den Stimmen innerhalb der Sprecher (*Stimmen, intra B, intra D*) statistisch signifikant unterscheidet. [f] weist einen F-Wert von $F = 0$ und dadurch einen extrem hohen p-Wert von $p = 0.9872$ auf, das heißt, bei diesem Frikativ ist die statistische Wahrscheinlichkeit sehr hoch, dass die Sprecher sich in dem Parameter COG nicht unterscheiden und die Hypothese abgelehnt wird.

Für [s] ist ein sehr signifikanter Unterschied zwischen den Sprechern B und D beim Parameter COG und auch zwischen den Stimmen innerhalb des Sprechers B erkennbar. Die Stimmen von Sprecher D unterscheiden sich nicht voneinander. Für [ʃ] ergibt sich durch die zweifaktorielle ANOVA ein signifikanter Unterschied zwischen den Sprechern B und D. Des Weiteren ist ein signifikanter Effekt innerhalb eines Sprechers vorhanden. Die einfaktorielle ANOVA für die Variationen innerhalb der Sprecher weist jedoch weder bei B noch bei D einen signifikanten Unterschied auf.

Auch für den Parameter Mom2 wurden nach dem selben Prinzip Varianzanalysen durchgeführt. Die Ergebnisse sind in Tabelle 5.6 auf Seite 68 dargestellt. *Sprecher*, *Stimmen*, *intra B* und *intra D* entsprechen den Bezeichnungen in Tabelle 5.5.

Die Sprecher B und D (*Sprecher*) unterscheiden sich hochsignifikant bei [f] im Parameter Mom2. Die ANOVA ergibt ebenfalls einen signifikanten Unterschied innerhalb des Sprechers B (*intra B*). Bei Mom2 von [s] zeigt sich dagegen kein signifikanter Unterschied zwischen den Sprechern, es gibt allerdings einen statistisch hochsignifikanten Einfluss zwischen den Stimmen innerhalb eines Sprechers (*intra B*). Bei [ʃ] treten signifikante Unterschiede zwischen B und D auf. Auch innerhalb eines Sprechers sind signifikante Abweichungen vorhanden, es ist jedoch weder anhand der Werte bei *intraB* noch *intraD* ersichtlich, von welchem Sprecher die Unterschiede verursacht werden.

		F-ratio	Pr(>F)	Sign.
[f]	Sprecher	12.12	0.0009	***
	Stimmen	3.58	0.0112	*
	intra B	5.73	0.0080	**
	intra D	0.57	0.5699	n.s.(.)
[s]	Sprecher	0.46	0.5007	n.s.(.)
	Stimmen	5.70	0.0004	***
	intra B	8.53	0.0007	***
	intra D	1.62	0.2126	n.s.(.)
[ʃ]	Sprecher	11.87	0.0017	**
	Stimmen	3.62	0.0159	*
	intra B	13.12	0.0005	***
	intra D	1.16	0.3391	n.s.(.)

Tabelle 5.6: Ergebnisse der ANOVA für Mom2

Fazit der Varianzanalysen

Die Hypothese, dass die Sprecher hinsichtlich der Parameter COG und Mom2 statistisch verschieden sind, musste teilweise abgelehnt werden: In zwei Fällen

traten keine statistisch signifikanten Unterschiede auf (COG von [f] und Mom2 von [s]). Bezüglich der intraindividuellen Unterschiede von Sprecher B und D waren des Öfteren statistisch signifikante Einflüsse erkennbar, die vor allem bei Mom2 bei Sprecher B zwischen den Stimmen HEI, IRM und JUE auftraten. Jedoch waren bei [f] und bei [ʃ] keine signifikanten intraindividuellen Unterschiede bei Sprecher B feststellbar. Die Stimmen KER, LOR, UNB von D unterschieden sich hinsichtlich der Parameter COG und Mom2 nie statistisch signifikant voneinander.

Kapitel 6

Diskussion

In diesem Kapitel werden die Ergebnisse der durchgeführten Analysen diskutiert (Abschnitt 6.1) und einige Aspekte auf die forensische Sprechererkennung bezogen (Abschnitt 6.2).

6.1 Ergebnis– und Methodendiskussion

Kapitel 5 erbrachte Ergebnisse, die in diesem Kapitel näher betrachtet und diskutiert werden müssen:

- Die Spektralmomente COG und Mom2 waren zwischen den Sprechern B und D nicht annähernd so unterschiedlich wie erwartet.
- Die Ergebnisse wichen hinsichtlich der verschiedenen Artikulationsorte der untersuchten Frikative bei [s] von der Literatur ab.
- Die Klassifikationen waren nur beim COG von [f] so erfolgreich, dass sie sich auch durch einen statistischen Test signifikant ausprägten.
- Prozentual ausgedrückt wurden, bis auf eine Ausnahme, mindestens 60 % der Werte korrekt klassifiziert
- Die Varianzanalysen erbrachten teilweise Ergebnisse, die mit den Resultaten der Klassifikation nicht in Übereinstimmung zu bringen waren.
- [f] war für eine Analyse der Stimmverstellung anhand der ausgewählten Parameter nicht brauchbar.

Der vorliegenden Arbeit lag die Hypothese zugrunde, dass die Sprecher B und D anhand der Spektralmomente identifiziert werden können. Die Basis dafür bildete die Annahme, dass verschiedene Sprecher sich durch unterschiedliche Ansatzrohre auszeichnen und gewisse artikulatorische Grenzen bei der Frikativproduktion (gerade der Sibilanten [s, ʃ]) eingehalten werden müssen, um nicht die phonetische Kategorie zu ändern. Dies wurde auch für Stimmverstellung vorausgesetzt, welche als eine Form extremer intraindividuelle Variation angesehen wurde (Abschnitt 2.1.1). Durch die angenommene Korrelation von Artikulation und spektraler Ausprägung sollten die Spektralmomente bis zu einem gewissen Grad sprecherunterscheidend sein.

Die Mittelwerte der Spektralmomente für die phonetischen Kategorien [f, s, ʃ] wiesen zwischen den beiden Sprechern nur relativ geringe Unterschiede auf oder waren sogar identisch, wie im Fall von [f]. Die weitere Betrachtung der Mittelwerte von den anderen Frikativen ergab keine Tendenzen innerhalb der Sprecher B oder D, die wiederum Hinweise auf eindeutige Unterschiede zwischen oder auch innerhalb der Sprecher ergeben hätten. Es waren jedoch Überschneidungen von Mittelwerten der Stimmen HEI, IRM, JUE von Sprecher B mit den Stimmen KER, LOR, UNB von Sprecher D vorhanden. Lediglich bei [ʃ] waren die Mittelwerte der Stimmen von Sprecher D für das COG höher als die Mittelwerte der Stimmen von Sprecher B. Für das zweite Spektralmoment waren bei den Stimmen von Sprecher B höhere Mittelwerte als bei Sprecher D festzustellen.

Interessanterweise waren die Standardabweichungen insgesamt bei Mom2 im Vergleich zum COG geringer. Dies deutet auf eine konsistente Produktion diffuser, beziehungsweise kompakter Spektren hin. Beim COG von [s] traten bei allen Stimmen beider Sprecher vergleichsweise hohe Standardabweichungen auf. Da [s] in Abbildung A.1 durch ein sehr flaches Spektrum auffiel, sind die hohen Standardabweichungen prinzipiell nicht sehr verwunderlich, da es sich dort um eine inkonsistente Position der Energiekonzentration handeln könnte (durch das statistische Maß COG ausgedrückt). Es scheint bei [s] der Fall zu sein, dass Einflüsse vorhanden waren, die nicht speziell mit der Artikulation zusammenhängen, beispielsweise Koartikulation oder die Position von [s] im Wort.

Die Koartikulation vor allem in Verbindung mit gerundeten Vokalen wirkt sich nach Tabain (2001) innerhalb der Sibilanten stärker auf [s] als auf [ʃ] aus. Da die Spektralmomente in der vorliegenden Analyse aus allen verfügbaren Frikativen der jeweiligen Kategorie ermittelt wurden, kann die Ursache tatsächlich in der Koartikulation zu finden sein. Lediglich vorkommende Frikative mit einer (au-

ditiv beurteilten) dentalen Qualität [s] wurden aus der Analyse ausgeschlossen, indem sie mit anderen Labels als [f, s, ʃ] markiert wurden. Vokalkontexte oder auch die Position des Frikativs im Wort (initial, medial, final) wurden ebenfalls nicht berücksichtigt. Die Position des Frikativs im Wort könnte insofern eine Rolle spielen, als dass vor allem mediale Frikative häufig vor allem in unakzentuierten Wörtern reduziert werden. Auch in dieser Hinsicht wurde das verwendete Sprachmaterial nicht eingeschränkt. [s] war im verwendeten Sprachmaterial vor allem in medialer und finaler Wortposition vertreten und kam in initialer Position nur als entstimmte Variante von /z/ vor, während [ʃ] nur in initialer Position auftrat. Die Annahme, dass es einen Grund gibt, der außerhalb der sprecherspezifischen spektralen Eigenschaften von [s] liegt, wird dadurch erhärtet, dass das hohe Mom₂ bei den Stimmen von B als auch von D auftrat, und [s] sich bei beiden Sprechern durch ein flaches Spektrum auszeichnete (siehe Abbildung A.1 und A.2). Dies wäre nicht der Fall, wenn die Produktion eines flachen Spektrums eine sprechertypische Eigenschaft von Sprecher B oder D wäre, hervorgerufen beispielsweise durch einen Sprachfehler, wie er bei Sprecher D vorliegt.

Einen wichtigen Bestandteil der Analyse zur Stimmverstellung stellte die automatische Klassifikation dar.

Zwei Aspekte müssen in Hinsicht auf die Klassifikation beachtet werden: Einerseits konnten die Sprecher B und D trotz überschneidender Mittelwerte und unterschiedlicher Standardabweichungen identifiziert werden. Die Betrachtung der absoluten Häufigkeit ergab häufig mehr korrekte Klassifikationen als falsche, mit Ausnahme von COG bei [f, s]. Diese Tatsache wurde verdeutlicht, indem die korrekten Klassifikationen als prozentuale Anteile der gesamten Zuordnungen ausgedrückt wurden. Hierbei zeigte sich, dass fast alle Zuordnungen der Testdaten auf die Trainingsmodelle zu mindestens 60% richtig waren. Eine Ausnahme bildete lediglich das COG von [f]. Ein Anteil korrekter Klassifikationen über der Zufallsgrenze von 50% ergab sich bei [s] nur durch die höhere Anzahl von Falschklassifikationen, die sich zwischen den Sprechern unterschied, weil die Anzahl der Frikative nicht einheitlich war.

Andererseits ergab der durchgeführte Proportionstest jedoch, dass im statistischen Sinne nur die Klassifikation bei COG von [ʃ] erfolgreich und somit nur dort eine Identifikation der beiden Sprecher anhand ihrer verstellten Stimmen möglich war. Diese beiden wichtigsten Ergebnisse der Klassifikation lassen sich folgendermaßen in Einklang bringen: Die mit Abstand schlechteste Klassifikation beim COG von [f] mit einem korrekten Anteil von 46,88 % führte im Proportionstest zu einem

nicht-signifikanten Ergebnis, dass durch eine Prüfgröße von $F = 0$ auffiel und, generell ausgedrückt, Gleichheit der Sprecher in diesem Parameter implizierte. Im Gegensatz dazu zeichnete COG von [ʃ] sich mit einem Anteil korrekter Klassifikationen von 83,33 % aus. Dieses Ergebnis spiegelte sich im Proportionstest durch Signifikanz in dem Parameter wider. Es ist zu vermuten, wenn die vergleichsweise geringe Anzahl n der verwendeten Frikative berücksichtigt wird, dass eine Erhöhung der verwendeten Frikative auch zu statistisch signifikanten Ergebnissen bei [s] im Proportionstest führen könnte. Allerdings wäre dann bei einer genügend großen Anzahl von Frikativen, beispielsweise [s], eine Feststellung der Identifikationsleistung von [s] verglichen mit [ʃ] nicht mehr aussagkräftig. Zudem könnte vermutet werden, dass eine zusätzliche Trennung der Frikative nach Vokalkontexten (und einer höheren Anzahl n) bei [s] zu anderen Ergebnissen hinsichtlich der Identifizierbarkeit der Sprecher durch verstellte Stimmen geführt hätte. Wichtiger als die statistische Signifikanz ist nach persönlicher Meinung jedoch, dass die Frikative [s] und [ʃ] besser als [f] in der Lage waren, zwischen den Sprechern B und D zu unterscheiden, obwohl Stimmverstellung vorlag, und dass das COG von [ʃ] tendenziell besser geeignet war, die Stimmen zu identifizieren, obwohl die gesamte Anzahl dieser phonetischen Kategorie mit $n = 36$ kleiner war als die von [s] mit $n = 87$. Vor allem bei der phonetischen Kategorie [f], die sich im Spektrum durch das Fehlen einer Energiekonzentration auszeichnet, war erkennbar, dass der Parameter COG (der prinzipiell nur eine statistische Größe darstellt), entsprechend der Erwartung, ungeeignet für eine erfolgreiche Klassifikation war.

In den Varianzanalysen traten, allgemein betrachtet, statistisch signifikante inter- und intraindividuelle Unterschiede auf. Bei COG von [ʃ] waren die interindividuellen Unterschiede größer als die intraindividuellen Unterschiede. Beim COG von [s] und Mom2 von [f] waren sowohl inter- als auch intraindividuelle Unterschiede vorhanden, die sich jedoch durch Signifikanz auf unterschiedlichen Konfidenzintervallen auszeichneten (interindividuell **, intraindividuell von Sprecher B * für COG [s], interindividuell ***, intraindividuell bei Sprecher B ** für Mom2 von [f]). Bei Mom2 von [s] und [ʃ] trat ein umgekehrtes Verhältnis auf; hier waren größere intraindividuelle als interindividuelle Abweichungen festzustellen. Lediglich bei COG von [f] waren weder inter- noch intraindividuelle Differenzen erkennbar. Beachtenswert ist es, dass sich trotz dieser inter- und intraindividuellen Differenzen in den Parametern COG und Mom2 mehr richtige als falsche Klassifikationen ergaben. Es trat jedoch nur bei COG von [ʃ] das von Wolf (1972) geforderte Verhältnis zwischen hoher interindividueller und geringer intraindividueller Variation auf, die

sich durch statistische Signifikanz gegenüber Nichtsignifikanz äußerte.

Aus den Varianzanalysen ergab sich eine weitere Tendenz bezüglich der Stimmverstellungsleistung der Sprecher B und D: Außer beim COG von [f] waren die Abweichungen zwischen HEI, IRM, JUE größer als die Variationen zwischen KER, LOR und UNB. Vor allem in Mom2 der untersuchten Kategorien [f, s, ʃ] wies Sprecher B hochsignifikante Abweichungen zwischen seinen Stimmen auf, während der gleiche Parameter für Sprecher D ausnahmslos nicht signifikante Ergebnisse erbrachte. Das deutet darauf hin, dass Sprecher B seine Stimmen tatsächlich unterschiedlicher voneinander produzierte als Sprecher D. Dieser in Abschnitt 4.1.1 beschriebene allgemeine auditive Eindruck konnte tendenziell anhand der untersuchten Parameter COG und Mom2 an drei verschiedenen Frikativkategorien bestätigt werden. Es ist jedoch an dieser Stelle nicht nachvollziehbar, ob die Ursache in dem Sprachfehler von Sprecher D liegt, der an den Frikativen auditiv nicht wahrgenommen werden konnte, aber zu einer Ausprägung in der Form von flacheren Frikativspektren führte, oder ob der Grund mit der vermehrten Übung in der Stimmverstellung zusammenhängt. Ein weiterer Grund könnte der Zweck der Stimmverstellung sein, indem die Zielsetzung, unterschiedliche Stimmen zu produzieren und sie im Rahmen einer Radioproduktion innerhalb von höchstens drei Minuten zu präsentieren, zur maximal möglichen Differenzierung der Stimmen führte. Dies sind allerdings nur Vermutungen, die nur anhand des Sprachsignals ohne Unterstützung von artikulatorischen Daten weder bestätigt noch abgelehnt werden können.

Zusammengefasst konnten die Analysen folgende Hypothesen bestätigen:

- Trotz einer vergleichsweise kleinen Anzahl von Frikativen war eine Klassifikation der Sprecher B und D durch jeweils drei verstellte Stimmen mit den Spektralmomenten tendenziell möglich, sodass in fast allen Fällen mindestens 60 % der Daten korrekt klassifiziert werden konnten und somit über der Zufallsgrenze von 50 % lagen.
- Die Stimmen von Sprecher B wurden abweichender voneinander produziert als die Stimmen von Sprecher D, was sich akustisch vor allem durch flachere Spektren bei Sprecher D ausdrückte.
- Trotz der hohen intraindividuellen Unterschiede von Sprecher B bestanden auch statistisch signifikante Unterschiede zwischen den Sprechern B und D.
- [ʃ] war besser zur Identifikation geeignet als [s]. COG bei [f] war für eine Identifikation nicht geeignet.

6.2 Konsequenzen

Aus den Ergebnissen der durchgeführten Analysen entstehen einige Konsequenzen in Bezug auf die forensische Sprechererkennung, die nun diskutiert werden. Es hat sich gezeigt, dass Stimmverstellung zu hoher intraindividuellem Variation eines Sprechers in den untersuchten Frikativen, vor allem in Mom2, führte. Diese große Variation könnte auf die häufige Produktion der gleichen verstellten Stimmen von Sprecher D zurückgeführt werden. Die geringe intraindividuelle Variation bei Sprecher D, die niemals eine statistische Signifikanz aufwies, ist entweder auf den vorhandenen Sprachfehler zurückzuführen oder darauf, dass seine Übung in der Stimmverstellung geringer ist. Trotz der geringeren intraindividuellen Variation in den untersuchten Parametern von Sprecher D waren die Sprecher mit dem verwendeten Verfahren anhand ihrer verstellten Stimmen überwiegend nicht identifizierbar.

Der Grund für die geringe Klassifikationsleistung ist demnach in den untersuchten phonetischen Kategorien zu sehen, die sich trotz der erwarteten Sprecherspezifität als nicht geeignet für eine Klassifikation herausstellten. Daraus lässt sich der Schluss ziehen, dass auch ein Klassifikationsverfahren, das auf der Anwendung von Entscheidungsgrenzen zwischen statistischen Verteilungen basiert, die Identifikation von Sprechern nicht verbessern kann, wenn die ausgewählten Parameter durch ihre Anfälligkeit gegenüber intraindividuellem Variation ungeeignet sind. Dabei spielt es keine Rolle, ob die Variation absichtlich produziert wurde, wie bei Stimmverstellung, oder durch organische Ursachen auftritt.

Dies wird dadurch bestätigt, dass in der automatischen Sprechererkennung von der Verwendung einzelner Untersuchungsparameter Abstand genommen wurde und stattdessen ganze Sets von Parametern untersucht werden:

„[...] research focuses on methods able to exploit efficiently whole sets of parameters for the comparison of the control recording and the disputed utterance. Thus, the advances are mainly due to improvement in techniques for making speakerdependent feature measures and models and do not derive from a better understanding of speaker characteristics of their means of extraction.“ (Meuwly 2000b, S. 1418)

Zusätzlich wird von Meuwly festgestellt, dass intraindividuelle Sprechervariabilität, in Form von Stimmverstellung oder Änderungen in der Stimme und dem Sprachverhalten im zeitlichen Verlauf, auch für die automatische Klassifikation ein Problem darstellt, weil keine einheitlichen Normalisierungstechniken in der

automatischen Sprechererkennung vorhanden sind, die zur Minimierung intraindividuelle und Maximierung interindividueller Variabilität dienen. Nützlich ist die automatische Sprechererkennung hauptsächlich, wenn zusätzlich andere Methoden verwendet werden, wie die auditiv-perzeptive oder die akustisch-phonetische Analyse des Sprachmaterials oder, wie empfohlen, nur eine ausreichende Anzahl an Untersuchungsparametern verwendet wird (Künzel 1987). Zudem ist es notwendig, einen gültigen Interpretationsrahmen zu erschaffen, das heißt, Referenzdatenbanken zu erstellen, anhand derer die Verteilungen günstiger Parameter in Bezug auf die mögliche Population beurteilt werden können (Meuwly et al. 2003). Dies wurde im Rahmen der vorliegenden Arbeit nicht getan, da eine einseitige Fragestellung zugrundelag, die sich lediglich auf die Bestätigung der Hypothese der Identifizierbarkeit der Sprecher bezog. In der Praxis wären vor allem auch die Irrtumswahrscheinlichkeiten der statistischen Auswertungen und somit die Falschidentifikationen stärker berücksichtigt worden. Zudem wurden von den Sprechern keine Referenzverteilungen erstellt, die direkt die intraindividuelle Variation berücksichtigt hätten. Der Ansatz, die intraindividuelle Variation zu betrachten, war lediglich durch die Varianzanalysen durchgeführt worden und entsprach teilweise den Ergebnissen der Klassifikation (COG bei [j]), teilweise waren aber auch erhebliche Unterschiede in den Varianzanalysen erkennbar, die sich in der statistischen Bewertung der Klassifikation nicht über einem Zufallsniveau ausprägten (Mom2 von [f]).

Nach wie vor sind die günstigen Parameter in der Sprechererkennungsforschung nicht bekannt, zumindest nicht unter der Voraussetzung, dass sie auch bei Stimmverstellung oder anderer intraindividuelle Variation anwendbar sein müssen. In der eigenen durchgeführten Analyse hat sich zudem ergeben, dass Frikative in ihren spektralen Eigenschaften trotz einer gewissen erforderlichen artikulatorischen Präzision bei Stimmverstellung derart variiert werden können und durch Koartikulation beeinflusst werden, dass Klassifikationen anhand spektraler Parameter, die aus phonetischen Segmenten ermittelt werden können, größtenteils erfolglos bleiben. Lediglich ein Untersuchungsparameter bei einer phonetischen Kategorie war zur Identifikation geeignet, und in einer praktischen Anwendung wäre diese Erkenntnis nicht anwendbar, sofern das Telefon verwendet würde, und der Bereich ab 3400 Hz nicht mehr zur Verfügung stünde. Dies bezieht sich auf die in Abschnitt 2.2 angesprochene Zielsetzung von Studien zur Stimmverstellung, aber auch anderen Bereichen. Wenige Studien existieren, die sich direkt mit der Untersuchung von Eigenschaften bei der Stimmverstellung beschäftigen, die auch für die

Erkennung von Stimmverstellung in der Praxis geeignet sind. Automatische Systeme sind heutzutage ebenfalls noch nicht geeignet, da sich diese Ansätze darauf konzentrieren, Sprecher unter idealen Bedingungen identifizieren zu können. Es scheint der Fall zu sein, dass die Sprechererkennungsforschung eher von der Wirtschaft gefördert wird, und dort Stimmverstellung selten vorkommt, stattdessen aber zuverlässige Systeme für die Sprecherverifikation entwickelt. Die Sprechererkennung im forensischen Bereich hat ebenfalls bei bestimmten Delikten kaum mit Stimmverstellung zu tun, in anderen Delikten kommt sie jedoch häufig vor (siehe Abschnitt 2.2). Es muss grundsätzlich davon ausgegangen werden, dass ein Täter seine Stimme und sein sprachlichen Verhalten verstellt, um unerkannt zu bleiben. Aus diesem Grund ist Forschung, die sich direkt auf die praktische Anwendung bezieht, wichtig (Masthoff 1996).

In der eigenen Analyse hat sich unter idealen Bedingungen eine intraindividuelle Variation bei Sprecher B ergeben, die sehr wahrscheinlich auf seine ständige, und über mehrere Jahre andauernde, Produktion der gleichen Stimmen zurückzuführen ist. Dies ist problematisch, da auch davon ausgegangen werden kann, dass ein Straftäter bei der „Planung“ seines Verbrechens die Stimmverstellung vorher übt. Wie lange ein Sprecher üben muss, um seine individuellen Merkmale zu verdecken, beziehungsweise ob dies gerade auch untersuchten Parameter betrifft, ist an dieser Stelle nicht zu klären. Es ist jedoch auffällig, dass der Sprecher D wesentlich geringere intraindividuelle Variation aufwies. Als Fazit bleibt, dass Stimmverstellung in Untersuchungen an segmentellen Kategorien, auch mit Unterstützung einer automatischen Klassifikation, zu Problemen in der Identifikation führt und somit die Angaben von Endres et al. (1971), McGlone et al. (1977) und anderen bestätigt werden können.

Kapitel 7

Zusammenfassung

Das Ziel der vorliegenden Magisterarbeit mit dem Titel *Eine akustisch-phonetische Analyse zur Stimmverstellung* war es, verschiedene Methoden der Sprechererkennung sowie die Problematik der Stimmverstellung als Vorbereitung auf eine eigene Untersuchung herauszuarbeiten.

Die hauptsächliche Schwierigkeit in der Sprechererkennung entsteht durch Variabilität der menschlichen Stimme und seiner individuellen Verwendung der Sprache. Das gesamtsprachliche Verhalten von Sprechern unterscheidet sich einerseits durch organische, andererseits durch erlernte Ursachen. Dadurch entstehen intra- als auch interindividuelle Sprecherunterschiede. Organische Faktoren wirken sich auf die Stimme und das gesamtsprachliche Verhalten durch die Komplexität der Sprachproduktion aus, indem es einerseits nicht möglich ist, in zwei aufeinanderfolgenden Äußerungen einen identischen subglottalen Druck, gleiche Muskelanspannungen und/ oder neuromuskuläre Prozesse zu produzieren. Andererseits zeichnen auch gerade diese artikulatorischen Eigenschaften unterschiedliche Sprecher aus und können dazu dienen, sie voneinander zu unterscheiden. Intraindividuelle Unterschiede werden erhöht durch die Zeit, die zwischen zwei Äußerungen vergeht, durch physiologische beziehungsweise psychologische Faktoren oder durch Stimmverstellung. Weitere Variation zwischen verschiedenen Sprechern entsteht auf der sprachlichen Ebene durch unterschiedliche Dialekte, Idiolekte und Soziolekte. Diese Variationen entstehen durch die Umgebung oder soziokulturelle Faktoren und werden als erlernte Ursachen der Sprechervariabilität bezeichnet.

Die forensische Sprechererkennung wird von beiden Formen der Sprechervariabilität beeinflusst, indem sie einerseits Sprechereigenschaften ausnutzt, die sich von anderen Sprechern unterscheiden, und die andererseits innerhalb eines Sprechers vergleichsweise stabil bleiben.

Diese Vorgehensweise ist jedoch schwierig, solange noch kein Verdächtiger vorhanden ist. Die Aufgabe besteht demnach zunächst darin, den Täter anhand seiner sprecherspezifischen Merkmale so zu beschreiben und zu charakterisieren, dass aufgrund dieses Sprecherprofils Verdächtige gefunden werden können.

Sprecherspezifische Merkmale können auf verschiedenen Ebenen beurteilt werden. In der heutigen forensischen Sprechererkennung dienen vor allem Merkmale aus dem Bereich der Sprache und Sprechweise als sehr wichtige Anhaltspunkte, unter anderem der Dialekt. Dieser Ansatz wird auch auditiv-perzeptiver Ansatz genannt.

Desweiteren, und ebenso wichtig, können aber auch akustisch-phonetische Merkmale eines Sprechers beurteilt werden, die zwar ebenfalls auditiv wahrnehmbar sein können, jedoch mit Hilfe instrumenteller Verfahren objektiviert werden und so ein höheres Maß an Entscheidungsstärke bieten, wie beispielsweise Formanten, die sich auditiv als unterschiedliche Vokalqualitäten auszeichnen.

Merkmale, die nicht mehr aus einem vorliegenden Sprachsignal selbst beurteilt werden können, sondern vor allem mit Hilfe automatischer Sprechererkennungssysteme ermittelt werden, sind beispielsweise LPC- oder Kepstralkoeffizienten. Durch Anwendung wahrscheinlichkeitstheoretischer Methoden werden daraus Modelle erstellt, welche den Verdächtigen in dem untersuchten Parameter beschreiben. Durch Ähnlichkeitsmessungen mit den Parametern des Täters, die auf statistischen Prinzipien basieren, wird die Identität eines Verdächtigen mit dem Täter bestätigt oder abgelehnt.

In der forensischen Sprechererkennung bestehen Schwierigkeiten, welche die ohnehin vorhandenen Komplikationen noch erhöhen, wie beispielsweise Sprachaufnahmen über das Telefon und/ oder andere Störgeräusche, oder eine sehr kurze Dauer des Sprachmaterials. Ein sehr großes Problem stellt auch die Stimmverstellung dar. Stimmverstellung tritt dann auf, wenn ein Täter davon ausgeht, dass seine Stimme nicht aufgezeichnet wird und er auf diese Art seine Identifikation verhindern kann. Gerade in Entführungen und Erpressungen kommt dies häufiger vor als beispielsweise bei Notrufmissbräuchen, bei denen der Täter sehr sicher sein kann, dass seine Stimme aufgenommen wird.

Die Folgen der Stimmverstellung auf das gesamtsprachliche Verhalten wurden in vergleichsweise wenigen Studien erforscht. In allen Studien kam jedoch ein Einfluss im Vergleich zum normalsprachlichen (unverstellten) Verhalten heraus. So änderten sich die Positionen und Bandbreiten von Vokalformanten, die Grundfrequenz, aber auch Frequenzverteilungen im Langzeitspektrum stark.

Die eigene Analyse wurde durchgeführt, um herauszufinden, ob unter idealen Studiobedingungen eine Identifikation zweier in Stimmverstellung trainierter Sprecher mit einem automatischen Verfahren an Parametern segmenteller Lautkategorien möglich ist. Für die Untersuchung wurden die Spektralmomente der Frikative [f, s, ʃ] ausgewählt. Frikative an sich sind für solch eine Analyse geeignet, weil sie leicht aus dem Sprachsignal zu extrahieren sind und ihre Spektren bestimmte Tendenzen aufweisen, die generell vergleichbar sind, sich jedoch auch durch interindividuelle Unterschiede auszeichnen. Die einzige Bedingung, die für eine Untersuchung der Frikative erfüllt sein muss, ist ein ausreichender Frequenzbereich. Dadurch ist die Sprecherspezifität im forensischen Bereich nicht relevant, weil dort häufig nur Sprache mit einem telefonbeschränkten Frequenzbereich von 300–3400 Hz zur Verfügung steht.

Es war innerhalb dieser Analyse von Interesse, herauszufinden, inwiefern sich zwei Sprecher anhand bestimmter Frikative identifizieren lassen, wenn ideale Bedingungen vorherrschen und nur verschiedene verstellte Stimmen der Sprecher vorliegen und die Sprecher zusätzlich in der Stimmverstellung trainiert sind.

Die Frikative sind gut geeignet, eine Analyse an ihnen durchzuführen, weil deren Spektren durch die Spektralmomente 1 und 2 sehr leicht und gut beschrieben werden können und keine aufwendigen Verfahren notwendig sind, um diese berechnen zu können. Es besteht die Annahme, dass es bei den Sibilanten einen Zusammenhang zwischen der Artikulation und der spektralen Ausprägung gibt. Gerade diese Frikativkategorien müssen jedoch sehr präzise artikuliert werden, da der alveolare und postalveolare Artikulationsort sehr nahe beieinander liegen.

Im Deutschen kommen als Sibilanten nur diese beiden Artikulationsorte vor und es wurden auch nur die stimmlosen Frikative untersucht. Als eine Referenz wurde der Frikativ [f] verwendet, da erwartet wurde, dass die interindividuellen Unterschiede hier geringer sind.

Aus diesen allgemeinen Tatsachen über die Frikative resultierend wurde als Hypothese aufgestellt, dass die Frikative [s, ʃ] durch ihre Spektralmomente dazu geeignet sind, die Sprecher B oder Sprecher D anhand sechs verschiedener verstellter Stimmen zu identifizieren [f] wurde zusätzlich verwendet, jedoch ohne die Erwartung, dass diese Kategorie zu einer Sprecheridentifikation beitragen könnte. Zusätzlich wurden durch das bestehende Training in der Stimmverstellung große intraindividuelle Unterschiede erwartet, von denen jedoch angenommen wurde, dass sie kleiner wären als die interindividuellen Variationen.

Die Basis für die Analysen war ein in der Phonetik häufig verwendeter Text, aus

dem Frikative extrahiert wurden, die nach auditiver sowie visueller Beurteilung eines Sonagramms segmentiert und als [f, s, ʃ] gelabelt worden waren. Eine wichtige Analyse war die Klassifikation der Spektralmomente, die auf einem Wahrscheinlichkeitsmodell aus der Hälfte der Sprecherdaten für jede untersuchte Kategorie [f, s, ʃ] basierte und getrennt für das erste und das zweite Spektralmoment durchgeführt wurde. Die Klassifikation ergab trotz geringer Frikativanzahlen für die verschiedenen Kategorien häufig mehr als die Hälfte an korrekten Zuordnungen der Testdaten zu dem jeweiligen Trainingsmodell. Tendenziell waren die Spektralmomente, berechnet an Frikativen, also für eine Identifikation der beiden Sprecher durch ihre verstellten Stimmen geeignet. Im Vergleich der untersuchten Frikative und der untersuchten Parameter durch einen statistischen Test stellte sich heraus, dass [ʃ] auf einem statistisch signifikanten Niveau am besten für eine Identifikation zu verwenden war, gefolgt von schlechteren Ergebnissen für die Parameter von [s] und [f].

Zusätzlich zu der Klassifikation wurden Varianzanalysen durchgeführt, mit deren Hilfe die inter- und intraindividuellen Unterschiede beurteilt werden sollten. Das Ergebnis dieser Analysen war, dass im zweiten Spektralmoment aller untersuchten Frikative die Stimmen von Sprecher B statistisch signifikant unterschiedlich voneinander waren und die Stimmen von Sprecher D im Gegensatz dazu keine signifikanten Ergebnisse aufwiesen. Trotzdem war eine statistisch signifikante interindividuelle Variation bei [f] und auch bei [s] festzustellen, das heißt, die Sprecher waren trotz starker intraindividuellen Unterschiede bei Sprecher B im Parameter Mom2 verschieden.

Als Konsequenzen für die forensische Sprechererkennung ergaben sich die Forderung nach einem einheitlichen Interpretationsrahmen, auch bei Stimmverstellung günstigen Parametern, deren Ermittlung jedoch immer noch ein Problem darstellt, sowie der Durchführung von Studien, die sich direkt auf die Anwendbarkeit in der Praxis beziehen.

Literatur

- Battaner, E., J. Gil und V. Marrero (2003). VILE: Acoustic Study of Inter and Intra Speaker Variation in Spanish. In *Actas del II Congreso de la Sociedad Española de Acústica Forense*, Barcelona, pp. 59–70.
- Broeders, A. P. A. (2001). Forensic speech and audio analysis 1998–2001. Paper presented at the 13th INTERPOL Forensic Science Symposium, Lyon, France, pp. 53–84.
- Broeders, A. P. A. und A. Amelvoort (1999). Lineup construction for forensic earwitness identification: a practical approach. In *Proceedings of the 14th International Congress of Phonetic Sciences ICPHS*, Volume 2, San Francisco, pp. 1373–1376.
- Bull, R. und B. R. Clifford (1984). Earwitness voice recognition accuracy. In G. Wells und E. Loftus (Eds.), *Eyewitness Testimony: Psychological perspectives*, pp. 92–123. Cambridge: Cambridge University Press.
- Bußmann, H. (1990). *Lexikon der Sprachwissenschaft* (2 ed.). Stuttgart: Kröner.
- Champod, C. und D. Meuwly (2000). The inference of identity in forensic speaker recognition. *Speech Communication* (31), 193–203.
- Clark, J. und C. Yallop (1995). *An introduction to phonetics and phonology*. Oxford, Cambridge: Blackwell.
- Endres, W., W. Bambach und G. Flösser (1971). Voice spectrograms as a function of age, voice disguise and voice imitation. *Journal of the Acoustical Society of America* 49(6), 1842–1848.
- Figueiredo de, R. M. und H. de Souza Britto (1996). A report on the acoustic effects of one type of disguise. *Forensic Linguistics* 3(1), 168–175.
- Fu, H., R. Rodman, D. Bitzer und B. Xu (1999). Classification of voiceless fricatives through spectral moments. In *Proceedings of the 5th International*

- Conference on Information Systems Analysis and Synthesis (ISAS)*, Skokie, pp. 307–311.
- Furui, S. (1995). Speaker recognition. In R. A. Cole, H. Uszkoreit, J. Mariani, A. Zaenen und V. Zue (Eds.), *Survey of the State of the Art in Human Language Technology*, pp. 42–49. Cambridge: Cambridge University Press.
- Gfroerer, S. (2003). Auditory–instrumental forensic speaker recognition. In *Proceedings of the 8th European Conference on Speech Communication and Technology*, Eurospeech 2003, Geneva, Switzerland.
- Gfroerer, S. (2004). Personal communication.
- Harrington, J. und S. Cassidy (1999). *Techniques in Speech Acoustics*. Dordrecht, Boston, London: Kluwer Academic Publishers.
- Hertrich, I. (1986). Experimentelle Untersuchungen zur individuellen Variabilität der menschlichen Sprechstimme. Technical report, Fakultät für Biologie der LMU München.
- Hollien, H. (1990). *The acoustics of crime - the new sciences of forensic phonetics*. New York: Plenum Press.
- Hollien, H. und W. Majewski (1977). Speaker identification by long–term spectra under normal and distorted speech conditions. *Journal of the Acoustical Society of America* 62(4), 975–980.
- Hollien, H., W. Majewski und E. Doherty (1982). Perceptual identification of voices under normal, stress and disguised speaking conditions. *Journal of Phonetics* 10, 139–148.
- Jakobson, R. (1960). Linguistics and Poetics. In T. A. Sebeok (Ed.), *Style in Language*, pp. 350–377.
- Jakobson, R., G. Fant und M. Halle (1952). *Preliminaries to Speech Analysis (MIT Acoustics Laboratory Technical Report, 13)*. Cambridge, MA: MIT Press.
- Jesus, L. M. T. und C. H. Shadle (2000). Parameterizing spectral characteristics of European Portuguese fricatives. In *Proceedings of the 5th Speech Production Seminar*, pp. 301–304.
- Kersta, L. G. (1962). Voiceprint identification. *Nature* 196, 1253–1257.
- Künzel, H. J. (1987). *Sprechererkennung – Grundzüge forensischer Sprachverarbeitung*. Heidelberg: Kriminalistik–Verlag.

- Künzel, H. J. (1989). Die Erkennung von Personen anhand ihrer Stimme. *Neue Zeitschrift für Strafrecht* 9, 400–405.
- Künzel, H. J. (1990). *Phonetische Untersuchungen zur Sprecher-Erkennung durch linguistisch naive Personen*. Zeitschrift für Dialektologie und Linguistik, Beihefte, Heft 69. Stuttgart: Franz Steiner Verlag.
- Kohler, K. J. (1995). *Einführung in die Phonetik des Deutschen - 2. Auflage*. Berlin: Erich Schmidt Verlag.
- Masthoff, H. (1985). Aktuelle Grenzen der automatischen Sprechererkennung in der Forensik. In J. P. Köster (Ed.), *Neue Tendenzen in der Angewandten Phonetik I*, pp. 95–104.
- Masthoff, H. (1996). A report on a voice disguise experiment. *Forensic Linguistics* 3(1), 160–167.
- Masthoff, H. (2000). Die Auswirkung von Stimmverstellung auf ausgewählte phonetische Merkmale. In B. Nolte (Ed.), *Anwendungen der Akustik in der Wehrtechnik*, pp. 58–66. Hamburg.
- Masthoff, H. (2004). Personal communication.
- McGlone, R. E., H. Hollien und P. Hollien (1977). Acoustic analysis of voice disguise related to voice identification. In *Proceedings of the International Conference on Crime Countermeasures*, Oxford BU113.
- Merlin, T., J. F. Bonastre und C. Fredouille (1999). Non directly acoustic process for costless speaker recognition and indexation. *International Workshop on Intelligent Communication Technologies and Applications*.
- Meuwly, D. (2000a). Sprechererkennung: Arbeit für Mensch oder Computer? *Crimiscope, publication de l'Institut de Police Scientifique et Criminologie de l'Université de Lausanne* (8).
- Meuwly, D. (2000b). Voice analysis In J. Siegel, P. Saukko und P. Knupfer (Eds.), *Encyclopedia of Forensic Science*, pp. 1413–1420. London: Academic Press.
- Meuwly, D. (2003). Decision schemes for forensic use of biometric technology. Unpublished.
- Meuwly, D. (2004). Personal communication.
- Meuwly, D., A. Drygajlo und A. Alexander (2003). Statistical methods and bayesian interpretation of evidence in forensic automatic speaker recogniti-

- on. In *Proceedings of the 8th European Conference on Speech Communication and Technology*, Eurospeech 2003, Geneva, Switzerland.
- Mokhtari, P. (1998). *An acoustic-phonetic and articulatory study of speech-speaker dichotomy*. Ph. D. thesis, School of Computer Science University College.
- Nittrouer, S. (1995). Children learn separate aspects of speech production at different rates: evidence from spectral moments. *Journal of the Acoustical Society of America* 97, 520–530.
- Nolan, F. (1983). *The phonetic bases of speaker recognition*. Cambridge, London, New York: Cambridge University Press.
- Nolan, F. (2001). Speaker identification evidence: its forms, limitations and roles. In *Proceedings of the Conference 'Law and Language: Prospect and Retrospect'*, Levi (Finnish Lapland).
- Ohala, J. (1983). Cross-language use of pitch: An ethological view. *Phonetica* 40, 1–18.
- Pelz, H. (1996). *Linguistik – Eine Einführung*. Hamburg: Hoffmann und Campe.
- Rodman, R. (1998). Speaker recognition of disguised voices. In *Proceedings of the Consortium on Speech Technology Conference on Speaker Recognition by Man and Machine: Directions for Forensic Applications*, COST250 Publishing Arm, Ankara, Turkey, pp. 9–22.
- Shadle, C. H. und S. J. Mair (1996). Quantifying spectral characteristics of fricatives. In *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP)*, Philadelphia, pp. 1521–1524.
- Tabain, M. (2001). Variability in Fricative Production and Spectra: Implications for the Hyper- and Hypo- and Quantal Theories of Speech Production. *Language and Speech* 44(1), 57–94.
- Tosi, O. (1979). *Voice identification: Theory and legal application*. Baltimore: University Park Press.
- URL (1). *The EMU Speech Database System*. <http://emu.sourceforge.net>.
- URL (2). *The R-Project*. <http://www.r-project.org>.
- URL (3). *SAMPA*. <http://www.phon.ucl.ac.uk/home/sampa/home.htm>.
- Van den Heuvel, H. (1996). *Speaker variability in acoustic properties of dutch phoneme realisations*. Ph. D. thesis, Katholieke Universiteit Nijmegen.

- Wolf, J. J. (1972). Efficient acoustic parameters for speaker recognition. *Journal of the Acoustical Society of America* 51, 2044–2056.
- Wright, R., S. Frisch und D. B. Pisoni (1996). Speech perception. In *Research on Spoken Language Processing, Progress Report No. 21*, Bloomington, Indiana University.
- Zetterholm, E. (1997). Impersonation: A phonetic case study of the imitation of a voice. In *Working Papers 46*, Department of Linguistics, Lund University.

Anhang A

Abbildungen

A.1 Voruntersuchung: Frikativspektren

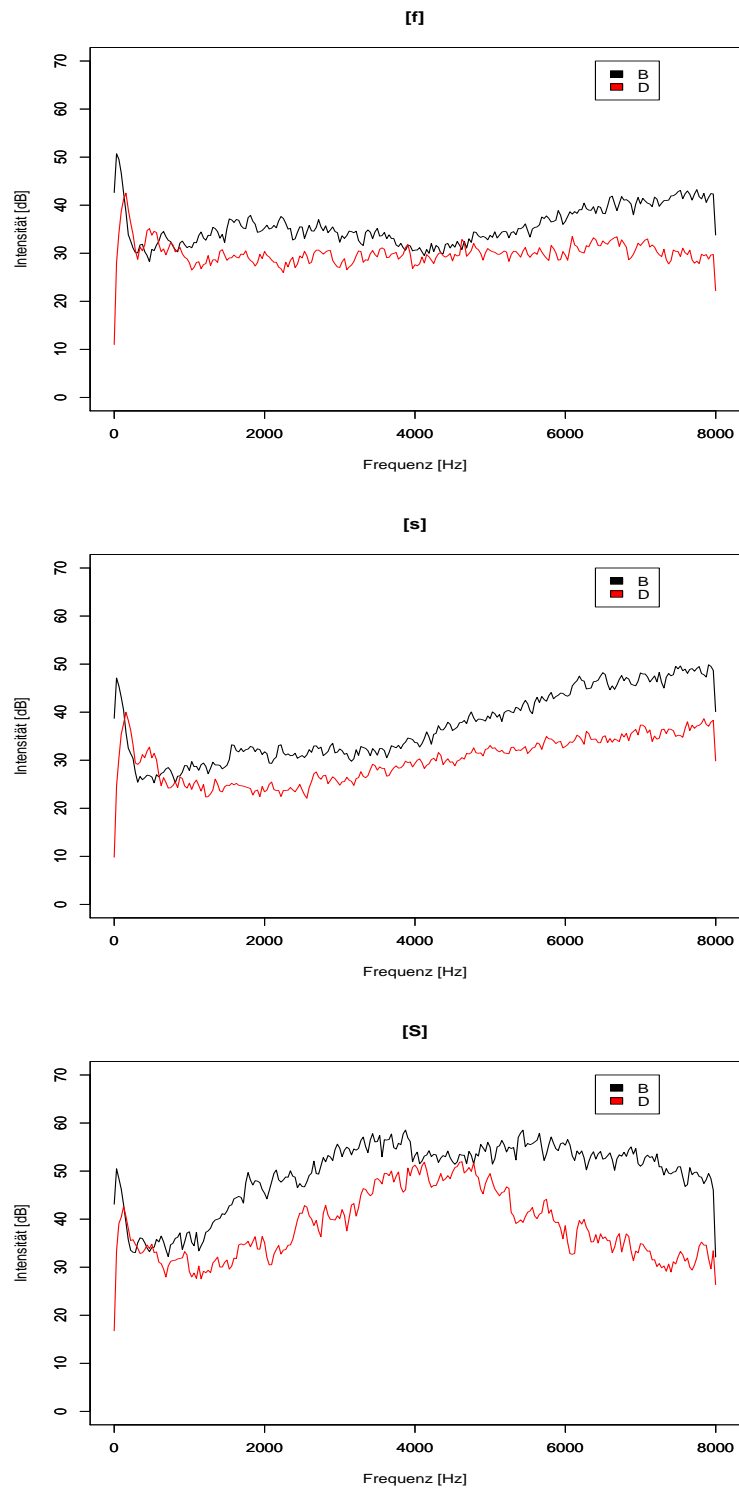


Abbildung A.1: Gemittelte Spektren über die Frikative [f, s, ʃ] für alle Stimmen von Sprecher B und D

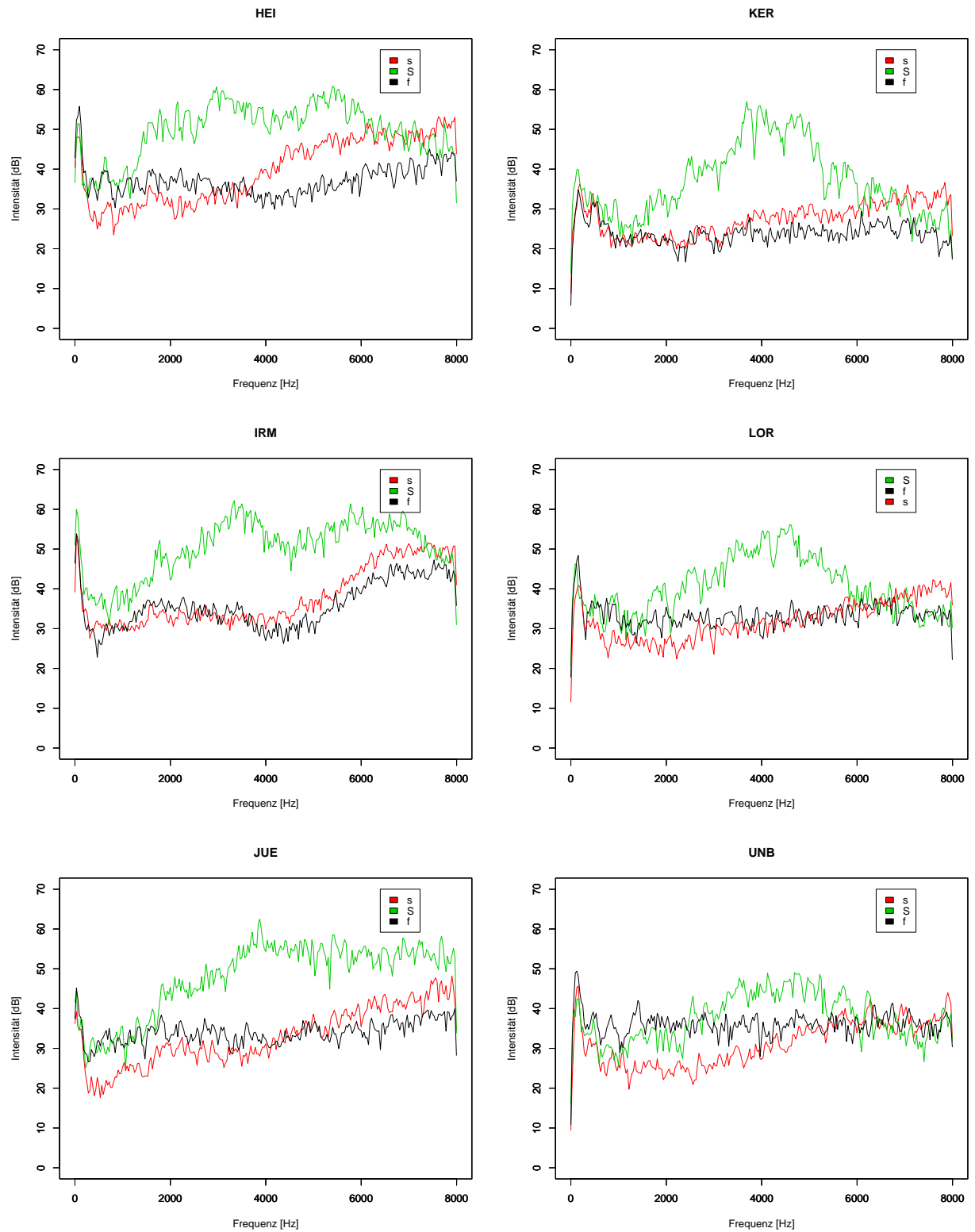


Abbildung A.2: Spektren der Frikative [f, s, ʃ] aller verstellten Stimmen von Sprecher B und D

A.2 Wahrscheinlichkeitsverteilungen

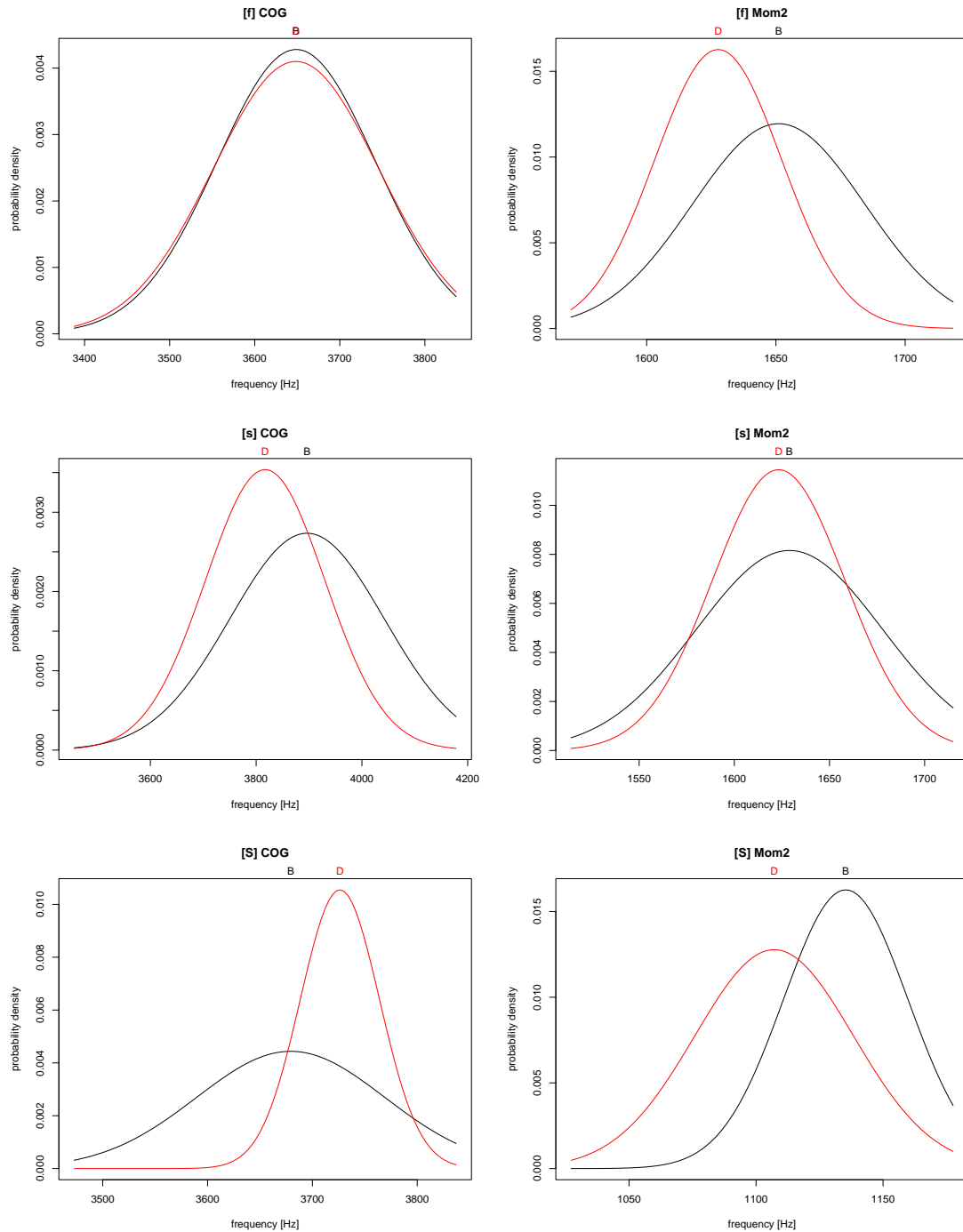


Abbildung A.3: Normalkurven von COG und Mom2 der einzelnen Stimmen für die Frikative [f, s, ʃ]

Anhang B

R-Scripts

Anmerkungen: Das Skript bezieht sich auf die Analysen am Beispiel von [s], entsprechende Vorgehensweisen für [f] und [ʃ]. Bei [ʃ] wurde der zu untersuchende Frequenzbereich in *moment()* variiert.

Zur Auswahl der Untersuchung von COG oder Mom2 wurde eine der Zeilen zu Beginn von Abschnitt B.2 auskommentiert.

B.1 Berechnung der Spektralmomente

```
#Segmentliste, Labels und Daten:
segs _ emu.query("Cmagi", "*" , "[Phonetic=s]")
speakers _ utt(segs)
speakers _ substring(speakers, 1, 2)
groups _ speakers
groups[groups %in% c("He", "Ju", "Ir")]_ "B"
groups[groups %in% c("Ke", "Lo", "Un")]_ "D"
dftvals _ emu.track(segs, "dft")
spec1 _ dcut(dftvals, 0.5, prop=T)
#Berechnung Spektralmomente, Zusammenbinden in Matrix:
m1 _ moment(spec1, samfreq=16000, low=1000, high=8000)
(# Bei [ʃ]: m1 _ moment(spec1, samfreq=16000, low=2000, high=7000) )
mom1 _ cbind(m1$first, m1$second)
sm1 _ cbind(mom1, speakers, groups)
```

B.2 Klassifikation

```

# Werte für COG
values _ as.numeric(sm1[,1])
# oder Werte für Mom2
values _ as.numeric(sm1[,2])
speakerlabs _ sm1[,4]
# Weist den Werten COG oder Mom2 „true“ oder „false“ zu → Auswahl als
Trainingsdaten („T“) oder Testdaten („F“)
logvec _ rep(c(T, F), length=nrow(sm1))
trainingdata _ values[logvec]
traininglabs _ speakerlabs[logvec]
testdata _ values[!logvec]
testlabs _ speakerlabs[!logvec]
# Trainingsphase (Erstellung Wahrscheinlichkeitsverteilungen und Entscheidungs-
grenzen)
tdat _ train(trainingdata, traininglabs)
# Testphase = Klassifikation
bdat _ classify(testdata, tdat)
# Erstellung Verwechslungsmatrizen
res _ confusion(testlabs, bdat)
#Proportionstest:
# Tatsächlich richtig klassifizierte Werte:
korrekt _ sum(diag(res))
# Theoretisch richtige Anzahl (=n/2)
theoretisch _ sum(apply(res, 1, sum)/2)
# Anzahl der gesamten Werte in der Matrix
n _ sum(res)
#Entscheidungsgrenze richtig vs. falsch klassifiziert
werte _ c(korrekt, theoretisch)
#Test:
prop.test(werte, c(n, n), alternative="greater")

```

B.3 ANOVA

```
#ANOVA 1 für COG:
#Zusammenfügen der zu untersuchenden Variablen (Sprecher, Stimmen, COG)
dat.df - data.frame(groups, speakers, m1$first)
#Varianzanalyse zwischen den verschiedenen Variablen:
summary(aov(m1$first ~ groups/speakers, dat.df))
#ANOVA 2 für die Unterschiede innerhalb des Sprechers B
which - groups=="B"
xB - m1$first[which]
xS - speakers[which]
dat.df - data.frame(xS ~ xB)
summary(aov(xB ~ xS, dat.df))
#ANOVA 3 für die Unterschiede innerhalb des Sprechers D
which - groups=="D"
xB - m1$first[which]
xS - speakers[which]
dat.df - data.frame(xS, xB)
summary(aov(xB ~ xS, dat.df))
```


Lebenslauf

Name: Ramona Lorenzen
Adresse: Schwarze Straße 6
24977 Langballig
Geburtsdatum: 27. November 1977
Geburtsort: Flensburg
Staatsangehörigkeit: Deutsch

Ausbildung:

1984-1988 Grund- und Hauptschule Sörup
1988-1997 Kurt-Tucholsky-Schule in Flensburg-Adelby
(Kooperative Gesamtschule, Zweig Gymnasium)
1997-1998 Studium Technik-Übersetzen an der FH Flensburg,
abgebrochen
1998-2004 Studium Phonetik und digitale Sprachverarbeitung,
Allgemeine Sprachwissenschaft und Psychologie
an der Christian-Albrechts-Universität Kiel

Akademische Lehrer:

Prof. Dr. K. J. Kohler (Phonetik und digitale Sprachverarbeitung)
Prof. Dr. J. M. Harrington (Phonetik und digitale Sprachverarbeitung)
PD Dr. A. P. Simspon (Phonetik und digitale Sprachverarbeitung)
Prof. Dr. U. Mosel (Allgemeine Sprachwissenschaft)
Dr. H. Fillbrandt (Psychologie)