Die Rolle phonetischer Information in der Sprechererkennung

Carola Schindler

2015

Die Rolle phonetischer Information in der Sprechererkennung

Carola Schindler

Inaugural-Dissertation
zur Erlangung des Doktorgrades
der Philosophie
an der Ludwig–Maximilians–Universität
München

vorgelegt von
Carola Schindler
aus Magdeburg
2015

Erstgutachter: Prof. Dr. Jonathan Harrington
Zweitgutachter: Prof. Dr. Phil Hoole
Tag der mündlichen Prüfung: 29.01.2016

Danksagung

Ich möchte mich ganz herzlich bei allen Menschen bedanken, die zu dieser Dissertation beitrugen, indem sie mich auf meinem Weg dahin unterstützten.
Ich danke Eva Reinisch von ganzem Herzen für ihr Engagement, ihre Begeisterung und ihre zahlreichen Ideen. Sie unterstützte mich stets mit konstruktivem Feedback und ihrer Erfahrung aus der psycholinguistischen Forschung. Ihr Wissen über die Methodik des Eye-Trackings sowie ihre Fähigkeiten im Bereich des wissenschaftlichen Schreibens befähigten mich zur Durchführung dieser Untersuchungen. Ich bin ihr sehr dankbar für ihre Geduld, Zeit und Energie, die sie mir schenkte.
Ich danke meinem Betreuer Jonathan Harrington für seinen Rat und seine Unterstützung. Er half mir mit seinen Kenntnissen über die Statistik bei der korrekten Auswertung meiner Daten. Außerdem wies er stets auf wichtige Aspekte des wissenschaftlichen Arbeitens und Schreibens hin.
Weiterhin möchte ich mich bei Christoph Draxler bedanken, der meine Arbeit in ihrer ersten Phase betreute und mich so gut wie möglich beriet.
Ich danke auch Phil Hoole, der als Zweitgutachter meine Arbeit gelesen und eingeschätzt hat.
Ein herzlicher Dank geht auch an Michael Jessen, der mir besonders zu Beginn meiner Arbeit mit Vorschlägen zu einem Forschungsthema und Tipps aus der Praxis der forensischen Phonetik zur Seite stand. Trotz seiner knapp bemessenen Zeit antwortete er mir immer ausführlich und motivierend auf meine E-Mails.
Ein großer Dank geht an den Sonderforschungsbereich 732 am Institut für maschinelle Sprachverarbeitung der Universität Stuttgart, der meine Dissertation ein Jahr lang finanziell und ideell mit einem Stipendium förderte. Ich danke vor allem Natalie Lewandowski, die mir von dieser Fördermöglichkeit erzählte und mich zu einer Bewerbung ermutigte. Außerdem danke ich Jagoda Bruni für ihre gute Organisation meiner Teilnahme an Veranstaltungen im Rahmen meines Stipendiums. Sie beantwortete stets schnell und freundlich meine zahlreichen organisatorischen Fragen. Ich bedanke mich auch bei Grzegorz Dogil für die Unterstützung meiner Bewerbung und seinen fachlichen Rat.
Ich bedanke mich auch bei all meinen Kollegen; besonders bei Raphael Winkelmann für seine Hilfe mit EMU, R und Latex; bei Uwe Reichel für seine Tipps in Perl; bei Thomas Kisler für seine Hilfe mit MAUS und bei Ulrich Reubold für seine Unterstützung beim Schreiben von Praat-Skripten.
Des Weiteren gilt mein Dank allen Teilnehmern des Doktoranden- und Post-Doktoranden-

Kolloquiums, die mir wertvolles Feedback zu meiner Forschung gaben.
Ich möchte auch allen Personen danken, die sich als Sprecher oder Probanden für meine Experimente zur Verfügung gestellt und meine teils langwierigen Untersuchungen mitgemacht haben.
Außerdem danke ich all meinen (Büro-)Kollegen für die schöne und fröhliche Zeit miteinander und das heitere Arbeitsklima.
Abschließend geht ein großes Dankeschön an meinen Ehemann Robert Schindler und meine Familie, die mich geduldig und aufmunternd auf meinem Weg begleiteten.

Abstract

Die gesprochene Sprache enthält neben den phonetischen bzw. lexikalischen Informationen, die den Inhalt einer Äußerung ausmachen, auch Informationen über den Sprecher. Beide Informationstypen interagieren miteinander, was dazu führt, dass manche Segmente mehr Informationen über einen Sprecher enthalten als andere und dass Wissen über den Sprecher dabei helfen kann, die phonetischen Informationen besser zu verarbeiten und somit eine Äußerung besser zu verstehen. Außerdem stellt sich die Frage, wie diese Informationen im Hinblick auf ein Sprachwahrnehmungsmodell (abstraktionistisch vs. exemplarbasiert) integriert werden.
Von diesem Stand ausgehend wird in dieser Arbeit der Einfluss der Segmente, insbesondere der Konsonanten, auf die Sprecherdiskrimination bzw. -identifikation untersucht. Dafür werden zunächst einige akustische Merkmale ausgewählter Konsonanten des Deutschen in einem Sprachkorpus analysiert. Es werden die ersten vier spektralen Momente der Laute gemessen und deren Sprecherspezifität bestimmt. Vor allem die Nasale /m/ und /n/ sowie die Frikative /f/ und /s/ offenbarten viele sprecherspezifische Merkmale.
Aufgrund der Annahme, dass sich diese akustisch gemessenen Merkmale auch perzeptiv in irgendeiner Form manifestieren müssen, wurde ein Sprecherdiskriminationsexperiment mit Hörern durchgeführt. In beiden Experimenten war das Sprachmaterial eine /aKa/-Sequenz. Im ersten Experiment enthielt der gesamte Stimulus Sprecherinformationen, während im zweiten Experiment nur der (statische Teil vom) Konsonant, aber nicht die Vokaletransitionen Sprecherinformationen enthielt. In beiden Untersuchungen zeigen sich Unterschiede in der Sprecherspezifität zwischen den verschiedenen Artikulationsmodi und -stellen, wobei die durchschnittliche Sprecherdiskriminationsrate im zweiten Experiment deutlich geringer ist als im ersten. Die Ergebnisse lassen darauf schließen, dass Nasale und Plosive viele ihrer Informationen in den Vokaltransitionen enthalten, während die Frikative mehr Informationen im (statischen Bereich des) Konsonanten besitzen.
Da die phonetischen und Sprecherinformationen miteinander interagieren, wurde im letzten Teil der Arbeit die zeitliche Koordination der Verarbeitung beider Informationstypen mittels eines Visual-World Eye-Tracking Experiments untersucht. Die Ergebnisse zeigen, dass die Hörer das Target mit großer Sicherheit identifizierten, aber dass mit steigender Anzahl an Sprechern (2 vs. 4 Sprecher) die Schwierigkeit der Targetidentifikation steigt. Im Fall von verschieden geschlechtlichen Sprechern wird zuerst das Geschlecht und dann der einzelne Sprecher erkannt. Außerdem wird nachgewiesen, dass die Sprecherinformationen tendenziell sogar früher verarbeitet werden als die phonetischen Informationen und selbst dann Verwendung finden, wenn phonetische Informationen allein zur Targetidentifikation ausreichend sind. In phonetisch ambigen Fällen werden die Sprecherinformationen verwendet, um diese Ambiguität zu verringern. Die Ergebnisse unterstreichen die Bedeutung von Sprecherinformationen in der Verarbeitung gesprochener Sprache und sprechen somit eher für ein episodisches, exemplarbasiertes Modell der Sprachwahrnehmung, welches Sprecherinformationen bereits zu einem frühen Zeitpunkt im Sprachverarbeitungsprozess integriert.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1. Einleitung

Wenn wir eine Person sprechen hören, sind wir in der Lage, nicht nur den Inhalt des Gesagten wahrzunehmen, sondern auch den spezifischen Stimmklang dieser Person. Da jeder Mensch sich sowohl durch anatomisch-physiologische als auch durch erlernte, verhaltensbasierte Faktoren in seiner Sprechweise unterscheidet, sind auch die Sprachlaute, die er produziert unterschiedlich. Daher sind wir oft in der Lage, Personen aufgrund ihrer Stimme zu erkennen und zu unterscheiden.
Die verschiedenen Merkmale einer Stimme variieren auf zwei verschiedene Weisen: Zum einen variieren sie zwischen unterschiedlichen Sprechern (Inter-Sprecher-Variation) und zum anderen variieren sie auch zwischen verschiedenen Äußerungen desselben Sprechers (Intra-Sprecher-Variation). Je stärker ein Stimmmerkmal zwischen Sprechern und je weniger es innerhalb eines Sprechers variiert, desto spezifischer ist es für den jeweiligen Sprecher und desto besser eignet es sich potenziell, um Sprecher voneinander zu unterscheiden. Das Sprechermerkmal kann dabei entweder ein akustisch messbares Merkmal sein, wie z.B. die Frequenz eines Vokalformanten, oder ein auditiv wahrnehmbares wie beispielsweise die Sprechtonhöhe oder die Stimmqualität eines Sprechers.
Es gibt viele Studien, welche die vokalischen Merkmale mit guten Ergebnissen auf ihre Sprecherspezifität untersucht haben (z.B. [177] [186] [187] [255]). Auch wenn diese sehr viele relevante Sprecherinformationen liefern können, so sind sie nicht unter allen Bedingungen - zum Beispiel im Fall von Telefonsprache - voll belastbar ([159] [42] [110]). Da die Konsonanten außerdem einen großen Anteil in unserer Sprache ausmachen (nämlich ca. 62 % [194]), wäre es gewinnbringend, das Sprecherdiskriminationspotenzial ihrer Merkmale eingehender zu analysieren. Einige Studien belegten bereits ihre sprecherspezifischen Eigenschaften (z.B. [101] [295] [16] [12] [142] [72]). Auf dieser Grundlage stellen die Sprecherinformationen der Konsonanten eine sinnvolle Ergänzung und Erweiterung der traditionell verwendeten Vokalmerkmale dar. Neben dem Fokus auf die Vokale besteht außerdem eine Konzentration auf die Englische Sprache bei der Suche nach sprecherspezifischen Merkmalen. Für die deutsche Sprache gab es bisher allerdings nur wenige Untersuchungen (z.B. [76] [179]).

Auch wenn sich einige Konsonanten ähneln, so unterscheiden sie sich doch in ihrer genauen Artikulation [25] und manche existieren im Englischen gar nicht (z.B. die Frikative /ç, x/). Daher lohnt sich ein genauerer Blick auf die lautspezifischen Sprechermerkmale der deutschen Sprache. Dafür sollen ausgewählte Konsonanten akustisch und perzeptiv auf ihre Sprecherspezifität untersucht werden. Neben dem Einfluss einer Telefonübertragung soll auch die Interaktion von phonetischen und Sprechermerkmalen betrachtet werden und ihrer zeitliche Koordination in der Sprachwahrnehmung.
Im Folgenden soll dargelegt werden, für welche Forschungs- und Anwendungsbereiche die gewonnen Erkenntnisse über Sprecherinformationen genutzt werden können.

1.1. Motivation

Merkmale, welche Informationen über die Identität eines Sprechers liefern, sind in verschiedenen Bereichen der (phonetischen) Forschung und Anwendung relevant. Die forensische Phonetik beispielsweise verwendet die stimmlichen Besonderheiten der Sprecher, um zu überprüfen, ob die Stimme eines Verdächtigen, mit der eines Täters übereinstimmt. Im Bereich der automatischen Sprecherverifikation wird anhand stimmlicher Merkmale der Anspruch auf eine bestimmte Identität eines Sprechers überprüft. Auf der anderen Seite wird in der Sprachwahrnemung untersucht, inwieweit die Wahrnehmung des Inhalts der Sprache mit den stimmlichen Merkmalen des Sprechers interagiert. Dies soll Aufschluss darüber geben, wie Sprache in unserem Gehirn repräsentiert ist und verarbeitet wird.
Für alle diese Forschungs-/Anwendungsbereiche sind sprecherspezifische Merkmale von Bedeutung. Daher sollen in dieser Arbeit Sprechermerkmale von verschiedenen Perspektiven und Anwendungsmöglichkeiten aus beleuchtet werden.

1.1.1. Sprechermerkmale in der Forensischen Phonetik

Die forensische Phonetik befasst sich vor allem mit Stimmanalysen und -vergleichen, um bei der Ermittlung von Straftätern in Kriminalfällen zu helfen. Dazu werden Sprachaufnahmen des Täters mit Sprachaufnahmen von Verdächtigen verglichen, um herauszufinden, ob beide Äußerungen von der gleichen Person gesprochen wurden oder von zwei verschiedenen. Für diesen Vergleich können verschiedene Merkmale herangezogen werden. Klassischerweise werden akustische und perzeptive Merkmale für den Vergleich verwendet. Bei der perzeptiven Analyse hören sich geschulte Phonetiker die zu vergleichenden Sprachaufnahmen an und beurteilen mit Hilfe von Skalen verschiedene Merkmale der Sprecher [98]. Durch

die Auswertung der Skalenwerte lässt sich anschließend ein Wert ermitteln, der aussagt, mit welcher Wahrscheinlichkeit es sich um den selben oder um verschiedene Sprecher handelt. Diese Einschätzung ist, bei allen Objektivierungsversuchen durch einheitliche Skalen, natürlich sehr subjektiv. Bei der akustischen Analyse werden mit Hilfe von spezieller Software (z.B. Praat) akustische Merkmale wie beispielsweise die Vokalformanten gemessen. Deren Frequenzen können bestimmt und mit den Werten anderer Sprechern vergleichen werden. Dieses Vorgehen ist objektiver als der Höreindruck einer Person, aber aufgrund von qualitativ schlechten Daten besteht die Gefahr von Messfehlern ([159] [42]).
Damit ein Sprechermerkmal einen Nutzen für die forensische Phonetik hat, muss die Inter-Sprecher-Variation gegenüber der Intra-Sprecher-Variation möglichst große sein. Die Aufgabe der Sprechererkennung ist es herauszufinden, ob die Unterschiede zwischen den Stimmproben wahrscheinlicher auf Inter- oder Intra-personeller Variation beruhen. Dabei gibt es aber fast kein Merkmal, das allein alle Sprecher diskriminieren könnte. Da es immer auch vom Sprecher abhängt, durch welches Merkmal er sich von einem anderen unterscheidet, benötigt man zur Diskriminierung mehrerer Sprecher auch mehrere Merkmale [254].

1.1.2. Sprechermerkmale zur automatischen Sprecheridentifikation- und verifikation

In der automatischen Sprecheridentifikation - und verifikation werden akustische Merkmale verwendet, die maschinell im Sprachsignal gemessen und extrahiert werden können. Oft werden dafür so genannte low-level akustische Merkmale verwendet, da sie einfach zu extrahieren, textunabhängig und einfach zu modellieren sind und eine geringe Datenmenge ausreicht. Der Nachteil besteht allerdings darin, dass sie sehr anfällig für Hintergrund- und Störgeräusche sind. Robuster wären prosodische oder lexikalische Merkmale höherer Ebenen, welche aber wesentlich aufwändiger zu verarbeiten sind [148]. Zu den am häufigsten verwendeten Merkmalen gehören die Mel-Frequency Cepstral Coefficients (MFCCs). Die einzelnen Koeffizienten geben jeweils über eine Eigenschaft des Cepstrums eines Sprechers Aufschluss. Diese Merkmale werden aus dem Sprachsignal extrahiert und zum Training der einzelnen Sprechermodelle verwendet ([148] [149]).
Sprecheridentifikation und -verifikation unterscheiden sich hinsichtlich der Fragestellung und des Entscheidungsfindungsprozesses. Bei der Verifikation gibt ein Sprecher eine bestimmte Identität vor, mit der dann seine Sprechprobe verglichen wird. Bei der Identifikation hingegen wird ein Sprachbeispiel von einem unbekannten Sprecher mit beliebig vielen anderen Sprecheridentitäten verglichen. Für die Entscheidungsfindung muss bei der Verifikation

nur ein Mustervergleich mit dem Muster der vorgegebenen Identität durchgeführt werden, während für die Identifikation ein Vergleich mit den Mustern aller (bekannten) Sprecher durchgeführt werden muss [249].
Der Bereich der automatischen Sprecherverifikation wird in dieser Arbeit aber nur insofern gestreift, als dass die untersuchten akustischen Merkmale auch für dieses Verfahren in Frage kämen.

1.1.3. Sprechermerkmale in der Sprachwahrnehmung

Die Sprachwahrnehmungsforschung beschäftigt sich mit dem Prozess der menschlichen Wahrnehmung und Verarbeitung von gesprochener Sprache. Eine gesprochene Äußerung enthält neben den inhaltlichen Informationen auch immer Informationen über die Stimme des Sprechers. Beide Informationstypen interagieren miteinander und können die Sprachwahrnehmung unterstützen oder behindern ([102] [103] [259]). Die Weise auf welche sich Sprechermerkmale auf die Wahrnehmung von Sprache auswirken, kann Aufschluss geben über die Art der Repräsentation von Sprache in unserem Gehirn und wie sie verarbeitet wird. Werden beispielsweise Worte mit all ihren spezifischen akustischen Unterschieden, die durch die verschiedenen Stimmen der Sprecher entstehen, abgespeichert, so wie es exemplartheoretische Modelle vorhersagen ([133] [103] [232] [135])? Oder werden akustische Details auf einem prälexikalischen Level herausgefiltert, sodass nur die reinen Wortinformationen übrig bleiben, wie es die abstraktionistischen Modelle behaupten ([184] [212] [213] [107])? Eine Antwort auf diese Frage versprechen Untersuchungen zur Koordination der Wahrnehmung und Verarbeitung von inhaltlichen und stimmlichen Informationen. Im dritten Experimentkapitel dieser Arbeit soll daher auf die zeitliche Koordination der Verarbeitung von phonetischen und Sprecherinformationen eingegangen werden.

1.2. Struktur der Dissertation

In dieser Arbeit sollen die stimmlichen Besonderheiten von Sprechern unter verschiedenen Blickwinkeln betrachtet werden. Dabei soll herausgearbeitet werden, welchen Nutzen wir aus sprecherspezifischen Merkmalen ziehen können; welches Erkenntnispotenzial sie besitzen und für welche Forschungs- und Anwendungsbereiche sie relevant sind.
In **Kapitel 2** soll zunächst dargestellt werden, durch welche anatomisch-physiologischen Faktoren Sprecher unterschiedliche Stimmmerkmale ausprägen. Außerdem sollen verschiedene Faktoren, wie Alter, Geschlecht, emotionaler Zustand, ect. und deren Auswirkungen

auf die Stimme eines Sprechers betrachtet werden. Die so entstehende sprecherspezifische Artikulation manifestiert sich ebenfalls im akustischen Sprachsignal, welches wir hören und messen können. Dieses Sprachsignal enthält auf unterschiedlichen Ebenen, die bei der Artikulation entstandenen sprecherspezifischen Merkmale. So findet man sowohl auf der globalen (oder auch suprasegmentalen) Ebene als auch in einzelnen Lauten (auf der segmentalen Ebene) Besonderheiten der Sprecher. Der Fokus dieser Arbeit wird auf den segmentalen Merkmalen liegen und darin besonders auf den Eigenschaften bestimmter Konsonanten, deren Charakteristika zuvor eingehend erläutert werden. Die akustischen Merkmale im Sprachsignal werden von Hörern auditiv wahrgenommen und verarbeitet. Daher stellt sich die Frage, wie mit den Sprecherinformationen im Signal bei der Sprachwahrnehmung verfahren wird. Dafür sollen zwei unterschiedliche theoretische Ansätze vorgestellt werden: der abstraktionistische und der episodische, exemplarbasierte.
Nach der theoretischen Einführung über die Entstehung, Beeinflussung und Wahrnehmung von sprecherspezifischen Merkmalen, soll in **Kapitel 3** untersucht werden, welche Merkmale im akustischen Sprachsignal ein hohe Sprecherspezifität aufweisen und sich somit potenziell zur Unterscheidung mehrerer Sprecher eignen. Der Fokus liegt dabei auf konsonantischen Merkmalen, deren Sprecherspezifität gemessen und mit der einiger vokalischen Merkmalen verglichen wird. Außerdem soll betrachtet werden, welchen Einfluss die Qualität des Sprachsignals auf die Sprecherspezifität der einzelnen Konsonanten hat. Dazu werden die Ergebnisse von über Mikrophon aufgenommenen Sprachaufnahmen mit über Telefon aufgenommenen verglichen.
Nachdem die Sprecherspezifität der akustischen Merkmale nachgewiesen wurde, soll in **Kapitel 4** die perzeptive Diskriminationsfähigkeit von menschlichen Hörern aufgrund konsonantischer Sprechermerkmale untersucht werden. Dazu soll ein Diskriminationstest durchgeführt werden, bei dem Hörer anhand von kurzen /aKa/-Stimuli Sprecher unterscheiden sollen. Dabei werden zwei Bedingungen getestet: einmal sind statische und dynamische Merkmale im Stimulus vorhanden und einmal nur statische (und keine dynamischen). Dank dieser Manipulation können neue Erkenntnisse über die Lokalisation von Sprecherinformationen im statischen Bereich des Konsonanten bzw. seinen dynamischen Transitionen zu den benachbarten Vokalen gewonnen werden.
In **Kapitel 5** soll mittels eines Visual-World Eye-Tracking Experiments analysiert werden, in welchem zeitlichen Zusammenhang die Verarbeitung von Sprecherinformationen zur Verarbeitung von sprachlich-inhaltlichen Informationen steht. Außerdem wird der Einfluss von Sprecheranzahl, Sprechergeschlecht und der stimmlichen Ähnlichkeit der Sprecher auf

den Identifikationsprozess untersucht. Auch die Auswirkung der phonetischen Ambiguität und der phonetischen Identität des wortinitialen Konsonanten soll getestet und ausgewertet werden.
In dem abschließenden **Kapitel 6** werden dann die neu gewonnen Erkenntnisse kritisch diskutiert, zusammengefasst und mögliche zukünftige Forschungsfragen umrissen.

2. Theoretische Grundlagen

2.1. Grundlagen der Phonation und Artikulation

Als Stimme werden die von den Stimmlippen produzierten Laute/Klänge bezeichnet, welche durch verschiedene Merkmale, wie Grundfrequenz, Lautstärke und Stimmqualität gekennzeichnet werden. Die Anatomie des Vokaltrakts sowie die Artikulationsweise eines Menschen formen gemeinsam das für eine Person spezifische Sprachsignal. Somit besteht das Sprachsignal immer aus anatomisch-bestimmten Merkmalen als auch aus verhaltensbasierten Merkmalen. Um zu verdeutlichen, wie diese beiden Faktoren auf die Stimme wirken, soll kurz der Aufbau und die Funktionsweise des Kehlkopfes und des Vokaltrakts erläutert werden.

2.1.1. Anatomie und Physiologie des Kehlkopfs

Der Kehlkopf (Larynx) bildet die Grenze zwischen der Luftröhre und dem Rachen- und Mundraum. Er stellt ein komplexes Ventil dar, dessen primäre Aufgabe es ist, das Eindringen von Nahrung in die Luftröhre zu verhindern. Seine Funktion für lautsprachliche Äußerungen besteht in der kontrollierten Stimmtonerzeugung - der sogenannten Phonation. Das Zusammenspiel der Kehlkopfmuskulatur mit dem Druck der ausgeatmeten Luft versetzt die Stimmlippen in Schwingung. Das Auftreten, die Geschwindigkeit, die Stärke und die Form der Schwingung bestimmen die Stimmhaftigkeit, die Stimmtonhöhe, die Lautstärke und die Stimmqualität.
Der Kehlkopf besteht hauptsächlich aus Knorpeln und Muskeln. Die drei wichtigsten Knorpel (in der Reihenfolge von unten nach oben) sind der Ringknorpel (Cricoid), der Stellknorpel (Arytenoid) und der Schildknorpel (Thyroid). Verbunden werden diese Knorpel durch drei verschiedene Muskeln (Muskelgruppen), den Cricoarytenoideus, den Thyroarytenoideus und den Cricothyroideus. Zwischen dem Schild- und dem Stellknorpel sind die Stimmlippen gespannt. Der Raum zwischen den Stimmlippen heißt Glottis.
Die Erzeugung des Stimmtons erfolgt durch muskuläre und elastische Kräfte der Stimmlip-

pen und durch aerodynamische Kräfte, ausgelöst von der durch die Glottis strömenden Luft. Die Kräfte wirken folgendermaßen: Am Anfang ist die Glottis geschlossen und die Stimmlippen sind gespannt. Durch den ansteigenden subglottalen Luftdruck wird die Glottis gesprengt und Luft beginnt durch den Spalt zu strömen. Aufgrund der engen Durchflussöffnung der Glottis wird die Fließgeschwindigkeit der Luft erhöht. Dies erzeugt, die senkrecht zur Fließrichtung wirkenden Bernoulli-Kräfte, welche durch ihre Sogwirkung die elastischen Stimmlippen wieder verschließen. Danach beginnt der gesamte Prozess wieder von vorn.
Wie schnell die Stimmlippen schwingen, ist maßgeblich von ihrer Länge abhängig. Je länger die Stimmlippen sind, desto langsamer schwingen sie und umgekehrt. Da Männer meist längere Stimmlippen (17-24 mm) haben als Frauen (13-17 mm), besitzen sie auch eine niedrigere Schwingungsrate der Stimmlippen und somit eine tiefere Grundfrequenz. Die Grundfrequenz für Männer liegt bei durchschnittlich 120 Hz, die der Frauen bei 230 Hz. Zusätzlich lässt sich die Schwingungsrate durch die muskuläre Einstellung der Stimmlippen und die Stärke des Ausatemdrucks variieren. So führt eine stärkere Spannung und/oder ein höherer Ausatemdruck der Stimmlippen zu einer erhöhten Schwingungsrate und somit zu einer höheren Grundfrequenz [235].
Auch die Qualität der Stimme, wird bei der Phonation im Kehlkopf bestimmt. Bei der Phonation entstehen Stimmimpulse, die von der Kehlkopfmuskulatur und dem Druck der ausgeatmeten Luft beeinflusst werden. Ob die Stimmlippen schwingen, wie stark, mit welcher Geschwindigkeit und in welcher Form bestimmt die Stimmhaftigkeit, die Lautstärke, die Tonhöhe und die Qualität der Stimme [239]. Die Stimmqualität umfasst alle charakteristischen Eigenschaften, die während des Sprechakts einer Person immer präsent sind. Dazu zählen alle laryngalen und supralaryngalen Merkmale, die bei der Gestaltung der Stimme mitwirken. Der Stimmqualität können verschiedene Funktionen zukommen. In manchen Sprachen dient sie der Bedeutungsunterscheidung. Im Allgemeinen aber gibt sie dem Hörer Aufschluss über physikalische, psychologische und soziale Merkmale des Sprechers. Es gibt neben der „normalen Stimme“ grob drei verschiedene Phonationstypen: die Knarrstimme, die behauchte Stimme und das Falsett. Die Knarrstimme wird durch sehr langsame und unregelmäßige Schwingungen der Stimmlippen gekennzeichnet, die durch ein starkes Zusammenpressen der Stimmlippen hervorgerufen werden. Werden die Stimmlippen dagegen sehr schwach zusammengepresst, strömt sehr viel Luft durch die Glottis und es entsteht die behauchte Stimme. Bei der Falsett-Stimme werden die Stimmlippen längst stark gedehnt, sodass sie dünner werden und dadurch schneller schwingen. Durch das schnelle Schwingen entsteht die für die Falsett-Stimme typische hohe Grundfrequenz [170].

Die Parameter des Kehlkopfes und der Glottis bestimmen die Phonation einer Äußerung. Da jede Äußerung mit der Phonation beginnt, wird das durch die Glottis erzeugte Geräusch auch als Quellsignal bezeichnet. Das Quellsignal breitet sich dann in Richtung Vokaltrakt aus, in dem es weiter modifiziert wird.

2.1.2. Anatomie und Physiologie des Vokaltrakts

Im Vokaltrakt wird das von der Glottis erzeugte Quellsignal durch die Artikulation modifiziert. Der Vokaltrakt liegt oberhalb der Glottis und umfasst den Luftraum, der durch Mund- und Nasenöffnung begrenzt ist. Dieser beinhaltet Artikulatoren (bewegliche Teile) und Artikulationsstellen (unbewegliche Teile), wobei davon ausgegangen wird, dass für eine Artikulationsstelle immer das gegenüberliegende artikulierende Organ verwendet wird. Zu den Artikulatoren zählen: die Zunge, die Lippen, der Unterkiefer, das Gaumensegel, das Zäpfchen, sowie in eingeschränkter Form der Rachen und die Glottis. Die Artikulationsstellen bestehen aus: Oberlippe, Schneidezähne, Alveolen, Palatum, Velum, Uvula, Pharynx und Glottis.
Neben dem Ort der Artikulation spielt auch der Artikulationsmodus eine entscheidene Rolle. Je nachdem, ob der Artikulator an einer Artikulationsstelle einen vollständigen oder teilweisen Verschluss oder nur eine Enge bildet, entstehen verschiedene Laute. Durch globale Veränderungen im Vokaltrakt, wie Kieferöffnung, Lippenvorstülpung oder -spreizung und die Lage und Höhe der Zunge entstehen Vokale. Wird eine lokale Enge gebildet, durch die die hindurchströmende Luft turbulent wird, bilden sich Frikative. Plosive werden durch einen vollständigen Verschluss an einer Artikulationsstelle gebildet, teilweise Verschlüsse rufen Laterale hervor und intermettierende Verschlüsse führen zu Trills. Durch ein Absenken des Velums können zusätzlich nasale Laute erzeugt werden [235].

2.1.3. Sprechermerkmale und ihre Einflussfaktoren

Die eingangs beschriebene Anatomie und Physiologie der menschlichen Stimme ist zwar prinzipiell für alle Menschen gleich, allerdings gibt es natürlich Unterschiede in den genauen Eigenschaften. Länge und Größe von Stimmlippen und Vokaltrakt können variieren, ebenso wie die Form des Velums und der Alveolen. Stimmliche Unterschiede, die aufgrund dieser Faktoren entstehen, bezeichnet man als anatomisch-physiologische Merkmale, da sie durch anatomische und physiologische Eigenschaften des Menschen bestimmt werden. Neben diesen Merkmalen gibt es aber auch die verhaltensbasierten, welche durch das individuelle

Sprechverhalten einer Person entstehen, wie zum Beispiel die Sprechgeschwindigkeit oder bestimmte Stimmqualitäten. Oft beeinflussen beide Faktoren, sowohl die physiologischen als auch die verhaltensbasierten, die stimmlichen Merkmale eines Sprechers. Auch wenn beispielsweise ein Großteil der Grundfrequenz durch die Länge der Stimmlippen bestimmt wird ([235] [106]), so gibt es auch immer noch kulturelle und/oder individuelle Merkmale, die ihre Lage und Variation ebenfalls beeinflussen ([228] [145]). Durch dieses Zusammenspiel der beiden Faktoren Physiologie und Verhalten kann oft nicht genau bestimmt werden, zu welchem Anteil ein Stimmmerkmal physiologisch oder verhaltensbasiert ist. Einige von diesen Faktoren, welche die Stimmmerkmale eines Sprechers beeinflussen, sollen hier im Folgenden diskutiert werden.

Alter

Kleinkinder (33-169 Wochen alt) besitzen eine wesentlich höhere Grundfrequenz als Erwachsene. Außerdem variiert diese stärker und weist einen größeren Frequenzbereich auf (400-1150 Hz) [144]. Auch weisen Kinderstimmen noch keine geschlechtsspezifischen Unterschiede auf [87]. Der Vokaltrakt von Kindern ist kleiner und unterscheidet sich anatomisch von dem Erwachsener. Zum einen liegt der Kehlkopf noch weiter oben und zum anderen schließt das Velum die Luftröhre ab, sodass Kinder in der Lage sind gleichzeitig zu schlucken und zu atmen. Aufgrund dieser anatomischen Differenzen können Kinder nicht genauso artikulieren wie Erwachsene. Allerdings sind sie dennoch in der Lage, akustisch gleich klingende Vokale zu produzieren, wie eine französische Studie [195].
Wenn Kinder wachsen, wächst auch ihr Vokaltrakt mit, sodass Jugendliche schon einen wesentlich größeren Vokaltrakt besitzen. Die deutlichsten Veränderungen macht die männliche Stimme durch. Ab ungefähr 13 Jahren beginnt die Grundfrequenz stark zu fallen und auch die Formanten sinken ab ([119] [200]). Am stärksten wächst bei den männlichen Jugendlichen während der Pubertät der pharyngale Abschnitt des Vokaltrakts [87]. Bei Mädchen ist eher ein gleichmäßiger Abfall der Grundfrequenz festzustellen, der ungefähr mit 13 Jahren beginnt und mit 19 beendet ist [65]. Ab einem Alter von ca. 15 Jahren kann man einen signifikanten Unterschied in der Grundfrequenz zwischen den Geschlechtern wahrnehmen.
Ist ein Mensch ausgewachsen, erreicht er die maximale Größe seines Vokaltrakts. Seine Grundfrequenz verändert sich nicht mehr bzw. nur sehr langsam. Je größer ein Mensch ist, desto größer ist auch sein Vokaltrakt und je länger seine Stimmlippen, wobei die akustischen Effekte der beiden Größen voneinander unabhängig sind [87]. Außerdem erreicht

man im Erwachsenenalter seine maximale Sprechrate. Man spricht mit weniger Pausen als Jugendliche oder ältere Menschen [200].
Mit zunehmendem Alter verändert sich die Stimme weiter. Zwar langsamer als im Jugendalter, aber die Veränderung hält an. So beobachteten [233], dass die Grundfrequenz bei Frauen mit zunehmendem Alter erst fällt und später wieder steigt. Auch bei männlichen Sprechern ist häufig dieser Effekt zu beobachten [248]. Außerdem weisen die Stimmen älterer Menschen mehr Jitter und Shimmer auf und der Stimmumfang sowie die Harmonizität werden geringer [200].

Geschlecht

Eines der augenfälligsten Merkmale eines Sprechers ist sein Geschlecht. Wir sind (fast) immer in der Lage männliche und weibliche Sprecher zu erkennen und zu unterscheiden. Typische Grundfrequenzwerte für Frauen liegen laut [285] bei 210 Hz und für Männer bei 120 Hz, wobei sich diese Werte, wie in schon im vorangegangenen Abschnitt beschrieben, mit dem Alter ändern können. Während bei Frauen die Grundfrequenz bis zur Menopause stabil bleibt und erst danach absinkt ([276] [226]), sinkt sie bei Männern zunächst (am stärksten während der Pubertät) bis ins Alter von ungefähr 35 Jahren, bleibt dann konstant und beginnt ab 55 Jahren wieder zu steigen ([120] [226]). Diese Effekte werden durch das Verhältnis der Hormone Östrogen und Testosteron bestimmt, welches sich im Laufe des Lebens verändert. Eine interessante Feststellung von [285] ist, dass Männer eine höhere Variation in der Grundfrequenz von Sprecher zu Sprecher (Inter-Sprecher-Variation) aufweisen als Frauen. Als mögliche Erklärung führen sie an, dass bei Männern zwei Faktoren wirken, die Unterschiede hervorrufen können und bei Frauen nur einer. Bei Männern sowie Frauen wirkt sich die Körpergröße auf die Grundfrequenz aus. Allerdings kommt bei Männern auch der Testosteronpegel zum Tragen, da dieser beeinflusst, wie stark das Wachstum des laryngalen Teils des Vokaltrakts während der Pubertät ist. Dadurch gibt es bei Männern zwei Quellen für Variabilität, was die Variabilität insgesamt erhöht [285].
Nun wird die Grundfrequenz aber nicht nur durch physiologische Merkmale geprägt, sondern auch durch verhaltensbasierte. Zum Beispiel bestimmt die jeweilige Kultur und Sprache sehr stark, was als männliche und was als weibliche Stimme wahrgenommen wird. So zeigten [145], dass männliche Mandarin-Sprecher eine höhere mittlere und maximale Grundfrequenz sowie einen größeren Grundfrequenzbereich aufwiesen als männliche Englisch-Sprecher. Obwohl sich die Stimmen der Männer physiologisch nicht unterschieden, sprachen die Mandarin-sprechenden Männer in Spontan- und Lesesprache stets höher als die englischen Sprecher.

Dieser Unterschied muss also von sprachlichen und/oder kulturellen Besonderheiten hervorgerufen worden sein. In einer anderen Studie von [228] wurden zweisilbige Wörter von amerikanischen und französischen Männern und Frauen verglichen. Während französische Frauen mit einer hohen Grundfrequenzvariation sprachen, hatten die amerikanischen Frauen einen geringeren F0-Bereich. Beide Frauengruppen wiesen eine behauchte Stimme auf, während nur die amerikanischen Männer mit einer gepressten Stimme sprachen, nicht aber die französischen. Dies deutet darauf hin, dass die behauchte Phonation der Frauen physiologisch bedingt ist, nicht aber die gepresste Stimme der amerikanischen Männer. [180] demonstrierten in ihrer Untersuchung, dass japanische Frauen mit einer sehr hohen Grundfrequenz sprechen, um besonders höflich zu sein, während bei englischen Sprechern beide Geschlechter eine höhere Grundfrequenz für höfliche Äußerungen nutzten.
Diese Studien deuten an, von welch vielfältigen Faktoren die Grundfrequenz eines Sprechers abhängen kann und dass die Unterschiede zwischen Männern und Frauen, mehr Ursachen haben als nur die anatomisch-physiologischen.

Sprechsituation

Die Sprechsituation bezeichnet den sozialen Kontext einer Interaktion und wird von verschiedenen situativen Faktoren beeinflusst, wie beispielsweise der sozialen Beziehung der Gesprächspartner, dem Ort, der Gesprächsabsicht und dem Thema. Die Sprechsituation, in der sich ein Sprecher gerade befindet, kann sich auf seine Stimme und Sprechweise auswirken. So sprechen Menschen zum Beispiel in formellen (z.B. geschäftlichen) Situationen eher monoton mit einer geringen Grundfrequenzvariation [242]. Je informeller eine Situation ist, desto häufiger kommt es dagegen zur Reduktion von Lauten [74]. Andererseits kann besondere Höflichkeit in manchen Sprachen durch eine höhere Grundfrequenz ausgedrückt werden [180].

Emotionen

Während Alter und Geschlecht eines Sprechers eher einen langfristigen Einfluss auf die Sprechermerkmale haben, können Emotionen temporäre Veränderungen erzeugen. Emotional deprimierte, traurige oder beschämte Menschen sprechen mit einer sehr geringen Grundfrequenzvariation. Wohingegen eine erhöhte F0-Variation einen erregten emotionalen Zustand des Sprechers markiert, wie Überraschung, Interesse, Freude; aber auch Verachtung und Ärger ([75] [90] [291] [261] [34]). [169] untersuchten anhand eines bedeutungslosen Wortes fünf verschiedene emotionale Zustände und deren Auswirkungen auf die Grundfrequenz,

den Schalldruckpegel, den intra-oralen Druck und den glottalen Schwingungsverlauf. Die verschiedenen emotionalen Zustände riefen unterschiedliche Werte in der Grundfrequenz, dem Schalldruckpegel als auch im glottalen Schwingungsverlauf hervor.
[136] führten eine Untersuchung zu den Auswirkungen von Emotionen auf die Stimmqualität durch. Sie ließen Probanden ein Computerspiel spielen, dessen Schwierigkeit sie manipulierten, sodass die Probanden unterschiedliche Emotionen erlebten während sie spielten. Der Grundfrequenzbereich war im neutralen und glücklichen Zustand am größten und bei negativen Gefühlen gering. Die Grundfrequenz war umso höher, je erregter der emotionale Zustand der Person war. Jitter trat vor allem dann auf, wenn der Proband sich glücklich fühlte. Fühlte er sich angespannt oder neutral, waren die Werte am geringsten. Die Jitter-Werte korrelierten mit der Grundfrequenz und zeigten für gestresste Probanden geringe Werte. Die Energie des Signals war am größten, wenn die Probanden glücklich, irritiert oder angespannt waren. Im deprimierten und gelangweilten Zustand war die Energie am geringsten. Der Glottis-Verschluss-Quotient „GOQ“ (Verhältnis der Zeit, in der die Glottis geschlossen ist, zur Gesamtzeit eines Schwingungszyklus) zeigte, dass im erregten Gefühlszustand (glücklich, ängstlich, angespannt) die Glottis schneller schließt. Insgesamt lassen sich alle gemessenen Effekte dadurch erklären, dass bei erregten Gefühlszuständen die Muskeln im Kehlkopf stärker angespannt sind.
[19] argumentieren jedoch, dass die akustischen Merkmale im Sprachsignal weniger den emotionalen Zustand des Sprechers wiedergeben, als vielmehr Instrument sind, um beim Hörer eine bestimmte Emotion und damit ein bestimmtes Verhalten auszulösen. Fest steht jedoch, dass Emotionen sich auf irgendeine Weise im Sprachsignal manifestieren und damit die stimmlichen Merkmale eines Sprechers beeinflussen.

Stress

Auch wenn externer Stress sich ebenfalls in emotionalen Reaktionen der Sprecher äußert, soll er hier als Auslöser besonders starker (negativer) Emotionen separat betrachtet werden.
[131] verglich die Ergebnisse verschiedener Studien zum Einfluss von Stress auf unsere Stimme und Sprache. In den Studien wurden verschiedene Arten von Stress untersucht. Manchmal sollten die Versuchspersonen nur eine bestimmte Emotion, wie zum Beispiel Ärger spielen, in anderen Fällen befand sich die Person tatsächlich in einer lebensbedrohlichen Situation. Einige mussten schwierige Aufgaben lösen und andere erlebten eine berufliche Bedrohung. Allgemein lässt sich sagen, dass unabhängig von der konkreten Stressform, die Versuchspersonen eine erhöhte durchschnittliche Grundfrequenz aufwiesen. Befanden sich

die Personen in Lebensgefahr, so wurde in einigen (aber nicht allen) Studien außerdem ein erhöhter Jitterwert oder der Anstieg der Variationsbreite verzeichnet. In zwei Studien wurde auch ein stärkerer Stimmtremor beobachtet. Spielten die Versuchspersonen die Emotion "Furcht", erhöhten sich ihre Jitterwerte. Bei der Emotion "Ärger" veränderte sich außer der durchschnittlichen Grundfrequenz nichts, außer in einer Studie, die auch eine erhöhte Standardabweichung von F0 feststellte. Wurden die Personen kognitivem Stress ausgesetzt, indem man sie schwierige Aufgaben lösen ließ, so wurden in manchen Studien vereinzelte andere Veränderungen festgestellt, wie zum Beispiel eine zentralere Lage der Formanten, eine Veränderung der Variationsbreite oder der Amplitude. Setzte man die Teilnehmer einer beruflich bedrohlichen Situation aus, so konnte man in zwei der Studien einen Anstieg der Standardabweichung der Grundfrequenz beobachten und in zwei anderen Fällen eine Veränderung der Jitter-Werte. In drei Fällen veränderte sich die Amplitude der hohen Frequenzen [131].
Auch wenn sich in den Untersuchungen bestimmte Effekte wiederholen, lässt sich kein Phänomen erkennen, außer vielleicht das Ansteigen der durchschnittlichen Grundfrequenz, das immer auftrat, sobald eine Person in eine stressige Situation geriet.

Krankheiten

Erkrankungen im Bereich des Kehlkopfes und der Glottis beeinflussen vor allem die Stimmqualität. Das Sprachsignal der Betroffenen kann sich auf zwei verschiedene Weisen vom "normalen" Zustand unterscheiden. Einmal kann sich die "Aperiodizität" des Signals erhöhen, was bedeutet, dass die Schwingungen der Stimmlippen unregelmäßig, sprich unterschiedlich lang sind. Zum anderen kann sich das "Rauschen" beim Sprechen erhöhen, wenn die Stimmlippen nicht richtig schließen. Beide Größen und ihre jeweilige Ausprägung geben Aufschluss über die etwaige Erkrankung des Patienten [196].
Neben solchen chronischen Krankheiten, gibt es natürlich viele Erkrankungen, die sich nur zeitweise auf die Stimme und das Sprachsignal auswirken wie zum Beispiel eine normale Erkältung. Die oft einhergehende Blockierung des Nasenraums verändert den Klang der Stimme [254].

Drogen

Rauchen Rauchen hat ebenfalls einen Effekt auf die Stimme. [99] zeigten, dass Frauen, die rauchen, im Schnitt eine tiefere Grundfrequenz haben. Während Nicht-Raucherinnen eine durchschnittliche Grundfrequenz von 183 Hz besaßen, hatten Raucherinnen eine durch-

schnittliche F0 von 164 Hz. Außerdem erhöhte sich ihr Grundfrequenzbereich, sprich ihre Grundfrequenz variierte in einem größeren Bereich als bei Nicht-Raucherinnen. [270] zeigten auch für männliche Raucher eine signifikante Verringerung der Grundfrequenz gegenüber Nicht-Rauchern. Neben der Grundfrequenz werden auch Merkmale der Stimmqualität durch das Rauchen beeinflusst. [104] zeigten, dass auch nach einer relativ kurzen Zeit des Rauchens (< 10 Jahre) Perturbationsparameter wie Jitter bei beiden Geschlechtern signifikant ansteigen. Bei Frauen verringerten sich eher die Grundfrequenzparameter wie durchschnittliche, höchste und niedrigste F0, während sich bei Männern eher eine Erhöhung der Stimmtremorparameter wie Frequenz- und Amplitudentremorintensität zeigte. Dabei entschied die Anzahl der gerauchten Zigaretten pro Tag über das Ausmaß der Veränderung. Allerdings ist der Effekt auf die Grundfrequenz möglicherweise nicht dauerhaft. [201] zeigten, dass sich nach einer Rauchpause von 40 Stunden die Grundfrequenz der Raucher erhöhte, während sich die Grundfrequenz der Nicht-Raucher im selben Zeitraum nicht veränderte. Wahrscheinlich bilden sich die rauchbedingten Veränderungen an den Stimmlippen wieder zurück, sodass die Sprecher nach einer Zeit des Nicht-Rauchens vermutlich wieder die gleichen oder ähnliche Stimmwerte aufweisen wie Nicht-Raucher.

Alkohol Auch der Konsum von Alkohol hat einen Effekt auf die Stimme. Zwar ist nicht bekannt, ob er einen andauernden Effekt auf die Stimme hat, aber definitiv einen kurzzeitigen. So konnte gezeigt werden, dass betrunkene Sprecher (min. 0,5 Promille) eine signifikant höhere Grundfrequenz aufwiesen als in nüchternem Zustand ([46] [29]). Dieser Effekt war für Männer und Frauen gleichermaßen stark und zeigte sich auch durchgehend bei unterschiedlichen Sprechstilen (Lesesprache, Spontansprache, Kommando- und Kontrollsprache) [29]. Außerdem erhöhte sich die Dauer der gelesenen Sätze [46], während die Artikulationsgenauigkeit aufgrund der Beeinträchtigung der sensomotorischen Fähigkeiten, sank [234]. Eine Studie von [21] untersuchte den Einfluss von Alkohol auf linguistische und paralinguistische Merkmale. Es zeigte sich, dass die Sprache der Probanden in betrunkenem Zustand stockender wurde. Sie machten mehr Häsitationen, die Häsitationen dauerten länger, mehr lange Pausen, sprachen Wörter falsch aus, verlängerten Laute oder unterbrachen Wörter.

Stimmverstellung

Wenn ein Sprecher nicht erkannt werden möchte, kann er seine Stimme verstellen, sprich sein stimmlich-sprachlichenes Verhalten verändern. Dabei gibt es nach [253] verschiedene

Bereiche, in denen er seine Stimme verändern kann. Er kann sie phonatorisch verändern, indem er beispielsweise seine natürliche Stimmlage erhöht oder vertieft, behaucht oder gepresst (z.B. [179]) spricht oder flüstert. Der Sprecher kann seine Stimme aber auch phonemisch verstellen, indem er ungewöhnliche Allophone verwendet wie z.B. bei der Imitation eines Dialekts oder Akzents (z.B. [204]) oder hypernasaliert (z.B. [40]). Auf der prosodischen Ebene könnte er probieren sein Sprechtempo zu manipulieren (z.B. [40]) oder einzelne Segmente zu kürzen oder zu längen. Als weitere Möglichkeit steht noch die physische Verformung des Vokaltrakts zur Verfügung wie beispielsweise durch Nase zuhalten, Kiefer zusammenpressen oder mit einem Fremdkörper im Mund sprechen.
Wie effektiv die Stimmverstellung ist, hängt maßgeblich von den Fähigkeiten des Sprechers ab, seine Stimme, Sprache und Sprechweise stärker zu verändern als es durch eine normalverteilte Variation zu erwarten wäre [118]. So wies [40] nach, dass professionelle Sprecher ihre Stimme wirkungsvoller verstellen konnten als Laien, sodass sie von Hörern schwerer erkannt wurden.

2.1.4. Der Einfluss des Übertragungskanals

Neben den Faktoren, die sich direkt auf die menschliche Stimme auswirken, existieren weitere Einflüsse, die das Sprachsignal beeinflussen. Dazu zählen zum Beispiel der Übertragungskanal, Hintergrundgeräusche oder die maschinelle Veränderung des Sprachsignals. Unter all diesen Faktoren spielt wahrscheinlich der Übertragungskanal die wichtigste Rolle. Wenn mit einem schlechten Gerät aufgenommen oder die Sprache über das Telefon übertragen wurde, wird ihre Qualität beeinträchtigt. Über den Telefonkanal werden nur die Frequenzen von ca. 300 bis 3400 Hz übertragen. Besonders die Frikative leiden häufig unter einer Übertragung per Telefon, da sie ihren Frequenzschwerpunkt häufig oberhalb von 3400 Hz haben.

Telefonkanal

Festnetztelefon Eine Untersuchung von [159] beschäftigte sich mit dem Einfluss des Telefonkanals auf Vokalformanten. Durch die Übertragung des Sprachsignals per Telefon erhielt man vor allem für den ersten Formanten (F1) abweichende Messwerte. Künzel führt diese Phänomene auf den verwendeten Algorithmus zurück. Als Schlussfolgerung warnt er vor dem “telephone effect” bei der Analyse von Vokalformanten [159]. Auch [210] stellten fest, dass das Telefon einen Einfluss auf das Sprachsignal hat und die Identifikationsfähigkeit von Personen senkt. Beim Vergleich von zwei über Telefon aufgenommenen

Sprechern, wurde eine unschuldige Person von naiven Hörern häufiger als Täter identifiziert als der Täter selbst. Bei der automatischen Sprechererkennung verursacht die verminderte Qualität der Sprachaufnahmen durch die Telefonübertragung ebenfalls eine Verringerung der Erkennungsrate [250].

Mobiltelefon [42] untersuchten, ob auch das Mobiltelefon einen solchen Einfluss hat und ob er sich von dem des Telefonskanals unterscheidet. Ihre Ergebnisse zeigten, dass die Veränderungen, die in den Formanten hervorgerufen wurden, sehr variabel waren. Auch sie ermahnen zur Vorsicht bei der Interpretation von Messwerten des Sprachsignals, wenn es per Mobiltelefon übertragen wurde.

Hintergrundgeräusche

Ein weiterer Störfaktor bei der Sprechererkennung sind Hintergrundgeräusche. Dabei sind sehr variable oder sprachähnliche Geräusche störender und schwieriger zu kompensieren als gleichmäßige. [225] zeigten mit ihrer Untersuchung zum Einfluss von sprachähnlichen Störgeräuschen auf die Wahrnehmung von Vokalen und Plosiven, dass bei Vokalen hauptsächlich F2 und und der mittlere Frequenzbereich des Spektrums gestört wurden. Die Hörer konnten aber durch den Informationsgehalt von F1 und die restlichen Informationen von F2 die Vokale relativ gut identifizieren. Plosive, deren Burst durch Störgeräusche überlagert war, konnten dennoch gut erkannt werden. Das lässt vermuten, dass Hörer bei der Plosividentifikation auf andere Merkmale, möglicherweise Formanttransitionen, zurückgreifen [225]. Auch wenn sich die Untersuchung auf die Lautidentifikationsfähigkeit bezog, so ist doch anzunehmen, dass eine Störung der Vokalformanten sich ebenfalls auf die Sprecheridentifikation auswirken wird. Viele Studien widmeten sich der Frage, wie man Hintergrundgeräusche neutralisieren kann, sodass sie die (automatische) Sprechererkennung nicht (oder nur wenig) beeinträchtigen (z.B. [256]).

2.1.5. Weitere Störfaktoren

Nicht nur der Telefonkanal kann einen verändernden Einfluss auf die Sprechweise haben, sondern auch die Situation des Telefonierens. Die Tatsache, dass ein Sprecher telefoniert und nicht mit einer im Raum anwesenden Person spricht, verändert sein Sprechverhalten. [85] beobachteten eine nicht-signifikante Tendenz der Vokalformanten sich zu extremeren Positionen im Sinne der Artikulation zu bewegen, sprich die Sprecher hyperartikulierten die Laute.

[81] analysierte den Einfluss von Gesichtsmasken auf die akustischen Merkmale von Frikativen. Die spektralen Werte der Frikative wurden durch die Masken signifikant verändert. Zwei verschiedene Einflüsse der Gesichtsmasken können zu diesem Effekt geführt haben: einmal können sie die Frikative akustisch gedämpft haben und andererseits können sie auch die Bewegungsfreiheit der Artikulatoren des Sprechers eingeschränkt haben, sodass dieser nicht „normal“ sprechen konnte.
[132] untersuchten, wie Sprecherverifikationssysteme auf maschinell veränderte Sprache reagieren. Abhängig vom verwendeten Algorithmus, konnten manche Systeme durch transformierte Sprache getäuscht werden.
Dies ist nur eine kleine Auswahl an möglichen Störfaktoren, um einen Eindruck zu vermitteln, welche Faktoren bei der Identifikation eines Sprechers einen negativen Einfluss haben können.

2.2. Von der Artikulation zur Akustik

Die akustische Phonetik beschreibt den Zusammenhang zwischen Sprachschall und den physiologischen sowie artikulatorischen Aspekten der Sprechorgane. Die akustischen Korrelate der Anatomie und Physiologie des Sprechapparats bilden das Sprachsignal. Dabei stehen Artikulation und Akustik in keinem linearen Zusammenhang. Das bedeutet, das eine kleine Änderung bei der Artikulation eine große Wirkung auf die Akustik haben kann und umgekehrt.
Auf welche Weise die Quelle (Glottis) und der Filter (Vokaltrakt) zusammenwirken, um das Sprachsignal zu erzeugen, lässt sich anhand des Quelle-Filter-Modells von [77] gut darstellen. Zwei Faktoren beeinflussen laut diesem Modell die Eigenschaften des Sprachschalls. Das ist zum einen die Glottis (Quelle), die sowohl periodischen, stimmhaften Rohschall erzeugen kann, wenn sie schwingt, als auch stimmlosen, rauschenden Rohschall, wenn sie geöffnet ist und nicht schwingt. Der Rohschall wird dann im Vokaltrakt (Filter) durch Verengungen an bestimmten Stellen modifiziert. Quelle und Filter werden im Modell vereinfachend als unabhängig voneinander betrachtet und ergeben durch ihre Kombination den vollständigen Sprachschall. Dazu werden das Spektrum des Quellsignals und des Filters miteinander multipliziert. Als Ergebnis erhält man das Spektrum des Sprachsignals.
Die Quelle-Filter-Theorie erklärt zwar, wie durch das Zusammenwirken von Glottis und Vokalttrakt das Sprachsignal mit seinen akustischen Eigenschaften entsteht, aber es ist noch unklar, welche Merkmale des Vokaltrakts zu dem endgültigen Sprachsignal führen. Ein

Grundlagenmodell zur Darstellung dieses Zusammenhanges, stellt das neutrale Rohr dar, welches die Entstehung der Vokalformanten erklärt. Dabei wird der menschliche Vokaltrakt vereinfachend als ein 17 cm langes Rohr mit einem Durchmesser von 4 bis 5 cm angenommen, welches auf der Seite der Quelle geschlossen ist und auf der anderen Seite offen. Die Wände des Rohres werden als schallhart angenommen. Tritt in diesem System ein Impuls auf und breitet sich aus, so wird er an allen verschlossenen Stellen reflektiert bis er schließlich an der einzigen offenen Stelle entweicht. Aufgrund der Geometrie des Systems entstehen Resonanzen in Form von stehenden Wellen. Dabei markieren die Stellen der größten Auslenkung (der Wellenbäuche) die Amplitude der Schwingung. Jede der Schwingungen hat eine andere Wellenlänge und somit auch eine andere Frequenz. Da der menschliche Vokaltrakt aber natürlich nicht schallhart und die Glottis nicht immer verschlossen ist, werden durch die Resonanz Frequenzbänder statt einzelner Frequenzen verstärkt, die sich als Formanten zeigen. Die Formanten werden beginnend mit F1 fortlaufend nummeriert [235].
In einem erweiterten Modell von [283] wird das Rohr nicht mehr als querschnittsflächengleich angenommen, sondern mit einem verengten oder erweiterten Abschnitt. Je nach Position der Verengung oder Erweiterung, verändern sich auch die entstehenden Formanten. Ob eine Verengung oder Erweiterung an einer bestimmten Stelle die Formanten im Vergleich zu dem neutralen Rohr absenkt oder anhebt, zeigt der Formantverschieber. So hat zum Beispiel eine Verengung im vorderen Teil des Rohres die gleiche Wirkung auf die Formanten wie eine Erweiterung im hinteren Teil. Betrachtet werden in diesem Modell nur die ersten drei Formantfrequenzen, da nur diese willkürlich beim Sprechen beeinflusst werden können [235].
Neben diesen modellhaften Verfahren, kann die Übertragungsfunktion des Signals (und somit die Form des Vokaltrakts) durch Formeln mathematisch beschreiben werden. Eine dieser Methoden wird Pole-Zero-Modell genannt, wobei die Polstellen die Wellenmaxima und die Nullstellen die Wellenknoten bezeichnen. Bei einem Vokal hätte die Übertragungsfunktion demzufolge ihre Maxima genau in den Frequenzbereichen, in denen die Formanten des Vokals liegen. Diese Übertragungsfunktion ist für jeden Sprecher individuell, da sich die Form des Vokalstrakts zwischen den Menschen unterscheidet. Deshalb kann das Pole-Zero-Modell zur Unterscheidung zweier Sprecher verwendet werden, was auch schon praktiziert wurde (siehe z.B. [72]).

2.2.1. Suprasegmentale Merkmale

Suprasegmentale Merkmale sind globale Merkmale, welche die gesamte Äußerung eines Sprechers durchziehen. Da sie nicht Untersuchungsgegenstand dieser Arbeit sind, sollen

sie hier nur kurz dargestellt werden als Alternative zu den segmentalen Merkmalen (siehe Abschnitt 2.2.2).
Zu den suprasegmentalen Merkmalen zählen alle prosodischen Merkmale (z.B. Wort- und Satzakzent, Sprechtempo, Rythmus und Pausen), die sich durch eine systematische Variation der Dauer, der Grundfrequenz und der Amplitude manifestieren [230]. Aber auch die von [170] beschriebenen Stimmqualitäten (behauchte Stimme, Knarrstimme, Falsett) gehören zu den suprasegmentalen Merkmalen der menschlichen Stimme [208]. Prosodische Merkmale können ebenfalls Bedeutungen transportieren und sind genauso sprach-spezifisch wie die Laute einer Sprache. Dabei können sie sowohl linguistische als auch extra-linguistische Informationen transportieren. Beispielsweise werden Wortgrenzen durch den Grundfrequenzverlauf markiert, aber auch die Meinung oder Befindlichkeit des Sprechers kann in seiner Sprechmelodie Ausdruck finden.
Die Unterschiede in Grundfrequenz und Amplitude entstehen hauptsächlich durch glottale Merkmale: (1) durch die Spannung der Stimmlippen und (2) durch den subglottalen Luftdruck. Die Erhöhung des subglottalen Drucks und der Stimmlippenspannung führen beide zu einer Erhöhung der Grundfrequenz und der Lautstärke. Werden nun einzelne Laute mit einer höheren Grundfrequenz und Lautstärke gesprochen, nehmen wir sie als betont war. Aber nicht nur eine Erhöhung der spektralen Merkmale kann die Wahrnehmung einer Betonung auslösen; auch die Veränderung temporaler Merkmale wie die Erhöhung der Dauer eines Lautes lassen ihn betont wirken. Steht ein Vokal allerdings in einer unbetonten Position, dann wird er im Deutschen reduziert, sowohl in seiner Dauer als auch in seiner Qualität. Der unbetonte Vokal wird zentralisiert ausgesprochen, sodass seine Formanten mehr denen des Neutralvokals [ə] entsprechen. Das heißt, auch die Stellung der Artikulatoren beeinflusst die Wahrnehmung von Betonung.
So lässt sich zusammenfassen, dass prosodische Merkmale sowohl durch Veränderungen an der Quelle (Glottis) verursacht werden, was Veränderungen in Grundfrequenz und Amplitude bewirkt, als auch durch die Koordination der Artikulatoren und der Form des Vokaltrakts, was Veränderungen in Dauer und Formantwerten verursacht [230].

2.2.2. Segmentale Merkmale

Im Gegensatz zu den globalen suprasegmentalen Merkmalen, beschränken sich die segmentalen Merkmale auf ein Segment bzw. einen Laut. Diese phonetischen Informationen geben ebenfalls Aufschluss über die Identität des Sprechers. Sprecher können sich laut [208] auf der segmantalen Ebene auf verschiedene Weise unterscheiden. Zum Ersten können zwei

Sprecher unterschiedliche Phonemsysteme haben, indem sie zum Beispiel verschiedenen Dialekten angehören. Zweitens können sie Differenzen in ihrer Phonotaktik aufweisen. Dies ist der Fall, wenn die Phoneme der beiden Sprecher in unterschiedlichen Kontexten erlaubt oder verboten sind. Eine dritte Möglichkeit besteht darin, dass Sprecher individuelle Aussprachevarianten der Phoneme verwenden und der vierte Typ beschreibt die Art auf die sich Phonemsysteme zweier Sprecher unterscheiden können (z.B. in einem bestimmten Allophon oder im Grad der Koartikulation). Zu guter Letzt kann aber auch die Betonung auf der suprasegmentalen Ebene einen Einfluss auf die Aussprache von Phonemen haben, wie zum Beispiel auf die Vokalqualität, Lautstärke, Dauer und Grundfrequenz [208].
Um die Möglichkeiten, in denen sich Laute zwischen Sprechern voneinander unterscheiden können, zu verdeutlichen, sollen zunächst die Vokale und Konsonanten in ihren spezifischen Eigenschaften genauer betrachtet werden.

Vokale

Vokale enthalten als stimmhafte Laute auch immer die Grundfrequenz des Sprechers, welche neben den Segment-spezifischen Eigenschaften (z.B. Formanten) wichtige Informationen zur Identität des Sprechers liefert. Dennoch sollen Vokale hier nur kurz betrachtet werden, da der Fokus dieser Arbeit auf den Konsonanten liegt. Allerdings sind die Vokale für die Wahrnehmung der Konsonanten insofern wichtig, dass viele Konsonanten einen Teil ihrer Informationen in den Transitionen zu den umliegenden Vokalen enthalten. Daher sollen ihre Produktion und ihrer wichtigsten Merkmale hier kurz dargestellt werden.
Alle Vokale beginnen mit einem gemeinsamen, an der Glottis produzierten Signal. Dieses Signal breitet sich über den Rachen in den Mund und schließlich in die Luft außerhalb aus. Die Form des Vokaltrakts (Rachen und Mundraum) bestimmen dabei, welcher Vokal am Ende entsteht. Der einfachste aller Laute, ist der Neutralvokal [ə]. Der Vokaltrakt nimmt bei der Artikulation dieses Vokals keine vorgegebene Form an, sondern bleibt in der neutralen Position. Nimmt man den Vokaltrakt als Röhre mit einer Länge von 17,5 cm an (durchschnittliche Vokaltraktlänge eines erwachsenen Mannes), dann entstehen durch die Resonanzen des Vokaltrakts Formanten, welche im Spektrum des Vokals sichtbar werden. Die erste Formantfrequenz (F1) liegt beim [ə] bei ca. 500 Hz, F2 bei 1500 Hz, F3 bei 2500 Hz und jede weitere (F4, F5, ect.) im Abstand von 1000 Hz. Die Position der Formanten, besonders von F1 und F2, hängen stark von der Form des Vokaltrakts ab, wie der Position der Lippen, der Zunge, des Rachens und des Kiefers. Die Lage von F3 charakterisiert nur bestimmte Sprachlaute und die höheren Formanten (ab F4) bleiben relativ konstant in

ihrer Frequenz. Sie hängen vor allem von der Länge des Vokaltrakts ab, weniger von den produzierten Sprachlauten.
Insgesamt hängt die Lage der Formantfrequenzen von 3 Faktoren ab: der Länge des Vokaltrakts sowie Ort und Enge der Konstriktion. Bei den vorderen Vokalen [i, e, ɛ, æ, a] hat die Zunge eine konstante vordere *Zungenlage*, nur die *Zungenhöhe* verändert sich. Vom [i] zum [a] senken sich die Zunge und gleichzeitig auch der Kiefer immer weiter ab. Bei den hinteren Vokalen [u, o, ɔ, ɑ] ist die Zungenlage weiter nach hinten verschoben. Die Zungenhöhe verläuft in ähnlichen Abstufungen wie bei den vorderen Vokalen. Die Vokale [u, o] sind außerdem noch gerundet, sprich die Lippen werden leicht vorgeschoben und gerundet. Wie bereits erwähnt, stehen F1 und F2 in engem Zusammenhang zu der Form des Vokaltrakts. So bewirkt eine Konstriktion im pharyngalen Trakt einen Anstieg von F1, während eine Konstriktion im vorderen Teil des Vokaltrakts einen Abfall von F1 erzeugt. Gibt es eine Konstriktion im hinteren Vokaltraktteil, so fällt F2 ab, wohingegen eine Konstriktion der vorderen Zunge F2 ansteigen lässt. Je stärker dabei die jeweilige Konstriktion ist, desto stärker ist auch der Anstieg/Abfall von F1/F2. Eine Lippenrundung bewirkt für alle Formanten einen Abfall ihrer Frequenz und je stärker die Rundung, umso stärker auch der Abfall. Aufgrund dieser Korrelationen haben hohe, geschlossene Vokale (z.B. [i, u] einen niedrigen F1 von ca. 250 Hz. Wohingegen tiefe, offene Vokale (z.B. [æ, ɑ]) einen hohen F1 von ca. 700 Hz haben. Alle andere Vokale ordnen sich zwischen diesen Extremwerten ein, sodass die F1-Werte die Vokale in ihrer Höhe unterscheiden. F2 hingegen unterscheidet die Vokale in ihrer vorn/hinten-Lage, sodass vordere Vokale (z.B. [i, e]) einen hohen F2 von ca. 2150 bis 2000 Hz haben und hintere Vokale (z.B. [u, o]) einen niedrigen F2 von 800 bis 900 Hz. Wobei die Formantwerte bei den gerundeten Vokalen zusätzlich durch die Lippenrundung gesenkt werden. Auf diese Weise ergeben sich die spezifischen Formantmuster der einzelnen Vokale [230].
Im Deutschen werden die Vokale in lange, gespannte Vokale und in kurze, ungespannte Vokale unterteilt. Zu den langen, gespannten Vokalen gehören [iː, eː, ɛː, aː, yː, øː, uː, oː], während [ɪ, ɛ, a, ʏ, œ, ɔ, ʊ, ə, ɐ] zu den kurzen, ungespannten zählen. Die ungespannten Vokale werden tendenziell zentralisierter gesprochen, sprich die Zunge bewegt sich nicht so weit von der Position des Neutralvokals weg wie bei den gespannten Vokalen ([235]).

Konsonanten

Den Konsonanten wird gegenüber den Vokalen eine weniger wichtige Rolle in der Sprecheridentifikation zugeschrieben, weshalb viele Studien zu sprecherspezifischen Merkmalen

sich auf Vokale konzentrierten (z.B. [177] [187] [268]). Allerdings zeigten bereits einige Studien, dass auch die Konsonanten einige relevante Sprecherinformationen enthalten (z.B. [68] [12] [14] [142] [143]). Deshalb sollen an dieser Stelle zunächst die spezifischen Eigenschaften der verschiedenen Konsonanten erläutert werden. Da nur Konsonanten der Artikulationsmodi *Nasal*, *Plosiv* und *Frikativ* untersucht werden, sollen auch nur diese Lautgruppen beschrieben werden. Außerdem sollen die verschiedenen Artikulationsstellen, besonders die *labiale* und die *alveolare*, betrachtet werden. Der Artikulationsmodus eines Konsonanten wirkt sich vor allem auf seine konstanten spektralen Merkmale aus, welche sich durch starke Ein/Aus-Veränderungen im Spektrum bemerkbar machen. Die Artikulationsstelle hingegen bewirkt eine Veränderung der Form des Vokaltrakts, wodurch die Positionsinformationen eines Konsonanten mit transitionalen Änderungen im spektralen Muster assoziiert sind. Eine alveolare Konstriktion löst einen Anstieg der zweiten Formantfrequenz (F2) aus, wohingegen eine labiale Konstriktion einen Abfall von F2 bewirkt. Das Ausmaß der F2-Veränderung hängt dabei von der Stärke der Konstriktion ab. Das heißt, bei einem alveolaren Konsonanten steigt vom Vokal zum Konsonanten F2 an und fällt nach dem Konsonant zum nächsten Vokal hin wieder ab. Für einen labialen Konsonanten zeigt sich das gegenteilige Bild: Abfall von F2 von Vokal zu Konsonant und Anstieg vom Konsonant zum nächsten Vokal. F3 verhält sich dabei häufig parallel zu F2 [230]. Diese Muster der Formanten zeigen sich grundsätzlich bei allen Konsonanten, wobei die Geschwindigkeit und Länge der Transitionen sich zwischen den verschiedenen Artikulationsmodi unterscheiden kann.

Nasale Zu den Nasalen gehören die Laute /m, n, ŋ/, welche alle zu den stimmhaften Konsonanten gehören. Ihre Besonderheit besteht darin, dass bei ihrer Artikulation nicht nur der Mundraum verwendet wird, sondern auch der Nasenraum. Im Mundraum wird ein Verschluss gebildet, zum Beispiel mit der Zunge, sodass keine Luft den Mundraum verlassen kann. Gleichzeitig wird aber das Velum (Gaumensegel) abgesenkt, sodass der Rachen mit dem Nasenraum verbunden wird. Dadurch kann die Luft durch die Nase ausströmen und es entsteht ein nasaler Laut [254]. Die orale Verschlussgeste ist dabei ähnlich zu der der Plosive /b, d, g/, weshalb die Nasale auch häufig als Plosive mit dem Merkmal *Nasalierung* beschrieben werden. Aufgrund der ähnlichen Verschlussgeste teilen die Nasale ihr Merkmal der abrupten Transitionen mit den Plosiven. Nur die Öffnungsgeste ist bei den Nasalen etwas langsamer, da sich kein Luftdruck im Mundraum aufbaut.
Insgesamt sind Nasale recht leise (z.B. im Vergleich mit Vokalen), da einerseits durch den

oralen Verschluss keine Luft aus dem Mund strömen kann und weil der orale Vokaltrakt Antiresonanzen (FZ1, FZ2, ...) auslöst, die bestimmte Frequenzbereiche im Nasalspektrum stark dämpfen bzw. auslöschen [108]. Das Spektrum des Nasalgeräusches enthält daher vor allem niedrige Frequenzen bis ca. 300 Hz, ausgelöst durch die Hauptresonanz des großen Nasenraumes, der durch die kleinen Nasenöffnungen begrenzt wird. In diesem Bereich (zwischen 300 und 400 Hz) befindet sich auch der erste Nasalformant (N1) und höhere Nasalformanten im Abstand von etwa 800 Hz ([77] [88]). Über 800 Hz haben Nasale nur noch wenig Energie, sodass sich die Muster der höheren Formanten zwischen den Nasalen kaum unterscheiden lassen.
Im Spektrogramm werden Nasale durch ihre abrupten Übergänge zu den benachbarten Vokalen charakterisiert. Diese resultieren erstens aus dem plötzlichen Anfang und Ende der oralen Verschlussgeste und zweitens aus dem bereits zum Zeitpunkt der oralen Verschlussgeste weit geöffneten Velum. Die Abwärtsbewegung des Velums beginnt schon lange vor der oralen Verschlussgeste, wodurch der Nasenraum bereits geöffnet ist, wenn der orale Verschluss erreicht wird. Der Vorsprung und die Verzögerung der velaren Geste bei der oralen Verschluss- und Öffnungsgeste ist ca. 100 ms lang. Dies bewirkt, dass vorangehende und folgende Vokale ebenfalls (teilweise) nasaliert werden [230].
Die Form und Größe des Nasenraumes eines Sprechers ist für alle Nasale gleich. Daher unterscheidet sich das Spektrum der verschiedenen Nasale nur wenig. Allerdings gibt es ein paar Unterschiede durch die unterschiedliche Länge des verschlossenen Mundraums. Auch wenn der Rachen-Nasen-Trakt den Großteil des nasalen Geräusches transportiert, so beeinflusst der orale Vokaltrakt ebenfalls das Nasalgeräusch. Je nachdem wie lang der verschlossene orale Teil ist - am längsten bei /m/, kürzer bei n und am kürzesten bei /ŋ/ - ändert sich die Frequenz der Antiresonanz, die vom oralen Vokaltrakt im Rachen-Nasen-Geräusch ausgelöst wird. Je länger der orale Vokaltrakt ist, desto tiefer ist die Antiresonanzfrequenz. Sie liegt für /m/ bei ca. 800 Hz, für /n/ bei 1500 - 2000 Hz und für /ŋ/ bei ca. 5000 Hz oder mehr. Da das nasale Spektrum aber nur bis 500 Hz starke Amplituden hat und darüber sehr schwach ist, kann man diese Unterschiede in der Lage der Antiresonanzen nur schlecht wahrnehmen. Prominenter und wichtiger für die Wahrnehmung der Artikulationstelle der Nasale scheinen dagegen die Transitionen zu den angrenzenden Vokalen zu sein ([77] [94]). Die Formanttransitionen weisen die beschriebenen, für Konsonanten typischen Verläufe auf; nur dass in den Transitionen zu den angrenzenden Vokalen zusätzlich noch Nasalierung auftritt [230]. Auch wenn einige der Unterschiede zwischen den Artikulationsstellen der Nasale schwer wahrzunehmen sind, so müssen doch eindeutige Informationen vorhanden

sein, da Hörer in der Lage sind, die Nasale /m/, /n/ und /ŋ/ zu unterscheiden. Laut [182] steckt ein Teil der Ortsinformationen auch im konsonantischen Teil der Nasale. Spätere Studien zeigten, dass vor allem der Übergang von Nasal zu Vokal viele Informationen zur Wahrnehmung der Artikulationsstelle enthält (z.B. [247]). Wie viele Informationen im konsonantischen Teil und wie viele in den Transitionen stecken, hängt aber auch von den umliegenden Vokalen ab. Bei offenen und hinteren Vokalen (z.B. /ɑ, u/) enthalten die Transitionen viele Ortsinformationen, während sie bei dem hohen, vorderen Vokal /i/ nur wenige Informationen zur Artikulationsstelle enthalten [247]. [160] argumentieren, dass die Energieveränderung im Bereich von 1450 bis 2300 Hz von */n/* zum folgenden Vokal größer ist als von */m/*, weil der alveolare Nasal in diesem Bereich einen Antiformanten hat, der labiale aber nicht. Diese Differenz kann ebenfalls als perzeptives Unterscheidungsmerkmal für labiale und alveolare Nasale dienen.
Tendenziell wird demzufolge der Artikulationsmodus von den konsonantischen Merkmalen bestimmt, während die Artikulationsstelle hauptsächlich in den Transitionen zu den angrenzenden Vokalen kodiert wird.

Plosive Die Artikulationsgeste der Plosive ist der der Nasale sehr ähnlich. Ein Artikulator bildet einen oralen Verschluss an einer Stelle des Vokaltrakts. Aber im Gegensatz zu den Nasalen wird das Velum nicht abgesenkt, sondern bleibt oben und geschlossen. Dadurch kann während der Verschlussphase keine Luft den Mundraum verlassen. Die Atemluft aus der Lunge strömt aufgrund des subglottalen Drucks durch die Glottis (wo die Stimmlippen bei stimmhaften Plosiven schwingen; bei stimmlosen einfach geöffnet bleiben) und von dort in den Mundraum. Während dieser Verschlussphase ist bei stimmlosen Plosiven kein und bei stimmhaften Plosiven ein schwaches Geräusch der schwingenden Stimmlippen zu hören. Durch die einströmende Luft steigt der supraglottale Luftdruck an bis er genauso groß ist, wie der subglottale. Ab diesem Zeitpunkt kann keine weitere Luft mehr in den Mundraum strömen, und die Stimmlippen hören auf zu schwingen. Spätestens jetzt wird der Verschluss gelöst und erzeugt durch den hohen Luftdruck ein explosionsartiges Ausströmen der Luft, was zu dem typischen Burst führt. Der erste Vorgang bei dem Burst ist die Verschlusslösung, eine stufenweise Erhöhung des Schalldrucks. Dem folgt eine gedämpfte Schwingung der Resonanzfrequenzen (Formanten) abhängig von der Position im Vokaltrakt und seiner momentanen Form. Der Burst eines stimmhaften Plosives ist kurz, nur ca. 10 bis 20 ms. Aufgrund des höheren Luftdrucks erfolgt die Verschlusslösung bei Plosiven schneller als bei Nasalen. Außerdem erfordert die Verschlussgeste keine präzise Koordination, was sie

einfach und schnell umsetzbar macht. Dadurch haben Plosive die schnellsten, kürzesten und abruptesten Transitionen von allen Konsonanten [230].
Plosive können sowohl stimmhaft als auch stimmlos sein, je nach Stellung der Glottis. Sind die Stimmlippen geöffnet, strömt die Luft einfach hindurch ohne die Stimmlippen in Schwingung zu versetzen und der Plosiv ist stimmlos. Sind die Stimmlippen hingegen geschlossen, so sorgt der kontinuierliche Luftstrom für ein ständiges Öffnen und Schließen der Stimmlippen und der Plosiv wird stimmhaft. Wärend der oralen Verschlussphase nimmt der supraglottale Druck bei einem stimmlosen Plosiv schneller zu als bei einem stimmhaften. Zum Zeitpunkt der Verschlusslösung hat sich besonders bei den stimmlosen Plosiven ein hoher Luftdruck im Mundraum aufgebaut. Wenn der Verschluss nun gelöst wird, strömt die Luft explosionsartig aus. Da auch die Glottis zu diesem Zeitpunkt geöffnet ist, strömt auch weitere Luft aus der Luftröhre nach. Dadurch entsteht bei stimmlosen Plosiven ein stärkerer Burst gefolgt von der Aspiration - einer Art Friktion - die je nach Luftdruck unterschiedlich lange dauern kann. Die Verschluss- und die Öffnungsgeste beim Plosiv dauern ca. 50 ms mit einer Verschlussphase von ca. 100 ms dazwischen. Nach der Verschlusslösung dauert es einen Moment, bevor die Stimmlippen für den nachfolgenden Vokal wieder anfangen zu schwingen [230]. Diese Zeitspanne nennt sich *Voice Onset Time (VOT)* [176]. Stimmhafte Plosive haben eine kurze VOT von ca. 0 bis 20 ms (oder sogar eine negative), während stimmlose Plosive eine längere VOT von 30 ms oder mehr haben. Während die VOT den hauptsächlichen Unterschied zwischen den stimmhaften und den stimmlosen Plosiven ausmacht, gibt es aber noch weitere, kleinere Unterschiede. Zum Beispiel ist die Verschlussphase bei stimmlosen Plosiven etwas länger als bei stimmhaften und die Position des Kehlkopfes höher. Letzteres erhöht den Luftdruck im Vokaltrakt und dehnt die Stimmlippen, wodurch bei der Verschlusslösung die Luft schnell ausströmt und die Stimmlippen etwas schneller schwingen. Das führt dazu, dass Vokale nach stimmlosen Plosiven eine leicht höhere Grundfrequenz haben als nach stimmhaften. Während der Verschlussphase der stimmhaften Plosive ist der Kehlkopf etwas tiefer, was dazu dient den Vokaltrakt zu verlängern [230]. Denn ein längerer Vokaltrakt kann mehr Luft aufnehmen, sodass der Luftdruck im Mundraum nicht so schnell das gleiche Druckniveau wie der subglottale Druck erreicht. Dadurch kann die Luft länger durch die Glottis strömen und die Stimmlippen schwingen lassen, wodurch eine längere Stimmhaftigkeit des Plosives ermöglicht wird. Außerdem ist die Höhe von F1 zwischen stimmhaften und stimmlosen Plosiven nach dem Burst unterschiedlich. Während des vollständigen oralen Verschlusses ist F1 auf seinem niedrigsten Wert, sodass er beim Burst (bei der Verschlusslösung) steigen muss. Da bei stimmhaften Plosiven die Stimmlip-

penschwingungen viel eher beginnen, ist ein Großteil der Transitionen stimmhaft und der Beginn von stimmhaftem F1 ist häufig bedeutend tiefer [84]. Demzufolge steigt F1 zwar bei beiden Plosiven nach dem Burst an ([275] [77]), aber bei den stimmhaften ist F1 deutlich tiefer als bei den stimmlosen Plosiven [108].
Die Informationen über die Artikulationsstelle stecken bei den Plosiven, wie eingangs allgemein für Konsonanten beschrieben, hauptsächlich in den Vokaltransitionen. Besonders die Artikulationsstellen *labial* und *alveolar* lassen sich durch den Verlauf von F2 (und F3) bestimmen. Schon in den 1950er Jahren zeigten [173] und [61] in Perzeptionsexperimenten mit künstlich erzeugten Stimuli, dass der „Locus“ von F2 die Artikulationsstellen der Plosive unterscheidet. Der „Locus“ eines Formanten beschreibt dabei seinen scheinbaren Startwert. Während /b/ am besten bei einem F2-Locus von 720 Hz von den Hörern wahrgenommen wurde, lag diese Frequenz für /d/ bei 1800 Hz. Nur für /g/ ließ sich kein eindeutiger Locus bestimmen. Für vordere Vokale lag er bei ca. 3000 Hz, aber für hintere Vokale konnte kein eindeutiger Locus bestimmt werden. In natürlich-sprachlichen Stimuli konnte dieser invariante „Locus“ für F2 aber nicht gefunden werden [108]. Aber nicht nur F2, sondern auch F3 kann zur Bestimmung der Artikulationsstelle verwendet werden; besonders zur Unterscheidung der alveolaren und velaren Artikulationsstelle [78]. So verlaufen die Formanten F2 und F3 in einem Vokal nach einem velaren Plosiv näher zusammen als in einem alveolaren [236].
Auch wenn die Formanttransitionen die meisten Informationen über die Artikulationsstelle enthalten, so stecken doch auch in dem Burst der Plosive einige Ortinformationen [108]. [269] zeigten in Perzeptionexperimenten, dass auch der Burst Informationen zur Artikulationsstelle des Plosives enthält; allerdings unterschiedlich viele in Abhängigkeit vom jeweiligen Plosiv und dem Vokalkontext. Beispielsweise ist der Burst von labialen Plosiven schwächer als von alveolaren wegen der geringeren Länge des vorderen Vokaltrakts [220].
Zusammenfassend lässt sich sagen, dass genau wie bei den Nasalen, der Artikulationsmodus von den konsonantischen Merkmalen bestimmt wird und die Artikulationsstelle hauptsächlich von den Transitionen.

Frikative Frikative ähneln in ihrem generellen Ablauf der Artikulation von Öffnungs-und-Verschluss-Kreislauf den Nasalen und Plosiven. Aber anders als bei diesen, wird bei den Frikativen kein vollständiger Verschluss des Mundraums gebildet, sondern nur eine Konstriktion [230]. Beispielsweise wird mit der Zunge am Gaumen eine Engstelle gebildet, durch die Luft strömt. Dadurch, dass die Lücke sehr klein ist, wird die laminare

Luftströmung zu einer turbulenten Strömung umgewandelt. Auf diese Weise entsteht das typische Geräusch eines Frikatives [109]. Alle Frikative teilen das durchgängige zufällige Geräusch; charakterisiert durch ein Spektrum mit annähernd gleicher Lautstärke auf allen Frequenzen. Zu diesem zufälligen Geräusch gesellt sich bei stimmhaften Frikativen eine teilweise Periodizität. Der stimmhaft-stimmlos-Kontrast ist bei Frikativen ähnlich wie bei Plosiven. Für stimmlose Frikative sind die Stimmlippen der Glottis weit geöffnet und für stimmhafte geschlossen. Allerdings schwingen die Stimmlippen auch bei einem stimmhaften Frikativ nicht immer. Oft werden nur intervokalische Frikative wirklich stimmhaft, sprich mit schwingenden Stimmlippen, produziert. Außerdem haben stimmlose Frikative (genau wie stimmlose Plosive) eine längere Verschlussphase als stimmhafte. Stimmhafte Frikative sind normalerweise etwas schwächer in ihrer Intensität als stimmlose Frikative, da die Stimmlippen geschlossen gehalten werden, so dass die Luft nicht so schnell durch die Glottis strömen kann. Dies führt zu einem langsameren Luftstrom und somit zu einem schwächeren Friktionsgeräusch.
Die Transitionen der Frikative zu den umliegenden Vokalen sind etwas länger und langsamer als die der Nasale und Plosive. Eventuell liegt das daran, dass die Zunge (oder ein anderer Artikulator) eine bestimmt geformte Konstriktion herstellen muss. Die Bildung dieser Konstriktion verlangt eine präzise Regulation, welche eine kontrollierte und langsame Bewegung voraussetzt. Diese langsamere Bewegung der Artikulatioren führt dann zu ebenfalls langsameren Transitionen.
Der turbulente Luftstrom, der zu dem Friktionsgeräusch führt, erzeugt ein Spektrum mit zufälligen Fluktuationen der Amplitude und deckt einen großen Frequenzbereich ab. Das ist das Quell-Geräusch der Frikative. Der Ort der Konstriktion und die Form und Größe des vorderen Vokaltrakts (zwischen Konstriktion und Lippen) haben einen Frequenz-filternden Effekt auf das Quell-Geräusch. Der hintere Hohlraum (zwischen Glottis und Konstriktion) hat wenig Einfluss auf das Frikativspektrum. Die Länge des vorderen Teils des Vokaltrakts bestimmt die Lage des Frequenzschwerpunktes des Frikatives. Im Sonagramm kann man sehr gut beobachten, wie mit abnehmender Länge des vorderen Vokaltraktteils, der Frequenzschwerpunkt (schwärzeste Stellen im Sonagramm) immer weiter steigt. So hat /h/ den niedrigsten Schwerpunkt und /f/ den höchsten, während alle anderen Frikative dazwischen liegen (/h/ 1000 Hz, /ʃ/ 3000 Hz, /s/ 4000 Hz und /f/ von 4500 bis 7000 Hz) [230]. Da bei dem Frikativ /f/ kaum ein vorderer Vokaltraktteil existiert, ist sein Spektrum sehr diffus und variabel ohne starke Resonanzschwerpunkte und generell wenig Energie. Die Sibilanten /s/, /ʃ/ hingegen haben mehr Energie in höheren Frequenzen. Nicht nur weil sie einen

größeren vorderen Hohlraum haben, sondern vor allem weil der Luftstrom auf die Zähne trifft, was eine hoch-frequente und energiereiche Turbulenz erzeugt [272]. Im deutschen reihen sich auch noch der palatale /ç/ und der velare Frikativ /x/ mit ein. [134] beschrieb, dass der Frequenzschwerpunkt von /ç/ etwas tiefer liegt als bei /ʃ/ (nämlich auf der Höhe von F3 des angrenzenden Vokals) und auch die Form des Spektrums leicht differiert. Der Schwerpunkt von /x/ liegt noch etwas tiefer, auf der Höhe von F2 des Nachbarvokals [108]. Die genannten Werte der Frequenzschwerpunkte gelten nur für durchschnittliche männliche Sprecher; für Frauen und Kinder würden sich die Frequenzschwerpunkte entsprechend ihrer Vokaltraktlänge nach oben verschieben. Außerdem ist wieder der Einfluss der umliegenden Vokale zu berücksichtigen, welcher sich besonders auf den Frikativ /h/ auswirkt, indem er die Lage seines Frequenzschwerpunktes verschiebt [230].
Insgesamt werden die Frikative in ihrem Artikulationsmodus ebenfalls durch die konsonantischen Merkmale gekennzeichnet. Genau wie die Nasale und Plosive enthalten sie auch Informationen über die Artikulationsstelle in ihren Transitionen. Hinzu kommen aber noch die Frequenzschwerpunkte, welche ebenfalls die Artikulationsstelle kodieren. Dadurch lassen sich Frikative auch unabhängig von ihren Transitionen besser identifizieren als Nasale und Plosive ([207]).

2.3. Die Rolle von Sprechermerkmalen in der Sprachverarbeitung

Nachdem die Entstehung von sprecherspezifischen Merkmalen und deren Auswirkung auf die Akustik besprochen wurden, soll nun der Fokus auf die menschliche Wahrnehmung und Verarbeitung des Sprachsignals gelegt werden, und welche Rolle die Sprechermerkmale dabei spielen.
Die zentrale Aufgabe in der Spracherkennung besteht für den Hörer darin, eine Bedeutung aus dem akustischen Signal eines Sprechers zu extrahieren. Da diese Aufgabe zu schwierig und breit ist, um von einer Disziplin abgedeckt zu werden, soll der hier beschriebene Ansatz sich auf den Prozess des „lexikalischen Zugangs“ beschränken [112]. Es gibt grundlegend unterschiedliche Annahmen darüber, wie bei der Verarbeitung des Sprachsignals mit extralinguistischen (oder auch indexikalischen Merkmalen), wie zum Beispiel Informationen über die Identität des Sprechers, umgegangen wird. Während die abstraktionistische Sichtweise davon ausgeht, dass das Sprachsignal zunächst gefiltert wird, sodass alle nicht-linguistischen Merkmale herausgefiltert werden, gehen exemplarbasierte, episodische Modelle davon aus,

dass alle Exemplare (z.B. eines Wortes) mit ihren indexikalischen Informationen zusammen abgespeichert werden. Diese zwei unterschiedlichen Sprachwahrnehmungsmodelle sollen im Folgenden vorgestellt werden.

2.3.1. Abstraktionistische Sprachwahrnehmungsmodelle

Die *abstraktionistischen* Sprachwahrnehmungsmodelle gehen von einer möglichst frühen Abstraktion von den akustischen Merkmalen eines Sprachsignals auf invariante Einheiten (z.B. Phoneme) aus. Bei diesem Prozess werden alle Informationen, die nicht zur Phonem- oder Worterkennung dienen, verworfen. Zwecks eines Überblicks sollen hier exemplarisch einige abstraktionistische Sprachwahrnehmungsmodelle vorgestellt werden.
Eines der bekanntesten Modelle, ist das sogenannte „TRACE"-Modell von [184]. Laut diesem Modell werden physikalische akustische Merkmale, phonemische und semantische Informationen verwendet, um das gehörte Wort auf eine Wort-Repräsentation im mentalen Lexikon abzubilden. Das Modell postuliert drei verschiedene Ebenen der Spracherkennung: auditorische Merkmale, Phoneme und Wörter. Die auditorischen Merkmale, die wir mit dem Hören der Wörter aufnehmen, werden bestimmten Phonemen zugeordnet und die Phoneme dann wiederum bestimmen Wörtern. Dabei können die Informationen der verschiedenen Ebenen miteinander interagieren, sprich Informationen der höheren Ebenen können sich auch auf untere Ebenen auswirken. Insgesamt können sich die Informationen in drei verschiedenen Richtungen bewegen: *auditorische Merkmale –> Phoneme –> Wörter*, *Wörter –> Phoneme –> auditorische Merkmale* und können *innerhalb jedes Levels untersucht werden*. Hört man zum Beispiel Stimmhaftigkeit und Friktionsgeräusche und schließt auf die Phoneme /b/ und /r/, vermutet man vielleicht das Wort „Brezel". Hört man die Phoneme aber im Kontext des Satzes „Ich sehe nichts ohne meine Br...", kann diese Information dazu verwendet werden, das Wort „Brezel" zu verwerfen und das Wort als „Brille" zu erkennen. Da stets die akustischen Merkmale auf eine bestimmte Phonemkategorie abgebildet werden ohne individuelle Merkmale des Sprechers mit einzubeziehen, werden diese im Prozess heraus gefiltert. Sind sie auch in den akustischen Merkmalen noch vorhanden, so werden sie bei der Abbildung der Merkmale auf konkrete Phoneme verworfen, sodass sie im weiteren Spracherkennungsprozess keine Rolle mehr spielen.
Während die „Top-Down"-Prozesse im Modell dabei helfen, den semantischen und syntaktischen Kontext einzubeziehen, sind ihre Effekte manchmal zu stark. Ein zu starker Einfluss von der Wortebene auf die Phonemebene kann die Erkennung von falschen Aussprachevarianten verhindern. Außerdem kann theoretisch jeder Teil im Sprachsignal ein neues Wort

beginnen, weil das Modell rhythmische Strukturen ignoriert, welche die Wortsegmentation beeinflussen [55].
Da das TRACE-Modell Ergebnisse neuerer Studien nicht erklären konnte, schlug [212] das „Shortlist"-Modell als Alternative vor. Es ist im Gegensatz zu dem interaktiven TRACE-Modell ein modulares Modell mit einem reinen „Bottom-up"-Ansatz. „Top-Down"-Informationen von der Wort- auf die Phonemebene seien redundant, da alle wichtigen lexikalischen Einschränkungen auch vollständig innerhalb einer Ebene operieren könnten. Das Shortlist-Modell nimmt eine Abfolge von Phonemen als Input und erstellt auf dieser Basis eine „Shortlist" mit Wortkandidaten, welche miteinander konkurrieren. Die Liste der Wörter wird permanent erneuert mit allen neuen phonemischen Informationen, die eintreffen. Nicht mehr passende Kandidaten werden aussortiert. Der Satzkontext beeinflusst die lexikalische Verarbeitung während dieses Selektionsprozesses. Durch diesen, erst spät erfolgenden, Einfluss von Kontextinformationen, wird ihnen ein geringes Gewicht beigemessen, wohingegen den Bottom-Up-Informationen eine hohe Bedeutung beigemessen wird. Passt der kontextuell geeignete Kandidat nicht mehr zu den eintreffenden akustischen Informationen, wird seine Aktivierung vermindert. Das Shortlist-Modell ist in der Lage, das gleiche Ergebnis wie das TRACE-Model zu produzieren ohne „Top-Down"-Prozesse zu verwenden.
[214] festigen ihren Standpunkt, dass „Top-Down"-Feedback unnötig ist, im „Merge"-Modell, welches phonemische Entscheidungen treffen soll. Dabei fließen prälexikalische Informationen ohne Feedback in das Lexikon. Da phonemische Entscheidungen auf der Verschmelzung von prälexikalischer und lexikalischer Informationen beruhen, sagt das Merge-Modell die lexikalische Einbindung in phonemischen Entscheidungen sowohl in Wörtern als auch Unsinn-Wörtern voraus. Dabei nutzt das Modell die Kompetition zwischen verschiedenen lexikalischen Hypothesen. [214] stützen die Gültigkeit ihres Modells mit Computer-Simulationen, welche die gute Eignung von modularen Modellen für die Spracherkennung bestätigen.
Eine Studie von [181] mit einem realistischeren, größeren Lexikon an Wörtern zeigte aber, dass „Top-Down"-Feedback sehr wohl die Worterkennung beschleunigen kann; besonders im Fall von störenden Hintergrundgeräuschen. Es gibt viele widersprüchliche Studien zu der Frage, ob „Top-Down"-Informationen für die Worterkennung hilfreich und notwendig sind [185], ober ob sie überflüssig bzw. sogar hinderlich sind ([216] [192]). Allerdings konnte das TRACE-Modell bereits erfolgreich für unterschiedliche Phänomene in der Sprach- und Worterkennung angewendet werden (z.B. [4] [57] [58]).

Ein aktuelles abstraktionistisches Sprachwahrnehmungmodell wurde von [107] vorgeschlagen als eine optimierte Version des TRACE-Modells. Es besteht aus einer zeit-spezifischen Phonemebene, einer zeit-invarianten Zeichenketten-Ebene (time-invariant string kernel level (TISK)) und einer zeit-invarianten Wort-Ebene. Das TISK-Modell ist ebenfalls interaktiv und erreicht die gleiche Leistung wie das TRACE-Modell bei weit weniger Rechenaufwand [107].
Den *abstraktionistischen* Modellen ist gemeinsam, dass sie alle von prälexilalischen Repräsentationen ausgehen, die befreit sind von subphonetischer Variation und allen Informationen, die nicht für eine kategoriale Unterscheidung benötigt werden. So gibt es meist einen Prozess der „Sprechernormierung“, in welchem akustische Merkmale normiert werden, sodass nur ein Set an Merkmalen für jede Kategorie benötigt wird; unabhängig vom Sprecher. Zum Beispiel variieren Formantfrequenzen systematisch mit dem Sprechergeschlecht ([229] [115]) und können mit Hilfe der Werte anderer Formanten normiert werden. Allen abstraktionistischen Modellen ist gemeinsam, dass sie ein minimales (einzelnes) Set an Merkmal-Kategorie-Abbildungen anstreben. Sie sind informatorisch sehr effizient, da sie alle Variabilität zwischen Situationen in einem kompakten Set von sublexikalischen Merkmal-Kategorie-Abbildungen vereinigen, was eine effiziente Generalisierung zwischen lexikalischen Einheiten ermöglicht. Allerdings ist die Normierung schwierig (manchmal unmöglich), da viel der Variabilität nicht aufgrund von festen (z.B. physiologischen) Faktoren verursacht wird, sondern eher von stilistischen, welche erlernt werden müssen [150].

2.3.2. exemplartheoretische Sprachwahrnehmungsmodelle

Die *episodischen* oder *exemplartheoretischen* Modelle wurden ursprünglich in der Psychologie zur Modellierung von Perzeption und Kategorisierung eingeführt. Inzwischen haben sie auch Einzug in die Sprachwahrnehmung gehalten, wo sie besonders die Effekte von akustischen Details wie bestimmten Sprechermerkmalen auf Sprachverarbeitungsprozesse erklären sollen. Einige wichtige Modelle sollen hier kurz vorgestellt werden.
[260] stellten als erste eine Gedächtnistheorie auf, die man in heutiger Ausdrucksweise als episodische (oder Exemplar-) Theorie bezeichnen würde. Die Theorie nahm an, dass jede Erfahrung, zum Beispiel die Wahrnehmung eines gesprochenen Wortes, eine einzigartige Spur (einen Eintrag) im Gedächtnis hinterlässt. Bei der Wahrnehmung eines neuen Wortes werden alle Einträge entsprechend ihrer Ähnlichkeit zu dem Stimulus aktiviert. Der am stärksten aktivierte Eintrag verbindet das neue Wort mit dem gespeicherten Wissen, was die Erkennung des Wortes bedeutet. Die Herausforderung bestand aber darin, eine Abstrak-

tion aus einer Kollektion von idiosynkratischen Einträgen zu erzeugen. Eine Lösung für dieses Problem lieferte [97] mit seiner Idee der gemischten Erinnerung. Eine solche Erinnerung resultiert aus der simultanen Aktivierung unterschiedlicher Gehirnbereiche. Diese Idee floss in die erste episodische Theorie mit ein, unter der Annahme, dass Abstraktion während des Abrufens von zahllosen, teilweise redundanten Einträgen entsteht. Während viele Sprachwahrnehmungstheorien davon ausgehen, dass es eine bestimmte Anzahl an kanonischen Repräsentationen gibt, auf die irgendwie durch ein variables, geräuschvolles Signal zugegriffen wird; schlägt [103] ein exemplarbasiertes Modell in Anlehnung an [97] und [260] vor mit speziellem Blick auf das mentale Lexikon. Er zeigte durch die Anwendung des „MINERVA 2“-Modells [117] auf experimentelle Sprachdaten (aus [102]), dass sich die Modellvorhersagen und die Experimentdaten sehr gut deckten [103].
Mit besonderer Berücksichtigung der zeitlichen Komponente der Sprache, präsentiert [133] ein weiteres exemplarbasiertes Modell der Sprachwahrnehmung. In den drei Ebenen des Modells wird zunächst das Sprachsignal in eine Sequenz von auditorischen Spektren umgewandelt, deren Ähnlichkeitsgrad zu den gespeicherten Spektren gemessen wird. Im Falle, dass das eintreffende Spektrum keinem der abgespeicherten (ausreichend) ähnelt, wird es separat als neuer Eintrag gespeichert. Während die Architektur des Modells eher dem „ALCOVE“-Modell („Attention Learning Covering Map“) von [158] ähnelt, ist die Nutzung von Exemplaren eher an das „GCM“ („Generalized Context Model“) von [217] angelehnt. Aufgrund der detaillierteren akustischen Informationen können Wörter früher und besser unterschieden werden als in abstraktionistischen Modellen wie TRACE [184] und Shortlist [212]. [133] schließt mit der Vermutung, dass abstrakte phonologische Strukturen nur im Moment der Worterkennung auftauchen und gleich wieder verschwinden; dass sie nur dadurch zustande kommen, dass Teilmengen von lexikalischen Einheiten, die eine abstrakte Entität ausmachen, implizit durch auditorische/perzeptuelle Gemeinsamkeiten verbunden werden. Außerdem ist ein exemplarbasiertes Modell in der Lage, Eigenschaften des Sprachwahrnehmungsprozesses von menschlichen Hörern nachzubilden. So können Hörer Wörter schneller erkennen, wenn der Sprecher möglichst prototypisch für sein Geschlecht ist. Bei einem nicht-prototypischen Sprecher verlängert sich die Erkennungszeit. Den gleichen Effekt zeigte auch ein Exemplar-Resonanz-Modell von [69], was für das hohe Potential von Exemplar-Modellen (mit einem Resonanz- oder ablaufinvarianten Mechanismus) spricht [135].
[231] schlägt ein Sprachwahrnehmungs- und Produktionsmodell mit einer Langzeit-Repräsentation von Wörtern, die mehr phonetische Details enthalten, vor. Mittels Exemplar-Theorie wird

implizites Wissen über die Wahrscheinlichkeitsverteilungen von phonologischen Elementen und Wort-spezifischen phonetischen Mustern modelliert. Die Produktionsziele der spezifischen phonologischen Elemente werden durch die stärkere Aktivierung von Exemplaren, die mit dem aktuellen Wort assoziiert sind, bestimmt. Sowohl in der Sprachproduktion als auch in der Wahrnehmung werden drei Ebenen der kognitiven Verarbeitung angenommen: Lexeme, phonologische Kodierung und quantitatives Wissen über phonetische Resultate. Auf diese Weise strukturierte Modelle sind in der Lage, die Produktivität der Sprachverarbeitung und die Existenz von allgemeinen Effekten zu erfassen. Allerdings spielen allophonische Details, die systematisch mit Wörtern variieren, eine größere Rolle als bisher angenommen. Einige dieser Effekte lassen sich nur dadurch erklären, dass man eine Assoziation von einzelnen Wörtern mit phonetischen Verteilungen annimmt. Diese Ergebnisse können mit Hilfe von exemplarbasierten (Sprachproduktions-)Modellen integriert werden. Individuelle Wörter können das Set an Exemplaren bestimmen, die als Produktionsziel dienen. Dabei wird das phonetische Resultat von Faktoren wie persönlicher Identität, kontextuellen und sozio-stilistischen Einflüssen geprägt.
Auch der Spracherwerb von Kindern lässt sich durch ein exemplartheoretisches Modell erklären. Dafür schlägt [232] stellungsabhängige Varianten, statt der sonst oft angenommenen Phoneme, für die Repräsentationsebene vor. Ihre Ergebnisse unterstützen die Annahme, dass der Spracherwerb durch zwei Zusammenflüsse an Informationen ermöglicht wird. Einmal durch die Beziehung von Oberflächen-Statistiken zu positionaler Kontrastivität in Wörtern, was es Kindern ermöglicht, seltene Transitionen als Wortgrenzen anzunehmen. Und zweitens die Beziehung von akustischen Merkmalen zu Diphon-Statistiken, was die phonetischen Möglichkeiten mit der Komplexität von Grammatik und Lexikongröße verbindet.
Auch wenn die Exemplar-Modelle in ihrer genauen Architektur voneinander unterscheiden, so ist ihnen doch eine einheitliche Idee gemeinsam. Die episodischen, exemplartheoretischen Modelle postulieren, dass Sprachwahrnehmung eine direkte Abbildung von detaillierten, akustischen Informationen auf linguistische Einheiten ist, wobei jede akustische Variation und alle indexikalischen Informationen erhalten bleiben. Es gibt keinen Sprechernormierungsprozess, da alle akustischen Informationen, zum Beispiel eines Wortes oder eines Vokals, separat abgespeichert werden und der Erkennungsprozess auf Basis einer ganzheitlichen akustischen Ähnlichkeit abläuft. Durch das Abspeichern von ausreichend großen Einheiten (z.B. Wörtern) reichen die akustischen Informationen in jedem Exemplar aus, um einzelne Sprecher unterscheiden zu können. Für unbekannte Sprecher werden neue Exemplare durch den Vergleich mit allen anderen Exemplaren von ähnlichen Sprechern erkannt. Die-

ser Prozess erlaubt Generalisierung zwischen verschiedenen Sprechern. Exemplar-Modelle streben ein maximales Set an Merkmal-Kategorie-Abbildungen an, nämlich eine für jede Beobachtung. Dadurch sind sie unendlich flexibel und können jedes informative Set an Merkmalen aufnehmen, einfach weil sie alle berücksichtigen. Weil sie aber jedes akustische Detail aufnehmen, können sie nur schwierig auf unbekannte Wörter generalisieren. Außerdem brauchen sie einen zusätzlichen Mechanismus, um Exemplare auf der Basis von indexikalischen Informationen zu markieren oder zu gewichten [150].

3. Akustische Analyse ausgewählter Konsonanten

3.1. Akustische Sprechermerkmale in Konsonanten und Vokalen

In der Einführung wurde beschrieben, welche akustischen Besonderheiten die einzelnen Vokale und Konsonanten charakterisieren. Aufgrund dieser einzigartigen Struktur besitzt jeder Laut spezifische akustische Merkmale. Diese Merkmale enthalten neben Informationen über die Identität des Lautes auch Informationen über die Identität des Sprechers. Einige dieser akustischen Merkmale sollen im Folgenden vorgestellt und diskutiert werden.

3.1.1. Akustische Sprechermerkmale in Vokalen

Im Fall der Vokale wurden hauptsächlich ihre Formanten für die Sprechererkennung verwendet. Mehrere Studien zeigten deren hohes Potential, Sprecher aufgrund der enthaltenen akustischen Merkmale unterscheiden zu können ([177] [187] [255]).
[177] untersuchte die Sprecherspezifität des zweiten und dritten Formanten (F2 und F3) von 11 Australischen Monophtongen. Es zeigte sich, dass vor allem F2 und F3 des vorderen Vokals /ɪ/, aber auch der anderen vorderen Vokale /ɪ, ɛ, æ/, sprecherspezifisch sind. Vermutlich resultiert dieses Ergebnis aus der Lage der Formanten in höheren spektralen Regionen. Außerdem werden in der Produktion von vorderen hohen Vokalen mehr individuelle Strategien vermutet. Weiterhin zeigten sich die ungespannten Vokale als sprecherspezifischer als die gespannten, was wahrscheinlich an deren kürzerer Dauer und dem damit einhergehenden geringeren Raum für Intra-Sprecher-Variabilität liegt [177]. Aber nicht nur die vorderen Vokale können Sprecherinformationen enthalten. [187] untersuchten die dynamischen Formantmerkmale des Britischen /uː/ und zeigten, dass vor allem F2 (aber auch F1) viele Informationen zur Sprecheridentität lieferte. Die guten Ergebnisse erklären

sie dadurch, dass die dynamischen Merkmale sowohl anatomische Unterschiede der Sprecher in ihrem Vokaltrakt als auch individuelle artikulatorische Strategien einfangen. [255] zeigte schließlich, dass sich die akustischen Merkmale von australischen Vokalen auch zur automatischen Sprecherdiskrimination eigneten. Die Merkmale von fünf gespannten und sechs ungespannten Vokalen von 171 Sprechern konnten fusioniert die Sprecher mit einer Fehlerrate von 10 % bis 14 % richtig unterscheiden.
Im Gegensatz zu Einzelvokalen enthalten Vokal-Diphtonge noch mehr dynamische Formantinformationen. [186] zeigten, dass dynamische Formantmerkmale von F1, F2 und F3 des Australischen Diphtongs /aɪ/ fünf männliche Sprecher zu 88 - 95 % richtig klassifizieren konnten.
Vokalformanten transportieren unter guten Bedingungen viele Sprecherinformationen. Verschiedene Untersuchungen zeigten jedoch ihre hohe Abhängigkeit von der Qualität der Sprachaufnahmen. Waren die Aufnahmen über Telefon oder Handy entstanden, wurden die Formanten auf unberechenbare Weise verändert. Der verwendete Algorithmus sowie die Einstellungen bei der Formantanalyse hatten einen signifikanten Einfluss auf die Ergebnisse [159] [42] [110]. Weiterhin ist eine erfolgreiche Sprecheridentifikation abhängig von der Länge der Vokalsegmente [73], wobei eine sichere Identifikation bei kürzeren Abschnitten schwieriger wird.

3.1.2. Akustische Sprechermerkmale in Konsonanten

Neben den Vokalen enthalten auch die Konsonanten in ihren akustischen Merkmalen Sprecherinformationen. Aufgrund der unterschiedlichen akustischen Besonderheiten der Konsonantenklassen sollen diese hier getrennt voneinander betrachtet werden.

Nasale

Durch die Zuschaltung des Nasenraums bei der Artikulation von Nasalen ergeben sich sprecherunterscheidende Faktoren [265]. Eine Studie von [278] zeigte mit Hilfe eines 3-D-Modells des Nasenraums, wie sich die Anatomie der verschiedenen Nasen- und Nebennasenräume auf die akustischen Merkmale von nasalen Konsonanten auswirkt. Die charakteristischen Pol- und Nullstellen im Spektrum der Nasale wurden sowohl durch die Asymmetrien zwischen dem linken und rechten Nasenraum ausgelöst, als auch durch seinen komplizierten Aufbau [278].
Ein besonderer Vorteil des Nasenraums eines Sprechers besteht darin, dass er nicht wil-

lentlich verändert werden kann. Im Gegensatz zu dem plastischen und leichtverformbaren Mund- und Rachenraum ist der Nasenraum eher fest. Die einzig beweglichen Teile sind die Nasenflügel. Aufgrund dieser Eigenschaften könnten nasale Konsonanten wie /m, n/ beständige Merkmale zur Sprecheridentifikation enthalten. Dank der relativen Festigkeit des Nasenraums ist keine große Intra-Sprecher-Variabilität möglich und wegen der komplexen Struktur des Naseninnenraums sollte die Inter-Sprecher-Variabilität hoch sein. Somit erfüllen Nasale die wichtigsten Parameter zur Sprecheridentifikation [254]. Externe Einflüsse, wie zum Beispiel eine Erkältung, beeinflussen den Nasenraum nur temporär.
Bereits in den 60er und 70er Jahren gab es Untersuchungen zur Diskriminationsfähigkeit von deutschen Phonemen, in denen Nasale sehr gut abschnitten (z.B. [101] [295]). [101] sahen sie sogar als die besten Segmente zur Sprecheridentifikation, besonders den alveolaren Nasal /n/. Sie wandten die Spektralmerkmale der Nasale in einem automatischen Sprecherverifikationssystem an und erhielten sehr gute Erkennungsraten von bis zu 97 % [208]. Allerdings wurden in diesen Untersuchungen isoliert gesprochene Laute oder Segmente aus gelesener Sprache verwendet, sodass sich die Ergebnisse nicht auf Spontansprache übertragen lassen. [237] analysierten die akustischen Merkmale von nasalen Konsonanten und nasalierten Vokalen und überprüften ihre Relevanz für die Sprecheridentifikation. Demnach eignen sich nasale Konsonanten besser als nasale Vokale für diese Aufgabe, da ihre akustischen Eigenschaften stabiler sind. Die Lautmerkmale der nasalierten Vokale könnten aber zur Erkennung einer hyper- oder hyponasalen Sprechweise fungieren [237]. Bei der Verwendung von phonetischen Merkmalen zur Sprecheridentifikation ist es sinnvoll, Phoneme mit einer hohen Informationsdichte zu verwenden, da dann eine geringere Anzahl von Segmenten benötigt wird. [16] stellten fest, dass die Auswahl der Phoneme wichtiger ist als die Quantität des Sprachmaterials. Wurden Vokale für die Sprechererkennung verwendet, ergab sich die höchste Diskriminationsrate. Das lag aber auch daran, dass die Vokale den größten Anteil (40 %) im Sprachmaterial ausmachten. Verwendeten sie von allen Lautklassen gleich viel Sprachmaterial, so konnten die Sprecher am besten durch Nasale unterschieden werden. Die zeigt, dass besonders bei geringen Sprachdatenmengen Nasale sehr viele Informationen zur Sprecheridentität liefern können.
In aktuelleren Untersuchungen wurden einige Merkmale der Nasale, wie Dauer, Formanten [142], Pol- und Nullstellen [72] untersucht. Die Studien lieferten gute Ergebnisse und bilden die Basis zur Identifikation weiterer sprecherspezifischer Merkmale der Nasale.

Frikative

Mehrere Untersuchungen [206] [105] [266] zeigten, dass Sprecher Frikative mit verschiedenen akustischen Parametern, wie Variationsbreite, Artikulationsstelle, Konstriktionslänge und Zungenhöhe realisieren. Diese unterschiedlichen Artikulationsstrategien machen sich höchst wahrscheinlich auch im akustischen Sprachsignal bemerkbar, sodass eine relativ hohe Inter-Sprecher-Variabilität zu erwarten wäre. Tatsächlich bestätigten bereits einige Untersuchungen diese Annahme. In der Studie von [68] eigneten sich Merkmale aus Cepstren von Frikativen mit einer Fehlerrate von 29 % (nach den Nasalen und Vokalen) am drittbesten für die (automatische) Sprecherdiskrimination. Dabei stellte sich besonders der stimmlose alveolare Frikativ /s/ als sehr informativ für die Sprecheridentität heraus. Gleiches zeigte auch [143] für die spektralen Momente von /s/. Auch im Versuch von [15] wurde mittels der Frikative eine Sprecheridentifikationsrate von 44,6 % erreicht, womit sie abermals hinter den Vokalen und Nasalen die höchsten Werte erbrachten.
Auch in der Lautklasse der Frikative scheint es deutliche Unterschiede in der Sprecherspezifität ihrer akustischen Merkmale zu geben. Dabei scheint die Artikulationsstelle des Frikatives in starkem Maße seinen Informationsgehalt zu beeinflussen. Vor allem der alveolare Frikativ /s/ zeigte sich in vielen Studien als sehr geeignet zur Sprecheridentifikation. Aber auch der post-alveolare Frikativ /ʃ/ erwies sich als sehr sprecherspezifisch, sodass mit seiner Hilfe sogar Sprecher, die ihre Stimme verstellten, unterschieden werden konnten. Aber nicht nur die alveolaren Frikative bieten ein hohes Potential an Sprecherinformationen. [203] zeigte, dass die labiodentalen Frikative eine große Inter-Sprecher-Variabilität in ihrer Artikulation aufweisen, welche sich auch im akustischen Sprachsignal bemerkbar machen sollte. Daher lohnt ein genauerer Blick auf die Lautgruppe der Frikative auf der Suche nach sprecherspezifischen akustischen Merkmalen.

Plosive

Plosive erwiesen sich in der Sprechererkennung häufig als die Laute mit den wenigsten Informationen über die Identität des Sprechers ([68] [15], ect.). Begründet liegt das möglicherweise in ihrer sehr kurzen Dauer, der Verschlussphase, wo keine akustischen Informationen vorhanden sind und dem sehr kurzen Burst, der wenig Raum für sprecherindividuelle Merkmale bietet.
[141] testete verschiedene Laute aus Äußerungen zur Identifikation von Sprechern. Die Plosive /t, b/ schnitten dabei am schlechtesten ab; interessanterweise gleichauf mit dem Frikativ /s/. Auch [41] stellten fest, dass sich von allen untersuchten Lauten (/x, r, s, p/) der Plosiv

am schlechtesten zur Sprecherunterscheidung eignet. In der Untersuchung von [68] konnte mit Hilfe von cepstralen Plosivmerkmalen nur eine geringe Sprecherdiskriminationsrate mit einer Fehlerrate von 38,1 % erreicht werden. Damit schnitten die Plosive am schlechtesten von allen untersuchten Lautkategorien ab. Auch in der Studie von [15] wurde mit Plosiven eine Identifikationsrate von nur 10,47 % erzielt. Die Untersuchung von [76] mit Sprachdaten des Verbmobil-Korpus' zeigte den Plosiv /p/ in Punkto Sprecherspezifität ebenfalls auf dem letzten Platz. Auch in einer Studie von [114] erwiesen sich die Plosive ebenfalls als wenig hilfreich für die Sprechererkennung. In einem mit Phonemmerkmalen trainierten Modell zur Sprecherklassifikation ließen die Plosive nur eine Fehlerrate von 30,9 % zu und waren damit schlechter geeignet als alle anderen Laute [140].
Die Untersuchungen von [82] zeigten allerdings, dass sich auch Plosive - besonders der alveolare Plosiv /t/ - ebenfalls zur Sprecherdiskrimination eignen können. Dieses Ergebnis steht allerdings im Gegensatz zu den meisten anderen Studien.

3.1.3. Ziele und Hypothesen

In der Sprechererkennung - sowohl der menschlichen als auch der maschinellen - erwiesen sich akustische Merkmale von Vokalen als besonders geeignet ([162] [27] [177] [222] [187] [255] [289] [268]). Allerdings tragen auch Konsonanten relevante Informationen über die Identität eines Sprechers, wobei die Lautgruppen der Nasale und Frikative herausragen ([101] [27] [12] [142] [143]). Da Konsonanten bisher weniger gut auf ihren Gehalt an Sprecherinformationen untersucht wurden, soll der Fokus der Arbeit auf dieser Lautgruppe liegen. Außerdem wurden in Studien, die Konsonanten bereits untersuchten, hauptsächlich englisch-sprachliche Laute betrachtet ([142] [143]). Obwohl die deutsche Sprache eine große Vielfalt an Frikativen (/f, v, s, z, ʃ, ç, x, h/) bietet, gab es bisher nur sehr wenige Untersuchungen zu den Konsonanten des Deutschen (z.B. [76] [179]). Daher ist es sinnvoll, diese Lautgruppe bei der Suche nach geeigneten Sprechermerkmalen ins Auge zu fassen.
In dieser Arbeit sollen die deutschen Konsonanten auf ihren Informationsgehalt zur Sprecheridentität untersucht werden. Dabei soll der Fokus vor allem auf den Lauten liegen, die sich in der Vergangenheit bereits als hilfreich zur Sprecheridentifikation erwiesen haben, nämlich den Nasalen und Frikativen. Die Plosive sollen dabei als die am wenigsten geeigneten Konsonanten teilweise als Referenzpunkt zum Vergleich der Leistungsfähigkeit anderer Konsonanten dienen. Unter diesem Aspekt sollen auch die Vokale als die am bestgeeignetsten Laute zur Sprecheridentifikation kurz betrachtet werden. Allerdings wird auf diesen Lautgruppen nicht das Hauptaugenmerk dieser Untersuchung liegen.

Da den meisten Studien englisches Sprachmaterial zugrunde lag, wurden die hinteren Frikative /ç, x, h/ ebenso wie die stimmhaften Frikative in den bisherigen Studien nur selten betrachtet. Die hinteren Frikative könnten aber aufgrund ihrer längeren Passage durch den Vokaltrakt des Sprechers viele akustische Informationen über dessen Form enthalten. Die stimmhaften Frikative /v, z/ könnten durch ihre zusätzliche Periodizität Informationen über die Grundfrequenz des Sprechers enthalten. Allerdings haben sie durch den langsameren Luftstrom auch weniger Energie (siehe Abschnitt 4.2.2), was ihrer Sprecherspezifität entgegensteht.
Die Artikulationsstelle der Frikative lässt sich durch ihre spektralen Merkmale gut identifizieren [93], d.h. ihr parametrisiertes Spektrum in Form der spektralen Momente enthält viele Informationen über den Laut. Da anzunehmen ist, dass Bereiche mit einer hohen Dichte an akustischen Informationen sowohl phonetische als auch Sprecherinformationen enthalten, ist es nicht abwegig, die gleichen Parameter wie zur Analyse phonetischer Merkmale auch zur Analyse der Sprechermerkmale zu verwenden. In der Tat wurde im Bereich der automatischen Sprechererkennung bereits gezeigt, dass die Mel-Frequency Cepstral Coefficients (MFCC), die schon lange in der Spracherkennung Anwendung finden, sich auch zur automatischen Sprechererkennung sehr gut eignen ([148] [149]). In der Untersuchung von [143] wurde auch bereits gezeigt, dass sich die spektralen Momente von /s/ zwischen Sprechern relativ stark unterscheiden. Daher sollen als akustische Merkmale die ersten vier spektralen Momente (Schwerpunkt, Varianz, Schiefe, Wölbung) verwendet werden.
Für die Nasale wurden bisher meist Formanten und Antiformanten bzw. Pol- und Nullstellen als charakteristische Merkmale analysiert ([142] [72]). Nasale werden insgesamt von ihrem niedrigen Spektrum dominiert, was aber je nach der Artikulationsstelle des Nasals differiert. Da anzunehmen ist, dass sich diese spektralen Unterschiede auch in den spektralen Momenten manifestieren, sollen diese näher untersucht werden.
Wie bereits erwähnt, sollen die Ergebnisse mit der Sprecherspezifität von Vokalen verglichen werden. Dazu sollen zwei Varianten zum Einsatz kommen. (1) Einerseits sollen konventionelle Formantwerte als charakteristische Vokalmerkmale verwendet werden. (2) Andererseits sollen für die Vokale ebenfalls die spektralen Momente gemessen werden. Auch wenn es keine Beweise gibt, dass sich Hörer bei der Vokalwahrnehmung an den Spektralmomenten orientieren, sollen diese Parameter hier Zwecks einer besseren Vergleichbarkeit mit den Konsonanten zusätzlich analysiert werden.
Weiterhin soll untersucht werden, wie sich eine Qualitätsminderung durch die Übertragung des Sprachsignals über den Telefonkanal auf die akustischen Merkmale der Konsonan-

ten auswirkt. Es soll ermittelt werden, bei welchen Lauten die Sprecherspezifität vom eingeschränkten Frequenzgang am wenigsten reduziert wird.

3.1.4. Methode

Die Verbmobil-Sprachdatenbank

Das Verbmobil-Korpus ist eine Dialog-Sprachdatenbank mit einer Sammlung von deutschen, amerikanischen und japanischen Dialogen. Die von 1993 bis 1996 aufgenommenen Daten umfassen insgesamt 885 Sprecher mit 1422 Aufnahmen. Der Korpus wurde in einer Zusammenarbeit des Instituts für Phonetik und Sprachliche Kommunikation in München mit dem Institut für Kommunikationsforschung und Phonetik in Bonn und dem Institut für Phonetik und digitale Sprachverarbeitung in Kiel erstellt. Die Sprecher der Dialoge wurden stets in separaten Räumen des Tonstudios aufgezeichnet. Die Abtastrate beträgt 16000 Hz. Das Ziel des Dialogs der Sprecher war immer, ein Meeting oder eine Reise zu planen [113].

Auswahl der Sprachdaten

Als Sprachdaten wurden die annähernd spontanen Dialoge der Sprecher ausgewählt, da sie der natürlichen Sprache am nächsten kommen. Der Verbmobil-Korpus wurde ausgewählt, weil er neben den hochqualitativen Studioaufnahmen per Mikrofon auch simultane Aufnahmen über Telefon enthält. Auf diese Weise kann später auch der Einfluss des Telefonkanals auf die Sprecherspezifität der akustischen Merkmale untersucht werden ohne Aufnahmen aus verschiedenen Sessions miteinander vergleichen zu müssen, was zusätzliche Variation verursachen könnte. Da für die forensische Anwendung vor allem Männerstimmen relevant sind, wurden zunächst die Dialoge der 49 männlichen Sprecher verwendet.

Die spektralen Momente

Die spektralen Momente stellen eine Parametrisierung des Spektrums dar (z.B. [91]). Ursprünglich stammt die Methode aus dem Bereich statistischer Verteilungen, wo manchmal die Momente zur Beschreibung eines Histogramms angewendet werden. Dabei ist x das Intervall einer Histogrammklasse, f der Zähler für die Anzahl der Zeichen einer Intervallklasse und i der Index für die Momente m_i ($i = 1, 2, 3, 4$). Somit können die spektralen Momente auf folgende Weise berechnet werden:

$$m_1 = \frac{\sum x \cdot f}{\sum x} \tag{3.1}$$

$$m_2 = \frac{\sum x(f - m_1)^2}{\sum x} \tag{3.2}$$

$$m_3 = (\frac{\sum x(f - m_1)^3}{\sum x}) \cdot m_2^{-1.5} \tag{3.3}$$

$$m_4 = ((\frac{\sum x(f - m_1)^4}{\sum x}) \cdot m_2^{-2}) - 3 \tag{3.4}$$

Bei der Berechnung von *spektralen* Momenten, wird das Spektrum als ein Histogramm betrachtet, als eine Verteilung der Amplituden über die verschiedenen Frequenzen. Daher wird x das Frequenzintervall und f der Amplitudenwert der gegebenen Frequenz in dB [108]. Das erste spektrale Moment m_1 beschreibt den Frequenzschwerpunkt bzw. den Mittelwert des Spektrums eines Lautes, d.h. die Frequenz der größten Energie (der größten Amplitude). Das zweite Moment m_2 beschreibt die Varianz der Energie und drückt aus, wie stark die Energie im Spektrum verteilt ist [108]. Das dritte Moment m_3 bringt die Schiefe zum Ausdruck, d.h. ob das Spektrum asymmetrisch ist. Wenn die Werte der Verteilung vom Mittelwert (Frequenzschwerpunkt) weiter in Richtung der hohen Werte reicht als zu den niedrigen, dann ergibt sich für die Schiefe ein positiver Wert. Das vierte Moment m_4 stellt die Wölbung des Spektrums in Bezug zur Normalverteilung dar, welche eine Wölbung von 3 hat. In einigen Formeln wird aber auch der Wert 3 substrahiert, damit 0 der Wölbung der Normalverteilung entspricht. Ist die Energieverteilung stärker gewölbt als die Normalverteilung, wird der Wert positiv (bzw. > 3) und im anderen Fall negativ (bzw. < 3) ([205]).

Durch die Bandlimitierung von Spektren ist das dritte spektrale Moment m_3 mit dem ersten Moment m_1 korreliert. Außerdem sind in der Regel m_2 und m_4 korreliert, auch wenn dies nicht zwangsläufig der Fall ist [108].

In vielen Studien (besonders zur automatischen Sprechererkennung) wurden erfolgreich kepstrale Koeffzienten (z.B. MFCC) verwendet [148] [111] [284]). Allerdings ist gerade bei höheren Koeffizienten häufig unklar, in welcher Relation dieses akustische Merkmal zur Artikulation des Lautes steht. Gerade mit Hinblick auf die forensische Anwendung ist es wichtig, ein einfaches, aber dennoch wirkungsvolles Maß zur Sprecherunterscheidung zu finden. So werden häufig schlechtere Ergebnisse in Kauf genommen, zugunsten der besseren Erklärbarkeit (z.B. vor Gericht) [254]. In dieser Arbeit soll versucht werden, akustische Sprechermerkmale zu finden, deren Zusammenhang zur Artikulation verständlich ist und die trotzdem eine hohe Sprecherdiskriminationsfähigkeit besitzen.

Spektrale Momente können vielfältige akustische Merkmale einfachen [281]. So können sie sowohl Frikative voneinander unterscheiden ([137]) als auch Merkmale des Sprechers (z.B. Geschlecht) ([92]). Außerdem lassen sie sich prinzipiell auf alle Laute anwenden (auch wenn sie für Frikative besonders charakterisierend sind), im Gegensatz zu beispielsweise Formanten, welche nur bei stimmhaften Lauten gemessen werden können. Daher sollen in dieser Arbeit spektrale Momente auf ihre Sprecherspezifität untersucht werden.

Analyse der Sprachdaten

Zunächst wurden die Sprachaufnahmen aller männlichen Sprecher sowohl über Mikrophon als auch über Mobiltelefon ausgewählt. Daraus ergaben sich insgesamt 49 männliche Sprecher, die alle aus der Münchner Umgebung stammten. Insgesamt liegen damit 4156 Äußerungen vor, die 34814 Nasale, 46955 Frikative und 81584 Vokale enthielten. Nicht jeder Sprecher hat die gleichen Äußerungen gesprochen, jedoch ist die Anzahl der Laute ähnlich. Da nicht alle untersuchten Laute gleichhäufig in der deutschen Sprache vorkommen, gibt es Unterschiede in der Anzahl der zur Verfügung stehenden Laute. Dieses Phänomen könnte man nur vermeiden, wenn man neue Sprachaufnahmen machen würde, bei denen man die Häufigkeit der Laute kontrolliert. Allerdings wäre das für spontansprachliche Äußerungen unmöglich.
Die Annotation der Daten erfolgte automatisch mit dem Münchner automatischen Segmentationssystem „MAUS“ ([262]). Die Sprachdaten wurden zur Weiterverarbeitung in das EMU Speech Database System eingelesen ([44]). Über die phonetische Annotation wurden die Konsonanten (/m, n, ŋ, f, v, s, z, ʃ, ç, j, x, h/) und Vokale (/aː, ɑ, i, ɪ, e, ɛ, ɛː, u, ʊ, o, ɔ, y, ʏ, œː, œ, ɐ, ə/) ausgewählt. Anschließend wurden mit dem Tkassp-Tool die Spektren der Nasale und Frikative berechnet mit $N = 512$ (was bei einer Abtastrate von 16000 Hz eine Fensterlänge von 0,032 s ergibt) und einer Frequenzauflösung von 40,0 Hz. Die Fensterverschiebung betrug 0,005 s und als Fensterfunktion wurde die Blackman-Funktion verwendet. Anschließend wurden die Spektraldaten nach R Version 2.13.0 ([241]) importiert und mit der Funktion *moments* die vier Spektralmomente zum zeitlichen Mittelpunkt des Segments berechnet. Die Ergebnisse dieser Berechnung wurden in eine Datentabelle geschrieben und gespeichert.
Die Vokalformanten wurden mit dem Snack-Tool (EMU pitch and formant tool) gemessen. Dafür wurde ein Kosinus-Fenster der Länge 0,049 s verwendet. In dem untersuchten Frequenzbereich wurden fünf Formanten erwartet, weswegen die LPC-Ordnung 15 betrug. Für die Analyse wurden später aber nur die ersten vier Formanten verwendet, damit die Anzahl

der untersuchten Merkmale für die Konsonanten und Vokale gleich ist. Als Präemphase wurde ein Faktor von 0,7 eingesetzt und der Nominalwert für F1 auf 500 Hz gesetzt, was ein durchschnittlicher Wert für Männer ist. Für jeden Formanten wurde der Wert zum zeitlichen Mittelpunkt in einer Datentabelle gespeichert.

Statistische Auswertung

F-ratio Von jedem der berechneten spektralen Momente soll nun bestimmt werden, wie sprecherspezifisch es ist. Sprecherspezifität zeichnet sich dadurch aus, dass das Merkmal zwischen zwei verschiedenen Sprechern (Inter-Sprecher-Variation) stark und zwischen den Äußerungen eines einzelnen Sprechers (Intra-Sprecher-Variation) möglichst wenig variiert. Eine Methode um geeignete Merkmale zu finden, ist die Varianzanalyse (ANOVA), welche die F-ratio (3.5) zwischen Merkmalen verschiedener Sprecher bestimmt. Je größer die F-ratio wird, desto spezifischer ist ein Merkmal für einen Sprecher.

$$F = \frac{\text{Inter-Sprecher-Variation}}{\text{Intra-Sprecher-Variation}} \tag{3.5}$$

Der Term der Inter-Sprecher-Variation beinhaltet sowohl die Variation, die auf den unterschiedlichen stimmlichen Eigenschaften der Sprecher beruht, als auch die, welche durch individuelle Schwankungen während einer Äußerung bzw. zwischen zwei Äußerungen des gleichen Sprechers entsteht. Um zu sehen, ob die Unterschiede zwischen den Sprechern systematisch sind und nicht nur auf zufälligen oder individuellen Unterschieden beruhen, wird durch diese (Intra-Sprecher-Variation) geteilt. Diese Verhältnis (F-ratio) gibt dadurch einen guten Eindruck, inwieweit die vorhandenen Unterschiede auf Unterschieden zwischen Sprechern beruhen [89].
Ein Nachteil der F-ratio ist allerdings, dass bei multimodalen Klassen oder Klassen mit gleichem Mittelwert die Aussagekraft reduziert werden kann [43]. Dieses Problem tritt bei allen Maßen auf, die hauptsächlich durch die Mittelwertdifferenz der Klassen bestimmt werden. Optimale F-ratios sind also nur für normalverteilte Klassen mit unterschiedlichen Mittelwerten zu erwarten. Daher wurden vor der Analyse die Daten stichprobenartig auf Normalverteilung getestet, wobei sie annähernd normalverteilt waren. Anders als zum Beispiel in [96] wurde in dieser Analyse nicht über Phonemklassen gemittelt, sondern für jede Klasse separat die F-ratio berechnet. Unter Berücksichtigung der genannten Kritikpunkte, gibt die F-ratio einen schnellen, intuitiven Überblick über die unterschiedliche Eignung der Phoneme zur Sprechererkennung.

Die F-ratio kann mit folgender Formel berechnet werden:

$$F = \frac{\frac{n}{m-1}\sum_{j=1}^{m}(\mu_j - \bar{\mu})^2}{\frac{1}{m(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij} - \mu_j)^2} \tag{3.6}$$

wobei n die Anzahl der Äußerungen ist, m die Anzahl der Sprecher und x_{ij} der Wert des Parameters in der $i$ten Äußerung des $j$ten Sprechers; μ_j ist der Mittelwert der $j$ten Gruppe und $\bar{\mu}$ der Mittelwert aller Daten [209]. Die Berechnung der F-ratio erfolgte in R mit einer Varianzanalyse (ANOVA). Dazu wurde die Funktion *aov* [127] verwendet.

Inter- und Intra-Sprecher-Variation Inter- und Intra-Sprecher-Variation haben unterschiedliche Ursachen. Während Inter-Sprecher-Variation hauptsächlich durch sprecherabhängige physiologische, demografische und soziale Faktoren verursacht wird, hängt Intra-Sprecher-Variation von linguistischen, emotionalen und anderen situativen Effekten ab. Da die Inter-Sprecher-Variation kontextunabhängig ist und für eine große Sprecherpopulation gilt, kann sie für eine Äußerung als konstant angenommen werden. Die Intra-Sprecher-Variation hingegen ist variabel während einer Äußerung und verlangt vom Hörer eine dynamische Anpassung [296].
Da die Inter-Sprecher-Variation die Unterschiede zwischen den Sprechern aufzeigt, die aufgrund physiologischer Differenzen vorhanden sind (und die Variation innerhalb eines Sprechers), ist sie für die Erforschung von Sprecherunterschieden relevanter als die Intra-Sprecher-Variation. Manche Studien konzentrieren sich daher ausschließlich auf die Inter-Sprecher-Variation ohne die mögliche Intra-Sprecher-Variation miteinzubeziehen ([187] [183]). Diese Untersuchungen gelten hauptsächlich der Erforschung neuer sprecherunterscheidender Merkmale und geben einen ersten Hinweis auf deren Potential. Studien, die ausschließlich Intra-Sprecher-Variation betrachten (z.B. [30]), verfolgen häufig das Ziel, diese durch (durch geeignete Methoden) zu minimieren.
Aufgrund der unterschiedlichen Ursachen beider Variationen und um den Einfluss der Inter- bzw. der Intra-Sprecher-Variation auf die F-ratio bestimmen zu können, ist eine separate Analyse beider Varianzen sinnvoll. Beispielsweise könnte sich ein Merkmal besonders stark zwischen den Sprechern unterscheiden (und dadurch eine hohe F-ratio erreichen) und ein anderes Merkmal eine sehr geringe Varianz zwischen verschiedenen Äußerungen des gleichen Sprechers haben. Die Studien von [143] und [211] zeigen, dass Inter- und Intra-Sprecher-Variation je nach Merkmal unabhängig von einander variieren können. Um herauszufinden, wie stark ein F-ratio-Wert auf der Inter- bzw. der Intra-Sprecher-Variation beruht, sollen

beide Varianzen auch separat betrachtet werden.

3.1.5. Ergebnisse

Konsonanten

Zunächst wurden die Konsonanten in Gruppen unterteilt (/m, n, ŋ, f, v, s, z, ʃ, ç, j, x, h/) und für jede der einzelnen Gruppen wurde das Verhältnis von Inter-Sprecher und Intra-Sprecher-Variation berechnet. Das jeweilige *spektrale Moment* war somit die abhängige Variable und der *Sprecher* die unabhängige. Auf diese Weise ergab sich für jeden Konsonanten und jedes Spektralmoment ein F-ratio-Wert, der angibt, ob die Inter-Sprecher-Variabilität signifikant größer ist als die Intra-Sprecher-Variabilität. Die F-ratios der einzelnen Parameter wurden anschließend nach Konsonant gruppiert in einem Balkendiagramm grafisch dargestellt (siehe Abbildung 3.1).

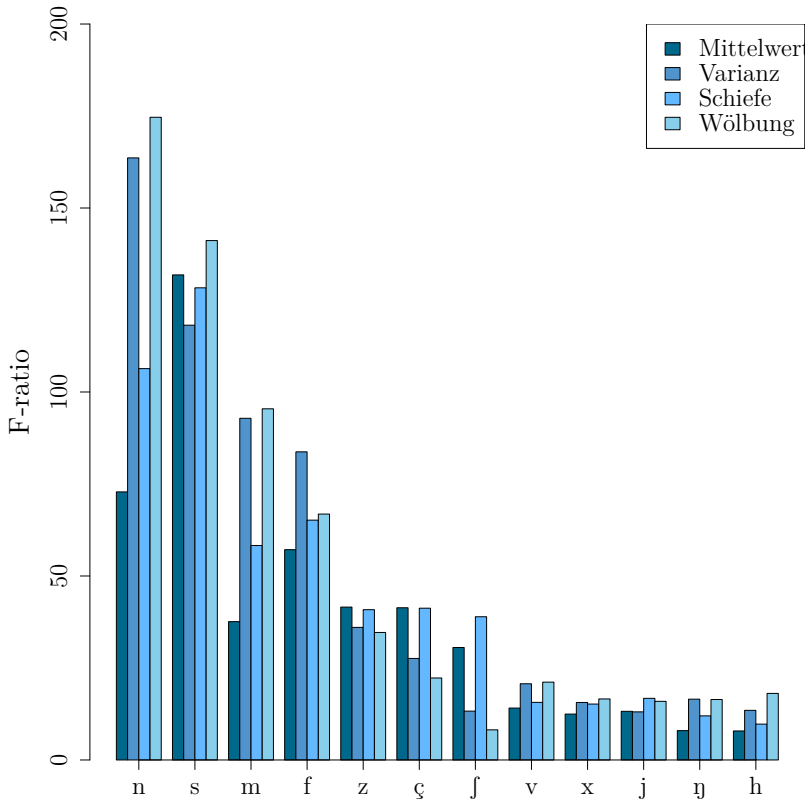


Abbildung 3.1.: Die F-ratios der Spektralmomente in Abhängigkeit vom Konsonant

Zur detaillierteren Übersicht über die Rangfolge der Parameter und ihrer Signifikanz,

wurden die F-ratios ebenfalls in einer Tabelle zusammengefasst (siehe Tabelle 3.1). Es zeigte sich, dass die spektralen Momente der einzelnen Konsonanten sehr unterschiedliche F-ratios erzielten, was für ihre unterschiedliche Sprecherspezifität spricht. Die Nasale /m, n/ und die Frikative /f, s/ erreichten mit Abstand die höchsten Werte, während die stimmhaften und die hinteren Frikative als auch der Nasal /ŋ/ wesentlich niedrigere F-ratios aufwiesen. Allerdings wird in der ausführlicheren Darstellung der Tabelle sichtbar, dass nicht immer alle Parameter eines Konsonanten besser sind als die eines anderen. So erwiesen sich beispielsweise die Varianz und die Wölbung von /n/ als am sprecherspezifischsten, während die Schiefe und der Mittelwert des Nasals auf Platz 7 und 11 lagen.

Rang	Parameter	F-ratio	p
1	m4 /n/	F[48,22642] = 174,66	< 0,001 *
2	m2 /n/	F[48,22642] = 163,62	< 0,001 *
3	m4 /s/	F[48,12752] = 141,15	< 0,001 *
4	m1 /s/	F[48,12752] = 131,82	< 0,001 *
5	m3 /s/	F[48,12752] = 128,31	< 0,001 *
6	m2 /s/	F[48,12752] = 118,15	< 0,001 *
7	m3 /n/	F[48,22642] = 106,35	< 0,001 *
8	m4 /m/	F[48,10066] = 95,43	< 0,001 *
9	m2 /m/	F[48,10066] = 92,85	< 0,001 *
10	m2 /f/	F[48,6542] = 83,74	< 0,001 *
11	m1 /n/	F[48,22642] = 72,85	< 0,001 *
12	m4 /f/	F[48,6542] = 66,83	< 0,001 *
13	m3 /f/	F[48,6542] = 65,18	< 0,001 *
14	m3 /m/	F[48,10066] = 58,28	< 0,001 *
15	m1 /f/	F[48,6542] = 57,15	< 0,001 *
16	m1 /z/	F[48,4332] = 41,55	< 0,001 *
17	m1 /ç/	F[48,4884] = 41,36	< 0,001 *
18	m3 /ç/	F[48,4884] = 41,24	< 0,001 *
19	m3 /z/	F[48,4332] = 40,84	< 0,001 *
20	m3 /ʃ/	F[48,2108] = 38,92	< 0,001 *
21	m1 /m/	F[48,10066] = 37,59	< 0,001 *
22	m2 /z/	F[48,4332] = 36,03	< 0,001 *
23	m4 /z/	F[48,4332] = 34,65	< 0,001 *
24	m1 /ʃ/	F[48,2108] = 30,56	< 0,001 *

25	m2 /ç/	F[48,4884] = 27,59	< 0,001 *
26	m4 /ç/	F[48,4884] = 22,28	< 0,001 *
27	m4 /v/	F[48,7010] = 21,15	< 0,001 *
28	m2 /v/	F[48,7010] = 20,71	< 0,001 *
29	m4 /h/	F[48,3597] = 18,09	< 0,001 *
30	m3 /j/	F[48,2757] = 16,75	< 0,001 *
31	m4 /x/	F[48,2532] = 16,59	< 0,001 *
32	m2 /ŋ/	F[48,1959] = 16,53	< 0,001 *
33	m4 /ŋ/	F[48,1959] = 16,43	< 0,001 *
34	m4 /j/	F[48,2757] = 15,94	< 0,001 *
35	m3 /v/	F[48,7010] = 15,65	< 0,001 *
36	m2 /x/	F[48,2532] = 15,63	< 0,001 *
37	m3 /x/	F[48,2532] = 15,20	< 0,001 *
38	m1 /v/	F[48,7010] = 14,11	< 0,001 *
39	m2 /h/	F[48,3597] = 13,49	< 0,001 *
40	m2 /ʃ/	F[48,2108] = 13,27	< 0,001 *
41	m1 /j/	F[48,2757] = 13,22	< 0,001 *
42	m2 /j/	F[48,2757] = 13,07	< 0,001 *
43	m1 /x/	F[48,2532] = 12,47	< 0,001 *
44	m3 /ŋ/	F[48,1959] = 11,98	< 0,001 *
45	m3 /h/	F[48,3597] = 9,74	< 0,001 *
46	m4 /ʃ/	F[48,2108] = 8,18	< 0,001 *
47	m1 /ŋ/	F[48,1959] = 7,97	< 0,001 *
48	m1 /h/	F[48,3597] = 7,88	< 0,001 *

Tabelle 3.1.: Die Sprecherspezifität der Spektralmomente der Konsonanten (* markiert signifikante Werte)

Vokale

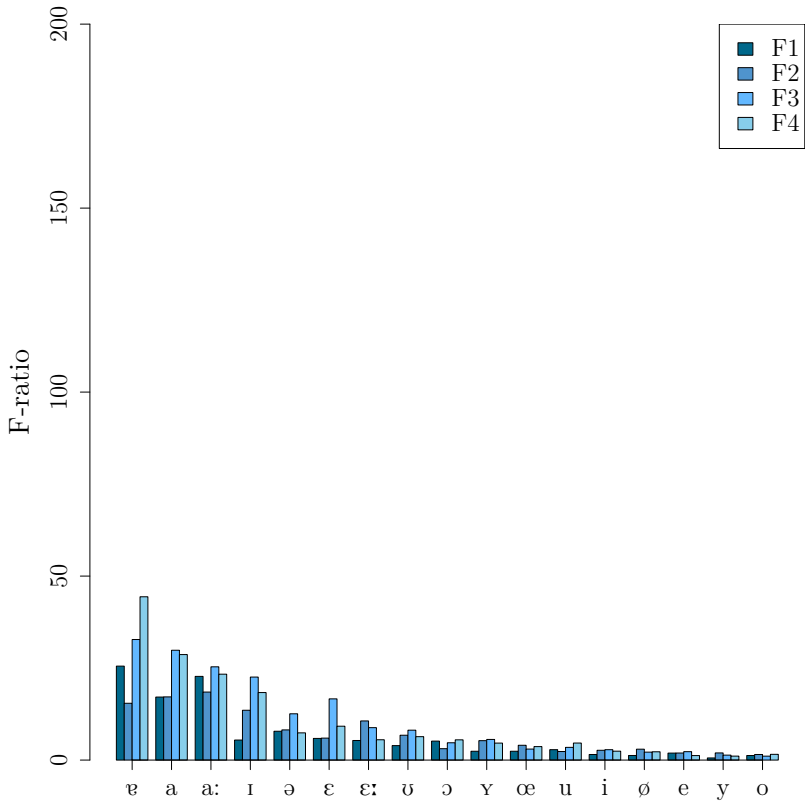


Abbildung 3.2.: Die F-ratio der Vokalformanten in Abhängigkeit vom Vokal

Die statistische Analyse der Vokale erfolgte analog zu der der Konsonanten. Nachdem die Vokale in einzelne Gruppen unterteilt wurden (/aː, ɑ, i, ɪ, e, ɛ, ɛː, u, ʊ, o, ɔ, y, ʏ, œː, œ, ɐ, ə/), wurde für jeden Formantwert (F1, F2, F3, F4) jedes Vokals der Einfluss des Sprechers berechnet. Dies erfolgte mittels einer Varianzanalyse mit der Funktion *aov* ([127]) mit dem jeweiligen *Formantwert* (z.B. F1) als abhängige Variable und dem *Sprecher* als unabhängige.

Rang	Parameter	F-ratio	p
1	F4 /ɐ/	F[48,11525] = 44,38	< 0,001 *
2	F3 /ɐ/	F[48,11525] = 32,76	< 0,001 *
3	F3 /ɑ/	F[48,9976] = 29,85	< 0,001 *
4	F4 /ɑ/	F[48,9976] = 28,67	< 0,001 *
5	F1 /ɐ/	F[48,11525] = 25,54	< 0,001 *
6	F3 /aː/	F[48,8403] = 25,34	< 0,001 *

7	F4 /aː/	F[48,8403] = 23,34	< 0,001 *
8	F1 /aː/	F[48,8403] = 22,74	< 0,001 *
9	F3 /ɪ/	F[48,8895] = 22,56	< 0,001 *
10	F2 /aː/	F[48,8403] = 18,48	< 0,001 *
11	F4 /ɪ/	F[48,8895] = 18,35	< 0,001 *
12	F2 /ɑ/	F[48,9976] = 17,17	< 0,001 *
13	F1 /ɑ/	F[48,9976] = 17,13	< 0,001 *
14	F3 /ɛ/	F[48,4871] = 16,63	< 0,001 *
15	F2 /ɐ/	F[48,11525] = 15,44	< 0,001 *
16	F2 /ɪ/	F[48,8895] = 13,54	< 0,001 *
17	F3 /ə/	F[48,7598] = 12,58	< 0,001 *
18	F2 /ɛː/	F[48,1949] = 10,63	< 0,001 *
19	F4 /ɛ/	F[48,4871] = 9,20	< 0,001 *
20	F3 /ɛː/	F[48,1949] = 8,81	< 0,001 *
21	F2 /ə/	F[48,7598] = 8,20	< 0,001 *
22	F3 /ʊ/	F[48,3378] = 8,13	< 0,001 *
23	F1 /ə/	F[48,7598] = 7,84	< 0,001 *
24	F4 /ə/	F[48,7598] = 7,39	< 0,001 *
25	F2 /ʊ/	F[48,3378] = 6,75	< 0,001 *
26	F4 /ʊ/	F[48,3378] = 6,36	< 0,001 *
27	F2 /ɛ/	F[48,4871] = 5,97	< 0,001 *
28	F1 /ɛ/	F[48,4871] = 5,88	< 0,001 *
29	F3 /ʏ/	F[48,1772] = 5,61	< 0,001 *
30	F4 /ɛː/	F[48,1949] = 5,52	< 0,001 *
31	F4 /ɔ/	F[48,2726] = 5,50	< 0,001 *
32	F1 /ɪ/	F[48,8895] = 5,44	< 0,001 *
33	F1 /ɛː/	F[48,1949] = 5,32	< 0,001 *
34	F2 /ʏ/	F[48,1772] = 5,27	< 0,001 *
35	F1 /ɔ/	F[48,2726] = 5,16	< 0,001 *
36	F3 /ɔ/	F[48,2726] = 4,70	< 0,001 *
37	F4 /u/	F[48,354] = 4,63	0,013 *
38	F4 /ʏ/	F[48,1772] = 4,61	< 0,001 *
39	F2 /œ/	F[47,548] = 4,02	< 0,001 *
40	F1 /ʊ/	F[48,3378] = 3,93	< 0,001 *

41	F4 /œ/	F[47,548] = 3,66	< 0,001 *
42	F3 /u/	F[48,354] = 3,45	0,36
43	F2 /ɔ/	F[48,2726] = 3,09	< 0,001 *
44	F3 /œ/	F[47,548] = 2,99	< 0,001 *
45	F2 /œː/	F[46,261] = 2,96	< 0,001 *
46	F1 /u/	F[48,354] = 2,83	< 0,001 *
47	F3 /i/	F[48,554] = 2,82	< 0,001 *
48	F2 /i/	F[48,554] = 2,66	< 0,001 *
49	F4 /i/	F[48,554] = 2,42	< 0,001 *
50	F1 /œ/	F[47,548] = 2,40	< 0,001 *
51	F1 /ʏ/	F[48,1772] = 2,39	< 0,001 *
52	F2 /u/	F[48,354] = 2,33	0,019 *
53	F3 /e/	F[42,315] = 2,28	< 0,001 *
54	F4 /œː/	F[46,261] = 2,25	< 0,001 *
55	F3 /œː/	F[46,261] = 2,16	< 0,001 *
56	F2 /y/	F[33,51] = 1,94	0,016 *
57	F2 /e/	F[42,315] = 1,92	< 0,001 *
58	F1 /e/	F[42,315] = 1,90	< 0,001 *
59	F4 /o/	F[48,1446] = 1,56	< 0,001 *
60	F1 /i/	F[48,554] = 1,52	0,017 *
61	F2 /o/	F[48,1446] = 1,51	< 0,001 *
62	F3 /y/	F[33,51] = 1,34	0,17
63	F1 /œː/	F[46,261] = 1,25	0,15
64	F1 /o/	F[48,1446] = 1,24	< 0,001 *
65	F4 /e/	F[42,315] = 1,22	0,17
66	F4 /y/	F[33,51] = 1,07	0,40
67	F3 /o/	F[48,1446] = 1,07	< 0,001 *
68	F1 /y/	F[33,51] = 0,58	0,95

Tabelle 3.2.: Die Sprecherspezifität der Vokalformanten (* markiert signifikante Werte)

Genau wie die Konsonanten unterschieden sich auch die Vokale in ihrem Grad an Sprecherspezifität (siehe Abbildung 3.2). Dabei erreichten nicht alle Vokale eine signifikante Differenz zwischen ihrer Inter-Sprecher- und ihrer Intra-Sprecher-Variation (siehe Tabelle 3.2). Die höchsten Werte erzielten die offenen Vokale /ɐ, ɑ, aː/, gefolgt von eher zentraleren

und am schlechtesten schnitten die gerundeten ab. Die genauen F-ratios der einzelnen Parameter mit zugehörigen Signifikanzwerten wurden tabellarisch dargestellt. Ähnlich wie auch bei den Konsonanten zeigte sich für die Vokale hier ebenfalls, dass die Sprecherspezifität nicht nur vom Vokal abhängt, sondern auch vom Formanten. So belegten F4 und F3 des Vokals /ɐ/ die vorderen Plätze, während F1 auf Platz 5 und F2 auf Platz 15 lagen.

Um die Sprecherspezifität der Vokale besser mit der von Konsonanten vergleichen zu können, wurden ebenfalls die spektralen Momente der Vokale analysiert. Mit der gleichen Methode wie bei den Konsonanten erfolgte eine Varianzanalyse *aov* ([127]) mit dem jeweiligen *spektralen Moment* als abhängiger und *Sprecher* als unabhängiger Variable.

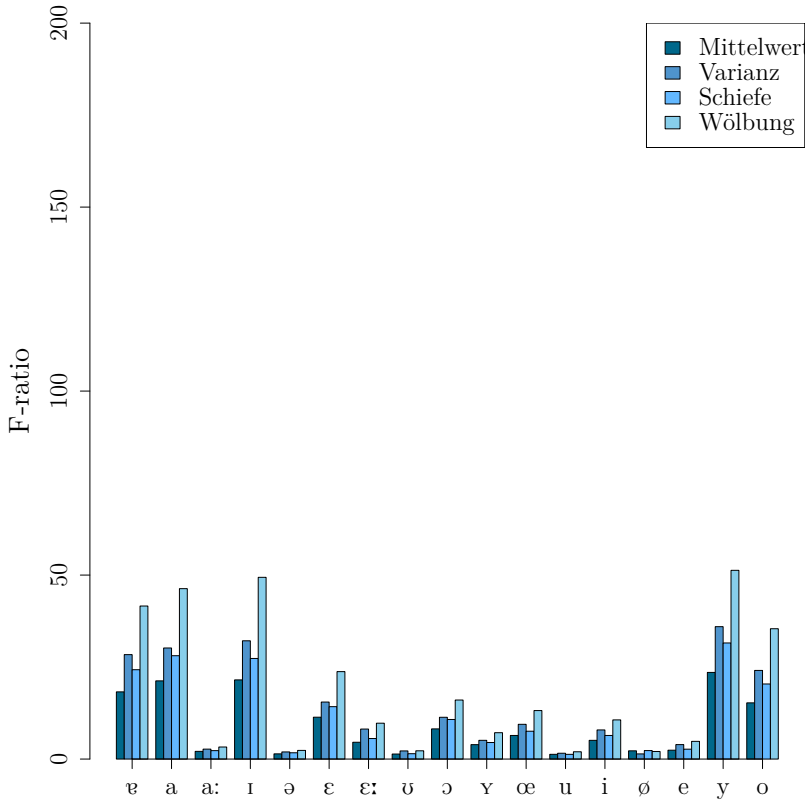


Abbildung 3.3.: Die F-ratio der Spektralmomente in Abhängigkeit vom Vokal

Auch wenn in dieser Analyse für die Konsonanten und die Vokale die gleichen akustischen Parameter, nämlich die Spektralmomente, verwendet wurden, reichte die durchschnittliche Sprecherspezifität der Vokale nicht an die der Konsonanten heran (siehe Abbildung 3.3). Allerdings waren die F-ratios der Spektralmomente höher als die der Vokalformanten (siehe Abbildung 3.2). Es zeigte sich aber auch, dass sich jedes der beiden akustischen

Merkmale offenbar für unterschiedliche Vokale eignet. Während zum Beispiel die F-ratios der Vokalformanten für /aː/ recht hoch waren, erzielte dieser Laut bei den Spektralmomenten nur sehr geringe Werte. Bei anderen Lauten wiederum verhielt es sich genau umgekehrt. In Tabelle 3.3 sind die detaillierten F-ratio-Werte samt den Signifikanzen dargestellt. Die Wölbung (m_4) von /ɐ/ wies die höchste Sprecherspezifität auf, gefolgt von /ɪ, ɑ, aː/.

Rang	Parameter	F-ratio	p
1	m4 /ɐ/	F[48,11525] = 51,29	< 0,001 *
2	m4 /ɪ/	F[48,8895] = 49,39	< 0,001 *
3	m4 /ɑ/	F[48,9976] = 46,31	< 0,001 *
4	m4 /aː/	F[48,8403] = 41,59	< 0,001 *
5	m2 /ɐ/	F[48,11525] = 35,98	< 0,001 *
6	m4 /ə/	F[48,7598] = 35,42	< 0,001 *
7	m2 /ɪ/	F[48,8895] = 32,14	< 0,001 *
8	m3 /ɐ/	F[48,11525] = 31,55	< 0,001 *
9	m2 /ɑ/	F[48,9976] = 30,18	< 0,001 *
10	m2 /aː/	[48,8403] = 28,38	< 0,001 *
11	m3 /ɑ/	F[48,9976] = 28,09	< 0,001 *
12	m3 /ɪ/	F[48,8895] = 27,34	< 0,001 *
13	m3 /aː/	F[48,8403] = 24,28	< 0,001 *
14	m2 /ə/	F[48,7598] = 24,09	< 0,001 *
15	m4 /ɛ/	F[48,4871] = 23,76	< 0,001 *
16	m1 /ɐ/	F[48,11525] = 23,55	< 0,001 *
17	m1 /ɪ/	F[48,8895] = 21,51	< 0,001 *
18	m1 /ɑ/	F[48,9976] = 21,25	< 0,001 *
19	m3 /ə/	F[48,7598] = 20,40	< 0,001 *
20	m1 /aː/	F[48,8403] = 18,26	< 0,001 *
21	m4 /ɔ/	F[48,2726] = 16,05	< 0,001 *
22	m2 /ɛ/	F[48,4871] = 15,49	< 0,001 *
23	m1 /ə/	F[48,7598] = 15,27	< 0,001 *
24	m3 /ɛ/	F[48,4871] = 14,24	< 0,001 *
25	m4 /ʊ/	F[48,3378] = 13,18	< 0,001 *
26	m1 /ɛ/	F[48,4871] = 11,38	< 0,001 *
27	m2 /ɔ/	F[48,2726] = 11,35	< 0,001 *
28	m3 /ɔ/	F[48,2726] = 10,76	< 0,001 *

29	m4 /ʏ/	F[48,1772] = 10,64	< 0,001 *
30	m4 /ɛː/	F[48,1949] = 9,75	< 0,001 *
31	m2 /ʊ/	F[48,3378] = 9,44	< 0,001 *
32	m1 /ɔ/	F[48,2726] = 8,20	< 0,001 *
33	m2 /ɛː/	F[48,1949] = 8,16	< 0,001 *
34	m2 /ʏ/	F[48,1772] = 7,91	< 0,001 *
35	m3 /ʊ/	F[48,3378] = 7,57	< 0,001 *
36	m4 /u/	F[48,354] = 7,16	< 0,001 *
37	m3 /ʏ/	F[48,1772] = 6,42	< 0,001 *
38	m1 /ʊ/	F[48,3378] = 6,41	< 0,001 *
39	m3 /ɛː/	F[48,1949] = 5,58	< 0,001 *
40	m2 /u/	F[48,354] = 5,10	< 0,001 *
41	m1 /ʏ/	F[48,1772] = 5,09	< 0,001 *
42	m4 /œ/	F[47,548] = 4,81	< 0,001 *
43	m1 /ɛː/	F[48,1949] = 4,57	< 0,001 *
44	m3 /u/	F[48,354] = 4,50	0,031 *
45	m2 /œ/	F[47,548] = 3,93	< 0,001 *
46	m1 /u/	F[48,354] = 3,92	0,062
47	m4 /i/	F[48,554] = 3,28	< 0,001 *
48	m2 /i/	F[48,554] = 2,70	< 0,001 *
49	m3 /œ/	F[47,548] = 2,69	< 0,001 *
50	m1 /œ/	F[47,548] = 2,41	< 0,001 *
51	m4 /e/	F[42,315] = 2,35	< 0,001 *
52	m3 /œː/	F[46,261] = 2,31	< 0,001 *
53	m3 /i/	F[48,554] = 2,29	< 0,001 *
54	m4 /o/	F[48,1446] = 2,24	< 0,001 *
55	m1 /œː/	F[46,261] = 2,24	< 0,001 *
56	m2 /o/	F[48,1446] = 2,21	< 0,001 *
57	m1 /i/	F[48,554] = 2,09	< 0,001 *
58	m4 /œː/	F[46,261] = 2,06	< 0,001 *
59	m4 /y/	F[33,51] = 1,98	0,014 *
60	m2 /e/	F[42,315] = 1,94	< 0,001 *
61	m3 /e/	F[42,315] = 1,71	< 0,001 *
62	m2 /y/	F[33,51] = 1,58	0,069

63	m3 /o/	F[48,1446] = 1,46	< 0,001 *
64	m1 /e/	F[42,315] = 1,41	0,054
65	m2 /œː/	F[46,261] = 1,40	0,056
66	m1 /o/	F[48,1446] = 1,36	< 0,001 *
67	m1 /y/	F[33,51] = 1,29	0,21
68	m3 /y/	F[33,51] = 1,28	0,21

Tabelle 3.3.: Die Sprecherspezifität der Spektralmomente der Vokale (* markiert signifikante Werte)

Inter-und Intra-Sprecher-Variation

Zur genaueren Betrachtung der Zusammensetzung der F-ratios, wurden die Sprecher-Variabilitäten einzeln analysiert. Diese Analyse sollte Aufschluss geben, ob eine der Variabilitäten besonders stark zur gesamten Sprecherspezifität beiträgt oder ob sich der Beitrag zwischen den einzelnen Konsonanten unterscheidet. Da in diesem Fall kein Verhältnis gebildet wird (wie bei der F-ratio), welches die einzelnen Größen normiert, wurde zuvor separat eine Normierung der einzelnen Parameter durchgeführt. Mit der Formel

$$norm(x) = \frac{(x - min(x)}{max(x) - min(x)} \tag{3.7}$$

wurde jeder einzelne Wert jedes spektralen Moments normiert, sodass danach alle Spektralmomente im Bereich zwischen 0 und 1 lagen. Zur Berechnung der Variabilitäten wurde ebenfalls die Funktion *aov* ([127]) verwendet, nur dass diesmal von den Ergebnissen nicht die F-ratio verwendet wurde, sondern die *Quadratsumme* der Sprecher (Inter-Sprecher-Variabilität) und der Residuen (Intra-Sprecher-Variabilität).
Es zeigte sich, dass die Intra-Sprecher-Variation der einzelnen Sprecher mit Werten von unter 0,02 sehr gering ist (siehe Abbildung 3.5). Obwohl es Unterschiede in der Intra-Sprecher-Variabilität zwischen den einzelnen Lauten gab, lagen doch alle Werte in einem engen Bereich. Die geringste Variabilität wies der Nasal /n/ auf, dessen Werte fast durchgehend unter 0,01 lagen. Im Gegensatz dazu waren die Werte der Inter-Sprecher-Variation viel größer und ähnelten viel stärker den F-ratio-Werten (siehe Abbildung 3.4). Die Werte liegen im Bereich zwischen 0,1 und 1,2. Die höchste Inter-Sprecher-Variation zeigte der Frikativ /s/, aber auch /f/ und die Nasale /m, n/ lagen im oberen Bereich.

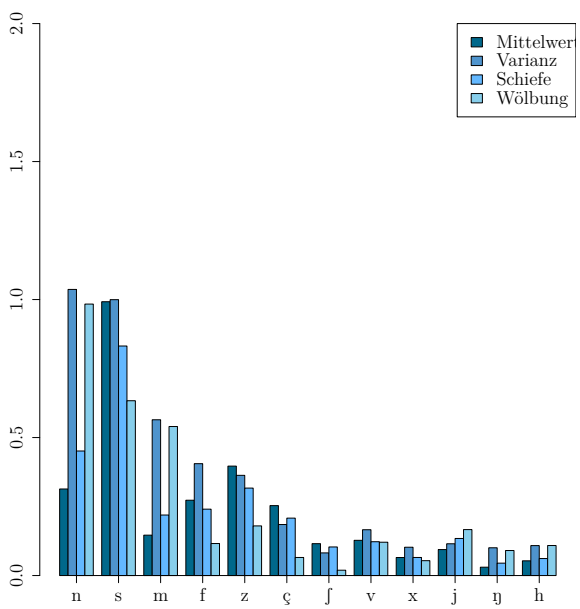


Abbildung 3.4.: Die Inter-Sprecher-Variation der spektralen Momente von Nasalen und Frikativen

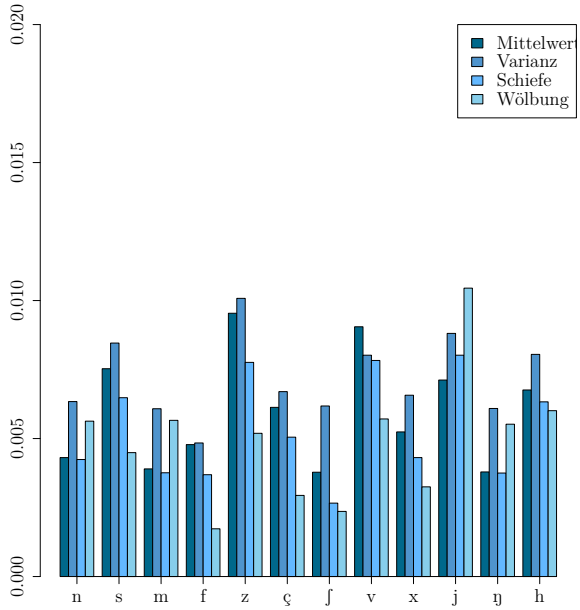


Abbildung 3.5.: Die Intra-Sprecher-Variation der spektralen Momente von Nasalen und Frikativen

3.1.6. Diskussion

Nasale und Frikative

In dieser Untersuchung wurden die spektralen Momente der Nasale und Frikative des Deutschen im Hinblick auf ihre Sprecherspezifität analysiert. Als Sprachmaterial dienten annähernd spontansprachliche Dialoge aus dem Verbmobil-Korpus und als Maß für die Sprecherspezifität der Konsonanten wurde die F-ratio (Verhältnis von Inter- und Intra-Sprecher-Variation) verwendet.

Die Ergebnisse zeigen, dass die Spektralmomente aller untersuchten Konsonanten sprecherspezifisch sind, d.h. die Inter-Sprecher-Variation ist immer signifikant größer als die Intra-Sprecher-Variation. Dabei scheinen die Nasale /m, n/ sowie die Frikative /f, s/ die meisten sprecherspezifischen Merkmale zu enthalten. Damit entsprechen die Ergebnisse den Erwartungen aus der Literatur. Die Nasale und der Frikativ /s/ wurden häufig als hilfreich zur Sprecheridentifikation bzw. -diskrimination beschrieben ([101] [295] [68] [27] [12] [142] [143]). Die hohe Sprecherspezifität der Nasale lässt sich artikulatorisch durch die Zuschaltung des Nasenraumes erklären. Dieser ist anatomisch-physiologisch nicht variabel, sodass er keine

große Intra-Sprecher-Variation zulässt und zwischen verschiedenen Sprechern unterschiedlich, wodurch eine hohe Inter-Sprecher-Variation erzeugt wird. Die guten Ergebnisse von /s/ lassen sich dadurch erklären, dass Sibilanten viel Energie im Spektrum aufweisen. Nicht nur weil sie einen großen vorderen Hohlraum haben, sondern vor allem weil der Luftstrom auf die Zähne trifft, was eine hoch-frequente und energiereiche Turbulenz erzeugt. Der Frikativ /f/ hingegen besitzt kaum einen vorderen Vokaltraktteil, sodass sein Spektrum sehr diffus und variabel ohne starke Resonanzschwerpunkte ist und generell wenig Energie hat [272]. Wahrscheinlich ist daher seine Sprecherspezifität geringer.
Wenn man bedenkt, dass der große vordere Hohlraum beim /s/ zu einer hohen Sprecherspezifität beiträgt, dann ist es verwunderlich, dass die hinteren Frikative /ç, j, x, h/, welche einen noch größeren vorderen Vokaltraktteil haben, so viel weniger sprecherspezifisch waren. Allerdings sind sie aber weniger energiereich [230] und haben wahrscheinlich daher eine geringere Sprecherspezifität.
Auffällig ist auch, dass die stimmlosen Frikative durchgehend sprecherspezifischer waren als die stimmhaften Frikative. Dies könnte an der schwächeren Intensität stimmhafter Frikative gegenüber stimmlosen liegen. Da die Stimmlippen geschlossen gehalten werden müssen, um ein Schwingen zu ermöglichen, kann die Luft nicht so schnell durch die Glottis strömen. Dies führt zu einem langsameren Luftstrom und somit zu einem schwächeren Friktionsgeräusch [230]. Eventuell sind deshalb die stimmlosen Frikative /f, s/ sprecherspezifischer als die stimmhaften Frikative /v, z/.
Die höchste Sprecherspezifität weisen die vorderen Konsonanten, der labialen und alveolaren Artikulationsstelle auf (/m, f, n, s/). Dabei erreichen die alveolaren Laute stets höhere Werte als die labialen. Dieser Effekt wurde bereits in der Literatur beschrieben (z.B. [6]) und soll daher in der nächsten Analyse eingehender betrachtet werden (siehe Abschnitt 3.2).

Vergleich von Konsonanten und Vokalen

Die Ergebnisse zeigen eindeutig, dass nicht nur Vokale, sondern auch Nasale und Frikative viele wichtige akustische Informationen zur Sprecheridentität enthalten. In diesem Experiment stellen sich die Konsonanten sogar als sprecherspezifischer dar als die Vokale. Die relativ schlechten Ergebnisse der Vokale widersprechen den Ergebnissen aus der Literatur (z.B. [177] [186] [187] [255]), die Vokale als sehr sprecherspezifisch beschreiben. Daher sollen zunächst einige technische Faktoren, welche sich verfälschend auf die Vokalformanten ausgewirkt haben könnten, diskutiert werden. (1) Durch die automatische Segmentierung

können Segmentgrenzen etwas ungenauer sein als bei einer manuellen Segmentation. Die Abweichungen sollten sich aber in Grenzen halten [263]. (2) Als zweiter Punkt muss erwähnt werden, dass die automatische Messung der Vokalformanten mit dem Snack-Tool in EMU ebenfalls nicht hundertprozentig zuverlässig ist. Die manuelle Überprüfung wäre nur mit unvertretbar hohem Aufwand möglich gewesen. Außerdem wurde eine sehr große Datenmenge analysiert, sodass die meisten Daten korrekt sein müssten und ein paar kleine Fehler tolerieren. (3) Theoretisch sind Formantwerte normalverteilt, aber durch Fehler in der automatischen Messung könnten sich Abweichungen ergeben haben. Eine Nicht-Normalverteilung der Daten könnte sich dann wiederum auf die Varianzanalyse ausgewirkt haben, welche normalverteilte Daten annimmt. Eine stichprobenhafte Kontrolle der Normalverteilung der Formantwerte zeigte in wenigen Fällen eine Abweichung. Der Großteil der Daten war (annähernd) normalverteilt.
Abgesehen von diesen technischen Faktoren, könnte die Ursache auch in dem gewählten akustischen Merkmal liegen. Denn im Gegensatz zu den Formanten, wird bei der Messung der spektralen Momente das gesamte Spektrum einbezogen. Deshalb stehen bei diesem Merkmal mehr Informationen zur Verfügung, woraus sich eine höhere Sprecherspezifität ergeben könnte. Aber selbst wenn wir allen diesen Faktoren einen gewissen Einfluss zugestehen, so ist der Unterschied zwischen den Ergebnissen der Konsonanten und der Vokale auffällig groß. Es ist nicht anzunehmen, dass eine solch eindeutige Differenz von möglichen kleinen Ungenauigkeiten ausgelöst wurde. Außerdem ist es auch nicht undenkbar, dass bestimmte Konsonanten sprecherspezifischer sind als Vokale. In manchen Studien ([68] [16]) erwiesen sich beispielsweise die Nasale als sprecherspezifischer als Vokale.

Vergleich gespannter und ungespannter Vokale

Die Ergebnisse zeigen, dass die ungespannten Vokale tendenziell sprecherspezifischer sind als die gespannten. Man könnte denken, dass ein gespannter (langer) Vokal mehr akustische Merkmale zur Sprechererkennung aufweisen müsste, da er mehr Raum für Variation bietet. Allerdings deuten die Ergebnisse genau in die andere Richtung. Dieses Ergebnis steht aber im Einklang mit vorangegangenen Untersuchungen. So zeigte [289], dass die Physiologie des Sprechers, und somit seine spezifischen Merkmale, sich stärker in unbetonten Silben auswirkt als in betonten. Dies könnte daran liegen, dass Vokale in unbetonten Silben kürzer und zentralisierter gesprochen werden, wodurch die individuelle Form des Vokaltrakts des Sprechers mehr Einfluss gewinnt. Da in unbetonten Silben stets ungespannte Vokale stehen, würde diese Erkenntnis diese Ergebnisse erklären und unterstützen. Außerdem könnten

die ungespannten Vokale aufgrund ihrer kurzen Dauer weniger Raum für Intra-Sprecher-Variation bieten, wodurch ihre Sprecherspezifität steigen würde [178].

Vergleich hoher und tiefer Vokale

In bisherigen Untersuchungen wurden meist die höheren Formanten der hohen Vokale wie /i/ als besonders sprecherspezifisch identifiziert (z.B. [177]). In dieser Untersuchung weisen jedoch die tieferen Vokale, genauer gesagt das /ɐ/, die höchste Sprecherspezifität auf. Dieses Ergebnis könnte dadurch entstanden sein, dass in dieser Untersuchung nur Sprachmaterial aus einer Aufnahme-Session untersucht wurde. [177] stellt nämlich fest, dass bei nur einer Aufnahme der offene Vokal /a/ eine recht hohe F-ratio aufwies. Verwendete man allerdings Material aus mehreren Aufnahme-Sessions, so waren die vorderen, hohen Vokale sprecherspezifischer [178], was die hohe Sprecherspezifität der tiefen, offenen Vokale erklären würde.

Vergleich der Inter-Sprecher- und Intra-Sprecher-Variation

Ein weiterer interessanter Punkt ist das Verhältnis der Inter-Sprecher-Variabilität zur Intra-Sprecher-Variabilität. Da die F-ratio das Verhältnis dieser beiden Größen angibt, bestehen zwei Möglichkeiten eine hohe F-ratio zu erhalten: Entweder ist die Inter-Sprecher-Variabilität sehr groß oder die Intra-Sprecher-Variabilität sehr klein. Entgegen der Vermutung, dass beide Möglichkeiten in den Sprachlauten vorkommen, war in den vorliegenden Ergebnissen stets die Inter-Sprecher-Variabilität der ausschlaggebende Faktor. Während die Intra-Sprecher-Variabilität nur in geringem Maße variierte, zeigte die Inter-Sprecher-Variabilität beinahe das gleiche Muster wie auch die F-ratio. Dies deutet darauf hin, dass die Inter-Sprecher-Variabilität einen wesentlich größeren Einfluss auf die F-ratio und somit auf die Sprecherspezifität hat als die Intra-Sprecher-Variabilität.

3.2. Sprecherspezifität in Abhängigkeit der Artikulationsstelle

3.2.1. Ziele und Hypothesen

Das vorgestellte Experiment zeigt, dass Nasale und Frikative eine große Sprecherspezifität aufweisen. Dabei fällt auf, dass vor allem die alveolaren Laute /n, s/ hohe Werte erzielen,

höhere als die labialen Konsonanten /m, f/. Das weist auf den entscheidenden Einfluss der Artikulationsstelle auf die Sprecherspezififtät eines Lautes hin. Ähnliche Ergebnisse zeigten bereits Studien für das Japanische (z.B. [6]).
Um genauer zu untersuchen, ob dieser Einfluss der Artikulationsstelle systematisch ist, sollen in diesem Teil Konsonanten verschiedener Artikulationsmodi (Nasale, Frikative, Plosive) untersucht werden. Außerdem werden sowohl die stimmhaften als auch die stimmlosen Varianten jedes Konsonanten betrachtet. Aufgrund der bisherigen Beobachtungen ist zu vermuten, dass alle alveolaren Konsonanten eine höhere Sprecherspezifität aufweisen werden als ihre labialen Gegenstücke.

3.2.2. Methode

Wie im vorhergehenden Experiment wurden die Sprachaufnahmen mit annähernd spontansprachlichen Dialogen von 49 Sprechern aus dem Verbmobil-Korpus als Sprachmaterial verwendet. Die Sprachsignalverarbeitung und die Berechnung der spektralen Momente erfolgte analog zu dem ersten Experiment. Anschließend wurden die Daten in einer Datentabelle gespeichert und in R grafisch und statistisch mit einer Varianzanalyse ausgewertet.

3.2.3. Ergebnisse

Um den Einfluss die Artikulationsstelle (labial und alveolar) auf die Sprecherspezifität eines Konsonanten zu ergründen, wurden zusätzlich zu den Nasalen und Frikativen auch Plosive untersucht. Es sollte herausgefunden werden, ob sich bei diesem Artikulationsmodus die gleichen Unterschiede zwischen den Konsonanten der labialen und der alveolaren Artikulationstelle zeigen würden.
Analog zu den vorangegangenen Experimenten wurden die Konsonanten gruppiert (/n, m, s, f, z, v, t, p, d, b) und ihre F-ratio durch eine Varianzanalyse *aov* [127] bestimmt. Dabei war das jeweilige *Spektralmoment* die abhängige und der *Sprecher* die unabhängige Variable. Es zeigten sich große Unterschiede zwischen den alveolaren und den labialen Konsonanten. Wenn man die F-ratios der spektralen Momente des alveolaren Konsonanten mit den Werten seines labialen Gegenstücks vergleicht, dann fällt auf, dass der alveolare Laut stets höhere F-ratio-Werte hat (siehe Abbildung 3.6). Dieses Ergebnis bestätigte auch eine Varianzanalyse mit der *F-ratio* als abhängige und der *Artikulationsstelle* (labial, alveolar) als unabhängige Variable. Die alveolaren Laute sind insgesamt sprecherspezifischer ($F[1,38] = 5{,}572$; $p = 0{,}0235$) als die labialen. Die Laute /m, n, f, s/ erzielten die höchsten Werte, wobei die

alveolaren Konsonanten /n, s/ bessere Ergebnisse erreichten als die labialen /m, f/. Die stimmhaften Frikative /z, v/ und die Plosive /p, b, t, d/ schnitten insgesamt schlechter ab, zeigten aber die gleichen Unterschiede zwischen alveolarem und labialem Konsonanten. Es zeigte sich aber auch, dass die F-ratio für alle Laute signifikant wurde ($p < 0{,}001$) (siehe Tabelle 3.4). Dennoch gab es große Unterschiede in Abhängigkeit vom Konsonant und der Artikulationsstelle. So fällt beim Betrachten der Tabelle auf, dass die ersten sieben Plätze alle von Parametern der alveolaren Konsonanten /n, s/ eingenommen werden, während die labialen Konsonanten /p, b/ komplett die niedrigsten Werte haben.

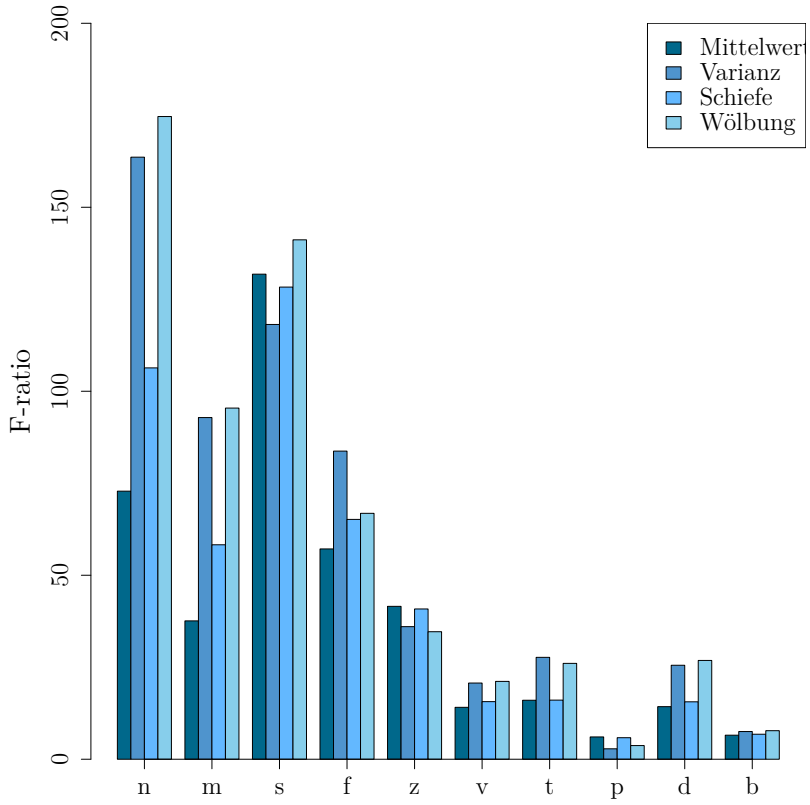


Abbildung 3.6.: Die F-ratio der Spektralmomente in Abhängigkeit vom Konsonant

Rang	Parameter	F-ratio	p
1	m4 /n/	F[48,22642] = 174,66	< 0,001 *
2	m2 /n/	F[48,22642] = 163,62	< 0,001 *
3	m4 /s/	F[48,12752] = 141,15	< 0,001 *
4	m1 /s/	F[48,12752] = 131,82	< 0,001 *

5	m3 /s/	F[48,12752] = 128,31	< 0,001 *
6	m2 /s/	F[48,12752] = 118,15	< 0,001 *
7	m3 /n/	F[48,22642] = 106,35	< 0,001 *
8	m4 /m/	F[48,10066] = 95,43	< 0,001 *
9	m2 /m/	F[48,10066] = 92,85	< 0,001 *
10	m2 /f/	F[48,6542] = 83,74	< 0,001 *
11	m1 /n/	F[48,22642] = 72,85	< 0,001 *
12	m4 /f/	F[48,6542] = 66,83	< 0,001 *
13	m3 /f/	F[48,6542] = 65,18	< 0,001 *
14	m3 /m/	F[48,10066] = 58,28	< 0,001 *
15	m1 /f/	F[48,6542] = 57,15	< 0,001 *
16	m1 /z/	F[48,4332] = 41,55	< 0,001 *
17	m3 /z/	F[48,4332] = 40,84	< 0,001 *
18	m1 /m/	F[48,10066] = 37,59	< 0,001 *
19	m2 /z/	F[48,4332] = 36,03	< 0,001 *
20	m4 /z/	F[48,4332] = 34,65	< 0,001 *
21	m2 /t/	F[48,15102] = 27,69	< 0,001 *
22	m4 /d/	F[48,9253] = 26,85	< 0,001 *
23	m4 /t/	F[48,15102] = 26,05	< 0,001 *
24	m2 /d/	F[48,9253] = 25,54	< 0,001 *
25	m4 /v/	F[48,7010] = 21,15	< 0,001 *
26	m2 /v/	F[48,7010] = 20,71	< 0,001 *
27	m3 /t/	F[48,15102] = 16,06	< 0,001 *
28	m1 /t/	F[48,15102] = 16,03	< 0,001 *
29	m3 /v/	F[48,7010] = 15,65	< 0,001 *
30	m3 /d/	F[48,9253] = 15,61	< 0,001 *
31	m1 /d/	F[48,9253] = 14,29	< 0,001 *
32	m1 /v/	F[48,7010] = 14,11	< 0,001 *
33	m4 /b/	F[48,3585] = 7,76	< 0,001 *
34	m2 /b/	F[48,3585] = 7,53	< 0,001 *
35	m3 /b/	F[48,3585] = 6,80	< 0,001 *
36	m1 /b/	F[48,3585] = 6,54	< 0,001 *
37	m1 /p/	F[48,1723] = 6,06	< 0,001 *
38	m3 /p/	F[48,1723] = 5,85	< 0,001 *

39	m4 /p/	F[48,1723] = 3,71	< 0,001 *
40	m2 /p/	F[48,1723] = 2,83	< 0,001 *

Tabelle 3.4.: Die Sprecherspezifität der Spektralmomente der labialen und alveolaren Konsonanten (* markiert signifikante Werte)

Um detaillierter zu untersuchen, welcher Variabilitätsanteil (Inter- oder Intra-Sprecher-Variabilität) für die Unterschiede sorgte, wurden beide Variabilitäten getrennt voneinander betrachtet. Dafür wurden die Spektralmomente vorher wieder normiert (wie in Abschnitt 3.1.5 beschrieben), sodass alle Werte im Bereich von 0 bis 1 lagen. Anschließend wurden mittels einer Varianzanalyse *aov* die *Quadratsumme* der Sprecher (Inter-Sprecher-Variabilität) und der Residuen (Intra-Sprecher-Variabilität) bestimmt.
Es zeigte sich, dass die Inter-Sprecher-Variabilität (siehe Abbildung 3.7) stets wesentlich größer war als die Intra-Sprecher-Variabilität (siehe Abbildung 3.8). Die alveolaren Konsonanten wiesen sowohl eine höhere Inter- als auch eine höhere Intra-Sprecher-Variation auf als die labialen.

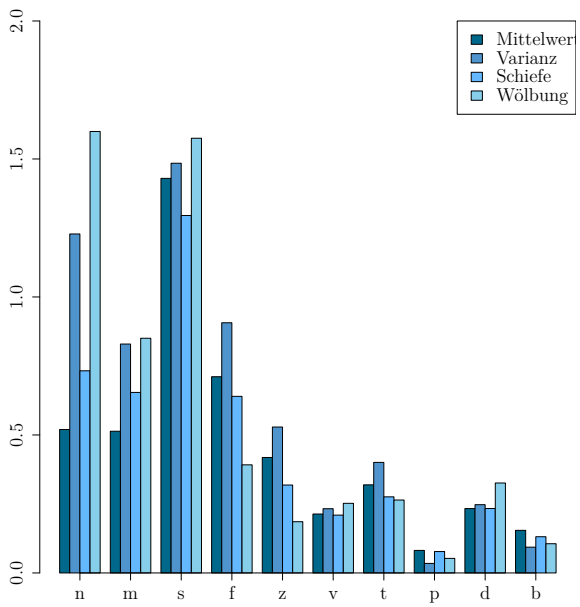


Abbildung 3.7.: Die Inter-Sprecher-Variation der spektralen Momenten von labialen und alveolaren Konsonanten

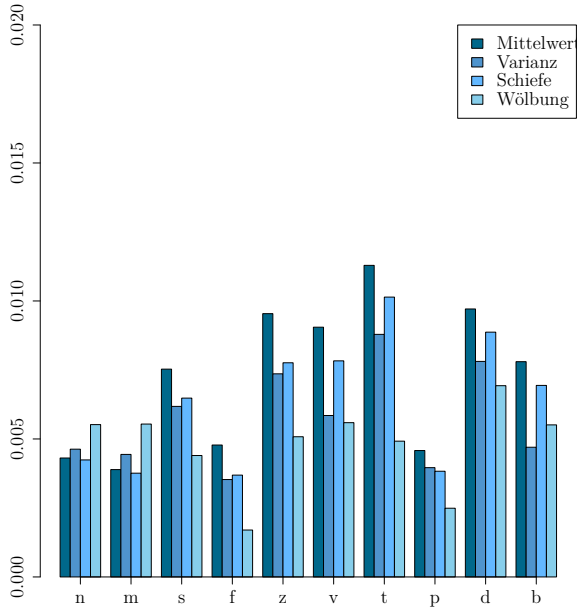


Abbildung 3.8.: Die Intra-Sprecher-Variation der spektralen Momenten von labialen und alveolaren Konsonanten

3.2.4. Diskussion

Die Hypothese, dass alveolare Laute eine höhere Sprecherspezifität aufweisen als labiale Laute, konnte bestätigt werden. Der alveolare Laut ist insgesamt signifikant sprecherspezifischer als sein labiales Gegenstück. Ein ähnliches Ergebnis hatte sich auch schon in einer Studie des Japanischen gezeigt [6]. Speziell der alveolare Frikativ /s/ hatte sich in mehreren Untersuchungen als besonders informativ über die Sprecheridentität erwiesen (z.B. [68] [143]). Dies zeigt, dass die neuen Ergebnisse zu Konsonanten des Deutschen mit anderen Studien in Einklang stehen.

Nun stellt sich die Frage, an welchen Merkmalen der alveolaren Konsonanten es liegt, dass sie sprecherspezifischer sind. Ein Aspekt ist sicherlich die unterschiedliche Anatomie der Alveolen, deren Form sich von Mensch zu Mensch unterscheidet. Da die Alveolen unbeweglich sind, können sie nicht zwischen den Äußerungen eines Sprechers variieren, sodass sie nichts (oder sehr wenig) zur Intra-Sprecher-Variation beitragen. Die einzige Quelle für Intra-Sprecher-Variation ist die Zunge als beweglicher Artikulator. Wie stark die artikulatorischen Bewegungen der Zunge tatsächlich variieren, müsste in einem weiterführenden Experiment geklärt werden.

Es gibt einige Hinweise darauf, dass besonders /s/ eine sehr komplexe Koordination der Artikulatoren erfordert. Beispielsweise ist das so genannte „Lispeln“ einer der häufigsten Sprechfehler [277], was beweist, wie schwierig dieser Laut zu produzieren ist. Da er ein hohes Maß an Präzision verlangt, um das gewünschte akustische Ergebnis zu erzeugen, bleibt nicht viel Spielraum für Intra-Sprecher-Variation. Außerdem trifft bei der Artikulation von /s/ der Luftstrom auf die Zähne, was eine hochfrequente und energiereiche Turbulenz hervorruft. Die starke Energie diese Geräusches macht sich auch im Spektrum des Lautes bemerkbar. Wenn man nun annimmt, dass laute Geräusche markanter sind und ihnen der Hörer deshalb mehr Informationen entnehmen kann (auch Informationen über die Identität des Sprechers), wäre das ebenfalls eine Erklärung für die höhere Sprecherspezifität von /s/. Der Frikativ /f/ hingegen besitzt weder ein Hindernis wie die Zähne noch einen nennenswerten vorderen Vokaltraktteil, sodass sein Spektrum sehr diffus und variabel ohne starke Resonanzschwerpunkte ist und generell wenig Energie hat [272]. Wahrscheinlich ist daher seine Sprecherspezifität geringer.

Die Nasale zeigen ein ähnliches Muster wie die Frikative /f, s/, denn auch hier ist der alveolare Laute sprecherspezifischer als der labiale. Durch die unterschiedliche Länge des verschlossenen Mundraums, gibt es Unterschiede in der Lage der Antiformanten von /m, n/. Je länger der orale Vokaltrakt ist, desto tiefer liegt die Antiresonanzfrequenz. Aufgrund

der Lage dieses Antiformanten, weist /n/ eine starke Energieveränderung zum folgenden Vokal im Bereich von 1450 bis 2300 Hz auf. Diese ist bei dem labialen Nasal */m/* schwächer ausgeprägt. Diese Merkmale, welche in erster Linie die Wahrnehmung der Artikulationsstelle beeinflussen, könnten auch Unterschiede in der Sprecherspezifität verursachen. Ein weiterer Parameter, der die Artikulationsstelle der Nasale kodiert, sind die Transitionen zu den angrenzenden Vokalen ([77] [94]). Diese könnten je nach Artikulationsstelle unterschiedlich viele Sprecherinformationen enthalten. Allerdings ist in diesem Fall aufgrund der automatischen Segmentierung nicht eindeutig, welcher Anteil der Transitionen möglicherweise dem Nasal oder den umliegenden Vokalen zugeordnet wurde.
Zur ausführlicheren Untersuchung der Auswirkungen der Artikulationsstelle auf die Sprecherspezifität der Konsonanten, wurden zusätzlich zu den Nasalen und Frikativen auch Plosive analysiert. Ihre Ergebnisse bestätigen den signifikanten Einfluss der Artikulationsstelle auf die Sprecherspezifität. Der alveolare Plosiv /t/ erweist sich als sprecherspezifischer als der labiale Plosiv /p/. Genau wie bei den Nasalen stecken bei den Plosiven viele Informationen über die Artikulationsstelle in den Transitionen. Durch den Verlauf von F2 und F3 lassen sich die labiale und die alveolare Artikulationsstelle voneinander unterscheiden [230]. Allerdings enthält auch der Burst der Plosive einige Ortsinformationen [108]. Beispielsweise ist der Burst von labialen Plosiven schwächer als von alveolaren wegen der geringeren Länge des vorderen Vokaltrakts [220]. Wie viele Informationen über die Artikulationsstelle der Burst aber genau beinhaltet, hängt auch von dem Vokalkontext ab [269]. Genau wie bei den Nasalen gilt hier, dass die Unterschiede zwischen den Artikulationsstellen auch Unterschiede im Grad der Sprecherspezifität verursachen könnten.
Wie erwartet, zeigen die Plosive die niedrigste Sprecherspezifität in der Analyse. Interessanterweise ergeben sich aber keine Unterschiede zwischen den stimmlosen und stimmhaften Plosiven, wohingegen bei den Frikativen die stimmlosen eindeutig sprecherspezifischer waren. Während stimmlose Frikative engergiereicher und wahrscheinlich aus diesem Grund sprecherspezifischer sind, scheint es zwischen den stimmhaften und stimmlosen Plosiven keine Unterschiede zu geben, die sich auf die Sprecherspezifität auswirkt.
Insgesamt zeigen alle untersuchten Konsonanten ein signifikantes Maß an sprecherspezifischen Merkmalen. Die Nasale und Frikative dind dabei sprecherspezifischer als die Plosive und die alveolaren Laute sprecherspezifischer als die labialen.

3.3. Der Einfluss des Telefonkanals auf die Sprecherspezifität

3.3.1. Ziele und Hypothesen

Die vorgestellten Ergebnisse der akustischen Analyse zeigen, dass nicht nur Vokale, sondern auch Konsonanten viele wichtige Sprecherinformationen enthalten. Die Sprachaufnahmen dieser Untersuchung besitzen alle Studio-Qualität. Leider liegt, besonders im Fall der forensischen Phonetik, oft kein qualitativ hochwertiges Sprachmaterial vor. Größtenteils sind die Aufnahmen über Festnetz oder Mobiltelefone übertragen und aufgezeichnet wurden. Da der Telefonkanal nur die Frequenzen zwischen ca. 300 und 3400 Hz überträgt, gehen dabei zweifellos akustische Informationen verloren. Da diese auch Sprecherinformationen enthalten, wird wahrscheinlich die Sprecherspezifität der Laute sinken, wenn sie über das Telefon übertragen werden.
Da die unterschiedlichen Lautgruppen ihren Energieschwerpunkt in unterschiedlichen Bereichen des Spektrums haben, werden sie vermutlich vom Telefonkanal unterschiedlich stark beeinflusst. Nasale haben zum Beispiel viel Energie im unteren Bereich des Spektrum (bis ca. 500 Hz), während sie in den oberen Frequenzen wenig Energie haben. Da ihre markantesten Frequenzbereiche übertragen werden, sollte sich der Telefonkanal nur wenig auf ihre Sprecherspezifität auswirken. Anders sieht es bei den Frikativen aus, welche viel Energie in den oberen Bereichen des Spektrums haben (über 3000 Hz). Dadurch werden ihre wichtigsten akustischen Informationen nicht übertragen, was vermutlich zu einer stärkeren Verringerung ihrer Sprecherspezifität führen wird.
Das Ziel dieser Untersuchung ist es, herauszufinden wie stark die Sprecherspezifität der Laute durch den Telefonkanal beeinträchtigt wird und welche Laute möglicherweise gar nicht oder nur wenig beeinflusst werden.

3.3.2. Methode

Wie im vorhergehenden Experiment wurden die Sprachaufnahmen mit annähernd spontansprachlichen Dialogen von 49 Sprechern aus dem Verbmobil-Korpus als Sprachmaterial verwendet. In diesem Fall wurden die simultan entstandenen und telefonisch übertragenen Aufnahmen verwendet.
Da die Mikrofon- und die Telefonaufnahmen zeitsynchron sind, konnten die Segmentationen und Annotationen der Mikrofondaten genutzt werden. Die anschließende Berechnung des

Spektrums und der spektralen Momente erfolgte mit den gleichen Parametern. Die Daten wurden ebenfalls in einer Datentabelle gespeichert und grafisch und statistisch mit einer Varianzanalyse ausgewertet.

3.3.3. Ergebnisse

Zur statistischen Analyse wurden die Laute wieder gruppiert (/m, n, ŋ, f, v, s, z, ʃ, ç, j, x, h/) und anschließend die F-ratio berechnet. Dafür wurde mit der Funktion *aov* [127] eine Varianzanalyse durchgeführt, wobei das jeweilige *spektrale Moment* die abhängige Variable war und der *Sprecher* die unabhängige. Als Ergebnis ergab sich somit für jedes Spektralmoment jedes Konsonanten ein F-ratio-Wert, der das Maß seiner Sprecherspezifität ausdrückt. Die Werte wurden in einem Balkendiagramm grafisch dargestellt (siehe Abbildung 3.10).

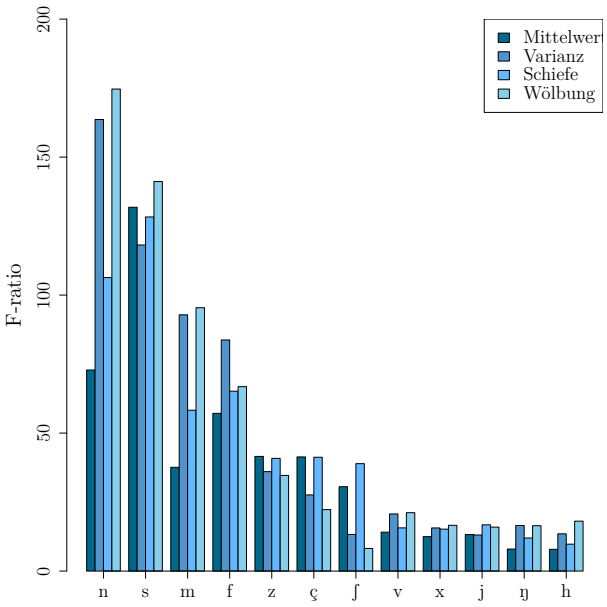


Abbildung 3.9.: Die F-ratio der Spektralmomente in Abhängigkeit vom Konsonant; Sprachaufnahmen über Mikrofon

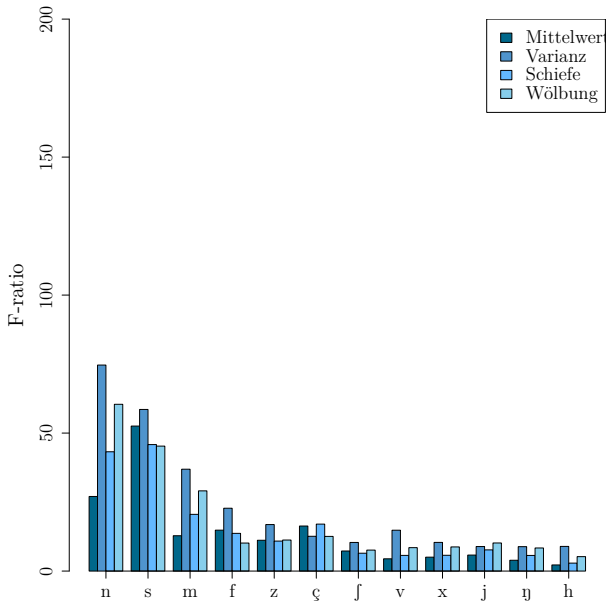


Abbildung 3.10.: Die F-ratio der Spektralmomente in Abhängigkeit vom Konsonant; Sprachaufnahmen über Telefon

Die Ergebnisse zeigen deutlich, dass sich an den Unterschieden in der Sprecherspezifität der Laute wenig geändert hat. Es sind immer noch die gleichen Laute (/n, s, m, f/ ect.) am

sprecherspezifischsten.
Die genauere statistische Auswertung der Daten in Form einer Tabelle (siehe Tabelle 3.5) zeigt deutlich, dass trotz der Übertragung durch den Telefonkanal das Verhältnis von Inter- zu Intra-Sprecher-Variation aller Konsonanten signifikant blieb. Auch wenn nach wie vor /n/ und /s/ die sprecherspezifischsten Konsonanten sind, so gibt es einige Unterschiede in der Sprecherspezifität der einzelnen spektralen Momente. Während F-ratios der Varianz (m_2) relativ größer geworden sind, scheinen die Werte der Wölbung (m_4) abgenommen zu haben.
Ein Vergleich der beiden Abbildungen (Abbildung 3.9 und Abbildung 3.10) zeigt, dass durch die Übertragung der Sprachdaten über den Telefonkanal bei den Lauten mit hohen F-ratio-Werten (z.B. /n, s, m, f/) die Sprecherspezifität auf ca. die Hälfte reduziert wurde. Die verschiedenen Artikulationsmodi wurden aber offenbar alle relativ gleich stark vom Telefonkanal beeinflusst. Nur bei den Lauten, die schon bei den Mikrofonaufnahmen niedrige F-ratios hatten (z.B. /x, j, ŋ, h/), wurden die Werte nicht mehr stark reduziert.

Rang	Parameter	F-ratio	p
1	m2 /n/	F[48,21581] = 74,65	< 0,001 *
2	m4 /n/	F[48,21581] = 60,43	< 0,001 *
3	m2 /s/	F[48,12124] = 58,56	< 0,001 *
4	m1 /s/	F[48,12124] = 52,54	< 0,001 *
5	m3 /s/	F[48,12124] = 45,84	< 0,001 *
6	m4 /s/	F[48,12124] = 45,28	< 0,001 *
7	m3 /n/	F[48,21581] = 43,23	< 0,001 *
8	m2 /m/	F[48,9616] = 36.91	< 0,001 *
9	m4 /m/	F[48,9616] = 29,06	< 0,001 *
10	m1 /n/	F[48,21581] = 27,03	< 0,001 *
11	m2 /f/	F[48,6259] = 22,77	< 0,001 *
12	m3 /m/	F[48,9616] = 20,55	< 0,001 *
13	m3 /ç/	F[48,4654] = 17,02	< 0,001 *
14	m2 /z/	F[48,4115] = 16,85	< 0,001 *
15	m1 /ç/	F[48,4654] = 14,82	< 0,001 *
16	m1 /f/	F[48,6259] = 14,80	< 0,001 *
17	m2 /v/	F[48,6659] = 14,80	< 0,001 *
18	m3 /f/	F[48,6259] = 13,67	< 0,001 *
19	m1 /m/	F[48,9616] = 12,80	< 0,001 *

20	m2 /ç/	F[48,4654] = 12,60	< 0,001 *
21	m4 /ç/	F[48,4654] = 12,56	< 0,001 *
22	m4 /z/	F[48,4115] = 11,25	< 0,001 *
23	m1 /z/	F[48,4115] = 11,17	< 0,001 *
24	m3 /z/	F[48,4115] = 10,87	< 0,001 *
25	m2 /x/	F[48,2434] = 10,39	< 0,001 *
26	m2 /ʃ/	F[48,1988] = 10,37	< 0,001 *
27	m4 /j/	F[48,2613] = 10,18	< 0,001 *
28	m4 /f/	F[48,6259] = 10,15	< 0,001 *
29	m2 /h/	F[48,3404] = 8,93	< 0,001 *
30	m2 /j/	F[48,2613] = 8,89	< 0,001 *
31	m2 /ŋ/	F[48,1858] = 8,84	< 0,001 *
32	m4 /x/	F[48,2434] = 8,75	< 0,001 *
33	m4 /v/	F[48,6659] = 8,48	< 0,001 *
34	m4 /ŋ/	F[48,1858] = 8,37	< 0,001 *
35	m3 /j/	F[48,2613] = 7,68	< 0,001 *
36	m4 /ʃ/	F[48,1988] = 7,60	< 0,001 *
37	m1 /ʃ/	F[48,1988] = 7,27	< 0,001 *
38	m3 /ʃ/	F[48,1988] = 6,47	< 0,001 *
39	m1 /j/	F[48,2613] = 5,80	< 0,001 *
40	m3 /x/	F[48,2434] = 5,72	< 0,001 *
41	m3 /v/	F[48,6659] = 5,68	< 0,001 *
42	m3 /ŋ/	F[48,1858] = 5,66	< 0,001 *
43	m4 /h/	F[48,3404] = 5,23	< 0,001 *
44	m1 /x/	F[48,2434] = 5,04	< 0,001 *
45	m1 /v/	F[48,6659] = 4,44	< 0,001 *
46	m1 /ŋ/	F[48,1858] = 3,92	< 0,001 *
47	m3 /h/	F[48,3404] = 2,88	< 0,001 *
48	m1 /h/	F[48,3404] = 2,22	< 0,001 *

Tabelle 3.5.: Die Sprecherspezifität der Spektralmomente der Konsonanten in Telefonsprache (* markiert signifikante Werte)

3.3.4. Diskussion

In dieser Untersuchung sollte der Einfluss des Telefonkanals auf die Sprecherspezifität der Konsonanten bestimmt werden. Dafür wurden die Konsonanten aus dem telefonisch aufgenommenen Sprachmaterial akustisch analysiert und die Sprecherspezifität ihrer spektralen Momente berechnet. Es wurde vermutet, dass vor allem die Frikative, deren Energie im oberen Bereich des Spektrums liegt, unter der Telefonübertragung leiden würden. Im Gegensatz dazu wurde für die Nasale, deren Energie eher im unteren Bereich des Spektrum liegt, ein schwächerer Einfluss erwartet. Hier zeigte sich nun aber, dass die Frikative auch nicht stärker beeinträchtigt wurden als die Nasale. Das zeigt, dass offenbar auch die energiereichen, informativen Spektralanteile der Nasale durch den Telefonkanal reduziert wurden. Da es sich bei allen Sprechern um Männer handelte, welche normalerweise eine durchschnittliche Grundfrequenz von ca. 120 Hz [235] haben, liegt auch der erste Formant des Nasals recht tief (unter oder um die 300 Hz). Dadurch, dass das Sprachsignal erst ab ca. 300 Hz übertragen wird, wurden möglicherweise wichtige akustische Informationen des Nasals heraus gefiltert. Außerdem haben auch Nasale in den oberen spektralen Bereichen akustische Informationen (siehe Abschnitt 2.2.2). Auch wenn diese energetisch schwach sind, könnten sie trotzdem zur Sprecherspezifität des Nasals beitragen.
Die Frikative zeigen die erwartete Reduktion ihrer Sprecherinformationen. Da der Telefonkanal nur Frequenzen bis ca. 3400 Hz überträgt, werden die höheren und energiereichen Frequenzbereiche der Frikative (ab ca. 3000 Hz) abgeschnitten; wodurch sich ihre Sprecherspezifität verringert.
Die statistische Analyse zum Einfluss der Qualität des Sprachsignals auf die Sprecherspezifität der einzelnen Spektralmomente zeigte, dass sich der Telefonkanal vor allem auf den Mittelwert und die Schiefe auswirkt, während Varianz und Wölbung nicht signifikant beeinflusst werden.
Zusammenfassend kann gesagt werden, dass alle Konsonanten durch die Übertragung über den Telefonkanal akustische Informationen einbüßen. Da dies auch die Sprecherinformationen betrifft, reduziert sich die Sprecherspezifität der Konsonanten.

3.4. Generelle Diskussion

In dieser Studie wurde die Sprecherspezifität einiger Konsonanten des Deutschen untersucht. Aufgrund verschiedener Ergebnisse aus der Literatur ([101] [295] [68] [203] [254] [179] [16] [12] [15] [142] [143] [72]) wurde vermutet, dass die Konsonanten, neben den Vokalen, viele

Informationen zur Sprecheridentität enthalten. Zur Analyse wurden annähernd spontansprachliche Dialoge von 49 männlichen Sprechern aus dem Verbmobil-Korpus verwendet. Als akustische Merkmale wurden die vier spektralen Momente (Mittelwert, Varianz, Schiefe, Wölbung) ausgewählt. Es zeigte sich, dass die Laute zu einem sehr unterschiedlichen Grad sprecherspezifisch sind. Während die Nasale /m, n/ und die Frikative /f, s/ sich als sehr sprecherspezifisch erweisen, fallen die Werte der anderen Konsonanten geringer aus. Dennoch war die Differenz zwischen Inter- und Intra-Sprecher-Variabilität für alle Konsonanten signifikant. Das bedeutet, dass alle Konsonanten ein gewisses Maß an Sprecherinformationen enthalten.
Um die Sprecherspezififät der Spektralmomente der Konsonanten besser einschätzen zu können, wurden sie mit Vokalmerkmalen verglichen. Dazu wurden einmal die klassischerweise verwendeten Vokalformanten und einmal ebenfalls die spektralen Momente der Vokale berechnet. Da Vokale häufig zur Sprecheridentifikation verwendet werden und in der Literatur als sehr sprecherspezifisch beschrieben sind ([177] [187] [289] [268]), wurde vermutet, dass sie bessere oder ähnliche Werte wie die Konsonanten erzielen würden. Überraschenderweise erzielen die Vokale aber in dieser Untersuchung schlechtere Werte als die Konsonanten. Auch wenn die spektralen Momente der Vokale sich als leicht sprecherspezifischer erweisen als die Formanten, so schneiden auch sie im Vergleich zu den Konsonanten schlechter ab. Es wird vermutet, dass dieses Ergebnis durch technische Faktoren beeinflusst sein könnte, wie der automatischen Segmentierung, der automatischen Formantberechnung oder von einer daraus resultierenden Nicht-Normalverteilung der Formantwerte. Allerdings kann jeder dieser Faktoren nur minimal zu dem Ergebnis beigetragen haben, weshalb sie nicht diesen großen Unterschied zwischen Konsonanten und Vokalen erklären können. Zudem gibt es auch in der Literatur bereits Studien, die einigen Konsonanten mehr Sprecherspezifität zuschreiben als den Vokalen [68]; zumindest in Relation zur Anzahl der Segmente [16].
Nach der ersten Analyse zeigte sich die Tendenz, dass die Konsonanten der alveolaren Artikulationsstelle durchgehend sprecherspezifischer waren als die der labialen. Daher wurde eine weiterführende Untersuchung mit einer größeren Anzahl an Konsonanten und verschiedenen Artikulationsmodi (Nasale, Frikative, Plosive) durchgeführt. Der erste Eindruck bestätigte sich; innerhalb eines Artikulationsmodus ist der alveolare Laut immer sprecherspezifischer als der labiale. Trotz aller Unterschiede zeigt sich aber auch wieder die prinzipielle Sprecherspezifität aller Laute in ihren signifikanten Unterschieden zwischen Inter- und Intra-Sprecher-Variation. Außerdem ist festzustellen, dass die Unterschiede in der Sprecherspezifität zwischen den alveolaren und labialen Lauten nicht aus einem systematischen

Unterschied in der Inter- oder Intra-Sprecher-Variabilität resultieren. Beide Variabilitäten sind für die alveolaren Konsonanten größer, aber die Inter-Sprecher-Variabilität in höherem Maß als die Intra-Sprecher-Variabilität. Das heißt, alveolare Laute weisen eher stärkere Unterschiede zwischen Sprechern auf, als dass sie innerhalb eines Sprechers weniger variieren als labiale Laute.
Als mögliche Erklärung für die größere Sprecherspezifität der alveolaren Laute wurden die artikulatorischen Bedingungen, die möglicherweise auch zu einer stärkeren Energie im Spektrum der Laute beitragen, diskutiert. Die höhere Energie und die somit größere Lautstärke macht diese Laute möglicherweise markanter und informativer, sodass sie mehr Informationen über die Sprecheridentität liefern können.
Im letzten Teil der Untersuchung wurde dann der Einfluss des Telefonkanals auf die Sprecherspezifität der einzelnen Konsonanten bestimmt. Es zeigt sich, dass alle Konsonanten gleichermaßen von der geringeren Frequenzbreite (300 - 3400 Hz) beeinträchtigt werden. Dabei scheint es keine Rolle zu spielen, in welchen Bereichen des Spektrums die Laute ihre höchste Energie und somit vermutlich ihren größten akustischen Informationsgehalt aufweisen. Die Sprecherspezifität der Konsonanten reduziert sich ungefähr auf die Hälfte. Dennoch zeigen sich immer noch signifikante Differenzen zwischen Inter- und Intra-Sprecher-Variabilität. Das unterstreicht, dass Konsonanten durchaus geeignet sind, gute Sprecherinformationen zu liefern.
In allen Experimenten zeigte sich, dass die F-ratios tendenziell größer wurden, wenn mehr Exemplare eines Lautes vorhanden waren. Das liegt daran, dass wenn die Datenmenge in einer Varianzanalyse zunimmt, die Intra-Sprecher-Variation abnimmt, nicht aber die Inter-Sprecher-Variation. Wenn die Mittelwerte unterschiedlich sind und die Datenmenge größer wird, nimmt die F-ratio tendenziell immer weiter zu [100]. Dadurch haben die Laute mit mehr Exemplaren vermutlich etwas höhere Werte geliefert als andere Laute. Um dieser Vermutung nachzugehen, müsste eine Analyse mit gleichvielen Exemplaren pro Konsonant durchgeführt werden. Aber auch wenn die F-ratios für diese Konsonanten dann niedriger wären, dürfte die allgemeine Tendenz ähnlich sein. Eine weitere Möglichkeit wäre, ein anderes Maß als die F-ratio zu verwenden, bei der dieser Effekt nicht auftreten kann (z.B. Likelihood ratio, Diskriminanzanalyse, etc.). Trotz ihrer Limitierung, hilft die F-ratio einen schnellen ersten Überblick über die Sprecherspezifität akustischer Sprechermerkmale zu gewinnen.

3.5. Zusammenfassung

Dieses Kapitel widmete sich der Fragestellung, wie hoch die Sprecherspezifität ausgewählter Konsonanten im Deutschen ist. Als akustische Merkmale werden dafür die Spektralmomente (Mittelwert, Varianz, Schiefe, Wölbung) herangezogen. Es zeigt sich, dass vor allem die Nasale /m, n/ und die Frikative /s, f/ viele Sprecherinformationen enthalten. Die anderen Nasale und Frikative erreichen zwar ebenfalls signifikante Werte, liegen aber deutlich niedriger. Es lässt sich feststellen, dass die Vokale weder in ihren Formanten noch in ihren Spektralmomenten mehr Sprecherinformationen enthalten als die Konsonanten. Mögliche Gründe für dieses unerwartete Ergebnis wurden zusammengetragen und bewertet.
Außerdem offenbart sich, dass alveolare Konsonanten sprecherspezifischer sind als labiale. Die möglichen Ursachen dafür liegen wahrscheinlich in der spezifischen Anatomie und Physiologie des Vokaltrakts als auch in der höheren spektralen Energie der alveolaren Laute.
Die letzte Untersuchung beweist den starken Einfluss der Qualität des Sprachsignals auf die Sprecherspezifität der Konsonanten. Wird das Sprachsignal statt mit dem Mikrofon per Telefon übertragen und aufgenommen, so verringert sich die Sprecherspezifität der Laute auf ca. die Hälfte. Dabei gibt es wenig Unterschiede zwischen den verschiedenen Artikulationsmodi. Allerdings wirkt sich der Telefonkanal nicht auf alle akustischen Merkmale gleichstark aus.

4. Perzeptive Diskrimination und Identifikation von Sprechern

Wie bereits gezeigt wurde, gibt es verschiedene Ebenen auf denen sich Sprechermerkmale befinden können: auf subsegmentaler, segmentaler und auf suprasegmentaler Ebene. Aber auch innerhalb der einzelnen Ebenen gibt es veschiedene Bereiche, verschiedene Segmente, die ein unterschiedliches Maß an Sprechermerkmalen enthalten können. Klassischerweise wird den Vokalen häufig eine große Rolle in der Sprechererkennung zugeordnet. Allerdings zeigten schon einige Studien, dass auch viele der Konsonanten über ein großes Potential zur Sprechererkennung verfügen. Diese verschiedenen Bereiche und Ebenen sollen in diesem Kapitel unter dem Aspekt der menschlichen perzeptiven Identifikation und Diskrimination von Sprechern betrachtet werden. Dabei soll besonders auf Merkmale eingegangen werden, die für perzeptive Studien verwendet wurden. Außerdem wird auf den Unterschied zwischen Sprecheridentifikation und -diskrimination eingegangen und auf die Identifikation von bekannten und unbekannten Sprechern. Schließlich soll noch der Einfluss der Dauer des Sprachmaterials auf die Sprechererkennungsleitstung von Hörern betrachtet werden, bevor das eigene Experiment vorgestellt wird.

4.1. Suprasegmentale Merkmale

Die suprasegmentalen Merkmale beschreiben Eigenschaften des Sprechers, die sich durch das gesamte Sprachsignal ziehen, also nicht nur auf einzelne Abschnitte einer Äußerung beschränkt sind. Dazu zählen zum Beispiel die Grundfrequenz eines Sprechers von ihrem Mittelwert bis zu ihrem globalen Verlauf (Prosodie und Intonation) und Mikroschwankungen in der Periodizität, auch Jitter genannt. Die Grundfrequenz wurde vielfach als mögliches sprecherunterscheidendes Merkmal mit sehr unterschiedlichen Ergebnissen untersucht. Bezeichnete [52] die Grundfrequenz als ein sehr konstantes Merkmal eines Sprechers, so sagte [17], dass die Grundfrequenz eines Sprechers genauso stark variiere wie die verschie-

dener Sprecher. Auch neuere Studien verwenden verschiedene Merkmale der Grundfrequenz zur Sprechererkennung wie die mittlere F0 und ihren Verlauf ([37] [171]) oder den F0-Bereich [254].
Auch wenn sich die mittlere Grundfrequenz leicht und einfach wahrnehmen lässt, eignet sie sich nur eingeschränkt zur Sprecheridentifikation. Da die meisten Sprecher eine Grundfrequenz im mittleren Bereich haben, Männer ca. 120 Hz und Frauen ca. 230 Hz [235], ist dieser Wert oft nicht sehr sprecherspezifisch. So haben beispielsweise ca. 65 % aller (schwedischen) männlichen Sprecher eine Grundfrequenz von 100 bis 130 Hz [175]. Das bedeutet, dass nur 35 % der Sprecher extreme Grundfrequenzwerte aufweisen, anhand derer man sie von anderen Sprechern unterscheiden könnte. Fällt ein Sprecher in diese 35 %, ist die durchschnittliche Grundfrequenz ein gutes Merkmal zu seiner Identifizierung, für 65 % der Sprecher eignet sich dieses Merkmal aber nicht, da es sich nicht stark genug vom Populationsdurchschnitt abhebt.
Spezifischer für einen Sprecher ist häufig der Verlauf der Grundfrequenz, sprich die Intonation und Prosodie einer Äußerung [79]. Während dieser Verlauf natürlich auch eine lexikalische bzw. semantische Bedeutung hat, indem beispielsweise ein Wort in einem Satz betont oder eine Frage markiert wird, indem die Grundfrequenz am Satzende steigt; so ist der präzise Verlauf abhängig von gelernten Mechanismen des Sprechers.
Ein weiteres Merkmal der Grundfrequenz beschreibt ihre Regelmäßigkeit bzw. Unregelmäßigkeit, sprich, ob jede Schwingung der Stimmlippen gleich lang ist oder ob sich ihrer Längen stark voneinander unterscheiden [288]. Je unterschiedlicher die Länge der Schwingungen, desto größer die Unregelmäßigkeit und desto höher der Jitterwert. Der Jitter in einer Stimme manifestiert sich perzeptiv in der wahrgenommenen Stimmqualität eines Sprechers. Außerdem wird die Stimmqualität auch durch den sogenannten Shimmer beschrieben, welcher Unregelmäßigkeiten in der Lautstärke quantifiziert. Jitter und Shimmer tragen zusammen sehr stark zu der wahrgenommenen Stimmqualität eines Sprechers bei. Eine Stimme mit hohen Jitter- und Shimmer-Werten würde von Hörern beispielsweise als sehr behaucht, rau oder heiser wahrgenommen werden [80]. Auffällig hohe Werte können dabei auch als pathologisch klassifiziert werden [288]. Beide Merkmale werden sowohl von anatomisch-physiologischen als auch von erlernten Faktoren bestimmt. So können beispielsweise kleine Knötchen auf den Stimmlippen zu Unregelmäßigkeiten in der Schwingung führen. Andererseits wird die Stimme manchmal absichtlich verändert, um sich sozial mehr Akzeptanz oder Ansehen zu verschaffen. Bei amerikanischen Frauen ist derzeit eine Knarrstimme (creaky voice) mit positiven Attributen belegt [297], sodass viele Frauen mit einer gepressten

Stimme sprechen. Diese Stimmmodifikation führt zu langsameren und unregelmäßigeren Schwingungen der Stimmlippen und somit auch zu erhöhten Jitterwerten [32].

Neben den genannten Merkmalen, welche alle Merkmale der Frequenz sind, gibt es auch temporale Merkmale im Sprachsignal. Diese beschreiben beispielsweise, wie schnell jemand spricht [156], wie viele Pausen er macht [154] und wie das Verhältnis von Vokal- und Konsonantdauern ist [62]. Nicht alle temporalen Merkmale können perzeptiv wahrgenommen werden. Manche lassen sich nur akustisch messen und analysieren. Allerdings sind Hörer in der Lage, Sprecher anhand von temporalen F0-Merkmalen (wie z.B. einem normierten oder monotonen F0-Verlauf) über dem Zufallsniveau zu unterscheiden [63].
Auch wenn Häsitationen genau genommen nicht zu suprasegmentalen Merkmalen gehören, sollen sie hier als eher global auftretende Phänomene unter diesem Punkt Erwähnung finden. Ob Sprecher in Pausen Füllwörter wie „äh“oder „ähm“ verwenden und welche Füllwörter sie bevorzugen, kann Aufschluss über die Identität des Sprechers geben ([67] [38]). Auch wenn es zwischen verschiedenen Aufnahmezeitpunkten Variation in dem Sprecherverhalten geben kann, so wurde doch eine generelle individuelle Strategie beim Gebrauch von Häsitationen nachgewiesen [38].

4.2. Segmentale Merkmale

Im Gegensatz zu den globalen suprasegmentalen Merkmalen beschränken sich die segmentalen Merkmale auf ein Segment bzw. einen Laut. Diese phonetischen Informationen geben ebenfalls Aufschluss über die Identität des Sprechers. [246] zeigten, dass Hörer in der Lage sind, allein anhand von phonetischen Informationen in Sinuswellen-Nachbildungen natürlich-sprachlicher Sätze Sprecher zu erkennen. Diese Nachbildungen enthielten keine Informationen über die Stimmqualität eines Sprechers, sodass nur die segmentalen phonetischen Informationen übrig blieben. Genauso wie die Grundfrequenz aber nicht zur Unterscheidung aller Sprecher geeignet ist, kann auch ein Segment nicht alle Sprecher gleich gut diskriminieren. Jeder Laut bringt durch seine spezifischen Merkmale andere Sprechermerkmale zum Tragen. So lassen sich die einen Sprecher besser durch /u/, die anderen besser durch /i/ unterscheiden [39]. Es gibt allerdings Laute, die sich im Durchschnitt besser zur Sprecheridentifikation eignen als andere. In der klassischen Sichtweise wird davon ausgegangen, dass vor allem Vokale sprecherspezifische Merkmale enthalten (wie Formanten, Grundfrequenz, Stimmqualität) [6], während Konsonanten hauptsächlich für die Spracher-

kennung relevant sind [172]. Allerdings zeigten bereits einige Untersuchungen, dass auch Konsonanten einen wichtigen Teil zur Sprechererkennung beitragen (z.B. [12] [142] [143]).

4.2.1. Vokale

Vokale enthalten als stimmhafte Laute viele Informationen, da sie außer ihren segmentspezifischen Eigenschaften auch immer die Grundfrequenz des Sprechers enthalten. Die Grundfrequenz und die Formanten der Vokale sind zum Beispiel die primären Parameter zur Identifikation des Geschlechts eines Sprechers [116]. Die genaue Lage der Vokalformanten, aber auch ihre Bandbreite [31] und Transitionen [268] beinhalten weitere Sprechereigenschaften. Dabei wird klassischerweise davon ausgegangen, dass die unteren Formanten (F1 und F2) vor allem phonetische und lexikalische Informationen und die höheren Formanten (F3 und F4) vor allem die idiosynkratischen Merkmale des Sprechers tragen [171]. Dabei hängt es aber auch immer vom Sprecher ab [171] und vom Vokal ([3] [8]), welche Parameter die meisten Informationen zur Sprecheridentität liefern.
Besser als orale Vokale eignen sich nasalierte Vokale zur Sprechererkennung. Aufgrund der Physiologie des Sprechers ergeben sich unterschiedliche Resonanzräume, welche sich in einer unterschiedlichen Formantstruktur manifestieren. Durch die „Zuschaltung“ des Nasenraums bei der Artikulation von nasalierten Vokalen wirken sich noch mehr individuelle physiologische Merkmale auf die Charakteristiken des Lautes aus. Daher enthalten nasalierte Vokale - besonders die hohen Vokale /i/ und /e/ - noch mehr Sprecherinformationen als orale [9].

4.2.2. Konsonanten

Auch wenn Vokale aufgrund ihrer höheren Energie, Dauer und Sonorität mehr Sprecherinformationen enthalten als Konsonanten [6], ist letztere Lautgruppe nicht zu vernachlässigen, denn auch Konsonanten tragen in ihrer akustischen Struktur relevante Merkmale, die Aufschluss über die Identität des Sprechers geben. Verschiedene Studien mit kurzen Konsonant-Vokalsilben zeigten, dass vor allem nasale und stimmhafte Laute in ihrem akustischen Signal viele Informationen über den Sprecher enthalten und sich somit auch zur perzeptiven Sprechererkennung eignen [10]. Neben Nasalen wie dem Laut /m/ eignet sich auch der alveolare Frikativ /s/ im Vergleich zu anderen Konsonanten (/l/ und /t/) überdurchschnittlich gut zur Sprechererkennung [14]. Die besondere Eignung der Nasale kann in ihrer velaren Artikulationsgeste begründet liegen, denn die spektralen Änderungen während

der Velumsgeste scheinen habituell und/oder physiologisch vom Sprecher abzuhängen [8]. Es zeigte sich, dass allgemein die alveolaren Konsonanten mehr Informationen als labiale enthalten [11]. Diese Ergebnisse zeigten sich sowohl bei der Identifikation von bekannten als auch von unbekannten Sprechern, wobei letztere allerdings niedrigere Erkennungsraten aufwies [7].
Es blieb jedoch unklar, ob sich die Spechermerkmale tatsächlich in den Konsonanten befanden oder eher in den umliegenden Vokalen. Diese Verteilung der Merkmale unterscheidet sich zwischen den verschiedenen Konsonanten. Während bei Nasalen der vokalische und der konsonantische Silbenteil wichtig war, lag der Fokus bei den Plosiven und Frikativen eher auf den Vokaltransitionen [13]. Außerdem spielt auch die Position des Lautes innerhalb einer Silbe eine Rolle. Silben mit Konsonanten im Onset waren sprecherspezifischer als vokalinitiale Silben. Nasale halfen bei der Sprecheridentifikation sowohl in Onset- als auch in Coda-Position [11].

4.3. Identifikation und Diskrimination von bekannten und unbekannten Sprechern

4.3.1. Perzeptive Sprecheridentifikation und -diskrimination

Während die Sprecherdiskrimination bedeutet, dass ein Hörer zwei (oder mehr) Sprecher voneinander unterscheiden kann, beinhaltet die Sprecheridentifikation, dass der Hörer in der Lage ist, einen bestimmten Sprecher zu identifizieren (indem er z.B. seinen Namen nennt). Beide Verfahren werden in perzeptiven Untersuchungen verwendet, aber die Sprecheridentifikation setzt voraus, dass die Hörer alle Sprecher kennen. Das bedeutet, dass sie entweder bereits vor dem Experiment mit den Sprechern bekannt sein müssen oder dass die vor dem Experiment in einer Trainingsphase mit den Sprechern vertraut gemacht werden müssen. Beide Varianten bergen Schwierigkeiten: Sucht man Sprecher aus, die bereits alle Hörer vorher kennen, muss man darauf achten, dass sie möglichst allen Hörern gleich gut bekannt sind. Durch diese Einschränkung fällt die Anzahl der Sprecher dann meist recht klein aus [12]. Eine Möglichkeit, dem entgegen zu wirken, ist die Verwendung von Stimmen berühmter Personen (z.B. [167] [166]). Sie sind einem breiten Publikum bekannt und es gibt relativ viele. Allerdings kann man auch in diesem Fall nicht vollständig kontrollieren, wie gut ein Hörer diese berühmten Personen kennt. Der Vorteil besteht aber darin, dass es weniger zeitaufwändig ist, da man die Hörer nicht erst trainieren muss. Die zweite Möglichkeit, bei

der zuvor alle Hörer die Sprecherstimmen erlernen hat den Vorteil, dass man jeden Sprecher nehmen kann und dass nach dem Training tatsächlich alle Hörer jeden Sprecher gleich gut kennen. Allerdings ist es je nach Anzahl der Sprecher ein langwieriger Prozess, den Hörern die Sprecherstimmen beizubringen.
Auch wenn beide Methoden auf den gleichen Grundlagen, nämlich den spezifischen Stimmmerkmalen eines Sprechers beruhen, laufen beide Prozesse auf unterschiedliche Weise ab. So schreibt [165], dass die Diskrimination von unbekannten Sprechern in anderen Gehirnarealen stattfindet als die Identifikation von bekannten Sprechern. Menschen, deren rechte Gehirnhälte geschädigt war, konnten bekannte Sprecher nicht mehr erkennen, wohingegen sie unbekannte Sprecher weiterhin gut unterscheiden konnten. Auf der anderen Seite löste eine Schädigung von beiden Gehirnhälften eine eingeschränkte Diskriminationsfähigkeit von unbekannten Sprechern aus.

4.3.2. Identifikation bekannter und unbekannter Sprecher

Sowohl der Prozess der Identifikation als auch der Diskrimination spielen in der perzeptiven Sprechererkennung eine wichtige Rolle. Die Sprecherdiskrimination bildet quasi die Grundlage für die Identifikation, da sie zeigt, in welchen Dimensionen sich Sprecher prinzipiell unterscheiden können. Diese Ergebnisse geben dann Hinweise darauf, welche Merkmale sich für die Sprecheridentifikation nutzen lassen würden [254].
Typischerweise werden Identifikationsstudien mit vorher bekannten Sprechern durchgeführt und Diskriminationsstudien mit unbekannten Sprechern. Wie bereits erwähnt, gibt es aber auch die Möglichkeit, Hörer mit zuvor unbekannten Sprechern vertraut zu machen, sodass sie zu einem gewissen Grad bekannt werden. Bei der Unterscheidung zwischen der Identifikation bekannter und unbekannter Sprecher steht (in diesem Fall) der Ausdruck „bekannt“ für die vor dem Experiment bekannten Sprecher, während „unbekannt“ für die erst im Experiment erlernten Sprecher steht. Auch wenn diese Variante in Perzeptionstudien weniger häufig vorkommt, wurde sie bereits in Untersuchungen verwendet (z.B. in [290] und [7]). Dabei wird den Hörern Sprachmaterial der Sprecher präsentiert, welches sich vom Testmaterial unterscheidet. Nach einer bestimmten Anzahl an Wiederholungen oder wenn sich die Hörer sicher fühlen, wird das Training beendet. Oft folgt anschließend noch ein Übungsteil, wo die Hörer die Sprecher identifizieren sollen und nach jeder Antwort Feedback erhalten, ob sie den Sprecher richtig erkannt haben. Dieser Schritt hilft sicher zu stellen, dass die Hörer die Sprecher zu einem ausreichend hohen Prozentsatz identifizieren können. Soll also die Identifikation von Sprechern getestet werden, so kann zuvor eine aufwändigere

Vorbereitung notwendig sein. Entweder müssen Sprecher ausgewählt werden, die möglichst allen Hörern gleichermaßen bekannt sind oder die Hörer müssen auf die Sprecherstimmen trainiert werden.

4.4. Einfluss der Dauer des Sprachmaterials

Je nach Fragestellung wird Sprachmaterial unterschiedlicher Länge für die perzeptive Sprecherdiskrimination bzw. -identifikation verwendet. Betrachtet man verschiedene Studien (siehe [167] als Übersicht), so wird deutlich, dass je länger das Sprachmaterial war, desto besser konnten die Sprecher von den Hörern erkannt werden ([39] [168] [163]). Während bei sehr kurzen Stimuli, bestehend aus nur einem Laut (meistens einem Vokal) die Identifikationsrate der Hörer zwischen 40 und 50 % lag, stieg sie bei Silben (bestehend aus zwei Lauten) schon auf 80 bis 90 % und erreichte bei Stimuli, die einen ganzen Satz beinhalteten oder mehrere Sekunden dauerten, fast 100 % [167]. Diese Ergebnisse zeigen, dass je mehr Sprachmaterial vorhanden ist, desto mehr Informationen sind über den Sprecher darin enthalten und desto besser können die einzelnen Sprecher von den Hörern erkannt werden. Mit abnehmendem Umfang des Sprachmaterials gewinnt der phonetische Inhalt an Bedeutung (siehe Abschnitt 4.2.1 und 4.2.2). Soll eine möglichst hohe Identifikations- bzw. Diskriminationsrate der Sprecher erreicht werden, so muss entweder ausreichend langes Sprachmaterial zur Verfügung stehen oder die zu verwendenden Laute müssen mit großer Sorgfalt ausgewählt werden. Dies gilt besonders, wenn sich die zu unterscheidenden Sprecher sehr ähnlich sind oder ein große Anzahl von Sprechern unterschieden werden soll.

4.5. Experiment 1: Sprecherdiskrimination anhand von statischen und dynamischen Informationen

4.5.1. Hypothesen und Ziele

In der Literatur wurde bereits gezeigt, dass nicht nur Vokale (z.B. [171] [31] [268]), sondern auch Konsonanten sprecherspezifische Merkmale enthalten, die für die Unterscheidung von Sprechern relevant sind ([101] [295] [68] [12] [14] [15] [142] [143] [82]). Dabei gibt es jedoch Unterschiede im Grad des Informationsgehalts zwischen den Konsonanten. So sehen die meisten Studien Nasale und Frikative in Punkto Sprecherspezifität auf den vorderen Plätzen, während Plosive fast immer als nicht sehr informativ dargestellt werden ([68] [14] [15] [82]).

Außerdem wurde gezeigt, dass die Artikulationsstelle der Konsonanten ebenfalls einen Einfluss auf die Sprecherspezifität hat. Beides hat auch die akustische Analyse in Kapitel 3 ergeben. [11] zeigten für das Japanische, dass die untersuchten alveolaren Laute stärker zur Sprecheridentifikation beitrugen als die labialen.
Bisher gibt es aber noch keine Untersuchung von deutschen Sprachlauten. Da sich die Laute in ihrer Artikulation zwischen den Sprachen zwar ähneln, aber immer noch unterscheiden, ist unklar, ob für deutsche Laute das gleiche gilt wie für englische ([68] [14] [15] [82]) oder japanische ([12] [11]). Um den exakten Beitrag jedes Lauts zur Sprecherdiskrimination bestimmen zu können, wurde in dieser Untersuchung auf eine einheitliche Vokalumgebung bei der Realisierung der Konsonanten geachtet, sodass sich Unterschiede in der Sprecherdiskriminationsleistung direkt auf den jeweiligen Konsonanten zurückführen lassen. Die kurze Dauer des Stimulus sorgt außerdem dafür, dass die Sprecherdiskrimination ausreichend schwierig ist, um Unterschiede zwischen den Konsonanten zu zeigen.
Da es das Hauptanliegen diese Studie war, herauszufinden, welche konsonantischen Merkmale sich zwischen Sprechern unterscheiden, wurde ein Sprecherdiskriminationstest als Methode ausgewählt. In einem klassischen AX-Diskriminationsdesign mussten die Versuchsteilnehmer entscheiden, ob zwei Stimuli vom gleichen oder von zwei verschiedenen Sprechern stammten. Dabei sollte gezeigt werden, dass Hörer in der Lage sind, auf Grund von sehr kurzen und isolierten Stimuli („Unsinnwörtern“) Sprecher sicher zu unterscheiden. Es wurde angenommen, dass je nach Konsonant die Diskriminationsleistung der Hörer unterschiedlich hoch sein würde. Aufgrund der bisherigen Ergebnisse aus der Literatur und der vorangegangenen akustischen Analyse in dieser Arbeit wurden für den Artikulationsmodus der Nasale und Frikative höhere Diskriminationsraten erwartet als für die Plosive. Außerdem wurde angenommen, dass die Konsonanten der alveolaren Artikulationsstelle bessere Erkennungsraten hervorrufen als die labialen Konsonanten. Dabei stellte sich die Frage, ob der Einfluss der Artikulationsstelle für alle Artikulationsmodi gleich groß ist oder Interaktionen auftreten.

4.5.2. Methode

Versuchspersonen

30 Studenten der Ludwig-Maximilians-Universität nahmen gegen eine kleine Bezahlung an der Studie teil. Alle Teilnehmer waren deutsche Muttersprachler und hatten keine bekannten Höreinschränkungen.

Sprachmaterial

Zur Erstellung der Stimuli wurden zunächst 12 männliche Sprecher aus Bayern aufgenommen. Die Beschränkung auf diesen einen Dialekt erfolgte, da im Deutschen viele unterschiedliche Dialekte existieren, die sich auf die Aussprache einzelner Laute auswirken können. Um eine einheitliche Aussprache zu gewährleisten, wurden nur Sprecher des Bayrischen ausgewählt. Alle sprachen die Stimuli /ama/, /ana/, /afa/, /ava/, /asa/, /aza/, /aʃa/, /apa/, /ata/ in Isolation mit 8 Wiederholungen in randomisierter Reihenfolge. Der Kontextvokal war dabei immer gleich, um für alle Konsonanten ein gleiches Maß an koartikulatorischem Einfluss zu gewährleisten. Der Vokal /a/ wurde ausgewählt, weil er am stärksten geöffnet und somit am weitesten von allen konsonantischen Lauten entfernt ist. Somit garantiert er ähnlich weite Wege der Artikulatoren zu den Artikulationsstellen der Konsonanten.
Die Aufnahmen wurden in einem schallisolierten Raum mit einem Studiomikrofon und der Software SpeechRecorder [64] bei einer Abtastrate von 44100 Hz durchgeführt. Die 9 verständlichsten Sprecher wurden ausgewählt und jeweils 4 Stimuli pro Konsonant manuell ausgesucht. Kriterien für die auditive Beurteilung der Stimuli waren deren Deutlichkeit und Ähnichkeit in der Intonation. Anschließend wurden alle Stimuli in ihrer Lautstärke normiert und in Praat [35] geschnitten und wie folgt bearbeitet.

Design

Zur Verringerung der Sprecherdiskrimination auf der Basis von rein akustischen Signalunterschieden mittels des echoischen Gedächtnisses [202], wurde zwischen die Stimuli 500 ms rosa Rauschen geschnitten ([51] [292]). Durch die kurze Unterbrechung waren die Hörer gezwungen, sich ein Sprechermodell zu konstruieren und zu merken, wodurch die Diskriminationsaufgabe schwieriger wurde. Rosa Rauschen entsteht durch die Filterung von weißem Rauschen, die es dem menschlichen Gehör anpassen soll. Im Gegensatz zu weißem Rauschen, dessen Energie über die Frequenz gleichverteilt ist, enthält rosa Rauschen gleiche Energie pro Oktave. Daher klingt es für menschliche Ohren natürlicher und angenehmer [282]. Deshalb wird es in Perzeptionsstudien bevorzugt verwendet (z.B. in [14] und [82]).
Da Unterschiede zwischen den Konsonanten sichtbar gemacht werden sollten, durfte die mittlere Erkennungsleistung nicht zu hoch sein. Wenn alle Hörer alle Sprecher zu 100 % erkennen, werden keine Unterschiede sichtbar. Daher mussten die Stimuli so manipuliert werden, dass die Diskriminationsaufgabe schwerer wurde und keiner der Konsonanten eine Erkennungsrate von 100 % erreichte. Um herauszufinden, wie stark und in welchen Parametern die Stimuli dafür bearbeitet werden mussten, wurden 3 kurze Pilotstudien mit einer

geringen Anzahl an Versuchsteilnehmern nach dem beschriebenen Design durchgeführt. 3 Versuchspersonen hörten die „originalen“ Stimuli mit nur normierter Amplitude. Bei diesem Versuch lag die durchschnittliche Diskriminationsrate bei 95 %. 3 weitere Versuchspersonen hörten die Stimuli mit normierter Amplitude und einer zusätzlich vereinheitlichten Grundfrequenz sowie einer flachen Intonationskontur. Dadurch gab die Grundfrequenz der Sprecher und ihre Intonation keine Informationen über ihre Identität preis. Durch diese Manipulation konnte die Erkennungsrate auf 89 % gesenkt werden. 3 weitere Versuchspersonen hörten die wie eben beschrieben manipulierten Stimuli, deren Vokaldauern zusätzlich einheitlich auf 120 ms gesetzt wurden. Die Diskriminationsleistung sank dadurch weiter auf 85 %. Vor Beginn des Hauptexperiments wurden die Anfangs- und Endvokale noch einmal gekürzt, diesmal auf 50 ms. Diese Stimuli wurden dann für das eigentliche Experiment benutzt.
Damit bei jedem Vergleich jede Antwort gleich wahrscheinlich war, musste es genauso viele same-speaker wie different-speaker Vergleiche geben. Da es naturgemäß mehr Möglichkeiten für Vergleiche zwischen verschiedenen Sprechern als zwischen gleichen gibt, wurden 4 verschiedene Versionen eines Stimulus von jedem Sprecher ausgewählt. Dies sicherte auch eine Unterscheidung der Stimuli aufgrund von Sprachverarbeitung und nicht aufgrund auditorischer Unterschiedserkennung [14]. Bei 9 Sprechern ergaben sich 36 different-speaker Vergleiche und durch die Wiederholungen der Stimuli konnten ebenfalls 36 same-speaker Vergleiche erstellt werden, sodass die Anzahl ausgeglichen war. Mit einem Perl-Skript wurden die Stimuli randomisiert und gepaart. Dabei wurde darauf geachtet, dass jeder Vergleichstyp (same- oder different-speaker) höchstens 3 mal hintereinander vorkam.

Durchführung

Das Experiment wurde mit der Software E-Prime 2.0 [238] nach dem klassischen AX-Design erstellt. Das heißt, es wurden in jedem Trial zwei Stimuli präsentiert und die Aufgabe der Hörer war es zu entscheiden, ob sie zweimal den gleichen Sprecher oder zwei verschiedene gehört haben. Zur Vermeidung von Störgeräuschen wurde das Experiment in einem schallisolierten Raum durchgeführt. Auf dem Bildschirm des Computers wurde zunächst ein weißes Kreuz in der Mitte angezeigt und mit der Wiedergabe des zweiten Stimulus wurden dann zwei Kästen eingeblendet. Auf der linken Seite stand „gleicher Sprecher“ und auf der rechten Seite „anderer Sprecher“. Ab dem Ende des ersten Stimulus musste der Hörer so schnell wie möglich die „1“ (gleicher Sprecher) oder die „0“ (anderer Sprecher) auf der Tastatur zu drücken. Das Experiment wurde in Blöcke gegliedert, wobei in jedem Block immer nur ein Konsonant auftrat. Jeder der 9 Blöcke enthielt dann 36 same-

speaker und 36 different-speaker Vergleiche, sodass jeder Hörer insgesamt 648 Stimuli-Paare beurteilen musste. Nach jedem Block gab es die Möglichkeit, eine Pause zu machen und anschließend selbständig fortzufahren.
Nach dem Experiment haben die Teilnehmer einen Fragebogen zu ihren persönlichen Daten ausgefüllt. Insgesamt dauerte das Experiment ca. 45 Minuten.

4.5.3. Statistik - Diskriminationsfähigkeit und Antworttendenz

Laut [223] werden Diskriminationsentscheidungen von zwei Komponenten beeinflusst: Von der perzeptiven Diskriminationsfähigkeit (auch Sensitivität genannt) und der Antworttendenz. In der Signalerkennung werden diese Komponenten durch d' (d-Prime) und β (beta) geschätzt.
Wenn eine Versuchsperson die Aufgabe erhält, zu entscheiden, ob zwei Stimuli gleich sind oder verschieden, hängt ihre Antwort von diesen zwei Komponenten ab. Jedes Stimuluspaar gehört entweder der Kategorie „gleich“ oder „verschieden“ an. Als Antwort kommen vier verschiedene Möglichkeiten in Frage (siehe Tabelle 4.1).

	Antwort	
Trial	verschieden	gleich
verschieden	Hit	Miss
gleich	False Alarm	Correct Rejection

Tabelle 4.1.: Auswertungsabelle der Antworten

Erkennt die Versuchsperson zwei verschiedene Stimuli als verschieden, zählt die Antwort als Hit. Antwortet sie mit „gleich“ obwohl die Stimuli verschieden sind, ist es ein Miss. Sind beide Stimuli gleich und der Hörer antwortet mit „verschieden“, ist es ein False Alarm (FA) und sind sie tatsächlich gleich eine Correct Rejection (CR). Die vier möglichen Ergebnisse sind jedoch nicht unabhängig voneinander:

$$Hit + Miss \;=\; Gesamtanzahl\ der\ „verschieden“ - Trials \tag{4.1}$$

$$FA + CR \;=\; Gesamtanzahl\ der\ „gleich“ - Trials \tag{4.2}$$

Deshalb berechnet man häufig nur die zwei unabhängigen Größen der Hit Rate und der FA Rate:

$$Hit\ Rate\ =\ \frac{Hits}{Hits + Misses} \tag{4.3}$$

$$FA\ Rate\ =\ \frac{FA}{FA + CR} \tag{4.4}$$

Die Gesamtanzahl der richtigen Antworten berechnet sich wie folgt:

$$\%richtig\ =\ \frac{Hits + CR}{Hits + Misses + FA + CR} \tag{4.5}$$

Da die Anzahl der „gleich“ und der „verschieden“-Trials gleich ist, kann die Prozentzahl an richtigen Antworten auch einfach über den Durchschnitt der Hit und FA Rate berechnet werden.
Wie bereits erwähnt, charakterisieren zwei Komponenten das Antwortverhalten einer Versuchsperson: Die perzeptive Diskriminationsfähigkeit und die Antworttendenz. Zur eingehenderen Erklärung dieser Komponenten, betrachten wir folgende drei Beispiele extremen Verhaltens:

- Ein Hörer (a) mit einer Hit Rate von 100 % und einer FA Rate von 0 % hat eine perfekte Diskriminationsfähigkeit und keine Antworttendenz.
- Ein Hörer (b) antwortet immer mit „verschieden“: Seine Hit und seine FA Rate liegen beide bei 100 %. Er weist keine Diskriminationsfähigkeit auf und eine sehr starke Antworttendenz.
- Ein Hörer (c) mit einer Hit und einer FA Rate von jeweils 50 % hat keine Antworttendenz und keine Diskriminationsfähigkeit, da seine Antworten dem Zufallsniveau entsprechen.

Um diese beiden Komponenten zu trennen, wird d’ berechnet, welches die Distanz zwischen den Verteilungen „gleich“ und „verschieden“ + Antworttendenz darstellt.

$$d'\ =\ z(Hit\ Rate) - z(FA\ Rate) \tag{4.6}$$

Ein hoher d’-Wert spricht somit für eine hohe Distanz zwischen diesen beiden Verteilungen und somit für eine gute Diskriminationsfähigkeit.
Um die Antworttendenz zu schätzen, wird β verwendet und kann folgendermaßen berechnet werden:

$$\beta = \exp \frac{-z(Hit\ Rate) \cdot z(Hit\ Rate)}{2} + \frac{-z(FA\ Rate) \cdot z(FA\ Rate)}{2} \tag{4.7}$$

Ein β von 1,0 würde in diesem Fall für keine Antworttendenz sprechen. Nähert sich der Wert 0,0 würde der Hörer eher „verschieden“ auf gleiche Stimuli antworten und hätte somit eine liberale Antworttendenz. Liegt β über 1,0, ist die Antworttendenz konservativ und der Hörer beurteilt verschiedene Stimuli eher als „gleich“. Letzteres tritt häufiger auf und ist intuitiv auch leichter zu verstehen. Wenn sich ein Hörer unsicher ist, ob es Unterschiede gibt, antwortet er eher mit „gleich“ [223].

4.5.4. Ergebnisse

Die Ergebnisse des Experiments wurden in einer Datentabelle gespeichert und in R Version 2.13.0 [241] importiert. Die Diskriminationsrate wurde zunächst in Prozent berechnet, um einen ersten Eindruck von den Leistungen der Versuchspersonen und den Unterschieden zwischen den Konsonanten zu gewinnen. Zur statistischen Analyse wurde die Diskriminationsfähigkeit in d'-Werten berechnet. Einerseits ist diese Einheit in der Literatur der Detektionstheorie verbreitet [223], andererseits kann man auf proportionalen Daten keine Varianzanalyse durchführen, da Effekte im mittleren Bereich weniger ins Gewicht fallen als Effekte an den Rändern (Nahe 0 oder 100 %) [23].

Die Hörer waren sehr gut in der Lage, die verschiedenen Sprecher zu unterscheiden, was sich in einer durchschnittlichen Diskriminationssrate von 83 % zeigte. Zur grafischen Darstellung wurden Balkendiagramme mit Fehlerbalken erstellt. Dabei zeigt der Balken den Mittelwert an und der Fehlerbalken das dazugehörige Schätzungsintervall, welches jeweils einem Standardfehler entspricht. Die Balkendiagramme zeigten deutliche Unterschiede zwischen den d'-Werten der verschiedenen Konsonanten (siehe Abbildung 4.1). Wenn sich zwei Fehlerbalken nicht überlappen, ist dies ein erster Hinweis darauf, dass sich zwei Größen wahrscheinlich signifikant voneinander unterscheiden. Entsprechend den deskriptiven Daten erreichten die Laute /n/, /ʃ/ und /t/ die höchsten d'-Werte und somit die besten Ergebnisse. Die übrigen Laute lagen leicht darunter und mehr oder weniger nah beieinander. Der durchschnittliche d'-Wert lag bei 2,11. Der Faktor Artikulationsstelle schien je nach Artikulationsmodus einen unterschiedlichen Einfluss zu haben. Während die Artikulationsstelle beim Artikulationsmodus der Frikative wenig Einfluss zu haben schien, beeinflusste sie die Nasale und Plosive deutlich (siehe Abbildung 4.2).

Zur statistischen Bestätigung der grafischen Eindrücke wurde eine Varianzanalye (ANOVA) in R mit der Funktion „ezANOVA"([20]) durchgeführt. Dabei war der *d'-Wert* die abhängige Variable, die *Versuchsperson* der Fallidentifikator und der *Konsonant* (/m, n, f, s, ʃ, v, z, p, t/) der unabhängige Faktor innerhalb der Versuchsperson. Die Analyse zeigte, dass sich der Faktor *Konsonant* signifikant auf den d'-Wert auswirkt ($F[8,232] = 5{,}00$; $p < 0{,}001$). Das heißt, dass sich die verschiedenen Konsonanten unterschiedlich gut zur Sprecherdiskrimination eignen.

Anschließend wurde der Einfluss von Artikulationsstelle und -modus und deren Interaktion auf die Diskriminationsfähigkeit der Hörer analysiert. Da sich der post-alveolare Frikativ /ʃ/ nicht den Artikulationsstellen „labial" oder „alveolar" zuordnen lässt, wurde er bei dieser Analyse ausgeschlossen. In der Varianzanalyse war die abhängige Variable wieder der *d'-Wert*, *Versuchsperson* der Fallidentifikator sowie *Artikulationsstelle* (labial, alveolar) und *Artikulationsmodus* (Nasal, stimmhafter Frikativ, stimmloser Frikativ, Plosiv) die unabhängigen Faktoren. Die *Artikulationsstelle* beeinflusste die Diskriminationsrate signifikant ($F[1,29] = 5{,}57$; $p = 0{,}03$), während der *Artikulationsmodus* keinen signifikanten Einfluss hatte ($F[3,87] = 2{,}18$; $p = 0{,}096$).

Die Interaktion zwischen *Artikulationsstelle* und *Artikulationsmodus* wurde ebenfalls signifikant ($F[3,87] = 5{,}53$; $p = 0{,}002$). Das bedeutet, dass ihr Einfluss nicht bei allen Konsonanten gleich groß war. Um zu ermitteln, in welchen Artikulationsmodi sich die alveolaren und labialen Konsonanten tatsächlich signifikant unterschieden, wurden paarweise *t*-tests durchgeführt. Da die Werte pro Versuchsperson miteinander verglichen wurden, kam ein gepaarter *t*-test zur Anwendung. Aufgrund der multiplen Vergleiche wurde eine Bonferroni-Korrektur zur Anpassung des p-Wertes durchgeführt. Das Signifikanzniveau lag daher bei $p = 0{,}0167$. Beim Vergleich der Artikulationsstelle zeigte sich, dass nur bei den Nasalen ($t = -2{,}79$; $df = 29$; $p = 0{,}009$) und den Plosiven ($t = -3{,}52$; $df = 29$; $p = 0{,}001$) die alveolaren Laute höhere d'-Werte erzielten als die labialen. Im Artikulationsmodus der Frikative hatte der Faktor *Artikulationsstelle* keinen signifikanten Einfluss.

Der Artikulationsmodus hatte in der Varianzanalyse keinen signifikanten Einfluss. Dennoch soll genauer betrachtet werden, ob es zwischen bestimmten Konsonanten verschiedener Artikulationsmodi vielleicht doch signifikante Unterschiede gibt. Mittels eines gepaarten *t*-tests wurden die Artikulationsmodi innerhalb einer Artikulationsstelle miteinander verglichen. Dabei zeigte sich, dass der alveolare Nasal /n/ signifikant höhere Werte erzielte als der alveolare Frikativ /z/ ($t = 3{,}50$; $df = 29$; $p = 0{,}002$); und der alveolare Plosiv /t/ höhere als der alveolare Frikativ /z/ ($t = -3{,}52$; $df = 29$; $p = 0{,}001$).

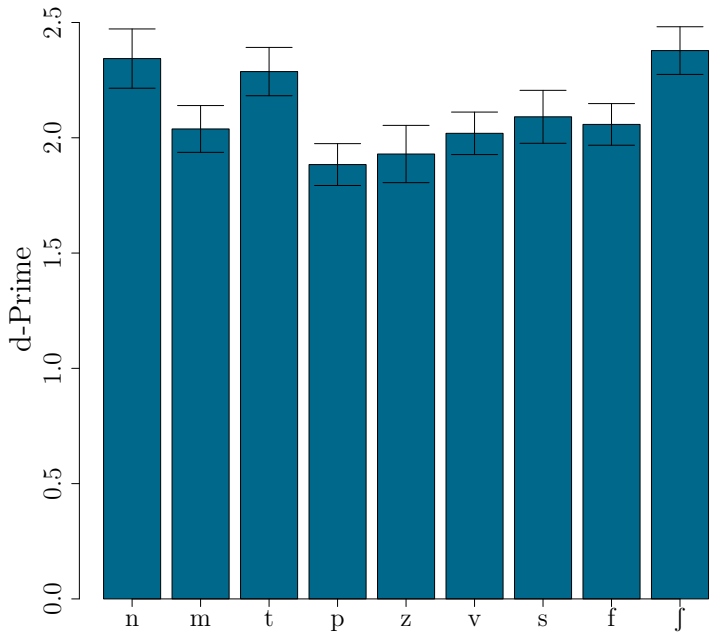


Abbildung 4.1.: d' (d-Prime) in Abhängigkeit vom Konsonant mit Fehlerbalken

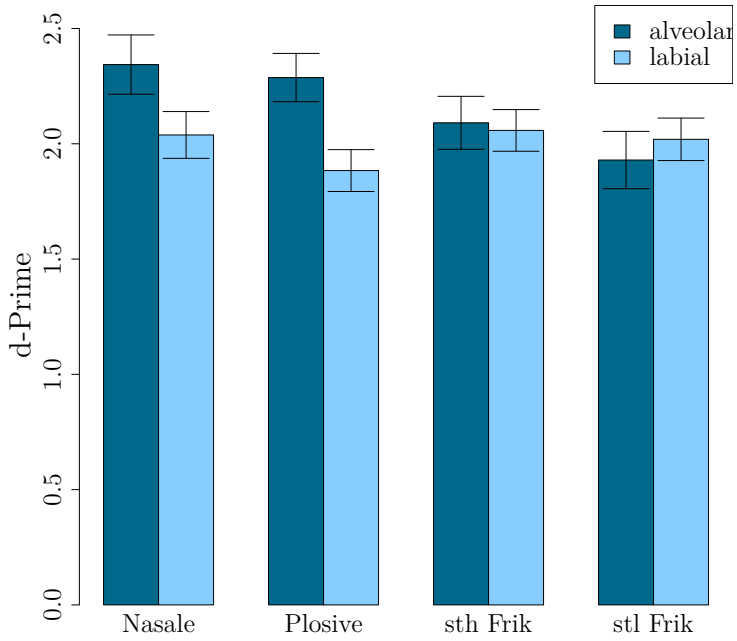


Abbildung 4.2.: d' (d-Prime) in Abhängigkeit von Artikulationsstelle und Artikulationsmodus des Konsonanten mit Fehlerbalken

Innerhalb der labialen Artikulationsstelle unterschieden sich die Artikulationsmodi nicht signifikant. Diese Ergebnisse zeigen, dass sich die Artikulationsmodi in der alveolaren Artikulationsstelle stärker unterscheiden als in der labialen. Die stimmhaften und stimmlosen Frikative unterschieden sich nicht signifikant voneinander.
Zusammenfassend ist festzustellen, dass sich bei Nasalen und Plosiven der Einfluss der Artikulationsstelle signifikant stärker zeigt als bei den Frikativen und dass sich die Artikulationsmodi der alveolaren Artikulationsstelle signifikant stärker unterscheiden als die der labialen.

4.5.5. Diskussion

Wie erwartet, konnten die Versuchspersonen die Sprecher mit einer großen Zuverlässigkeit voneinander unterscheiden, auch wenn das Sprachmaterial sehr kurz und limitiert war. Wie gut die Sprecher unterschieden werden konnten, beeinflusste auch der phonetische Inhalt des Stimulus. Während /n/, /ʃ/ und /t/ zu sehr hohen Erkennungsleistungen beitrugen, waren sie bei den anderen Konsonanten etwas geringer. Dass Nasale und (alveolare) Frikative viele Sprecherinformationen enthalten, ist aus der Literatur bekannt ([68] [179] [12] [14] [142] [143]). Überraschend war das gute Abschneiden von /t/, da Plosive normalerweise wenig Sprecherinformationen beinhalten [68]. Dass /t/ aber im Vergleich zu /p/ die Sprecher besser unterschied, ist im Einklang mit vorherigen Studien, die labiale Plosive allgemein als weniger markant und charakteristisch beschreiben als Plosive anderer Artikulationsstellen [146]. Das mag unter anderem daran liegen, dass labiale Plosive wegen der geringeren Länge des vorderen Vokaltrakts einen schwächeren Burst haben als alveolare Plosive [220]. Vielleicht befinden sich wichtige Informationen über den Sprecher nicht nur im statischen Bereich des Konsonanten, sondern auch in den dynamischen Transitionen zu den benachbarten Vokalen. [221] zeigten, dass Konsonanten und Vokale sich aufgrund von Koartikulation überlagern, wodurch sich beide Laute nicht klar voneinander trennen lassen. Gerade der Verlauf der Transitionen vom vorangehenden Vokal zum Konsonanten bestimmt bei Plosiven, welche die schnellsten und abruptesten Transitionen aufweisen, häufig die perzeptiv wahrgenommene Artikulationsstelle ([274] [230]). Gleiches gilt auch für die Nasale, die ein ähnliches Ergebnis zeigten wie die Plosive. Die alveolaren Nasale erwiesen sich als sprecherspezifischer als die labialen. Da sich die Artikulationsstellen aufgrund der statischen konsonantischen Merkmale bei beiden Lautgruppen nur minimal voneinander unterscheiden, liegt die Vermutung nahe, dass die Unterschiede in ihrer Sprecherspezifität auf ihren Transitionen beruhen. Da die Transitionen die wahrgenommene Artikulationsstelle

des Plosives oder Nasals bestimmen, ist es logisch, dass sich die Transitionen zwischen labialen und alveolaren Plosiven und Nasalen unterscheiden. So weist beispielsweise das /n/ eine starke Energieveränderung im Bereich von 1450 bis 2300 Hz am Übergang zum Folgevokal auf, da sich in diesem Bereich ein Antiformant befindet, was bei /m/ nicht der Fall ist [160]. Außerdem bewegt sich die Zungenspitze als Artikulator schneller als die Lippen [28]. Da die alveolaren Laute mit der Zungenspitze und die labialen mit den Lippen produziert werden, ist klar, warum die alveolaren Laute schnellere Transitionen aufweisen können. Diese schnelleren Transitionen wiederum rufen eine starke Energieveränderung hervor, welche viele akustische Informationen liefert.
Wenn die Transitionen so einen starken Einfluss auf die Lautwahrnehmung haben, ist anzunehmen, dass sie sich ebenfalls auf die Sprechererkennung auswirken, da phonetische und Sprecherinformationen in starkem Maße miteinander interagieren. Wenn also wenige akustische Unterschiede zwischen den statischen Merkmalen innerhalb der Konsonanten existieren, ist es wahrscheinlich, dass die Differenzen in der Sprecherspezifität in Bereichen liegen, die sich akustisch stärker unterscheiden, wie z.B. die Transitionen. Ein prinzipieller Unterschied zwischen labialen und alveolaren Transitionen sagt zwar noch nichts über Unterschiede in ihrer Sprecherspezifität aus, aber die Ergebnisse lassen vermuten, dass die Transitionen der alveolaren Nasale und Plosive mehr Sprecherinformationen enthalten als die der labialen. Dennoch lässt sich nicht ausschließen, dass auch die akustischen Unterschiede der statischen Merkmale innerhalb der Konsonanten zu den unterschiedlichen Werten zwischen alveolaren und labialen Lauten geführt haben. Wahrscheinlich tragen beide Informationstypen zu den gefundenen Differenzen in der Sprecherspezifität bei.
Frikative sind im Gegensatz zu Nasalen und Plosiven in ihrer Wahrnehmung nicht so stark von ihren Transitionen abhängig und sind deshalb „unabhängiger“ von ihrem Vokalkontext. Daraus folgt, dass die Informationen zur Lautidentifikation tatsächlich innerhalb des Frikatives liegen und weniger in seinen Übergängen zu den angrenzenden Vokalen. Das könnte wiederum an den langsameren Transitionen der Frikative im Vergleich zu Nasalen und Plosiven liegen. Andererseits bestimmt die Länge des vorderen Vokaltraktteils den Frequenzschwerpunkt des Frikatives. Dieser ist perzeptiv gut wahrnehmbar und gibt somit wichtige Informationen über die Artikulationsstelle preis (siehe Abschnitt 4.2.2). Aufgrund der Interaktion von phonetischen und Sprechermerkmalen sind deshalb die Sprecherinformationen ebenfalls an dieser Position zu vermuten. Demzufolge dürften die Transitionen der Frikative keinen (oder nur einen geringen) Einfluss auf den Sprecherinformationsgehalt haben. Da die Ergebnisse keinen signifikanten Unterschied zwischen den labialen und al-

veolaren Frikativen zeigten, scheinen die Transitionen tatsächlich keinen Einfluss auf den Sprecherinformationsgehalt der Frikative gehabt zu haben.
Vom Einfluss der Transitionen abgesehen, wären dennoch stärkere Unterschiede zwischen den labialen und alveolaren Frikativen zu erwarten gewesen. Eine mögliche Ursache für die Abwesenheit solcher Unterschiede liegt möglicherweise in der Amplitudennormierung der Stimuli. Da diese Normierung über alle Stimuli erfolgte, ohne die spezifische Lautstärke der unterschiedlichen Konsonanten zu berücksichtigen, könnte der Einfluss dieser Merkmale auf die Sprecherdiskriminationsfähigkeit des Lautes entfernt worden sein. Beispielsweise ist normalerweise der alveolare Frikativ /s/ lauter als der labiale Frikativ /f/ [272]. Es ist zu vermuten, dass Laute mit höherer Lautstärke stärker zur Sprechererkennung beitragen, da ihre Merkmale prominenter sind. Durch die Angleichung der Lautstärke können somit Unterschiede in der Sprecherdiskriminationsfähigkeit der Laute verloren gegangen sein. Das heißt, ein /f/ könnte von seiner Lautstärke her nun genauso gut zur Sprechererkennung beitragen wie ein /s/. Dies entspricht den Ergebnissen dieses Experiments. Für eine zukünftige Untersuchung wäre es ratsam, die Lautstärke getrennt nach Konsonanten zu normieren, sodass lautspezifische Amplitudenwerte erhalten bleiben.
Es lässt sich abschließend zusammenfassen, dass sowohl die Plosive als auch die Nasale stark von ihren Transitionen bestimmt werden, sodass diese wahrscheinlich auch großteils die Unterschiede in der Sprecherdiskriminationsfähigkeit zwischen labialen und alveolaren Lauten verursachen. Für die Frikative ergaben sich keine Unterschiede in der Sprecherspezifität, sodass weder die Transitionen noch die konsonantischen Informationen Unterschiede zwischen alveolaren und labialen Lauten hervorzurufen scheinen. Sprich, Laute mit der Lokalisation vieler akustischer Informationen in den Transitionen zeigen Unterschiede zwischen den Artikulationsstellen, während Laute, deren akustische Informationen eher im statischen Bereich des Konsonanten liegen, keine solchen Unterschiede besitzen. Somit würde die Art der Sprecherinformationen in den Lauten („statisch“ vs. „dynamisch“) die Abhängigkeit der Sprecherdiskriminationsfähigkeit von der Artikulationsstelle erklären.

4.6. Experiment 2: Sprecherdiskrimination anhand von statischen Informationen

4.6.1. Hypothesen und Ziele

Das erste Experiment bewies die gute Diskriminationsfähigkeit der Hörer. Zur Unterscheidung der Sprecher verwendeten die Hörer Informationen, die sich sowohl in den statischen Bereichen Konsonanten als auch in den dynamischen Bereichen der Vokaltransitionen befanden. Um detaillierter den Einfluss der Konsonanten auf die Sprechererkennung untersuchen zu können, sollte der Informationseinfluss der Vokale (und der Transitionen) vollständig neutralisiert werden. Zu diesem Zweck wurden neue Stimuli mit sprecherneutralen Vokalen erzeugt, sodass sich die Stimuli nur noch in ihren konsonantischen Anteilen unterschieden. Dadurch wurden die Hörer gezwungen, zur Sprecherdiskrimination ausschließlich die Sprecherinformationen in den Konsonanten (ohne Transitionen) zu verwenden.
Es wurde erwartet, dass die Erkennungsrate sinkt, da die Aufgabe wesentlich schwieriger war. Durch das Eliminieren der dynamischen Informationen sollte ein stärkeres Hervortreten der (statischen) konsonantischen Informationen erreicht werden. Es würde sich so besser zeigen, wie viele Sprecherinformationen die Konsonanten selbst liefern und inwiefern sich diese zwischen den Artikulationsmodi und der Artikulationsstelle unterscheiden. Es ist zu vermuten, dass die Nasale und Plosive, deren phonetische und Sprecherinformationen eher in den dynamischen Transitionen liegen, stärker in ihrer Sprecherspezifität sinken als die Frikative, deren Informationen eher im statischen Bereich des Konsonanten selbst liegen.

4.6.2. Methode

Versuchspersonen

31 Studenten der Ludwig-Maximilians-Universität nahmen gegen eine kleine Vergütung an der Studie teil. Alle Teilnehmer waren deutsche Muttersprachler und hatten keine bekannten Höreinschränkungen. Keiner der Hörer hatte zuvor an dem ersten Experiment oder an den Vorversuchen teilgenommen.

Sprachmaterial

Die Stimuli und ihre Vorverarbeitung (Normierung von Amplitude, Grundfrequenz, Grundfrequenzverlauf, Vokaldauer) waren die gleichen wie in Experiment 1. Um die Sprecherin-

formationen in den Vokalen zu eliminieren, wurden für den Vokalkontext die Vokale eines einzigen Sprechers verwendet. Aus den vorhandenen Aufnahmen wurde ein zehnter Sprecher ausgewählt, dessen Grundfrequenz im mittleren Bereich lag, damit er sich nicht zu stark von den anderen 9 Sprechern unterschied. Von diesem zehnten Sprecher wurden nur die Vokale verwendet und die Konsonanten entfernt. Um die Konsonanten herausschneiden zu können, wurden die Sprachsignale zunächst in Praat [35] segmentiert und gelabelt. Die Segmentgrenzen wurden durch auditive und visuelle (Oszillogramm, Sonagramm) Untersuchung an Stellen einer deutlichen Veränderung im Sprachsignal gesetzt. Bei den Frikativen waren das die Übergänge vom periodischen zum aperiodischen Signal und anders herum. Bei den stimmhaften Frikativen wurden außerdem der Beginn der Friktion und eine Verringerung der Intensität (im Vergleich zu den Vokalen) als Indikatoren einer Segmentgrenze betrachtet. Die Nasale zeichneten sich durch einen relativ starten Abfall der Intensität im Sprachsignal aus und zeigten ein anderes periodisches Muster als die Vokale. Die Anfangsgrenze der Plosive wurde zu Beginn der Verschlussphase, die sich durch akustische Stille auszeichnet, gesetzt und das Segmentende mit Einsetzen der Periodizität und Sichtbarwerden der Formanten des Folgevokals. Nach der Annotation der Sprachdaten, wurden die Konsonanten herausgeschnitten und in den Vokalkontext des zehnten Sprechers eingefügt. Dabei wurde auf den koartikulatorischen Kontext Rücksicht genommen und zum Beispiel das /s/ in den /a/-Kontext von /s/ geschnitten usw. Dadurch entstanden natürlichere Transitionen zwischen Vokalen und Konsonanten. Da die Vokalinformationen weitestgehend eliminiert werden sollten, wurden nicht vier verschiedene Vokal-Kontexte erstellt (wie in Experiment 1), sondern nur einer für beispielsweise alle /s/ eines Sprechers. Die Stimuli hörten sich trotz der starken Manipulation noch natürlich an. Der Übergang zwischen Vokalen und Konsonanten war kaum wahrnehmbar. Allerdings war der Unterschied zwischen den Stimuli nun auditiv sehr schwer wahrzunehmen.

Design und Durchführung

Das Design und die Durchführung entsprachen der in Experiment 1 beschriebenen Vorgehensweise (siehe Abschnitt 4.5.2 und Abschnitt 4.5.2).

4.6.3. Ergebnisse

Die gewonnenen Daten wurden analog zu Experiment 1 ausgewertet. In diesem Experiment lag die durchschnittliche Anzahl an richtigen Antworten deutlich niedriger. Die Versuchs-

personen erreichten im Durchschnitt eine Erkennungsrate von 62 %, was allerdings immer noch deutlich über dem Zufallsniveau von 50 % lag.
Ebenso wie in dem letzten Experiment wurden die prozentualen Werte wieder in d'-Werte umgerechnet. Der d'-Wert lag im Mittel bei d' = 0,76, wobei es aber starke Unterschiede zwischen den einzelnen Konsonanten gab. Die hohen Werte der Frikative /s, ʃ, z, f/ zeigen, dass anhand dieser Laute die Sprecher besser unterschieden werden können als durch die Nasale und Plosive (siehe Abbildung 4.3).
Für den Artikulationsmodus der Plosive und Frikative erreichten die alveolaren Laute stets höhere d'-Werte als die labialen (siehe Abbildung 4.4). Einzig bei den Nasalen zeigte sich ein umgekehrtes Bild.
Zur statistischen Überprüfung dieses Eindrucks wurde eine Varianzanalye (ANOVA) in R mit der Funktion „ezANOVA" [20] durchgeführt. Dabei war der *d'-Wert* die abhängige Variable, *Versuchsperson* der Fallidentifikator und *Konsonant* (/m, n, f, s, ʃ, v, z, p, t/) der unabhängige Faktor innerhalb der Versuchsperson. Die Analyse zeigte, dass sich der Faktor *Konsonant* signifikant auf den d'-Wert auswirkt ($F[8,240] = 14{,}6$; $p < 0{,}001$). Daraus folgt, dass sich die Konsonanten unterschiedlich gut zur perzeptiven Sprecherdiskrimination eignen.
Um herauszufinden, welchen Einfluss Artikulationsstelle und -modus auf die Sprechererkennung haben, wurde eine weitere Varianzanalyse durchgeführt. Da sich der post-alveolare Frikativ /ʃ/ nicht den Artikulationsstellen „labial" oder „alveolar" zuordnen lässt, wurde er bei dieser Analyse ausgeschlossen. Die abhängige Variable war wieder der *d'-Wert*, *Versuchsperson* der Fallidentifikator und *Artikulationsstelle* (labial, alveolar) und *Artikulationsmodus* (Nasal, stimmhafter Frikativ, stimmloser Frikativ, Plosiv) die unabhängigen Faktoren. Sowohl die *Artikulationsstelle* ($F[1,30] = 30{,}96$; $p < 0{,}001$) als auch der *Artikulationsmodus* ($F[3,90] = 15{,}98$; $p < 0{,}001$) hatten einen signifikanten Einfluss auf die Diskriminationsfähigkeit. Die signifikante Interaktion zwischen Artikulationsstelle und -modus ($F[3,90] = 8{,}56$; $p < 0{,}001$) deutet darauf hin, dass der Einfluss der Artikulationsstelle nicht in jedem Artikulationsmodus gleich groß ist.
Zur genaueren Betrachtung, welche Konsonanten sich im Einzelnen signifikant unterschieden, wurden die Laute paarweise in den Kategorien Artikulationsstelle, Artikulationsmodus und Stimmhaftigkeit getestet. Da die Werte pro Versuchsperson miteinander verglichen werden sollten, wurde ein gepaarter *t*-test verwendet. Wegen der multiplen Vergleiche wurde wieder eine Bonferroni-Korrektur zur Anpassung des p-Wertes durchgeführt. Das Signifikanzniveau lag daher bei $p = 0{,}0167$. Im Artikulationsmodus der Frikative erreichten die alveolaren

Laute eine signifikant höhere Diskriminationsfähigkeit als die labialen (stimmhafte Frikative: $t = -5{,}18$; $df = 30$; $p < 0{,}001$; stimmlose Frikative: $t = -3{,}37$; $df = 30$; $p = 0{,}002$). Die größte Differenz zwischen den beiden Artikulationsstellen wiesen die stimmhaften Frikative auf. Anschließen wurden die Artikulationsmodi innerhalb einer Artikulationsstelle untersucht. Der stimmlose alveolare Frikativ /s/ erreichte die höchsten d'-Werte und unterschied sich somit signifikant von /n/ ($t = -5{,}81$; $df = 30$; $p < 0{,}001$) und /t/ ($t = 4{,}78$; $df = 30$; $p < 0{,}001$). Der stimmhafte alveolare Frikativ /z/ unterschied sich signifikant von /n/ ($t = -4{,}09$; $df = 30$; $p < 0{,}001$). Der alveolare Nasal /n/ und der alveolare Plosiv /t/ unterschieden sich nicht signifikant voneinander.

Anschließend wurden die Konsonanten der labialen Artikulationsstelle miteinander verglichen, welche insgesamt weniger signifikante Unterschiede zeigten als die alveolaren Laute. Die geringsten Diskriminationsfähigkeiten ergaben sich für den stimmhaften Frikativ /v/ und den Plosiv /p/, die sich beide nicht signifikant voneinander unterschieden. Der labiale Nasal /m/ erreichte signifikant höhere d'-Werte als /v/ ($t = 2{,}76$; $df = 30$; $p = 0{,}01$). Der stimmlose Frikativ /f/ zeigte die höchste Diskriminationsfähigkeit und lag somit auch signifikant vor /p/ ($t = 4{,}39$; $df = 30$; $p < 0{,}001$). Der Faktor Stimmhaftigkeit verursachte einen signifikanten Unterschied zwischen den labialen Frikativen /f, v/ ($t = 4{,}33$; $df = 30$; $p < 0{,}001$), während er die alveolaren Frikative /s, z/ nicht voneinander abgrenzte. Somit scheint sich die Stimmhaftigkeit bei den labialen Frikativen stärker auf die Diskriminationsfähigkeit auszuwirken als bei den alveolaren.

Insgesamt bestätigte sich die Überlegenheit der alveolaren Artikulationsstelle in 3 von 4 Artikulationsmodi. Außerdem unterscheiden sich die Artikulationsmodi für die alveolare Artikulationsstelle stärker voneinander als für die labiale. Die Stimmhaftigkeit hingegen wirkt sich nur bei den labialen Frikativen signifikant auf die Diskriminationsfähigkeit aus, während sie bei den alveolaren keinen Unterschied hervorruft. Alles in allem zeigen die Frikative eine signifikant größere Diskriminationsfähigkeit als die Nasale und Plosive.

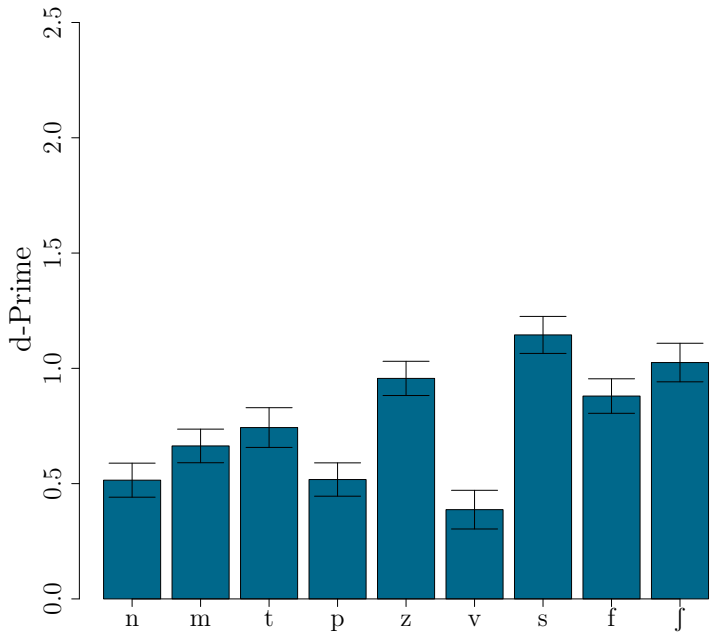


Abbildung 4.3.: d' (d-Prime) in Abhängigkeit vom Konsonant mit Fehlerbalken

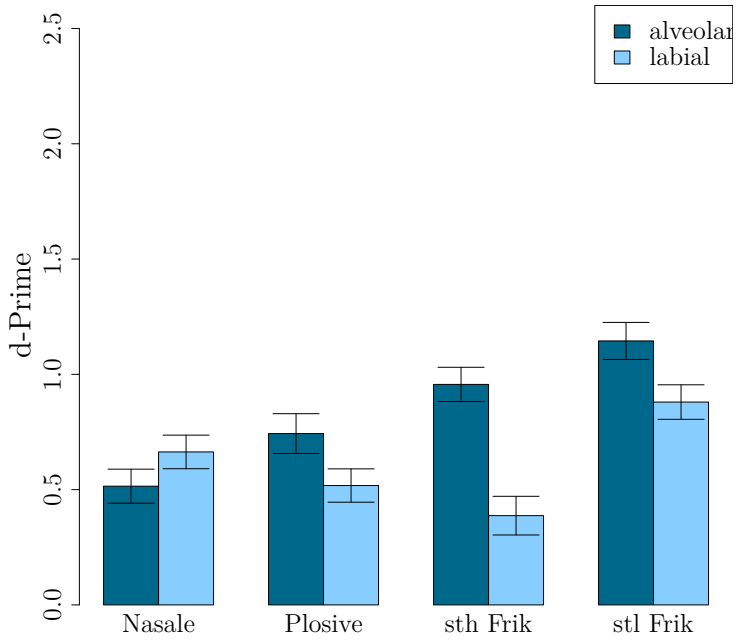


Abbildung 4.4.: d' (d-Prime) in Abhängigkeit von Artikulationsstelle und Artikulationsmodus des Konsonanten mit Fehlerbalken

4.6.4. Diskussion

Um die Unterschiede in der Sprecherspezifität der Konsonanten noch deutlicher hervortreten zu lassen, wurden die dynamischen Sprecherinformationen der Vokalübergänge eliminiert. Dadurch erhöhte sich die Schwierigkeit der Diskriminationsaufgabe und die Erkennungsrate sank auf 62 %. In Anbetracht der perzeptiv sehr ähnlichen Stimuli beweist dieser, immer noch über dem Zufallsniveau liegende Wert, die sehr gute Sprecherdiskriminationsleistung der Hörer.

Wie auch im Experiment zuvor, zeigte sich der Einfluss des phonetischen Inhalts des Stimulus auf die Sprechererkennungsrate. Die hohen Werte der Frikative /s, ʃ, z, f/, gegenüber den Nasalen und Plosiven, unterstreichen deren bessere Eignung zur Sprecherdiskrimination unter diesen Bedingungen. Der signifikante Einfluss des Artikulationsmodus zeigte deutlich, dass große Unterschiede im Sprecherinformationsgehalt zwischen den verschiedenen Lautkategorien bestehen. Dabei erwiesen sich die Frikative - vor allem die alveolaren - als signifikant diskriminationsfähiger als die Nasale und Plosive. Dass (alveolare) Frikative viele Sprecherinformationen enthalten, ist aus der Literatur bereits bekannt ([68] [143]). Frikative kodieren - im Gegensatz zu Nasalen und Plosiven - einen größeren Anteil an phonetischen Informationen in ihren Frequenzschwerpunkten, das heißt im statischen Bereich des Konsonanten selbst [207]. Nimmt man nun wieder an, dass sich an den Stellen der phonetischen Informationen auch Informationen über den Sprecher befinden, wird klar, warum Frikative auch ohne ihre Transitionen noch relativ viele Sprechermerkmale enthalten. Überraschend war allerdings, dass Nasale und Plosive ähnlich (schlecht) abschnitten, wobei Nasalen in der Literatur ein großer Nutzen für die Sprechererkennung zugeschrieben wird und den Plosiven ein sehr geringer ([12] [68]). Dieses schlechte Abschneiden lässt sich möglicherweise dadurch erklären, dass viele akustische Informationen bei Nasalen und Plosiven eher in ihren dynamischen Bereichen (Transitionen) und weniger im statischen Bereich des Konsonanten liegen (siehe Abschnitt 4.2.2). Dadurch, dass die Transitionen in diesem Experiment keine Sprecherinformationen enthielten, konnten nur die statischen konsonantischen Merkmale zur Sprecherdiskrimination verwendet werden. Der Sprecherinformationsgehalt dieser Konsonanten war somit geringer, was sich in einer geringeren Diskriminationsfähigkeit äußerte. Allerdings würde man für die Nasale dennoch eine höhere Sprecherdiskriminationsfähigkeit erwarten als für die Plosive. Eine weitere mögliche Ursache für deren niedrige Werte könnte ein Mismatch zwischen den nasalen Konsonanten und den nasalierten Vokalen des anderen Sprecher sein. Da in der Umgebung eines Nasals die Vokale häufig auch (zumindest teilweise) nasaliert werden (siehe Abschnitt 4.2.2), könnten

die auf diese Weise transportierten „Nasenrauminformationen“ des anderen Sprechers mit den Sprecherinformationen des nasalen Konsonanten interferiert haben. Da somit zwei verschiedenartige Informationen über die Physiologie des Nasenraumes des „Sprechers“ zur Verfügung standen, konnten die Hörer möglicherweise diese Informationen nicht voneinander trennen und somit den Sprecher des nasalen Konsonanten nicht so zuverlässig erkennen. Vermutlich aus diesen Gründen erreichten die Nasale nur eine geringere Diskriminationsfähigkeit als die Frikative und eine ähnliche wie die Plosive.
Alle Konsonantengruppen bis auf die Nasale zeigten einen signifikanten Unterschied in der Sprecherdiskriminationsfähigkeit zwischen labialen und alveolaren Lauten, wobei letztere stets höhere Erkennungsraten erzielten. Dieses Ergebnis weist darauf hin, dass die Frikative und Plosive Sprechermerkmale enthalten, die sich zwischen der labialen und der alveolaren Artikulationsstelle unterscheiden. Die Nasale hingegen scheinen in ihrem statischen konsonantischen Teil jedoch nicht ausreichend Informationen zu beinhalten, um die zwei Artikulationsstellen in Punkto Sprecherspezifität unterscheiden zu können.
Aufgrund der Ergebnisse des ersten Experiments wurde vermutet, dass die Art der Amplitudennormierung der Stimuli einen Teil der Lautunterschiede reduziert haben könnte. In diesem Experiment fanden sich aber trotz dieses möglichen Einflusses in fast allen Konsonantengruppen starke Unterschiede zwischen dem labialen und dem alveolaren Laut. Selbst wenn also lautspezifische Amplitudeninformationen neutralisiert werden, so bestehen dennoch signifikante Unterschiede zwischen den Lauten der beiden Artikulationsstellen. Sprich, wären die Lautstärkeinformationen vorhanden, würden diese den Unterschied eher verstärken als reduzieren. Einzig im Fall der Nasale könnten die fehlenden Amplitudeninformationen einen Nachteil für den alveolaren Nasal im Vergleich zum labialen hervorgerufen haben.
Auch in diesem Experiment zeigte sich die unterschiedliche Wichtigkeit von (statischen) innerkonsonantischen Sprecherinformationen. Während die Frikative viele Informationen im statischen Bereich des Konsonanten zu enthalten scheinen, weisen Nasale und Plosive dort weniger Sprecherinformationen auf. Somit scheinen die Sprecherinformationen abhängig vom Konsonanten in unterschiedlichen Bereichen („statischer Bereich“ vs. „dynamischer Bereich“) lokalisiert zu sein. Daher sollten Nasale und Plosive für einen maximalen Informationsgewinn mit ihren dynamischen Vokaltransitionen betrachtet werden, während bei Frikativen auch eine ausschließliche Betrachtung der statischen konsonantischen Informationen sinnvoll ist.

4.7. Generelle Diskussion

In den zwei Experimenten sollte untersucht werden, welche Rolle phonetische Konsonanteninformationen in der perzeptiven Sprecherdiskrimination spielen. Aufgrund der Literatur wurde vermutet, dass besonders Nasale und Frikative viele Informationen zur Sprecheridentität enthalten ([101] [295] [68] [12] [14] [15] [142] [143] [82]), wohingegen Plosive eher wenig Sprecherinformationen liefern ([12] [68]). Weiterhin wurde angenommen, dass die Konsonanten der alveolaren Artikulationsstelle mehr Sprecherinformationen beinhalten als die der labialen [11]. Sowohl die besondere Eignung der Nasale und Frikative als auch die der alveolaren Konsonanten hatte sich auch in den akustischen Analysen von Kapitel 3 dieser Arbeit gezeigt. Nun sollte die Frage beantwortet werden, ob sich die Effekte, die bereits im Englischen und Japanischen gefundenen wurden, auch für Konsonanten des Deutschen finden lassen und ob sich die Unterschiede zwischen alveolarer und labialer Artikulationsstelle bei verschiedenen Artikulationsmodi (Nasal, Frikativ, Plosiv) zeigen. Des Weiteren stellte sich die Frage, in welchem Bereich des Stimulus die Sprecherinformationen liegen - im statischen oder im dynamischen. So ist beispielsweise bekannt, dass Nasale und Plosive viele ihrer phonetischen Informationen, vor allem Informationen zur Artikulationsstelle, in ihren Transitionen zu den benachbarten Vokalen enthalten und nicht im konsonantischen Segment selbst (siehe Abschnitt 4.2.2). Frikative hingegen enthalten einen größeren Anteil ihrer phonetischen Informationen im statischen Bereich des Konsonanten und weniger in den Transitionen. Da anzunehmen ist, dass sich die phonetischen Informationen in den Bereichen des Lautes befinden, der insgesamt viele akustische Merkmale enthält, lag die Vermutung nahe, dass die Informationen über die Identität des Sprechers in den gleichen Bereichen zu suchen sind. Daher war anzunehmen, dass mit allen statischen und dynamischen Sprecherinformationen Nasale und Frikative ähnlich viel zur Sprecherdiskrimination beitragen. Die Plosive wurden aufgrund der Ergebnisse aus der Literatur als weniger informativ eingeschätzt, sollten aber dank der vorhanden dynamischen Vokakinformationen noch immer Sprecherinformationen liefern. Dabei sollten die alveolaren Konsonanten durchgehend informativer sein als die labialen. Für den Fall, dass nur die Konsonanten und nicht die Vokale mit ihren Transitionen Sprecherinformationen enthalten, wurde angenommen, dass die Frikative mehr Sprecherinformationen enthalten als die Nasale und Plosive, wobei die Nasale weiterhin informativer sein sollten als die Plosive. Die Unterschiede zwischen der alveolaren und labialen Artikulationsstelle sollten sich auch ohne die Transitionsinformationen zeigen.

Um die Unterschiede im Sprecherinformationsgehalt der Konsonanten sichtbar zu machen,

mussten die Sprecher ausreichend schwer unterscheidbar sein. Im Falle einer hundertprozentigen Erkennungsrate für alle Stimuli wären keine Unterschiede zwischen den verschiedenen Konsonanten erkennbar. Daher wurden die Stimuli sowohl in Amplitude, Grundfrequenz und Grundfrequenzverlauf als auch ihrer Vokaldauer normiert, um sie ähnlicher und somit schwerer unterscheidbar zu machen. Für das zweite Experiment wurden außerdem die Konsonanten der 9 zu unterscheidenden Sprecher in den Vokalkontext eines zehnten Sprechers geschnitten. Dabei wurde der Vokalkontext auf den Konsonanten abgestimmt, sodass die Transitionen natürlicher klangen. Durch diese Manipulation enthielten die Stimuli in den vokalischen Segmenten keinerlei Sprecherinformationen mehr und die Hörer mussten sich zur Sprecherdiskrimination ganz auf die Informationen im konsonantischen Segment verlassen. Während den Hörern im ersten Experiment dynamische und statische Informationen für die Sprecherdiskrimination zur Verfügung standen, konnten sich die Hörer im zweiten Experiment nur noch auf die statischen Konsonantinformationen verlassen, da die dynamischen Vokalinformationen nicht sprecherspezifisch waren. Dies führte wie erwartet zu einer geringeren Sprecherdiskriminationsrate im zweiten Experiment (62 %) im Vergleich zum ersten (83 %). Dennoch lag die Erkennungsrate in beiden Experimenten über dem Zufallsniveau. Daraus lässt sich schlussfolgern, dass die Hörer (bis zu einem bestimmten Grad) in der Lage waren, die Sprecher auch ausschließlich anhand der (statischen) Konsonantinformationen zu unterscheiden. Dieser Unterschied in der Diskriminationsleistung verdeutlicht aber auch, dass ein beträchtliches Maß an Sprecherinformationen in den Vokalen bzw. den Vokaltransitionen liegt.

Beide Experimente bewiesen, dass nicht alle Konsonanten gleich viele Sprecherinformationen beinhalten, sondern dass es signifikante Unterschiede gibt. Im ersten Experiment, in Anwesenheit der dynamischen Informationen, zeigten sich die alveolaren Nasale und Plosive als signifikant sprecherspezifischer als die Frikative und wiesen signifikante Unterschiede zwischen alveolarer und labialer Artikulationsstelle auf. Die Frikative unterschieden sich zwischen ihren Artikulationsstellen nicht in ihrer Sprecherdiskriminationsfähigkeit. Nasale und Plosive haben sehr schnelle und abrupte Transitionen, die mit einer starken Energieveränderung einhergehen. Gleiches gilt für alveolare Laute, welche mit der Zungenspitze artikuliert werden. Diese bewegt sich schneller als die Lippen [28], was schnellere und abruptere Transitionen und somit eine starke Energieveränderung bewirkt. Dynamische Merkmale wie diese Engergieveränderungen enthalten typischerweise viele akustische Informationen ([147] [95] [151]), die bei der Spracherkennung und vielleicht auch bei der Sprechererkennung hilfreich sind. Die Frikative hingegen haben eher langsame Transitionen,

da die Artikulation eine hohe Präzision verlangt und daher stärker kontrolliert werden muss. Aufgrund dieser langsameren Transitionen enthalten die Frikative vermutlich nicht so viele akustische Informationen in diesen Bereichen und somit auch weniger Sprecherinformationen als die Nasale und Plosive. Auch unterschieden sich die Artikulationsmodi stärker in der alveolaren Artikulationstelle als in der labialen.
Im zweiten Experiment, in Abwesenheit von Vokal- und Transitionsinformationen, waren die Frikative sprecherspezifischer als die Nasale und Plosive und unterschieden sich nun auch zwischen den Artikulationsstellen signifikant voneinander. Die Plosive behielten ihre Differenz zwischen alveolarer und labialer Artikulationsstelle bei, aber auf einem insgesamt geringeren Niveau, während die Nasale nicht nur in ihrer Diskriminationsfähigkeit reduziert wurden, sondern auch in ihrer Artikulationsstellendifferenz bis unter das Signifikanzniveau sanken. Die Artikulationsmodi unterschieden sich abermals in der alveolaren Artikulationsstelle stärker als in der labialen.
Das starke Absinken der Sprecherdiskriminationsfähigkeit der Nasale und Plosive lässt sich durch den Wegfall der dynamischen Transitionsinformationen erklären. Da beide Lautgruppen viele akustische Merkmale in ihren Vokalübergängen und weniger im statischen Teil des Konsonanten enthalten, lieferten sie im zweiten Experiment weniger Sprecherinformationen als im ersten. Unerwartet war, dass die Nasale in keinem der Experimente sprecherspezifischer waren als die Plosive, obwohl dies nicht den Berichten der Literatur entspricht ([12] [68]). Allein die hohe Sprecherspezifität von /t/ wurde in wenigen Studien bereits gezeigt (z.B. [82]). Welche akustischen Merkmale diese perzeptiv informatorische Wirkung ausmachen, sollte in weiteren Studien genauer analysiert werden.
Es scheint, dass der Sprecherinformationsgehalt bei Nasalen und Plosiven von den Transitionen abhängt. Die Frikative hingegen enthalten offenbar mehr Sprecherinformationen in ihrem statischen konsonantischen Teil, sodass ihre Sprecherspezifität zwar deutlich, aber weniger stark reduziert wird als die der Nasale und Plosive.
Allgemein zeigte sich die Tendenz, dass die alveolaren Konsonanten sprecherspezifischer sind als die labialen. Allerdings trat dieser Effekt im ersten Experiment nur bei den Nasalen und Plosiven und im zweiten nur bei den Plosiven und Frikativen auf. Es scheint, als hätten die Transitionen bei den Nasalen ebenfalls die Unterschiede in der Diskriminationsfähigkeit zwischen den alveolaren und labialen Lauten verursacht und mit ihrem Wegfall wurde diese Differenz neutralisiert. Bei den Frikativen hingegen scheinen erst durch die Entfernung der Transitionsinformationen die Unterschiede zwischen den Artikulationsstellen sichtbar zu werden. Ähnlich dem Prinzip, dass man bei zu hohen Sprecherdiskriminationsraten keine

Unterschiede zwischen den Konsonanten sehen kann, wurden die Unterschiede zwischen den Frikativen erst durch die Neutralisation der dynamischen Vokalinformationen und der damit verbundenen Erschwerung der Sprecherdiskrimination sichtbar. Einzig bei den Plosiven verändert sich die Differenz zwischen alveolaren und labialen Lauten nicht und war in beiden Experimenten signifikant. Das legt die Vermutung nahe, das Plosive ebenso viele Sprecherinformationen im statischen Bereich Konsonanten wie in den Vokaltransitionen enthalten.
Während die Ergebnisse der Frikative erwartbar und plausibel waren, lassen sich die unerwartet niedrigen Resultate der Nasale im zweiten Experiment (und vor allem im Vergleich zu den Plosiven) nicht einfach erklären. Wenn man bedenkt, dass sich laut der Literatur ([101] [295] [68] [12] [14] [142]) Nasale wesentlich besser zur Sprecheridentifikation eignen als Plosive, scheint dieses Ergebnis unpassend. Allerdings ist auch bekannt, dass Nasale und Plosive aufgrund ihrer Artikulation gleichermaßen stark von ihren Transitionen abhängen (siehe Abschnitt 4.2.2). Dahingehend wäre es also nicht verwunderlich, wenn sich die Neutralisation der Transitionsinformationen gleichermaßen stark auf beide Artikulationsmodi auswirken würde. Ein weiterer Aspekt für die geringe Sprecherspezifität der Nasale im zweiten Experiment könnte durch einen Mismatch der „Nasenrauminformationen“ entstanden sein. Die Informationen über den Nasenraum der zwei verschiedenen Sprecher könnten interferiert und so die Unterscheidung der Sprecher erschwert haben.
Wenn bei den Nasalen und Plosiven, wie vermutet, die meisten phonetischen und Sprecherinformationen in den dynamischen Transitionen zu den Vokalen liegen und sich der Unterschied zwischen alveolarer und labialer Artikulationsstelle hauptsächlich aufgrund dieser Informationen manifestiert, dann muss dieser Unterschied in Merkmalen dieser Transitionen begründet liegen. Um herauszufinden, wodurch der Unterschied in der Sprecherdiskriminationsfähigkeit zwischen labialen und alveolaren Konsonanten zu Stande kommt, müsste man diese Vokaltransitionen detailliert akustisch oder artikulatorisch untersuchen. Es ist zu vermuten, dass aufgrund der anatomischen Position der alveolaren Artikulationsstelle, die Transitionen zu den umliegenden Vokalen ausgeprägter sind als bei der labialen. Dies könnte in einer zukünftigen Studie überprüft werden.
Wie eingangs beschrieben, wurden die Stimuli akustisch manipuliert, d.h. ihre Amplitude, mittlere Grundfrequenz, Grundfrequenzverlauf und Vokaldauer wurden normiert. Diese Manipulation war nötig, um die Stimuli ausreichend schwierig zu gestalten. Bei einer zu hohen Sprecherdiskriminationsrate der Hörer hätte man sonst keine Unterschiede im Einfluss der Konsonanten beobachten können. Daher musste die Erkennungsrate durch eine Angleichung

der Stimuli abgesenkt werden. Durch die erwähnten akustischen Manipulationen könnten einige Laute an Sprecherinformationen eingebüßt haben. Unter „normalen“ Umständen wären sie noch informativer und würden noch stärker zur Sprechererkennung beitragen als es in diesem Experiment der Fall war.
Eine weitere Ursache für den starken Einfluss der Transitionen könnte der Vokalkontext der Konsonanten in den Stimuli gewesen sein. Als Kontextvokal wurde der offene Vokal /a/ gewählt. [247] zeigten, dass besonders bei offenen und hinteren Vokalen die Transitionen sehr viele Informationen über die Artikulationsstelle enthalten, während vordere Vokale wie /i/ nur wenige Informationen in den Transitionen aufweisen. Dieses Phänomen könnte erklären, weshalb die Transitionen in diesem Experiment eine so entscheidende Rolle spielten. In einem Folgeexperiment mit Konsonanten im /i/-Kontext könnte überprüft werden, ob der Kontextvokal einen signifikanten Einfluss auf den Informationsgehalt der Transitionen hat.
Insgesamt konnten die Hörer die Sprecher gut bis sehr gut unterscheiden. Das zeigt, dass die Konsonanten, ob mit oder ohne dynamische Informationen, viele relevante Informationen zur Sprecherdiskrimination tragen.

4.8. Zusammenfassung

Es wurden zwei Experimente zur perzeptiven Sprecherdiskrimination durchgeführt. Dabei mussten Hörer in einem AX-Task entscheiden, ob zwei /aKa/-Stimuli von einem oder von zwei Sprechern stammten. Im ersten Experiment stammte die gesamte /aKa/-Sequenz von einem Sprecher, sodass sowohl statische als auch dynamische Informationen über den Sprecher enthalten waren. Im zweiten Experiment wurden die sprecherspezifischen Merkmale der dynamischen Vokaleinformationen entfernt, sodass nur noch die statischen Bereiche der Konsonanten Sprecherinformationen enthielten. In beiden Experimenten wurden die Stimuli in ihrer Amplitude, Grundfrequenz, Grundfrequenzverlauf und Vokaldauer normiert, sodass die Sprecher schwieriger zu unterscheiden waren. Diese Manipulationen waren notwendig, um mögliche Einflüsse der Konsonanten besser sichtbar zu machen.
Wie erwartet, konnten die Hörer die Sprecher im ersten Experiment problemlos unterscheiden, während sie im zweiten Experiment größere Schwierigkeiten hatten. Dennoch lag auch im zweiten Teil die Sprecherdiskriminationsrate über dem Zufallsniveau. Dieses Ergebnis unterstreicht die wichtige Rolle der Vokale als Träger von Sprecherinformationen.
Es zeigten sich Unterschiede in der Diskriminationsfähigkeit der Konsonanten in Abhängigkeit von ihrem Artikulationsmodus und ihrer Artikulationsstelle. Dabei waren alveolare

Laute oft signifikant sprecherspezifischer als labiale. Je nachdem, ob die Konsonanten ihre phonetischen und Sprecherinformationen eher im (statischen) Konsonanten (wie die Frikative) oder eher in ihren (dynamischen) Transitionen zu den benachbarten Vokalen (wie die Nasale und Plosive) enthielten, zeigten sich die Unterschiede in der Sprecherspezifität zwischen den Artikulationsstellen eher beim Vorhandensein oder Nicht-Vorhandensein der Transitionsinformationen. Generell scheinen aber alle Konsonanten einen Teil ihrer Informationen in den Vokaltransitionen zu enthalten.
Diese Ergebnisse verdeutlichen, in welchem Maß und auf welche Weise konsonantische Merkmale Informationen zur Sprecheridentität liefern können. Die Resultate von englischen und japanischen Studien konnten für das Deutsche größtenteils bestätigt werden. Außerdem zeigte sich, dass der Konsonant nicht nur die Diskriminationsfähigkeit beeinflusst, sondern ebenfalls die Lokalisation der Sprecherinformationen im Konsonanten.
Offen bleibt, inwieweit die Normierungen (Amplitude, mittlere Grundfrequenz, Grundfrequenzverlauf, Vokaldauer) die Sprecherspezifität der untersuchten Laute gesenkt haben könnte und welche akustischen Merkmale den Unterschied in der Sprecherdiskriminationsfähigkeit zwischen labialen und alveolaren Konsonanten hervorrufen. Diese Fragen zu beantworten, könnte und sollte ein Anliegen von weiterführenden Studien sein.

5. Der zeitliche Zusammenhang zwischen Sprechererkennung und der Verarbeitung phonetischer Information

Dieses Kapitel widmet sich der zeitsynchronen Aufnahme und Verarbeitung von Informationen in gesprochener Sprache. Unsere Sprache enthält neben den inhaltlichen, lexikalischen Informationen auch immer nicht-lexikalische Informationen wie zum Beispiel Informationen über den Sprecher oder die Sprechsituation. In diesem Teil der Arbeit soll ein genauerer Blick auf die Verarbeitung von Sprecherinformationen und deren zeitliche Koordination zur Verarbeitung lexikalischer Informationen geworfen werden. Die Methode, die dies ermöglicht ist das Visual-World Paradigma welches mittels Eye-Tracking die Blickbewegungen von Personen auf einem Bildschirm verfolgt. Daher soll zunächst die Physiologie des menschlichen Auges, die Methodik des Eye-Tracking und das Visual-World Paradigma vorgestellt werden. Anschließend folgt eine Darstellung der aktuellen Literatur zur zeitsynchronen Worterkennung sowie die Einflüsse von Sprechermerkmalen auf selbige. Nach der Vorstellung einiger relevanter Studien, werden zwei Experimente besprochen, in ihrer Methodik, den Ergebnissen und deren möglichen Interpretationen. Abschließend erfolgt eine Diskussion über die Einordnung der gefundenen Resultate in die Theorie der Sprachwahrnehmung.

5.1. Stand der Forschung

5.1.1. Eye-Tracking

Das Eye-Tracking - oder auch Blickbewegungsverfolgung - ist ein Verfahren zur Beobachtung und Aufzeichnung von menschlichen Augen- und Blickbewegungen. Die physiologischen

Grundlagen des menschlichen Auges, als auch das technische Verfahren des Eye-Trackings sollen in diesem Abschnitt erklärt werden. Außerdem werden Parameter von Fixationen als Indikator für Aufmerksamkeit vorgestellt und erläutert.

Die Physiologie des Auges: Augen- und Blickbewegungen

Die menschlichen Augenbewegungen lassen sich grob in zwei Kategorien unterteilen: Die einen dienen dazu, die visuell interessanten Objekte in der Fovea, dem Punkt des schärfsten Sehens, zu stabilisieren (eigenkompensatorische Augenbewegungen) und die anderen dazu das relevante Objekt in dieser Position zu platzieren (zielgerichtete Augenbewegungen). Die zwei Formen der zielgerichteten Augenbewegungen sind die Sakkaden, welche Objektbilder mit höchster Geschwindigkeit von der Peripherie in die Fovea der Retina holen, die anderen sind langsame Folgebewegungen, welche die Fovea nachführen, wenn sich ein Sehobjekt bewegt [126].
Die Sakkaden, welche ca. 30 bis 50 ms dauern, sind ruckartige, schnelle Augenbewegungen, während denen weniger optische Informationen wahrgenommen werden, sodass man in diesen Phasen tatsächlich fast blind ist. Dieser Vorgang nennt sich sakkadische Suppression [138]. Sakkaden können als schnellste Augenbewegungen in Abhängigkeit von der Entfernung des Zielobjekts eine Geschwindigkeit von bis zu 900°/s erreichen, welche nicht willentlich variiert werden kann. Sakkaden können entweder beabsichtigt durchgeführt werden, um ein Objekt genauer zu betrachten, oder unbeabsichtigt von externen Faktoren ausgelöst werden, wenn sich etwas im peripheren Gesichtsfeld verändert. Die Latenzzeit zur Ausführung einer Sakkade beträgt ca. 200 ms. Wurde die Bewegung einmal ausgelöst, kann sie weder durch plötzlich auftauchende Stimuli noch durch kognitive Prozesse unterbrochen oder beeinflusst werden. Deshalb spricht man bei diesen Augenbewegungen auch von ballistischen Bewegungen [153].
Zu der Kategorie der eigenkompensatorischen Augenbewegungen gehören die bei der Fixation von Objekten auftretenden Mikrobewegungen wie Tremor, Drift und Mikrosakkaden. Von einer Fixation spricht man, wenn ein bestimmter Punkt im Raum - der Fixationspunkt - fokussiert wird. Fixationen dauern in der Regel zwischen 200 und 600 ms. Auch bei der Betrachtung eines festen Punktes, macht der Augapfel minimale Bewegungen, welche sich in drei Kategorien unterteilen lassen: Drift, Tremor und Mikrosakkaden. Die Drift bezeichnet ein kontinuierliches Abgleiten vom Fixationspunkt, während der Tremor kleinste Zitterbewegungen beschreibt. Beiden Bewegungsformen wird die Aufgabe zugeordnet die Netzhaut minimal zu verschieben, wodurch das eintreffende Bild nicht ständig auf dieselben

Rezeptoren der Netzhaut fällt. Erst dadurch wird eine Wahrnehmung möglich, da die Rezeptoren nur auf Veränderung reagieren. Für die Entstehung des Tremors könnten aber auch Instabilitäten in der Steuerung der drei Muskelpaare verantwortlich sein, welche für die Bewegungen des Auges zuständig sind. Den Mikrosakkaden wird die Aufgabe zugeschrieben, die durch Drift und Tremor entstandenen Abweichungen zu kompensieren und die Fixierung des Objekts wieder herzustellen. Sie können allerdings auch spontan auftreten und den Blick sogar vom Fixationsobjekt weg führen [138]. Die Mikrosakkaden unterscheidet von den Sakkaden vor allem die Größe der Bewegung und die Tatsache, dass Mikrosakkaden willentlich nur unterdrückt, aber nicht ausgelöst werden können [153]. Diese Bewegungen sind für ein scharfes Sehen notwendig, da der Schärfebereich des Auges (Fovea) bei nur 1° liegt. Alles was außerhalb dieses Bereiches, aber immer noch innerhalb von 100° liegt, wird unschärfer und blasser wahrgenommen [138].
Mit Hilfe des Eye-Tracking (auch Blickerfassung oder Okulographie genannt) kann man die Blickbewegungen einer Person verfolgen. Da Augenbewegungen sowohl durch die Wahrnehmung der visuellen Umgebung als auch durch kognitive Prozesse einer Person gesteuert werden, dienen sie in vielen wissenschaftlichen Bereichen als wertvolles Werkzeug. Außerdem bieten sie umfangreichere Daten über den gesamten Prozess einer Entscheidung als beispielsweise ein Klick auf einen Knopf oder eine Reaktionszeit und können somit die Frage nach den Ursachen und Einflüssen, die zu einer Antwort führten, klären [252].

Eye-Tracker: Methoden und Funktionsweisen

Die verschiedenen Verfahren des Eye-Trackings basieren auf den anatomisch-physiologischen Eigenschaften des Auges und können durch Eigen- oder Fremdbeobachtung oder durch technische Hilfsmittel erfasst werden. Dabei ist der Messbereich definiert als der Winkelbereich, in dem Augenbewegungen gemessen werden können. Der zeitliche Messbereich kennzeichnet die Dauer der Aufzeichnung, welche sowohl von den Speicherkapazitäten als auch vom Tragekomfort des Eye-Tracking-Systems abhängt. Die örtliche Auflösung ist von der Größe der kleinsten messbaren Veränderung abhängig, während die zeitliche Auflösung aus der Anzahl der Messwerte pro Zeiteinheit resultiert. Die örtliche Genauigkeit ist die Differenz zwischen der wahren Augenposition und dem vom System gemessenen Wert und kann nicht größer sein als die örtliche Auflösung. Die zeitliche Genauigkeit bezeichnet die Dauer zwischen der Erfassung einer Augenposition und der Bereitstellung der Messgröße [138].
In den Anfängen des Eye-Trackings wurden die Augenbewegungen direkt beobachtet und notiert. In den 1970er Jahren wurden dann neuere Verfahren entwickelt: (1) Bei den

Retinal-Nachbildern wurden durch starke Lichtreize Nachbilder auf der Retina erzeugt, deren Position dann verfolgt werden konnte. (2) Die Methode der Elektrookulogramme misst die elektrische Spannung zwischen Netzhaut und Hornhaut um die Augenbewegungen einzufangen. (3) Die Kontaktlinsenmethode setzt auf verspiegelte Kontaktlinsen, deren Reflexion dann von einer Kamera aufgezeichnet wird. (4) Bei der Search-Coil Methode werden Kontaktlinsen mit eingebetteten Spulen verwendet, die einem magnetischen Feld ausgesetzt werden. Die induzierte Spannung erlaubt dann die Berechnung der Augenbewegungen. (5) Die moderneren Formen des Eye-Trackings sind videobasiert. Mittels einer Kamera werden die Bewegungen des Auges aufgezeichnet. Eine der verbreitetsten Methoden basiert auf der Blickachsenmessung. Dabei wird aus der Differenz zwischen einem Lichtreflex (üblicherweise dem kornealen Reflex) und einem festen Punkt im Auge (zum Beispiel dem Mittelpunkt der Pupille) auf die Blickachse geschlossen. Der Lichtreflex wird durch einen schwachen Infrarot-Lichtstrahl erzeugt, dessen Reflexion dann von einer speziellen Kamera aufgezeichnet wird. Die Lage der Bezugspunkte wird durch Verfahren der automatischen Bildverarbeitung ermittelt. Der Vorteil diese Methode besteht darin, dass beide Bezugspunkte im Auge liegen und nur aus deren Differenz auf die Blickrichtung geschlossen wird. Eine Kopfbewegung hat demzufolge keinen Einfluss auf die Messung [138].

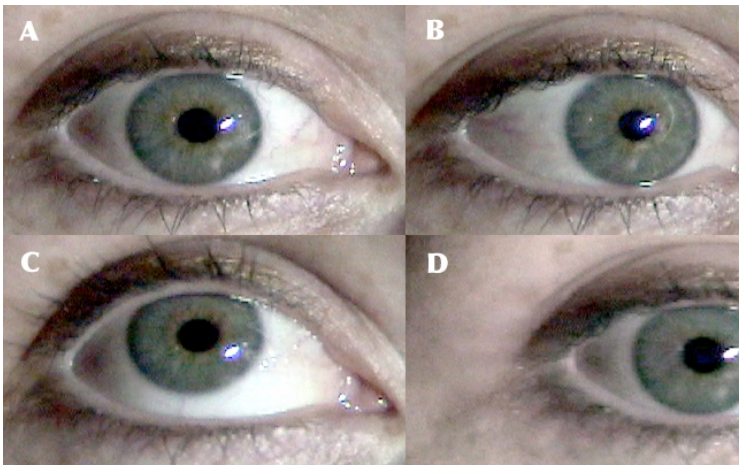


Abbildung 5.1.: Die **korneale Reflexion** erscheint als heller weißer Punkt rechts neben der Pupille (A). Die relative Position der Pupille und der kornealen Reflexion verändern sich mit der horizontalen (B) und der vertikalen (C) Bewegung des Auges. Das Verhältnis der Pupille und der kornealen Reflexion bleibt stabil, auch wenn sich der Kopf bewegt (D) [251].

Mobile und stationäre Systeme

In der Auslegung des Eye-Tracker wird zwischen mobilen und externen Systemen unterschieden. Die mobilen Systeme werden auch Head-mounted Eye-Tracker genannt und erlauben in erster Linie Mobilität. Augenkamera und Blickfeldkamera sind dabei an einem Kopfgerüst montiert, welches beim Probanden am Kopf fixiert wird. Mobile Eye-Tracker werden hauptsächlich für Feldstudien eingesetzt. Die Bewegung von Probanden bringt allerdings das Problem mit sich, dass die Daten bei der Aufzeichnung nicht-parametrisiert werden können. Möchte man die Daten statistisch auswerten, müssen die Videos manuell durchgesehen werden.
Anders sieht es bei den externen Systemen, den sogenannten Remote Eye-Trackern aus. Dabei werden die Messungen berührungslos durchgeführt und der Proband kann sich in einem gewissen Rahmen frei bewegen. Um die Kopfbewegungn der Probanden auszugleichen, werden verschiedene Techniken verwendet. Pan-Tilt-Systeme haben eine bewegliche Kamera, die den Kopfbewegungen des Probanden folgen kann. Tilting-Mirror Systeme arbeiten mit Spiegeln, welche ein Nachverfolgen des Auges ermöglichen, während die Kamera raumfest bleibt. Die Fixed-Camera-Systeme hingegen verzichten völlig auf bewegliche Teile und setzen statt dessen auf Bildverarbeitungsmethoden zum Ausgleich von Kopfbewegungen. Die Daten dieser Systeme sind parametrisierbar und können somit statistisch ausgewertet werden. Anwendung finden diese Verfahren meist in der Markt- und Medienforschung.
Eine noch präzisere Messung ermöglichen die Tower-Systeme, bei denen die Kamera ebenfalls raumfest ist und der Proband mittels einer Kinnstütze am Kopf fixiert wird. Diese Methode ist etwas unbequemer für die Probanden, wird aber in den Neurowissenschaften wegen ihrer höheren Genauigkeit verwendet [66]. Dank der unterschiedlichen Verfahren und ihrer verschiedenen Vorteile, wird Eye-Tracking heutzutage in verschiedenen wissenschaftlichen Bereichen eingesetzt, wie in der Markt-und Werbeforschung, der Psychologie, Medizin und in der Psycholinguistik für das Verständnis sowohl von geschriebener als auch gesprochener Sprache.

Fixationen als Indikator für Aufmerksamkeit

In vielen Studien wurden Fixationen, deren Dauer und Ort untersucht. Die Auswahl dieser Parameter aus der Menge aller möglichen, erfolgte aufgrund ihrer Beziehung zur Aufmerksamkeit. Zahlreiche Untersuchungen zeigen, dass im Allgemeinen eine Übereinstimmung von Fixationspunkt und dem Fokus der Aufmerksamkeit angenommen werden kann. Die Aufmerksamkeit führt das Auge durch Sakkaden zu seinem Zielort, den es dann fixiert.

Erscheint ein Reiz in der Peripherie des Blickfeldes, muss sich die Aufmerksamkeit erst vom aktuellen Fixationsort lösen, bevor sie sich dem neuen Zielort zuwenden kann. Hat sich die Aufmerksamkeit gelöst, bewegen sich die Augen mittels einer Sakkade zum neuen Fixationsort, an dem sie dann verweilen, so lange die Aufmerksamkeit auf diesem Punkt liegt. Allerdings müssen Fixationsort und der Fokus der Aufmerksamkeit nicht immer zusammenfallen. Der Mensch ist durchaus in der Lage einem Objekt in der peripheren Region des Blickwinkels seine Aufmerksamkeit zu schenken, während er einen anderen Punkt fixiert. Das würde dafür sprechen, dass sich die Aufmerksamkeit unabhängig von den Augen bewegen kann. Umgekehrt erscheint es aber zweifelhaft, dass sich die Augen bewegen können ohne mit einer Verlagerung der Aufmerksamkeit einher zu gehen.
Der am häufigsten genutzte Parameter ist die Fixationsdauer. Vorangegangene Untersuchungen legen nahe, dass die Dauer einer Fixation von der Schwierigkeit der Aufgabe abhängt. Daher wird häufig angenommen, dass die Dauer der Fixation auch der Dauer der Informationsverarbeitung des wahrgenommenen Objekts entspricht [138]. Diese sogenannte Mind-Eye Hypothesis wurde 1980 von [139] aufgestellt. Sie zeigen, dass Menschen beim Lesen so lange auf ein Wort schauen, bis sie es soweit wie möglich verarbeitet haben. Hat sich die Versuchsperson für eine Interpretationsvariante des Wortes entschieden, liest sie weiter. Obwohl nicht in jedem Fall davon ausgegangen werden kann, dass auf dem Fixationspunkt auch immer die Aufmerksamkeit liegt - Menschen können sehr wohl an etwas anderes denken, als an das worauf sie schauen (z.B. auf eine rote Ampel schauen, aber an ihr Abendessen denken) - ist anzunehmen, dass in einer experimentellen Situation im Labor die Personen ihre Aufmerksamkeit auf das Geschehen am Bildschirm richten.
[264] untersuchte, wie Menschen zu einer internen Repräsentation von der Realität gelangen und analysierte dafür Reaktionszeiten auf unterschiedliche auditive Stimuli, welche die Probanden anwiesen auf bestimmte geometrische Formen zu schauen. Sobald die Probanden ein Wort wahrnahmen, standen ihnen auch die semantischen Informationen zur Verfügung und sie schauten auf die entsprechende Form. Somit wurde die von [139] postulierte Mind-Eye-Hypothese, die sie für das Lesen aufgestellt hatten, auch in einem anderen Gegenstandsbereich empirisch bestätigt.
Diese Ergebnisse unterstützen das Modell der Prozessüberwachung, wonach die Fixationen kognitiv gesteuert werden. Somit würde sich die Verarbeitungsschwierigkeit auf die Fixationsdauer auswirken. Dieses Modell wird von vielen Studien unterstützt. Allerdings untersuchten die meisten davon nur das Fixationsverhalten bei Leseaufgaben. Außerdem lassen sich mit diesem Modell nicht die relativ konstanten Fixationsdauern erklären - in

Anbetracht der zu vermutenden stärkeren Variabilität der kognitiven Prozessdauern [138]. Dem Modell der kognitiv gesteuerten Fixationen steht das Modell des kognitiven Rückstandes gegenüber. Danach sind die Fixationsdauern so kurz, dass die Informationsverarbeitung durch kognitive Prozesse zwangsweise in einem zeitlichen Abstand ablaufen muss. Dieses Modell kann jedoch nicht die systematischen Unterschiede in der Fixationsdauer erklären [138].
Werden die Fixationen in Relation zu einem bestimmten Objekt oder Bereich betrachtet, spricht man auch von *areas of interest* (AOI). Dabei wird die absolute und relative Häufigkeit der Fixationen bestimmter Objekte oder Bereiche ermittelt. Außerdem wird auch häufig die Fixationsdauer gemessen und daraus die durchschnittliche Fixationsdauer eines Objekts berechnet. Daraus lassen sich dann ebenfalls die Variation und die Verteilung der Fixationsdauern pro Objekt ermitteln. Dank moderner Bildverarbeitungsverfahren lassen sich die häufig fixierten Bereiche des Bildschirms erkennen und markieren. Diese Gebiete werden als diejenige interpretiert, die von den Probanden mehr Aufmerksamkeit bekommen haben [138].
In verschiedenen Studien, welche die visuelle Wahrnehmung von Gemälden untersuchten, wurde gezeigt, dass Personen vor allem auf interessante und informative Bereiche eines Bildes schauen und leere oder einheitliche Flächen unbeachtet lassen. Neuere Studien versuchten den Grund für das Interesse an bestimmten Bereichen zu quantifizieren und ermittelten zwei maßgebliche Faktoren: Einerseits beeinflussen statistische Eigenschaften wie eine hohe Ortsfrequenz (Bilder mit niedriger Ortsfrequenz: unscharf und flächig; Bilder mit hoher Ortsfrequenz: detailreich und scharf) und lokaler Kontrast in hohem Maß die Fixationswahrscheinlichkeit. Diese „Bottom-up-Effekte“ können analysiert und zu einer mentalen Karte mit den prägnantesten Merkmalen eines Bildes kombiniert werden. Andererseits spielen „Top-Down-Prozesse“ wie persönliches Wissen, Erfahrung und Überzeugungen eine wichtige Rolle in der Bildwahrnehmung. Auch die Fragestellung zur Betrachtung des Bildes rief drastische Unterschiede zwischen den fixierten Bereichen hervor. Außerdem ziehen Dinge, die an einem unerwarteten Ort auftauchen die Aufmerksamkeit auf sich [252].

Visual-World und Printed-Word Paradigma

In diesem Abschnitt sollen zwei wichtige Paradigmen der Psycholinguistik, die häufig in Eye-Tracking Experimenten Anwendung finden, vorgestellt werden: das Visual-World - und das Printed-Word Paradigma. Beide Ansätze haben gemeinsam, dass Versuchsteilnehmer Äußerungen hören, während sie auf einen Bildschirm schauen. Dabei werden die Augen-

bewegungen der Teilnehmer für die spätere Analyse mittels Eye-Tracking aufgezeichnet. Der Unterschied zwischen den beiden Paradigmen besteht darin, dass beim Visual-World Paradigma Bilder (von Gegenständen) gezeigt werden, die auf ein bestimmtes Wort referieren, während beim Printed-Word Paradigma die Wörter in ihrer orthografischen Form präsentiert werden.
Das Visual-World Paradigma ist die ältere und „klassische“ Variante, welche schon von [47] erstmals verwendet wurde, aber sich erst nach der Studie von [280] etablierte. Typischerweise hören dabei Versuchspersonen kurze sprachliche Äußerungen, während sie auf Abbildungen (Gegenstände in Szenen [5], Gegenstände im Raster [123], ect.) auf einem Bildschirm schauen.
[47] zeigte erstmals die enge Verbindung von gesprochener Sprache und der visuellen Welt. Wurden den Versuchspersonen auditiv Sprachstimuli präsentiert, während sie auf dem Bildschirm verschiedene Bilder bzw. Objekte sahen, richteten sie ihren Blick auf die Objekte, die semantisch in möglichst enger Beziehung zu den gerade gehörten Worten standen.
Einen Schritt weiter gingen die Studien, die visuell initiierte lexikalische Kompetitor-Effekte in einfachen „Objekt-Bewegungs“-Tasks untersuchten, während derer die Augenbewegungen der Versuchspersonen mittels Eye-Tracking beobachtet und aufgezeichnet wurden. Bei der Ausführung von Aufgaben wie „Pick up the candle“ (Nimm die Kerze auf) wurden die Versuchsteilnehmer durch Wörter mit initial identischen Phonemen, wie „candy“ (Süßigkeit), dahingehend beeinflusst, dass sie neben dem Target auch kurzzeitig auf dieses Objekt (den lexikalischen Kompetitor) schauten. Gleiches, nur mit einem unterschiedlichen Zeitpunkt, gilt für Wörter mit phonetisch identischem Reim wie „handle“ (Griff), welche ebenfalls verstärkt fixiert wurden, bis die Probanden dann (ausschließlich) auf das richtige Target blickten [279]. Den Prozess, in dem zwei lexikalisch/phonologisch ähnliche Wörter miteinander konkurrieren, bis disambiguierende Informationen eintreffen, bezeichnet man als lexikalische oder phonologische Kompetition.
[193] und [123] bewiesen, dass nicht nur die bildlichen Repräsentationen von Wörtern phonologische Kompetition auslösen, sondern auch die orthografische Form von geschriebenen Wörtern. Damit begründeten sie das Printed-Word Paradigma. Auf diese Weise lassen sich Experimente einfacher gestalten, ohne auf die Eindeutigkeit einer Bildbedeutung oder die Abbildbarkeit eines Wortes achten zu müssen. Mit der Verwendung geschriebener Wörter lassen sich auch unerwünschte Nebeneffekte von Bildern, wie Form- oder Farbkompetition unterbinden und gewünschte Effekte wie phonologische Kompetition verstärken [123].
Sowohl das Visual-World als auch das Printed-Word Paradigma beruhen auf der Annahme,

dass visuelle Referenten wie zum Beispiel Bilder von Gegenständen, aber auch geschriebene Wörter in ihrer Bedeutung mit gehörten Dingen assoziiert werden. Dabei vergehen rund 200 ms bis die Versuchsperson mit ihrem Fixationsverhalten auf den akustischen Reiz reagiert [4], da die Planung einer Sakkade diese Zeit benötigt [153]. Das Phänomen der phonetischen Kompetition tritt sowohl bei visuellen (Bildern) als auch bei orthografischen (geschriebenen) Darstellungen von Wörtern auf. Eine Voraussetzung dafür, dass ein Wort mit einer visuellen Darstellung assoziiert werden kann, ist natürlich, dass der Hörer die Bedeutung des Bildes erkennt und versteht. Gegenstände lassen sich gut abbilden und erkennen, während abstrakte Wörter uneindeutiger sind. Es ist aber durchaus möglich, den Hörern die Bedeutung von Bildern beizubringen. So mussten in einigen Studien die Versuchspersonen völlig neue Wortkreationen lernen, die jeweils mit einer abstrakten Form assoziiert werden mussten (z.B. [48]). Dies gelang ihnen nach einer Übungsphase aber meist problemlos.

Warum Menschen auf Objekte schauen, die in einer Äußerung erwähnt werden, ist nicht restlos geklärt. Es wird aber vermutet, dass es einen Vorteil bringt die auditiven und die visuellen Informationen miteinander zu verknüpfen, um Dinge und Zusammenhänge schneller verstehen zu können. Da man visuelle Informationen nur aufnehmen kann, wenn man seinen Blick auf das relevante Objekt richtet, können die Blickbewegungen etwas über die Verarbeitung der visuellen, aber auch der synchron verarbeiteten auditiven linguistischen Informationen aussagen [125].

Es wurde gezeigt, dass diese Methode verschiedene Sprachprozesse, von auditiver Worterkennung bis hin zu grammatikalischer Mehrdeutigkeitsauflösung, erfassen kann. Ein großer Vorteil besteht darin, dass sie Hinweise auf das Geschehen vor einer bestimmten Wort- oder Satzregion geben kann. Außerdem kann die Methode relativ direkte Einblicke in die Interpretationen der Versuchspersonen zur wahrgenommenen Sprache geben, während das bei Verfahren, die auf Lesen beruhen, nur indirekt möglich ist. Viele interessante Fragen über Sprache können nur durch die Einbettung von Sprache in den Kontext einer realen Welt beantwortet werden. Ein weiterer Vorteil besteht darin, dass sich das Verfahren einfach anwenden lässt und für alle Alters- und Kompetenzgruppen verwendet werden kann. Allerdings ist es nicht immer simpel Augenbewegungen der Versuchspersonen zu höheren kognitiven Funktionen wie Sprache zuzuordnen.

Das Visual-World Paradigma lässt sich aber nicht nur für die Erforschung von phonetischen Effekten verwenden, sondern auch um neue Erkenntnisse über die Verarbeitung von stimmlichen Merkmalen von Personen zu gewinnen. So konnten Hörer nicht nur Wörter anhand

abstrakter Formen lernen und wiedererkennen, sondern auch die stimmlichen Merkmale von Sprechern [50]. Auf diese Weise könnte das in der Worterkennung bewährte Verfahren des Visual-World Paradigmas auch den Prozess der Sprechererkennung erhellen.

5.1.2. Online-Worterkennung

Die korrekte Erkennung von Wörtern in gesprochener Sprache ist ein essentieller Bestandteil in unserer Kommunikation. Da gesprochene Sätze in einer großen Zahl existieren, können wir nicht alle Varianten in unserem Gedächtnis speichern. Glücklicherweise bestehen alle möglichen Sätze aus einer begrenzten Anzahl an Wörtern aus denen sich die Bedeutung eines Satzes oder einer Äußerung zusammensetzt. Daher beginnt mit dem Erkennen der Wörter eines Satzes das Verstehen der gesamten Äußerung. Dabei bedarf es keiner absoluten, expliziten Identifikation eines Wortes. Viel mehr wird auf der Grundlage der wahrgenommenen akustischen Merkmale eine Auswahl an Wörtern in Betracht gezogen und deren Wahrscheinlichkeiten in dem gegebenen Kontext beurteilt. Sobald plausible Varianten eines Wortes zur Verfügung stehen, werden auch ihre grammatikalischen und semantischen Eigenschaften erfasst, sodass mögliche Interpretationen der gesamten Äußerung gebildet werden können [191]. Die bei der Online-Worterkennung ablaufenden Prozesse und auftretenden Effekte sollen in diesem Abschnitt näher erläutert werden.
Für den Worterkennungsprozess sind zwei Typen von Informationen relevant: segmentelle (Merkmale einzelner Segmente) und suprasegmentelle (prosodische Merkmale) Informationen, welche vom Hörer aus dem Sprachsignal extrahiert werden. Es wird inzwischen allgemein angenommen, dass in diesem Prozess ein prälexikaler Level existiert, der dabei hilft, das stark variable akustische Sprachsignal auf einer bestimmten lexikalischen Repräsentation abzubilden. Allerdings besteht noch keine einheitliche Meinung über die genaue Art dieses prälexikalen Levels und ob er zum Beispiel sprach-spezifisch [190] ist oder nicht. Die segmentellen Informationen geben Aufschluss über die Laute, die in einem Wort enthalten sind und sind somit die erste und wichtigste Größe in der Worterkennung. Bei der Verarbeitung von segmentellen Informationen sind Hörer sehr sensibel und intolerant: Sobald ein Laut nicht mit der mentalen Repräsentation eines Wortes übereinstimmt, wird es als mögliches Wort ausgeschlossen. Wurde bei dem niederländischen Wort „honing“ (Honig) ein Laut verändert zu „woning“ (Wohnhaus), konnte dieses Wort von den Hörern nicht mehr als „honing“ erkannt werden, auch wenn es sich nur in einem Segment unterschied [191].
[4] zeigten, dass während der Wahrnehmung von Sprache immer bestimmte Wörter aktiviert werden, nämlich alle, die phonetisch mit dem gehörten Wort zum gegebenen Zeitpunkt

übereinstimmen. Die Aktivierung eines Wortes wurde dabei durch die proportionale Fixationsdauer auf Bilder der benannten Gegenstände bestimmt. Sie konnten nachweisen, dass diese phonetische Übereinstimmung nicht unbedingt am Wortanfang liegen musste, sondern sich auch im Reim des Wortes befinden konnte. Denn sowohl „beetle“ (Käfer), als auch „speaker“ (Lautsprecher) wurden kurzzeitig aktiviert, wenn eine Versuchsperson das Wort „beaker“ (Becher) hörte. Da „beetle“ mit dem gegebenen Wort am Wortanfang übereinstimmte, wurde es gleichzeitig mit dem intendierten Wort aktiviert, bis mit der zweiten Silbe von „beaker“ dann ein Mismatch entstand und die Aktivierung des Wortes „beetle“ wieder zurück ging. Das Wort „speaker“ hingegen wurde erst gegen Ende des Wortes „beaker“ mitaktiviert, da sich beide Wörter im Reim phonetisch glichen. Die Aktivierung aufgrund des Reimes war schwächer als die aufgrund des Wortanfangs, hielt aber länger an. [193] bestätigten diese Ergebnisse. Sie bewiesen, dass selbst bei minimalem Mismatch (nur in einem phonetischen Merkmal) am Wortanfang oder -ende, das Target sofort gegenüber den Kompetitoren präferiert wurde. Wobei die stärkste phonologische Kompetition bei den Wörtern mit Mismatch am Wortende (z.B. „buffel“ (Büffel) vs. „buffer“ (Puffer) gefunden wurde und eine schwächere bei Wörtern, die sich am Wortanfang unterschieden und sonst gleich waren (z.B. „lotje“ (Lottoticket) vs. „rotje“ (Feuerwerkskörper)). Dabei schienen sich zweisilbige Wörter besser von einem Mismatch am Wortanfang zu „erholen“ als einsilbige. Dies führten die Forscher auf den größeren prozentualen Anteil, den ein unpassendes Phonem in einem einsilbigen, im Gegensatz zu einem zweisilbigen Wort, ausmacht zurück. Bei einem längeren Wort bliebe mehr Zeit, den anfänglichen Mismatch wieder durch positive übereinstimmende Informationen auszugleichen, sodass diese Wörter als Kompetitoren etwas stärker betrachtet wurden als kürzere [193]. Generell lässt sich aber festhalten, dass negative Informationen (ein Mismatch) immer einen stärkeren Effekt zu haben scheinen als positive Informationen (Übereinstimmung). Ein einzelner, nicht-übereinstimmender Laut kann ein Wort bei der Erkennung schon ausschließen, selbst wenn es mehr übereinstimmende Laute gab.

Entscheidend ist allerdings auch, ob durch den Mismatch eine neues Wort mit einer anderen Bedeutung entsteht (wie bei niederländisch „honing“ (Honig) und „woning“ (Wohnhaus)). Besitzt das Wort viele phonologisch ähnliche „Nachbarn“ so ist die Entstehung eines anderen sinnvollen Wortes wahrscheinlicher und die Identifikation des Wortes somit schwieriger. Entsteht durch den Mismatch nur ein nicht-existentes Wort (wie „cat“ (Katze) zu „gat“), so kann die leicht fehlerhafte Phonemsequenz noch der gespeicherten Version des Wortes zugeordnet und es somit erkannt werden [191].

Ein weiterer Faktor besteht in der Größe des Mismatches. Experimente mit kreuzweise ausgetauschten Phonemen, die dann nicht übereinstimmende koartikulatorische Informationen enthielten, zeigten, dass auch hier wiederum die Herkunft der Mismatchinformationen entscheidend war, und ob es sich bei den Herkunftswörtern um echte oder nicht existente Wörter handelte. Auch die Manipulation der VOT (voice onset time) bei stimmlosen Plosiven erzeugte mit zunehmender Reduktion einen immer stärkeren Mismatch [191]. [45] zeigten, dass sogar subphonemische Merkmale wie die Distribution der VOT von Plosiven die Worterkennung in gesprochener Sprache beeinflusst. Sie erstellten zwei /b/-/p/-Kontinua (z.B. „beach“ (Strand) vs. „peach“ (Pfirsich)), eines mit einer geringen und eines mit einer starken Variation der VOT. Bei einer geringen Variation sind die VOT-Merkmale sehr informativ für den initialen Plosiv, wohingegen sie durch eine wachsende Variation immer uninformativer werden bis die Merkmale bei einer vollständigen Überlappung beider Kategorien überhaupt keine Aussagekraft mehr hätten. Die wachsende Unsicherheit der Hörer in ihrem Urteil zeigte sich in einer höheren Fixationswahrscheinlichkeit des phonologischen Kompetitors gegenüber dem Target und offenbarte somit die Sensibilität der Hörer für die Verteilung feiner phonetischer Merkmale.
Bei der Erkennung von gesprochenen Wörtern verwenden Hörer alle ihnen zur Verfügung stehenden Informationen, sowohl temporale (wie VOT-Merkmale) als auch spektrale Informationen aus den Wörtern selbst und ihrem linguistischen Kontext. Dabei ging die Verarbeitung der spektralen Merkmale leicht, aber statistisch signifikant der Verarbeitung der temporalen Merkmale voran [245]. Der Einfluss eines solchen phonetischen Details richtet sich ebenfalls nach seinem Nutzen für eine lexikalische Unterscheidung [191].

Neben den segmentellen Merkmalen können auch suprasegmentelle Informationen den lexikalischen Zugang von Phonemfolgen beschränken. Das Betonungsmuster eines Wortes, in welcher Weise seine Silben sich in ihrer Akzentuierung unterscheiden, kann durch Unterschiede in der Grundfrequenz, Dauer und Lautstärke signalisiert werden. Verschiedene Studien zeigen, dass ein Mismatch in der Betonung der initialen Silben eines Wortes den lexikalen Zugang stören. Der Einfluss der Prosodie ist natürlich zwischen den Sprachen verschieden. Einerseits haben nicht alle Sprachen eine lexikalische Betonungsunterscheidung und andererseits ist in manchen Sprachen die Betonung in ihrer Position festgelegt, sodass sie vorhersagbar und somit wenig informativ für lexikalische Unterschiede ist. Aber auch in den Sprachen mit freier lexikalischer Betonung gibt es Unterschiede in der Sensitivität für prosodische Informationen. Beispielsweise scheinen niederländische Hörer Informationen aus

dem Betonungsmuster stärker zu verwenden als englische Hörer [53]. Begründet wird das dadurch, dass Betonungsänderungen im Englischen häufig eher mit segmentellen Änderungen (z.B. Vokalreduktion) einhergehen (noun „conduct“ vs. verb „conduct“). Außerdem tragen Dauerunterschiede zwischen ein- und mehrsilbigen Wörtern zur lexikalen Entscheidung bei. So ist die gleiche Silbe in einem einsilbigen Wort länger als in einem mehrsilbigen (z.B. „cap tucked“ vs. „captain“), was dabei hilft die Passgenauigkeit eines möglichen Wortes zu überprüfen ([60] [257]).
Es lässt sich festhalten, dass Hörer Informationen in Abhängigkeit von ihrer Fähigkeit zur lexikalischen Disambiguierung für die Worterkennung verwenden. Dabei spielt das akustische Signal die wichtigste und entscheidende Rolle, auch wenn es weitere Einflussgrößen gibt. So wird verständlich, weshalb in verschiedenen Sprachen oder Situationen verschiedene Arten von Signalinformationen verwendet werden [191].

Es ist zu beachten, dass es bei der gleichzeitigen Aufnahme von visuellen und auditiv sprachlichen Informationen, mehrere Level gibt, auf denen eine Beziehung zwischen beiden Informationen gefunden werden kann (sprachliche Level wie Phonologie, Semantik, aber auch nicht sprachliche wie Form und Farbe [123]). In der Psycholinguistik sind hauptsächlich die sprachlichen Level der Repräsentation relevant, wie phonologische (gesprochenes Wort „candle“ (Kerze) und Bild einer Kerze), semantische (gesprochenes Wort „piano“ (Klavier) und das Bild einer Trompete) oder assoziative („Computer“ (Computer) und „Mouse“ (Maus)) [124]. Dabei sind semantische Beziehungen auch ohne phonologische Übereinstimmung von zwei Stimuli vorhanden, sodass auch zwei phonologisch unähnliche Wörter semantische Kompetition auslösen können [122]. Allerdings gibt es auch nicht sprachliche Repräsentationen, die eine Beziehung zwischen zwei Informationsquellen herstellen können, wie z.B. die Form oder die Farbe eines Objektes. So schauten die Probanden beim Erklang des Wortes „snake“ (Schlange) auch auf das Bild eines Kabels oder bei dem Wort „lips“ (Lippen) auf das Bild einer Erdbeere. Die Fixationswahrscheinlichkeit eines visuellen Objekts spiegelt somit die Übereinstimmung von gespeichertem Wissen über visuelle Eigenschaften eines Wortreferenten (z.B. Schlange = lang und dünn; Lippen = rot) und den aus dem visuellen Objekt gewonnenen visuellen Merkmalen wieder [124].
Mehrere Studien zeigten, dass Probanden visuelle Objekte (still) benennen und ihnen so eine phonologische Repräsentation zuordnen. Allerdings spielt die Zeit, die die Teilnehmer für das Betrachten der visuellen Objekte am Bildschirm hatten, bevor das gesprochene Wort präsentiert wurde, eine wichtige Rolle. Die Benennung eines Objektes scheint länger zu

dauern als das Erkennen von visuellen (Form, Farbe, ect.) oder semantischen Übereinstimmungen. War die Zeitspanne zwischen der visuellen Präsentation der Bilder am Bildschirm und der auditiven Präsentation des Targetwortes nur 200 ms oder geringer, zeigten sich keine phonologischen, sondern nur Form- und semantische Kompetitoreffekte [123].

[191] kommt zu dem Schluss, dass es bei der Worterkennung in gesprochener Sprache drei verschiedene Repräsentationsarten gibt: prälexikalische, Wortform- und Wortbedeutungsrepräsentationen. Es ist umstritten, ob diese Kategorien tatsächlich existieren, da auch noch ungeklärt ist, in welcher Repräsentationsform sprachliche Informationen überhaupt in unserem Gehirn gespeichert werden. Die Prozesse der gesprochenen Worterkennung sind dagegen schon relativ gut erforscht. Es scheint, dass sprachliche Informationen nach ihrer Aufnahme sofort an die nächsten „Verarbeitungsstufen“ weitergeleitet werden. Sprich, akustische und phonetische Informationen werden sofort verwendet um Hypothesen über die Wortform und anschließend auch über die Wortbedeutung zu formulieren. Diese werden mit Zunahme der phonetischen Informationen aus dem Sprachsignal immer wieder aktualisiert und angepasst bis am Ende (meistens) nur noch eine mögliche Interpretation für ein Wort übrig bleibt.

Der Einfluss von Sprechercharakteristika auf die Worterkennung

Sprache enthält nicht nur linguistisch relevante phonetische und lexikalische Informationen, sondern auch Informationen über die Identität des Sprechers. Aufgrund dieser dualen Funktion interagieren beide Informationstypen in der Sprachperzeption miteinander ([49] [50] [54]). In mehreren Studien wurden die Auswirkungen von sprecherspezifischen Merkmalen auf die Erkennung von gesprochenen Wörtern untersucht. Dabei gibt es widersprüchliche Ansichten darüber, wie bei der Verarbeitung von gesprochener Sprache verfahren wird. Die zwei vorherrschenden Modelle sollen hier kurz vorgestellt werden: (1) Zum einen gibt es das Abstraktionsmodell (in verschiedenen Modifikationen), welches annimmt, dass es eine Art Normierungsprozess gibt, der alle nicht lexikalischen und somit für die Worterkennung bedeutungslosen Informationen, wie z.B. Unterschiede zwischen Sprechern (oder auch innerhalb eines Sprechers), heraus filtert. Nach dem Normierungsfilter bleiben nur die „reinen“ lexikalischen Informationen übrig, die dann auf einer abstrakten mentalen Wortrepräsentation abgebildet werden (z.B. [174] [184] [273] [215] [213]). (2) Auf der anderen Seite gibt es das Exemplar-Modell, welches die Annahme trifft, dass die mentalen Wortrepräsentationen, die wir in unserem Gehirn gespeichert haben, auch die nicht-lexikalischen Informationen

wie Sprechercharakteristika enthalten (z.B. [133] [103] [135] [231] [232]). Dadurch kann es natürlich nicht nur eine abstrakte gespeicherte Form für ein Wort geben, sondern es muss viele Exemplare von jedem Wort geben, welche alle die gleiche Bedeutung haben, aber unterschiedliche zusätzliche nicht-lexikalische Informationen. Inwieweit das jeweilige Modell die Ergebnisse der Experimente in diesem Kapitel erklärt, wird in der generellen Diskussion besprochen (siehe Abschnitt 5.4).
Die lexikalischen und nicht lexikalischen Informationen interagieren auf verschiedenen Ebenen miteinander. Wie [130] zeigen, hängen sowohl die lexikalische als auch die sublexikalische (oder auch prälexikalische) Ebene von Sprechercharakteristika ab, letztere allerdings schwächer. Aber auch die Wortverarbeitung auf höheren Ebenen, wie der semantischen und pragmatischen Ebene, wird von Sprecherinformationen beeinflusst. Sprach zum Beispiel eine eindeutig männliche Stimme einen Satz wie „Ich würde gern wie Britney Spears aussehen.“, so löste dieser Mismatch eine stärkere Reaktion im Ereignis-Korrelierten-Potential (EKP) aus, als wenn der Satz und die Sprechermerkmale übereinstimmten. [33]. Für das exemplarische Modell sprechen Resultate, nach denen Sprechermerkmale den Prozess der Worterkennung unterstützen oder behindern können (z.B. [102] [103] [259]). Wenn Versuchspersonen Wortlisten, die von mehreren Sprechern gesprochen wurden, hören, fällt es ihnen schwerer zu entscheiden, ob sie ein bestimmtes Wort schon einmal gehört haben, als wenn die Wörter nur von einem Sprecher gesprochen wurden [103]. Außerdem brauchen Hörer länger um ein Wort zu erkennen, wenn es in einem ihnen unbekannten regionalen Akzent gesprochen wurde [1]. Wenn sie sich allerdings auf einen Sprecher oder sogar eine Sprechergruppe und seine/ihre speziellen Sprachcharakteristiken eingestellt haben, können sie die Wörter dieser Sprecher schneller und besser erkennen [259]. Neue Wörter von bekannten Sprechern werden auch bei Anwesenheit von Störgeräuschen besser erkannt, als bekannte Wörter von unbekannten Sprechern [219]. Allerdings konnten Sprechermerkmale nicht von einer lexikalischen Einheit (Wörter, Sätze, ect.) auf eine andere generalisiert werden. Eine verbesserte Worterkennung zeigt sich nur nach dem Training der Hörer mit Wörtern; eine verbesserte Satzerkennung nur nach dem Training mit Sätzen [218].

Weiterhin belegen viele Perceptual Learning Studien den Einfluss von sprecherbezogener Variation auf die Worterkennung in gesprochener Sprache. So zeigen beispielsweise [71], dass die Kategoriengrenze zwischen /s/ und /f/ so verschoben werden konnte, dass Wörter mit dem gleichen ambigen Laut am Wortende bei einem Sprecher mit einem /s/ und bei dem anderen Sprecher mit einem /f/ wahrgenommen werden, was die Bedeutung der Wörter

verändert. Diese perzeptiv gelernte Kategorienveränderung ist für diesen einen Sprecher stabil, überträgt sich aber nicht auf andere Sprecher. Das bedeutet, dass die Hörer gelernt haben, dass dieser eine Sprecher diese spezielle Aussprachevariante von /s/ bzw. /f/ hat. Da das aber nicht heißt, dass andere Sprecher genauso sprechen, übertragen die Hörer ihr gelerntes Wissen von dem einen Sprecher nicht auf andere. Die stimmlichen Charakteristika des Sprechers entscheiden also darüber, welche Laute wie interpretiert werden. Auf diese Weise wird ein sprecherspezifisches Merkmal plötzlich bedeutungsunterscheidend in der Worterkennung. Das wiederum belegt, dass Sprecherinformationen aus dem Sprachsignal vor der Weiterverarbeitung nicht „herausgefiltert“ werden, sondern dass sie bestehen bleiben und mitverarbeitet werden und gegebenenfalls für die Wortidentifikation herangezogen werden können.

Allerdings gibt es auch bei diesem Effekt des perzeptiven Lernens lautabhängige Unterschiede. So scheinen unterschiedliche Aussprachevarianten bei Frikativen von den Hörern als sprecherspezifisch erachtet zu werden, während Aussprachevarianten bei Plosiven stärker generalisiert und auch auf neue Sprecher übertragen werden. [157] vermuten, dass dies an den unterschiedlichen Variationsdimensionen der beiden Lautkategorien liegt. Während sich Plosive (in ihrer Stimmhaftigkeit) entlang eines temporalen Kontinuums ihrer VOT unterscheiden, liegt der Unterschied bei Frikativen in ihrer spektralen Zusammensetzung. Im Falle eines temporalen Unterschiedes zwischen zwei Plosivproduktionen scheinen Hörer ihre Kategoriegrenzen immer dem jeweils zuletzt gehörten Plosiven anzupassen, unabhängig um welchen Sprecher es sich handelt. Spektral unterschiedliche Varianten, wie im Falle der Frikative, können dagegen anscheinend nebeneinander existieren und werden sprecherabhängig erlernt und gespeichert. Eine mögliche Erklärung für diese Differenzen liegt in der unterschiedlichen Informativität der beiden Lautkategorien für die Sprecheridentität. Frikative enthalten sprecherspezifische Informationen, Plosive hingegen nicht. Da mit den verschiedenen Frikativvarianten auch immer Unterschiede in den spektralen Merkmalen (welche sprecherspezifisch sind) einhergehen, werden diese Aussprachevarianten sprecherspezifisch behandelt. Die temporale Variation zischen den Plosiven kodiert allerdings keine Sprecherunterschiede, weshalb diese unterschiedlichen Varianten nicht mit einem bestimmten Sprecher assoziiert, sondern allgemein erlernt werden. Allerdings gibt es auch Studien, die dieser Erklärungshypothese widersprechen. In denen die VOT sehr wohl sprecherabhängig variierte und die Frikative sprecherambig waren und generalisiert wurden. Ein weiterer Unterschied zwischen den Plosiven und Frikativen in dem Experiment von [157] ist, dass die Plosive in ihrer Stimmhaftigkeit variieren (/d/ vs. /t/), wohingegen die Frikative sich

in ihrer Artikulationsstelle unterscheiden (/s/ vs. /ʃ/). Diese Tatsache könnte ebenfalls das unterschiedliche Verhalten im perzeptiven Lernen der beiden Lautkategorien begründen, genauso wie zahlreiche nicht-akustische Faktoren, wie zum Beispiel soziale Faktoren [157]. Zusammenfassend könnte man daraus schlussfolgern, dass spektrale Merkmale eher als sprecherspezifisch angenommen werden und temporale Merkmale eher als lautspezifisch. Allerdings sollte diese Erklärung nur als tendenzielle Richtung betrachtet werden, da es durchaus Ausnahmen gibt.
[56] untersuchten, ob diese Anpassung an einen neuen Sprecher auf der Anpassung des Sprachsignals an die vorhandenen mentalen Repräsentationen beruht oder auf einer Anpassung der Repräsentationen selbst. Sie argumentieren, dass (auch) letzteres der Fall sein muss, da die Hörer die Aussprachevariante eines gelernten Sprecherdialekts auch auf andere ähnliche Wörter übertragen, die von dem Sprecherdialekt gar nicht betroffen sind. Außerdem könnte ein gelernter Sprechereffekt abhängig von seiner Position in der Silbe sein und sich nicht auf andere Positionen übertragen [129].

Normalerweise sind in den bekannten Sprachen Sprechercharakteristika nicht bedeutungunterscheidend. Dass diese Fähigkeit aber erst erworben und erlernt werden muss, zeigen Forschungsergebnisse aus dem Erst- ([121] [22]) und Zweitsprachenerwerb ([36]). Kleinkinder oder Lernanfänger einer Fremdsprache können zunächst nur die Wortvarianten eines einzelnen Sprechers erkennen, nämlich die der Mutter oder des Sprachlehrers. Das gleiche Wort von einem anderen Sprecher enthält zu viele sprecherspezifische Merkmale, als dass sie in diesem Wort das gleiche wie das bekannte erkennen können. Erst wenn sie von vielen anderen Sprechern das gleiche Wort gehört haben, sind sie in der Lage, über sprecherspezifische Merkmale im Sprachsignal zu generalisieren. Erst dann können sie beurteilen, welche Informationen im Sprachsignal lexikalisch und somit bedeutungsunterscheidend sind und welche nicht [48].
Auch wenn Sprechermerkmale im Normalfall keine lexikalische Bedeutung tragen, so können sie doch dabei helfen, bestimmte Wörter schneller zu erkennen. Wenn der Hörer zum Beispiel gelernt hat, dass ein Sprecher immer bestimmte Wörter verwendet und ein anderer Sprecher andere Wörter, dann kann er anhand der stimmlichen Informationen des Sprechers bereits die Wahrscheinlichkeit für ein Wort abschätzen. So konnte [48] zeigen, dass ähnlich klingende Wörter (z.B. „sheep“ (Schaf) und „sheet“ (Bettlaken)) weniger lexikalische Kompetition hervorrufen, wenn sie von verschiedenen Sprechern (einem männlichen und einem weiblichen) statt vom gleichen Sprecher gesprochen werden. Auch fiktive Wörter, welche

die Versuchspersonen zu Beginn des Experiments erst lernen müssen, erzielen das gleiche Ergebnis.
Sprecher sind sehr schnell in der Lage, sich auf die spezifischen Aussprachevarianten eines neuen Sprechers einzustellen. Auf der Grundlage dieses neu erworbenen Wissens werden dann seine gesprochenen Wörter interpretiert. Anhand von zwei Sprechern, von denen einer eine dialektale Aussprachevariante von „bag“ (Tasche) hatte, wiesen [286] nach, dass je nachdem welchen Sprecher die Probanden hörten, sie die Standard- oder die Dialektvariante von „bag“ als das entsprechende Wort wahrnahmen. Wie gut und schnell sich Hörer auf ein spezielles Merkmal z.B. einen neuen Akzent einstellen können, ist abhängig von der Stärke des Akzents und dem Grad in dem der Sprecher mit diesem Akzent vertraut ist [294]. Der Prozess der Adaption läuft dabei automatisch und kann auch nach einer längeren Pause zwischen Trainings- und Testphase noch beobachtet werden [293].
Laut der Studie von [224] können Hörer sowohl implizit als auch explizit die sprecherspezifischen Merkmale verwenden, um Wörter und Sprecher besser identifizieren zu können. Dabei spielt es keine Rolle, ob die Stimmen von 2 oder von 20 Sprechern erlernt werden. Daraus lässt sich ableiten, dass die Aufnahme und Verwendung von Sprecherinformationen nicht bewusst von den Hörern gesteuert wird und daher nicht strategisch eingesetzt werden kann.

Nicht nur die spezifischen Merkmale innerhalb eines Lautes können die Wortwahrnehmung der Hörer beeinflussen. Auch sprecherspezifische Merkmale im Kontext (z.B. ein generell hoher F1 oder ein generell tiefer F1) können beeinflussen, welches Wort die Hörer eher wahrnehmen. Das zeigt, dass Hörer sich auf die Vokaltrakteigenschaften eines bestimmten Sprechers einstellen und jeden Laut im Verhältnis zu diesen Merkmalen interpretieren. Dadurch kann ein identischer Laut in zwei verschiedenen Kontexten als zwei verschiedene Laute wahrgenommen werden [267].
Wenn Hörer Phoneme eines unbekannten Sprechers hören, welche sich in bestimmten Frequenzen von ihren vorhandenen mentalen Repräsentationen unterscheiden, dann müssen sie sich erst auf Grundlage des Kontextes auf die veränderten Frequenzen einstellen. Dieses Phänomen zeigten erstmals [162] in ihrer Untersuchung von Vokalinformationen. Spielt man Hörern einen ambigen Laut in einem konstanten /b_t/ Kontext vor und manipuliert den ersten und/oder zweiten Formanten (F1 und F2) im Trägersatz, so wird der Laut von den Hörern unterschiedlich kategorisiert. Eine Manipulation der Formantfrequenzen der Vokale erzeugt grob gesagt perzeptiv den Eindruck eines längeren oder kürzeren Vokaltrakts und somit den eines anderen Sprechers. In Abhängigkeit von den Vokalformanten

des Trägersatzes wird der Ziellaut (ambiger Vokal) als ein anderer Laut wahrgenommen. Das heißt, die Hörer kompensieren den Effekt des Sprechers und interpretieren den Laut abhängig von den Sprechermerkmalen. Diesen Prozess bezeichnen [164], die weitergehende Untersuchungen zu dem Thema durchführten, als Sprechernormalisierung. Sie zeigten, dass wenn spektrale Bereiche, die zur Identifikation eines Lautes wichtig sind (z.B. die Lage von F3 bei der Unterscheidung von /d/ und /g/) im Kontextsatz manipuliert werden, der Laut dann als ein anderer wahrgenommen wird. Werden allerdings nur nicht-relevante Bereiche (wie z.B. F1) verändert, so hat dies keine Auswirkung auf die Lauterkennung der Hörer. Dieser Effekt kann nicht nur mit sprachlichen, sondern auch mit nicht sprachlichem Kontext (verschieden hohe Töne) reproduziert werden und ist dabei sogar stärker. Daher argumentieren [164], dass sich Hörer nicht auf sprecherspezifische Vokaltraktparameter einstellen, sondern auf die durchschnittliche Energie der Frequenzen (LTAS). Somit würde es keine Rolle spielen, ob die Unterschiede durch verschiedene Sprecher oder andere akustische Faktoren verursacht werden, was einen stärkeren Einfluss allgemeiner auditiver Prozesse in der Sprachwahrnehmung vermuten lässt.

5.1.3. Online-Sprechererkennung

Nachdem erläutert wurde, in welchem Maße Sprechermerkmale die Worterkennung in gesprochener Sprache beeinflussen, soll nun auch die Sprechererkennung an sich betrachtet werden. Jedoch gibt es nicht halb so viele Studien zur Online-Sprechererkennung wie zur Online-Worterkennung. Oft wurde die Sprechererkennung nur indirekt bei der Worterkennung mituntersucht. Aus diesen Untersuchungen lassen sich aber schon erste Informationen über den Prozess der Online-Sprechererkennung gewinnen. Eine der interessantesten Fragen ist dabei, in welcher zeitlichen Koordination die Sprechererkennung zur Worterkennung steht. Einige Studien haben sich dieser Frage mehr oder weniger detailliert gewidmet, brachten aber keine einheitlichen Ergebnisse hervor.

Einzelne Untersuchungen präsentieren die Verarbeitung der Sprecherinformationen eher als einen nachfolgenden Prozess zur Worterkennung. So sprechen die Ergebnisse von [188] für eine relativ späte Aufnahme und Verarbeitung von Sprecherinformationen. Nur bei einer ausreichend langen Zeitspanne können sich Sprechermerkmale manifestieren und mit lexikalischen Informationen interagieren. Der Fakt, dass die sprecherspezifischen Merkmale erst spät verarbeitet werden, wird als Indiz dafür gesehen, dass es zusätzliche Informationen sind, die zu den lexikalischen hinzukommen [48]. Ebenfalls in dieser Studie werden sprecherspezifische Informationen von den Hörern frühestens 500 ms nach Targetbeginn

zur Target-Disambiguierung verwendet. Sind die Wörter neu erlernt, setzt der Effekt noch später ein. Allerdings sind in diesem Fall auch die phonemischen Effekte langsamer, sodass die Relation die gleiche bleibt. [227] untersuchten die Phonem- im Vergleich zur Gender-Klassifikation und zeigten in ihrem Versuch, dass erstere letzterer voran geht. Dabei liefern die Vokale vor allem Informationen über die stimmlichen Merkmale, während die Konsonanten für die Phonemklassifikation entscheidend sind.
In anderen Studien wird wiederum von einer parallelen und gleich schnellen Verarbeitung von Sprecherinformationen berichtet. [152] untersuchten die Verarbeitungsprozesse von Stimm- und Wortinformationen mit Hilfe von Magnetoenzephalographie (MEG). Dabei zeigte sich, dass beide Arten von Informationen parallel und sehr schnell verarbeitet werden. Schon bevor die Wörter von den Hörern bewusst wahrgenommen werden, machen sich Sprecher- und Worteinflüsse in den neurophysiologischen Messungen bemerkbar. Und auch [50] zeigten in einer späteren Studie zur akustischen und semantischen Interpretation von Sprecherinformationen, dass Hörer zwei Sprecher ähnlich schnell unterscheiden können wie zwei Wörter. Hörer merken sich Sprecher sehr gut, besonders wenn sie explizit dazu aufgefordert werden. Aber auch wenn die Sprecherinformationen dabei helfen, die Aufgabe (z.B. das Erkennen bestimmter Wörter) besser und schneller zu absolvieren, verwenden die Hörer die stimmlichen Informationen der Sprecher. Ist keines von beidem der Fall, scheinen die Hörer die Sprecherinformationen zu ignorieren bzw. nicht zu verwenden.

Es lässt sich spekulieren, dass analog zur Studie von [50] (in der Hörer lernten, Sprecher mit abstrakten Formen zu assoziieren) Hörer auch in der Lage sein sollten, Fotos von Personen mit den Stimmen von bestimmten Sprechern zu assoziieren. Genauso wie bei der Präsentation von Wortreferenten, sollten die Hörer wenn sie die Stimme einer bestimmten Person hören, häufiger auf das Bild dieser Person schauen als auf die Bilder von anderen Personen. Statt einem phonetisch Kompetitor (ähnliches Wort) hätte man dann einen Sprecherkompetitor (ähnliche Stimme). Die Übertragung des Konzepts „Kompetitor“ aus dem lexikalischen Bereich in den Bereich der Sprechererkennung ist neuartig und daher potenziell problematisch. Es ist schwierig einzuschätzen, wann zwei Stimmen das gleiche Maß an Ähnlichkeit besitzen wie zwei Wörter. Sind zwei weibliche Sprecher sich so ähnlich wie zwei Wörter, die mit den gleichen zwei Lauten beginnen oder mit den gleichen drei? Das hängt vermutlich von der Ähnlichkeit der weiblichen Sprecher ab, z.B. wie nah ihre Grundfrequenz beieinander liegt. Da es in dem Bereich der Sprechererkennung bisher noch keine Erfahrungen mit der Konstruktion von Sprecherkompetitoren gibt, wird diese Arbeit

einen ersten Versuch wagen. Dabei sollen in zwei Experimenten unterschiedlich ähnliche Sprecher und Wörter als Kompetitoren verwendet werden.
Bei der Sprechererkennung gilt natürlich die gleiche Vorraussetzung wie bei der Worterkennung, nämlich dass der Hörer die abgebildete und gehörte Person kennen muss. Dies kann durch einen Trainingsprozess vor Beginn des Experiments gewährleistet werden. Im anschließenden Test ließe sich dann vermutlich Kompetition zwischen ähnlichen Sprechern beobachten. Sind sich zwei Stimmen akustisch ähnlich, könnten Hörer zunächst beide Sprecher in Betracht ziehen, bevor sie sich für einen entscheiden. Worin die akustische Ähnlichkeit besteht, kann vielfältig sein. Zum Beispiel wäre anzunehmen, dass sich Sprecher innerhalb eines Geschlechts ähnlicher sind als verschieden geschlechtliche Sprecher. Würde man also am Bildschirm die Bilder von zwei Männern und zwei Frauen sehen und anschließend eine weibliche Stimme hören, wäre zu erwarten, dass man zunächst auf die Bilder beider Frauen schaut (aber nicht oder nur sehr wenig auf die Bilder der Männer), bevor man sich eindeutig für eine der Frauen entscheidet. Der Grad der Ähnlichkeit würde dabei die Stärke der Sprecherkompetition beeinflussen. Da diese Sprecherkompetition auf der Ähnlichkeit des Geschlechts beruht, würde man in diesem Fall auch von Gender-Kompetition sprechen. Es wären aber auch Varianten von Sprecherkompetition innerhalb eines Geschlechts denkbar, zum Beispiel zwischen zwei Männern oder Frauen mit ähnlicher Grundfrequenz.

5.2. Experiment 1

5.2.1. Hypothesen und Ziele

Im Vergleich zur Worterkennung ist nur wenig über den Prozess der Sprechererkennung und seine zeitliche Koordinierung bekannt. Es ist unklar, ob bei der Sprechererkennung analog zur Worterkennung Sprecherkompetition im Verlauf des Identifikationsprozesses auftritt. Bei der Worterkennung wird bis zur Disambiguierung der Wörter auch auf den lexikalischen Kompetitor geschaut. In der Sprechererkennung sind natürlich bereits ab Wortbeginn disambiguierende Merkmale vorhanden. Dadurch könnten die Sprecher relativ schnell disambiquiert werden. Allerdings ist es für Hörer wahrscheinlich schwieriger verschiedene Sprecher zu identifizieren als Wörter, da Sprecher in unserem Alltag seltener unterschieden werden müssen als Wörter. Dennoch sind Hörer natürlich in der Lage (besonders bekannte) Sprecher schnell zu erkennen und zu unterscheiden. Deshalb wäre es interessant zu sehen, ob und wie stark Sprecherkompetition auftritt, zu welchem Zeitpunkt sie einsetzt und wie lange die Disambiguierung dauert.

Ein weiterer Interessenspunkt besteht darin, die zeitliche Relation der Sprechererkennung zur Worterkennung zu untersuchen. Nutzen die Hörer zuerst die phonetischen/lexikalischen Merkmale des Wortes, um dann den Sprecher erkennen zu können, oder verwenden sie die Sprechermerkmale in der Stimme, um dann das Wort besser verstehen zu können? Die zeitliche Reihenfolge des Auftretens der beiden Kompetitionstypen könnte helfen eine Antwort auf diese Frage zu finden. Abgesehen von dem zeitlichen Ablauf, könnte auch die Stärke der Kompetitoren etwas über die Schwierigkeit der Unterscheidung aussagen.
Das Visual-World Eye-Tracking Paradigma, das verwendet wurde um die Aufnahme von feinen phonetischen Details zu offenbaren [258], könnte auch Einblicke in die relative zeitliche Koordinierung der Erkennung eines Sprechers einer bestimmten Äußerung gewähren. Die Hörer führten einen Visual-World Eye-Tracking Task aus, in dem sie Wörter von zwei Sprechern hörten und dabei am Bildschirm Sprecher-Gegenstand Kombinationen - ein Bild von einem Sprecher kombiniert mit einen Bild von einen Gegenstand - betrachteten. Somit konnten zur Identifizierung des korrekten Referenten sowohl phonetische als auch Sprecherinformationen verwendet werden. In der *Sprecherbedingung* mussten die Hörer beide Informationen - phonetische und Sprecherinformationen - verwenden um den korrekten visuellen Referenten zu finden, da am Bildschirm jeder Gegenstand zweimal gezeigt wurde, aber kombiniert mit dem jeweils anderen Sprecher. In der *Gegenstandsbedingung* konnten Sprecherinformationen genutzt werden um den Erkennungsprozess zu beschleunigen, aber die alleinige Verwendung von phonetischer Information war ausreichend zur Lösung des Problems.
In beiden Bedingungen wurde Kompetition zwischen phonetisch ähnlichen Gegenständen erwartet, wie es bereits in vorangegangenen Studien mit dem Visual-World Paradigma gezeigt wurde [4]. Die kritische Frage in der aktuellen Studie war jedoch, ob auch Kompetition zwischen den Sprecherreferenten auftreten würde und in welcher zeitlichen Koordination zur Aufnahme der phonetischen Information diese stehen würde.

5.2.2. Methode

Versuchspersonen

24 Studenten der Ludwig-Maximilians-Universität München nahmen gegen eine kleine Vergütung an der Studie teil. Alle Teilnehmer waren deutsche Muttersprachler und zwischen 19 und 28 Jahren alt. Keiner der Teilnehmer gab eine Sprach- oder Hörstörung an.

Sprachmaterial

Es wurden 128 abbildbare, ein-bis dreisilbige deutsche Substantive ausgewählt, die mit einer Konsonant-Vokal-Sequenz (CV-Sequenz) begannen (siehe Anhang). Jeweils 16 Wörter begannen mit einem der folgenden Konsonanten: /p, t, b, d, f, s, m, n/ gefolgt von einem Vokal. Dabei wurde darauf geachtet, dass jedem Anfangskonsonanten genauso häufig ein bestimmter Vokal folgte. Alle Wörter wurden von zwei jungen männlichen deutschen Muttersprachlern produziert und aufgenommen. Zur Sprachaufnahme wurde die Software Speechrecorder verwendet [64]. Für jedes Wort wurde per Google Bildsuche ein passendes Bild ausgesucht, dass dieses Wort darstellt (z.B. ein Bild von einem Ball für das Wort „Ball“). Bei Unsicherheit darüber, welches Bild sich für die Darstellung eines Wortes am besten eignete, wurde die Einschätzung einer zusätzlichen Person eingeholt. Das Bild, wo die Ansichten beider Personen übereinstimmten, wurde dann ausgewählt. Außerdem wurden zwei Bilder von zwei jungen Männern ausgewählt, die die Sprecher repräsentieren. Dabei wurde auf ein ähnliches Alter und Gesichtsausdruck der Männer, einen ähnlichen Hintergrund sowie eine vergleichbare Bildqualität geachtet.

Akustische Merkmale der Sprecher

Für dieses Experiment wurden zwei männliche Sprecher ausgewählt, damit die Sprecher nicht zu einfach zu unterscheiden waren (wie bei einer Unterscheidung zwischen einem männlichen und einem weiblichen Sprecher). Diese Auswahl sollte auch sicherstellen, dass die Ergebnisse auf sprecherspezifischen und nicht auf geschlechtsspezifischen Merkmalen beruhen. Dennoch sollten sich die Stimmen aber nicht zu ähnlich sein, um von den Hörern sicher unterschieden werden zu können. Um dies sicher zu stellen, wurde eine akustische Analyse mit Praat ([35]) durchgeführt. Wichtige Merkmale zur perzeptiven Sprecherunterscheidung sind die Grundfrequenz ([208] [37]) und die durchschnittlichen Formantfrequenzen ([86] [246]), weshalb sich die akustische Analyse darauf konzentrierte.

Sprecher	F0	F0-SD	min. F0	max. F0	F0-Bereich	F1	F2	F3	F4
m1	134 Hz	30 Hz	111 Hz	204 Hz	93 Hz	532 Hz	1501 Hz	2693 Hz	3890 Hz
m2	115 Hz	40 Hz	85 Hz	222 Hz	137 Hz	600 Hz	1656 Hz	2628 Hz	3494 Hz

Tabelle 5.1.: Mittlere akustische Merkmale der Sprecher m1 und m2

Die Analyse zeigt, dass sich die beiden Sprecher in ihren Grundfrequenzmerkmalen leicht, aber nicht signifikant, unterscheiden (siehe Tabelle 5.1). Sprecher 1 hatte eine höhere

Stimmtonlage als Sprecher 2, aber eine geringere Variation. Dies wird deutlich, wenn man die Standardabweichung, die minimale und maximale F0 und den Grundfrequenzbereich betrachtet. Auch in den mittleren Formantwerten zeigen sich Unterschiede. Während Sprecher 1 in F1 und F2 niedrigere Formantfrequenzen hat, sind sie für F3 und F4 höher als die von Sprecher 2.

Design

Die Bilder der Sprecher und Gegenstände wurden zu zwei verschiedenen Typen von Displays kombiniert: In der *Sprecherbedingung* wurden beide Sprecher den selben zwei Gegenständen zugeordnet, zum Beispiel wurden beide Sprecher mit den Bildern *Fernrohr* und *Filzhut* angezeigt. Die Gegenstände begannen mit dem gleichen Konsonanten gefolgt von einem unterschiedlichen Vokal. In der *Gegenstandsbedingung* wurden vier verschiedene Gegenstände angezeigt. Sprecher 1 wurde mit zwei Gegenständen, beginnend mit zwei verschiedenen Konsonanten und Vokalen (z.B. *Fernrohr* und *Silber*) kombiniert. Sprecher 2 wurde dann mit Gegenständen der gleichen Anfangskonsonanten, aber einem anderen Folgevokal abgebildet (z.B. *Sessel* und *Filzhut*) (siehe Abbildung 5.2).

(a) Beispieldisplay für die Sprecherbedingung, welche die zwei Sprecher und die Gegenstände *Fernrohr* und *Filzhut* zeigt.

(b) Beispieldisplay für die Gegenstandsbedingung, welche die zwei Sprecher und die Gegenstände *Fernrohr*, *Silber*, *Sessel* und *Filzhut* zeigt.

Abbildung 5.2.: Beispieldisplays für die Gegenstands- und die Sprecherbedingung

Wenn Sprecher 1 der Targetsprecher war, dann war das andere Bild von Sprecher 1 der Sprecherkompetitor. Wenn *Fernrohr* das Targetwort war, dann war entweder das andere

Bild von *Fernrohr* (Sprecherbedingung) oder *Filzhut* (Gegenstandsbedingung) der phonetische Kompetitor. Der andere Sprecher mit dem phonetisch unähnlichen Wort wird im folgenden als Distraktor bezeichnet. In der Abbildung ist das Target stets links-oben, der Sprecherkompetitor links-unten, der phonetische Kompetitor rechts-unten und der Distraktor rechts-oben.
Die Aufgabe der Versuchspersonen war es, sich die aufgenommenen Wörter anzuhören und den korrekten Gegenstand und Sprecher, der ihn genannt hat, zu identifizieren. In der Sprecherbedingung, in der die Hörer die Sprecher mit den gleichen Gegenständen kombiniert sahen, waren sie gezwungen den Sprecher zu identifizieren um die richtige Bildkombination zu finden. In der Gegenstandsbedingung hingegen, in der vier verschiedene Gegenstände abgebildet wurden, konnten sie die richtige Antwort nur aufgrund der phonetischen Information finden.

Durchführung

Das Experiment wurde mit der Software Experiment Builder [271] erstellt und bestand aus drei Teilen. Im ersten Teil wurden die Versuchspersonen mit den Bildern und den Wörtern für die sie stehen vertraut gemacht. Dabei sahen sie alle Bilder am Bildschirm mit ihrer Bezeichnung darunter. Alle 128 Wörter wurden in randomisierter Reihenfolge präsentiert und die Teilnehmer konnten selbst das Tempo bestimmen, in welchem sie sich durch die Wörter klickten.
Im zweiten Teil lernten die Versuchspersonen, die Sprecherstimmen mit den zugehörigen Bildern zu assoziieren. Zunächst wurde jedes Sprecherbild in randomisierter Reihenfolge einzeln präsentiert, jeweils mit den gleichen 8 Wörtern. Die Wörter wurden so ausgewählt, dass jeder Initialkonsonant einmal vorkam. Anschließend wurden den Hörern Bilder von beiden Sprechern am Bildschirm gezeigt und sie mussten mit der Maus auf den Sprecher klicken, der ihrer Meinung nach das gehörte Wort gesprochen hat. Dabei wurden die gleichen 8 Wörter verwendet wie zuvor. Nach jedem Trial wurde der Versuchsperson Feedback gegeben, ob sie den Sprecher richtig erkannt hat, indem das Bild des richtigen Sprechers nach dem Klicken kurz stehen blieb. Somit konnte jeder Teilnehmer überprüfen, ob er sich die Stimmen der beiden Sprecher merken und sie korrekt identifizieren konnte.
Im dritten und Hauptteil wurden die Versuchspersonen an ein Eyelink 1000 System (SR Research) angepasst, welches ihre Augenbewegungen verfolgte und überwachte. In jedem Trial wurde einer der oben beschriebenen Bildschirme präsentiert. 1200 ms später wurde ein Wort über Kopfhörer abgespielt (bei weniger als 200 ms zum Bilder anschauen treten

keine/wenige phonetische Effekte auf [123]). Die Aufgabe der Hörer war es mit der Maus auf die passende Sprecher-Gegenstand Bildkombination zu klicken. Jeder Teilnehmer bearbeitete 128 randomisierte Trials, wobei eine Hälfte von Sprecher 1 gesprochen wurde, die andere von Sprecher 2; eine Hälfte kam aus der Sprecherbedingung, die andere aus der Gegenstandsbedingung. Da jedes der 128 Wörter von zwei Sprechern aufgenommen wurde, gab es insgesamt 256 Wörter, die sich entweder in ihrer Phonetik oder ihren Sprechermerkmalen unterschieden. Jedem Hörer wurden 64 der 256 Wörter als Targetwörter präsentiert. Dabei wurden immer 8 Hörern verschiedene 64 Targetwörter präsentiert, sodass alle 256 Wörter einmal vorkamen. Auf diese Weise blieb das Target immer unvorhersagbar für die Teilnehmer, auch wenn Bilder wiederholt wurden, um Antworten für andere Gegenstände, Sprecher oder Bedingungen zu erhalten.
Die Targetpositionen auf dem Bildschirm wurden gleichmäßig verteilt. Dafür wurde ein Perlskript ([243]) verwendet, welches innerhalb eines Sets von 8 Trials dafür sorgte, dass das Target auf jeder Position (oben-links, unten-links, oben-rechts, unten-rechts) genau zweimal vorkam. Die Positionen der anderen 3 Bilder wurden in zufälliger Weise angeordnet, wobei nicht auf gleichhäufiges Vorkommen innerhalb eines Trialsets geachtet werden konnte. 1000 ms nach der Antwort der Versuchsperson startete der nächste Trial und nach jedem zehnten wurde eine Abweichkorrektur durchgeführt.

5.2.3. Statistik

Die verwendeten statistischen Analyseverfahren sollen in diesem Teil kurz vorgestellt und erklärt werden. Dabei wurde die Beschreibung bereits an die Anforderungen und Fragestellungen dieses Experiments angepasst.

Mixed Models

Ein Mixed Model ist ein statistisches Modell mit gemischten Effekten, das sowohl Fixed Effects/Factors als auch Random Effects/Factors enthält. Dadurch können in einem Modell mehrere Faktoren und deren Effekte analysiert werden. Ein Mixed Model sieht von seiner Grundstruktur her folgendermaßen aus:

$$y = X\beta + Zu + \epsilon \tag{5.1}$$

Wobei y ein Vektor aus Beobachtungen ist, β ein Vektor aus Fixed Effekts, u ein Vektor aus Random Effects, ϵ ein Vektor aus zufälligen Fehlertermen und X und Z Matrizen mit

Regressoren, welche die Beobachtungen y mit β und u verknüpfen.
Der Einfachheit halber werden im folgenden die Notationsweisen aus der Statistik-Software R ([241]) verwendet.
Allgemein gibt es zwei Ansätze, ein Mixed Modell zu erstellen: den Design-Driven und den Data-Driven Approach. Der Design-Driven Approach geht von der Struktur des Experiments aus und leitet daraus mögliche Varianzfaktoren ab. Der Data-Driven Approach betrachtet stärker die Daten an sich und überprüft, welche Faktoren überhaupt einen signifikanten Einfluss auf das Modell haben, d.h. signifikant viel Varianz erklären. Beide Ansätze sind legitim und finden in der Forschung Verwendung.

Faktoren, die wiederholbar und unabhängig sind, gelten als Fixed Factors; Faktoren, die zufällig sind und nicht wiederholbar als Random Factors. Bei Random Factors geht man davon aus, dass sie einen Mittelwert von 0 haben und mit einer bestimmten Varianz um diesen Punkt herum schwanken. Fügt man einen Random Effect Term in ein Modell ein, legt man diesen Mittelwert auf das Intercept, das durch die Zahl 1 repräsentiert wird. Um beispielsweise Versuchspersonen mit unterschiedlichen Reaktionsgeschwindigkeiten auszugleichen, schreibt man *(1|Versuchsperson)*. Dadurch wird für Versuchspersonen mit schnellen Reaktionszeiten das Intercept etwas abgesenkt und für Versuchspersonen mit langsamen Reaktionszeiten angehoben. Somit wird der Einfluss von individuellen Unterschieden zwischen Versuchspersonen im Mixed Model ausgeglichen. Gleiches kann für andere Random Factors wie *Wort* oder *Trial* durchgeführt werden.
Jedoch können nicht nur die Reaktionszeiten der Versuchspersonen individuell verschieden sein. Es können auch Unterschiede im Verhalten der Versuchspersonen auftreten wie z.B. ein Lern- oder ein Ermüdungseffekt während des Experiments. Solche Differenzen kann man mit einem Random Slope von *Trial* über *Versuchsperson* ausgleichen, was dann als *(1+Trial|Versuchsperson)* notiert wird. Dabei beschreibt *Trial* die Größe, deren Verlauf unterschiedlich sein kann und *Versuchsperson* die Größe innerhalb derer der Verlauf variiert. Die 1 steht in diesem Fall für den Korrelationsparameter, welcher eine mögliche Korrelation zwischen dem Intercept und dem Slope von *Trial* beschreibt. Ist keine Korrelation vorhanden oder möchte man diese nicht mitberücksichtigen, schreibt man statt dessen *(0+Trial|Versuchsperson)*.
Der Parameter *Residual* in der Modellzusammenfassung beschreibt den zufälligen Fehler in der Verteilung. Jede Varianz, die nicht durch einen der definierten Random Factors erklärt werden kann, zählt zu diesem Faktor.

Fixed Factors hingegen sind die Faktoren, die im Experiment manipuliert werden. Sie variieren nicht zufällig, sondern ihr Einfluss sollte immer stabil und gleichgroß sein. Ein solcher Faktor könnte in einem Experiment beispielsweise die Unterscheidung zwischen verschiedenen Bedingungen sein. Diese Faktoren werden einfach mit *+ Bedingung* in dem Modell hinzugefügt [18] [24].

Ein Modell, das alle relevanten Faktoren enthält, wird als maximales Modell bezeichnet und gilt als optimal. Allerdings haben Mixed Models die Tendenz, mit größer werdender Komplexität, d.h. mit einer größeren Anzahl von Faktoren nicht mehr zu konvergieren (besonders in der aktuellen Version von R 3.2.0). Das bedeutet, dass in einer bestimmten Anzahl von Iterationsschritten keine Anpassung der Parameter an die Daten durchgeführt werden kann. In diesem Fall ist es notwendig, das Modell zu vereinfachen, indem man die Anzahl an Random Factors reduziert. Dabei sollte man jedoch mit Bedacht vorgehen, damit das Modell weiterhin aussagekräftig bleibt. Laut dem Design-Driven Approach wäre es besser, (zuerst) die Faktoren zu eliminieren, die den Einfluss des zu untersuchenden Fixed Factors erhöhen und die beizubehalten, die seinen Einfluss reduzieren. Dies bedeutet, dass man zuerst mögliche Korrelationen zwischen zwei Random Factors ignorieren sollte, indem man z.B. *(0+Trial/Versuchsperson)* statt *(1+Trial/Versuchsperson)* schreibt. Wenn dies noch nicht ausreicht, kann man anschließend die Random Intercepts für *Versuchsperson* ect. entfernen. Die Random Slopes hingegen wie z.B. *(0+Trial/Versuchsperson)* sollte man im Modell belassen. Dadurch wird die Gefahr eines Typ 1 Fehlers, nämlich einen signifikanten Effekt zu finden, wo keiner ist, verringert. Daher gelangt man auf diesem Weg zu einer eher konservativen Schätzung, was im Zweifelsfall der sicherere Weg ist [24].
Der andere Weg ist der Data-Driven Approach, nach dem es eher empfehlenswert wäre, die Random Factors zu entfernen, die bei einem Modellvergleich keinen signifikanten Unterschied erzeugen. Somit werden zuerst die Faktoren eliminiert, die wenig oder keine Varianz erklären, während die Faktoren, die signifikant viel Varianz erklären, im Modell verbleiben. Da in der aktuellen R-version (3.2.0) Random Slopes in einem Modell häufig dazu führen, dass es nicht mehr konvergiert, empfiehlt sich eher der letztere Ansatz.
Die aktuelle Version 1.1.7 vom lme4-Paket ([26]) gibt bei der Analyse von Mixed Models leider keine p-values aus, weil unklar ist, welches die genaue Anzahl an Freiheitsgraden ist. Für eine normale Analyse reicht aber auch der t-Wert aus. Möchte man für das Signifikanzniveau von 5 % testen, dann ist ein Ergebnis signifikant, wenn der t-Wert für den

Faktor über 2 oder unter -2 liegt [18]. Alternativ kann aber auch das Paket „lmerTest“ [161] verwendet werden, welches selbstständig p-Werte berechnet.

Jackknife-Methode

Die von [240] und [287] entwickelte Jackknife-Methode wird in der Statistik zum Resampling einer Stichprobe verwendet, um die Varianz und den Bias zu schätzen. Die häufigste Variante ist das delete-1-Jackknife, wobei immer für $n-1$ Werte der Stichprobe eine Schätzung (der Mittelwert) berechnet wird, aus denen sich dann die durchschnittliche Gesamtschätzung ergibt.

Dabei berechnet sich der Gesamtschätzwert $\bar{\theta}_{Jack}$ wie folgt:

$$\bar{\theta}_{Jack} = \frac{1}{n}\sum_{i=1}^{n}(\bar{\theta}_i) \tag{5.2}$$

Wobei $\bar{\theta}_i$ der Schätzwert der i-ten Beobachtung ist, n die Anzahl der Messwerte und i der i-te Messwert.

Die Varianz der Stichprobe berechnet sich daher wie folgt:

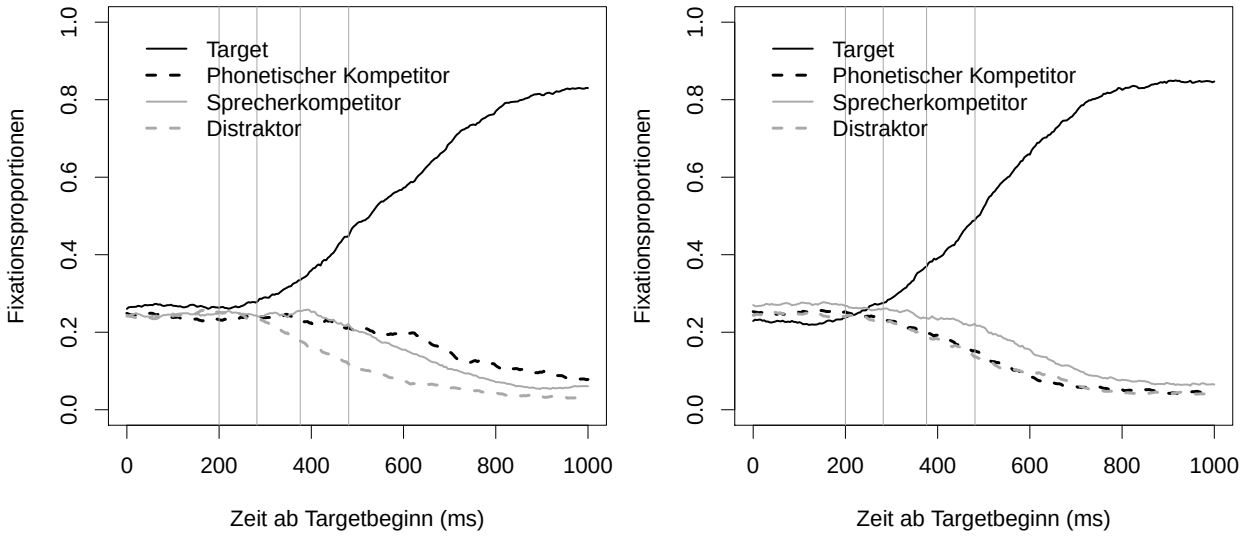
$$Var(\theta) = Var((\frac{\sum_{i=1}^{n}(X_i)}{n})) = \frac{\sigma^2}{n} = \frac{n-1}{n}\sum_{i=1}^{n}(\bar{\theta}_i - \bar{\theta}_{Jack})^2 \tag{5.3}$$

Dabei ist $\bar{\theta}_i$ der Schätzwert, der sich aus dem Weglassen der i-ten Beobachtung ergibt und $\bar{\theta}_{Jack}$ der Gesamtschätzwert für alle Messwerte. σ^2 steht für die Varianz, n für die Anzahl der Messwerte und X_i für den Mittelwert der i-ten Schätzung.

Angenommen, die Mittelwerte für alle Versuchspersonen sind ähnlich, dann ergeben sich zwischen den einzelnen Teilschätzungen nur geringe Unterschiede und somit eine geringe Varianz. Bei größeren Unterschieden würden sich diese in einer höheren Varianz bemerkbar machen. Auf diese Weise können auch für Daten, die (z.B. durch vorherige Durchschnittsbildung) keine „natürliche“ Varianz mehr besitzen, Größen wie die Varianz oder die Standardabweichung berechnet werden ([240] [287] [70] [198] [197]).

5.2.4. Ergebnisse

Die Ausgabedateien des Eye-Trackers wurden mit Hilfe eines angepassten Perl-Skripts [244] in eine Datentabelle mit Informationen (Versuchspersonenennummer, Targetbild, ect.) und eine Tabelle mit den zeitlich synchronisierten und kategorisierten Fixationspositionen umgewandelt. Beide Datentabellen wurden dann zur Auswertung in R Version 3.2.0 [241] eingelesen. Zur statistischen Analyse wurden die Pakete „lme4“ [26] und „lmerTest“ [161] verwendet. Die Ergebnisse des Experiments wurden in einer Datentabelle gespeichert und zur Analyse in R Version 3.2.0 [241] importiert.
Alle Teilnehmer wählten in durchschnittlich 98,6 % der Trials das richtige Bild aus. Für die nachfolgende Analyse wurden nur diese richtigen Trials verwendet. Abbildung 5.3 zeigt die Fixationsproportionen über die Zeit der vier Sprecher-Gegenstand Kombinationen. Schwarze volle Linien markieren das Target, schwarze gestrichelte Linien den phonetischen Kompetitor (der andere Sprecher kombiniert mit dem phonetisch ähnlichen bzw. identischen Gegenstand); graue volle Linien markieren den Sprecherkompetitor (der phonetisch unähnliche Gegenstand gepaart mit dem gleichen Sprecher) und graue gestrichelte Linien den Distraktor (anderer Sprecher kombiniert mit phonetisch unähnlichem Gegenstand).



(a) Sprecherbedingung

(b) Gegenstandsbedingung

Abbildung 5.3.: Fixationsproportionen in Abhängigkeit von der Zeit

Die vertikalen Linien repräsentieren den Wort-/Konsonantenbeginn, Konsonantenende, Vokalende und Wortende. Dabei sind alle Werte um 200 ms verschoben. Es wird angenom-

men, dass zwischen dem akustischen Signal und der Augenbewegung aufgrund dieses Signals 200 ms verstreichen [4]. Die Zeiten des Initialkonsonanten und Vokals wurden normiert, um einen besseren Vergleich zwischen verschiedenen Konsonanten zu ermöglichen.
Es wurden zwei separate Zeitfenster analysiert: T1 von 200 ms bis 500 ms (siehe Tabelle 5.2) und T2 von 500 ms bis 900 ms (siehe Tabelle 5.3). T1 wurde gewählt, um die Fixationen während der Wortverarbeitung widerzuspiegeln (durchschnittliche Wortdauer: 480 ms). T2 reichte vom Wortende bis zu dem Punkt, wo die Targetfixationen aufhörten zu steigen und die Kompetitorfixationen auf ein Minimum fielen.
Proportionale Daten sind multiplikativ (z.B. Fixation von Objekt A ist zweimal wahrscheinlicher als Fixation von Objekt B) und bringen das Problem mit sich, dass Effekte nahe 0 und nahe 100 % stärker gewichtet sind als im mittleren Bereich. Die meisten statistischen Analyseverfahren gehen jedoch von linearen Daten aus, wodurch es bei der statistischen Analyse zu verfälschten Ergebnissen kommen kann. Daher müssen die multiplikativen Fixationsproportionen erst in eine additive Skala umgerechnet werden. Dies wird durch eine logistische Transformation gewährleistet (siehe [23] [128]).
Anschließend wurden für diese logistisch transformierten Werte lineare Mixed-Effect Models berechnet. Die abhängige Variable war die Fixationspräferenz zwischen den verschiedenen Kompetitortypen (z.B. phonetischer Kompetitor - Distraktor). Als Fixed Factor enthielten die Modelle nur einen Intercept Term. Wenn sich dieser signifikant von Null unterschied ($t > 2$ oder $t < -2$), wurde einer der Kompetitoren stärker fixiert als der andere. Das Regressionsgewicht (b) gibt dabei die Richtung der Differenz an. Ist b positiv, so steigt die abhängige Variable mit einem Anstieg der unabhängigen Variable; ist b negativ, so sinkt die abhängige Variable mit steigender unabhängiger Variable. Der Betrag von b sagt aus, um wie viel die abhängige Variable steigt oder sinkt bei Veränderung der unabhängigen Variable um eine Einheit [18]. Als Random Factor wurde der Faktor *Versuchsperson* einbezogen. Ein zusätzlicher Random Slope über Versuchselemente zeigte bei einem Modellvergleich keine signifikante Änderung der Ergebnisse und wurde daher nicht ins Modell einbezogen.

Während T1 wurde das Target in beiden Bedingungen sehr früh gegenüber den Kompetitoren bevorzugt. In der Sprecherbedingung wurden sowohl der Sprecher- als auch der phonetische Kompetitor stärker fixiert als der Distraktor, unterschieden sich untereinander aber nicht. Das bedeutet, dass die Versuchspersonen zeitweise die anderen Referenten, die wenigstens in Sprecher- oder phonetischer Information übereinstimmten, beachtet haben. In der Gegenstandsbedingung wurde der Sprecherkompetitor stärker betrachtet als der

phonetische und der Distraktor, welche sich nicht unterschieden. Das heißt, die Hörer beachteten den Gegenstand, der mit dem Targetsprecher kombiniert war, trotz des phonetischen Mismatchs am Wortanfang.

	Sprecherbedingung			**Gegenstandsbedingung**		
Vergleich	b	t	p	b	t	p
phonetischer Komp. - Distraktor	0,31	2,90 *	0,004	0,06	0,60	0,55
Sprecherkomp. - Distraktor	0,35	3,13 *	0,005	0,33	2,87 *	0,009
Sprecherkomp. - phonetischer Komp.	0,03	0,29	0,78	0,27	2,32 *	0,03

Tabelle 5.2.: Ergebnisse der Sprecher- und der Gegenstandsbedingung in T1 (200 - 500 ms): Fixationspräferenzen zwischen den verschiedenen Kompetitoren (* markiert signifikante Werte)

	Sprecherbedingung			**Gegenstandsbedingung**		
Vergleich	b	t	p	b	t	p
phonetischer Komp. - Distraktor	0,76	5,96 *	$< 0{,}001$	0,001	0,02	0,99
Sprecherkomp. - Distraktor	0,46	7,06 *	$< 0{,}001$	0,36	4,81 *	$< 0{,}001$
Sprecherkomp. - phonetischer Komp.	-0,30	-2,17 *	0,04	0,36	5,13 *	$< 0{,}001$

Tabelle 5.3.: Ergebnisse der Sprecher- und der Gegenstandsbedingung in T2 (500 - 900 ms): Fixationspräferenzen zwischen den verschiedenen Kompetitoren (* markiert signifikante Werte)

In T2 waren die Ergebnisse ähnlich, mit zwei Ausnahmen: Erstens, die Targetfixationen waren in der Sprecherbedingung niedriger als in der Gegenstandsbedingung ($b = -0{,}51$; $t = -5{,}12$; $p < 0{,}001$); diese Differenz war in T1 nicht signifikant. Zweitens, in der Sprecherbedingung wurde der phonetische Kompetitor signifikant häufiger betrachtet als der Sprecherkompetitor. Scheinbar waren sich die Hörer früher sicher, welches Wort gesprochen wurde als darüber, wer es gesagt hatte. Die Ergebnisse lassen vermuten, dass die Versuchspersonen selbst nach Wortende noch einmal überprüft haben, ob das phonetisch ähnliche Wort (gepaart mit dem anderen Sprecher) das Target hätte sein können.
Zusätzliche Analysen zum Einfluss von Artikulationsstelle und -modus des Initialkonsonanten des Targets zeigten keine signifikanten Ergebnisse. Somit wurden keine Unterschiede zwischen den verschiedenen Anfangskonsonanten gefunden.

Vergleich der Sprecher- und der Gegenstandsbedingung

In der Sprecherbedingung waren die zwei Gegenstandswörter und somit der phonetische Kompetitor identisch. In der Gegenstandsbedingung hingegen überlappten die beiden Wörter nur im ersten Konsonanten, sodass nur dieser kurze gemeinsame Anfang eine phonetische Kompetition auslösen könnte. Das bedeutet, dass der phonetische Overlap zwischen den phonetisch ähnlichen Wörtern unterschiedlich groß war. In diesem Vergleich sollte untersucht werden, wie sich dieser Faktor des phonetischen Overlaps auf die Stärke der Kompetition auswirkte.

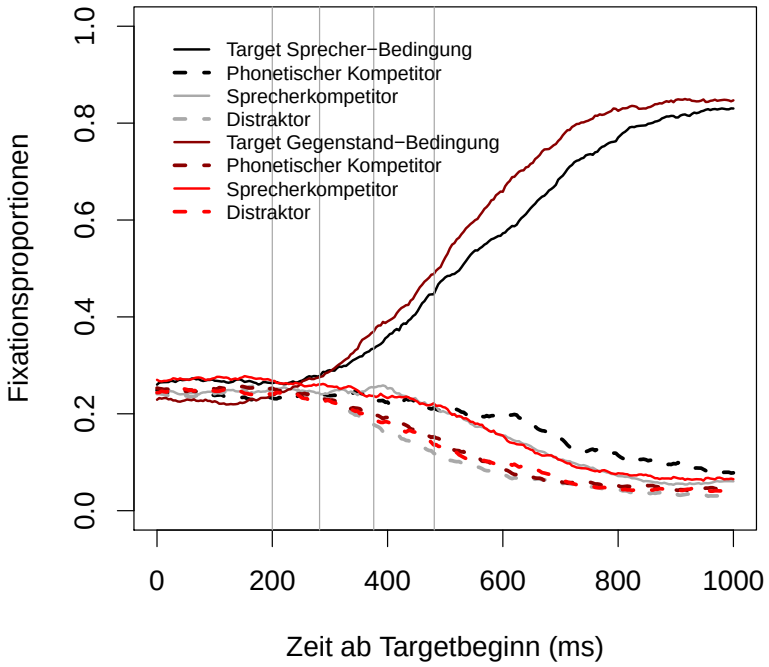


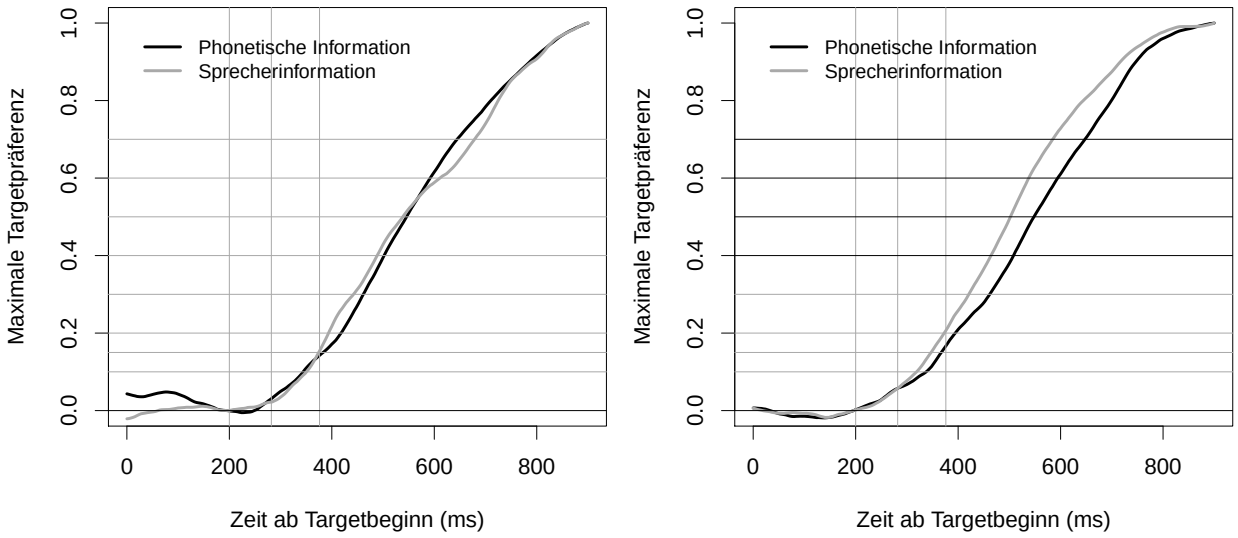
Abbildung 5.4.: Sprecher- und Gegenstandsbedingung: Fixationsproportionen in Abhängigkeit von der Zeit

Die beiden Sprecherkompetitoren aus den beiden Bedingungen unterschieden sich nicht voneinander (siehe Abbildung 5.4). Dies bestätigte auch ein *t*-Test über die Differenzen von Sprecherkompetitor und Distraktor in beiden Bedingungen (T1: $t = 0{,}12$; $p = 0{,}90$; T2: $t = 1{,}06$; $p = 0{,}29$). Die Differenzen zwischen phonetischem Kompetitor und Distraktor unterschieden sich zwischen den Bedingungen hingegen stark, sodass ihr Unterschied sowohl in T1 ($t = 3{,}24$; $p = 0{,}002$) als auch in T2 ($t = 5{,}71$; $p < 0{,}001$) signifikant wurde. Dies zeigt, dass der phonetische Overlap einen signifikanten Einfluss auf die phonetische Kompetition

hat, aber keinen auf die Sprecherkompetition.

Jackknife-Analyse

Um die zeitliche Verarbeitung von phonetischer und Sprecherinformation genauer zu untersuchen, wurde eine Maximaleffektanalyse mittels der Jackknife-Methode ([197] [189]) durchgeführt. Für dieses Verfahren wurden beide phonetische Kompetitoren (Targetwort und phonetisch ähnliches Wort) als auch beide Sprecherkompetitoren (beide Bilder des gleichen Sprechers) gemittelt. Dadurch erhält man einen Graphen für den Faktor Sprecher und einen für den Faktor Phonetik/Wort. Anschließend wurden Zeitpunkte berechnet, an denen die Fixation der Hörer auf Sprecher oder Wort einen bestimmten Prozentwert des Maximums erreichte. Die Prozentpunkte lagen bei 10, 15, 20, 30, 40, 50, 60 und 70 %, um einen möglichst großen Bereich abzudecken. Durch diese Analyse wird nicht vorrangig die Größe des Kompetitionseffekts betrachtet, sondern vor allem dessen Anstiegsgeschwindigkeit. Steigt ein Effekt schneller als der andere, lässt sich vermuten, dass die Informationen, die zu diesem Effekt führen, von den Hörern eher wahrgenommen und verarbeitet wurden. Die Zeitpunkte, zu denen die Prozentwerte des Maximumeffekts für Sprecher und Phonetik/Wort erreicht waren, wurden anschließend mit *t*-tests verglichen.



(a) Sprecherbedingung

(b) Gegenstandsbedingung

Abbildung 5.5.: Proportionen der maximalen Targetpräferenz in Abhängigkeit von der Zeit zwischen 200 und 900 ms.

Da die Fixationsdaten von einzelnen Versuchspersonen meist nicht zuverlässig sind, wurde für diese zeitliche Analyse die delete-1-Jackknife Methode verwendet (siehe Abschnitt 5.2.3). Da bei diesem Verfahren jede Versuchsperson $n-1$ mal zur Varianz beiträgt, musste der t-Wert entsprechend angepasst werden. Dies geschah auf folgende Weise:

$$t_{corrected} = \frac{t}{n-1} \tag{5.4}$$

In der Abbildung 5.5 sind die Ergebnisse dargestellt. In der Sprecherbedingung wurde zu keinem Prozentwert des Maximums eine signifikante zeitliche Differenz zwischen phonetischer und Sprecherinformation gefunden. Es ist zu beachten, dass in dieser Bedingung alle Wörter mit dem selben Konsonanten begannen, sodass die Sprecherinformation früher hätte genutzt werden können als die phonetische Information zur Identifikation des Targets. In der Gegenstandsbedingung ging der Effekt der Sprecherinformation dem der phonetischen Information voraus. An den Punkten 40 % (t = -2,40), 50 % (t = -2,56), 60 % (t = -3,01) und 70 % (t = -2,97) des Maximums wurde diese Differenz signifikant. Während im niedrigen Prozentbereich kein Unterschied gefunden werden konnte, wurde später das Maximum des Sprechereffekts eher erreicht als das Maximum des Phonetikeffekts.

5.2.5. Diskussion

Das Experiment widmete sich der Fragestellung, ob das Visual-World Eye-Tracking Paradigma, das für die zeitliche Verarbeitung von feinen phonetischen Details in der zeitsynchronen Worterkennung verwendet wurde, auch Einblicke in die relative zeitliche Koordinierung der Sprechererkennung liefen kann. In einer kombinierten Sprecher-/Worterkennungsaufgabe identifizierten die Teilnehmer die visuellen Referenten mit einer hohen Genauigkeit. Das Target wurde schnell erkannt, während sich die Kompetitoren hauptsächlich in ihrer Absinkgeschwindigkeit unterschieden als in ihrem Anstieg. Wie in Abbildung 5.4 zu sehen, stieg das Target in der Gegenstandsbedingung schneller als in der Sprecherbedingung. Das lässt vermuten, dass die Sprecherbedingung, in der die Hörer Sprecherinformationen verwenden mussten, um den richtigen Referenten zu finden, schwieriger war, als die Gegenstandsbedingung, in der das Target auch ohne Sprecheridentifikation gefunden werden konnte.
In der Gegenstandsbedingung waren die Sprecherinformationen nur optional zu nutzen, da die phonetischen Informationen (Wörter) allein zur Targetidentifikation ausreichten. Als alleinige Entscheidunsgrundlage eigneten sich die Sprecherinformationen nicht, da immer zwei gleiche Sprecher am Bildschirm angezeigt wurden. Dennoch wurde in dieser Bedingung

nur Sprecherkompetition und keine phonetische Kompetition gefunden. Dies zeigt, dass trotzt ihres Nachteils gegenüber den phonetischen Informationen die Sprecherinformationen (auch) genutzt wurden. Hätten die Hörer die Sprecherinformationen ignoriert, hätte es keine Sprecherkompetition geben dürfen.
In der Sprecherbedingung wurde sowohl phonetische als auch Sprecherkompetition gemessen, wobei beide am Anfang gleich stark waren und dann die Sprecherkompetition unter das Niveau der phonetischen Kompetition sank. Die Ergebnisse dieser Bedingung lassen zwei verschiedene Interpretationen zu: (1) Die stärkere phonetische Kompetition (im zweiten Zeitintervall) deutet darauf hin, dass die Hörer die Gegenstandswörter schwerer unterscheiden konnten als die Sprecher. Dies würde bedeuten, dass die Sprechermerkmale besser erkannt und verwendet werden konnten als die phonetischen Merkmale. (2) Andererseits könnte die starke phonetische Kompetition auch bedeuten, dass die Hörer sofort erkannten, welches das richtige Gegenstandswort war und anschließend abwechselnd auf die beiden verschiedenen Sprecher (die mit den identischen Gegenständen kombiniert waren) achteten. Zum Beispiel erkannten sie sofort, dass es zweimal den Gegenstand „Fernrohr“ gab und schauten dann zwischen den beiden „Fernrohr“-Bildern hin und her, um zu entscheiden, welches der gehörte Sprecher war. In diesem Fall könnte die Fixation auf die phonetisch identischen Gegenstandswörter nicht als phonetische Kompetition interpretiert werden, sondern müsste als Sprecherkompetition gewertet werden. Andersherum müsste dann auch die Sprecherkompetition eher als phonetische Kompetition gesehen werden. Denn dort betrachteten die Hörer die Bildkombinationen mit dem identischen Sprecher, aber zwei verschiedenen Gegenständen. Folglich erkannten sie den richtigen Sprecher recht schnell, sodass die Kompetition nach Wortende wieder absank. Allerdings sank die Sprecherkompetition auch nach Wortende nicht auf das Niveau des Distraktors ab. Daher ist anzunehmen, dass sich die Hörer nicht hundertprozentig sicher waren, ob sie das Wort richtig erkannt hatten und ihre Entscheidung bis zum Trialende immer wieder überprüften. In diesem Fall würden die Ergebnisse für eine schnelle Erkennung des Wortes und eine etwas langsamere Erkennung des richtigen Sprechers sprechen.

Während die Ergebnisse der Sprecherbedingung, verschiedene Interpretationen zulassen, bleibt für die Gegenstandsbedingung die Frage: Weshalb benutzten die Hörer die nur optionalen und unzureichenden Sprecherinformationen offenbar stärker als die phonetischen Informationen? Ein Grund für Sprecher- statt phonetischer Kompetition könnte sein, dass die phonetische Überlappung zu kurz war, um phonetische Kompetition auszulösen (nur

der Anfangskonsonant überlappte). Vorangegangene Studien haben gezeigt, dass sogar sub-phonemische Mismatches die phonetische Kompetition beeinflussen können [59] und hier könnten die nicht übereinstimmenden Vokale zwischen den phonetischen Kompetitoren diesen Mismatch bewirkt haben. Allerdings erklärt das immer noch nicht, warum der Kompetitor, der mit einem anderen Konsonanten beginnt, Sprecherkompetition zugelassen hat (der Targetsprecher wurde auch mit einem Gegenstand mit anderem Anfangskonsonant als das Target kombiniert).
Außerdem ist es eigenartig, dass obwohl die Wörter anscheinend so einfach zu unterscheiden waren, dass keine phonetische Kompetition auftrat, Sprecherinformationen zur Targetidentifizierung verwendet wurden. Diese waren, wie bereits erwähnt, nur optional zu nutzen und waren darüber hinaus nicht ausreichend für die Targetidentifizierung. Dennoch fokussierten sich die Versuchspersonen auf die Sprecher und deren Merkmale, was die höhere Sprecherkompetition belegt.
Betrachtet man die zeitliche Koordinierung von Sprecher- und Worteffekten, scheint die Sprechererkennung tatsächlich etwas schneller zu sein als die Worterkennung. Das heißt, die Hörer scheinen zuerst entschieden zu haben, wer der Sprecher ist und erst danach das Wort, das gesprochen wurde. In weiterführenden Experimenten müsste untersucht werden, ob die zeitliche Reihenfolge strategisch sein könnte, da es (in der Gegenstandsbedingung) weniger Sprecher am Bildschirm gab als Gegenstände. Laut einer Studie von [224] ist dies allerdings nicht allzu wahrscheinlich, da Hörer unabhängig von der Anzahl der Sprecher deren stimmliche Merkmale zur Targetidentifizierung verwendeten. Da dies die erste Untersuchung zu Sprecherkompetition in einem Visual-World Paradigma war, wurde die Unterscheidbarkeit der Sprecher maximiert: Die Verwendung von Sprecherinformationen in der Sprecherbedingung war entscheidend und die Sprecherbilder waren etwas größer als die Gegenstandsbilder. Außerdem wurden während des gesamten Experiments die gleichen zwei Sprecher gezeigt, wohingegen die Gegenstände zwischen den Durchgängen variierten. Da bisher wenig über den zeitlichen Ablauf der Sprecheridentifikation und ihr Verhältnis zur Wortidentifikation bekannt war, diente dieses Experiment vor allem der Exploration dieses Prozesses. In einem Nachfolgeexperiment soll untersucht werden, wie stabil der Sprechereffekt ist, wenn der Faktor Sprecher weniger prominent ist und ob durch eine stärkere phonetische Überlappung am Wortanfang die phonetische Kompetition steigt. Außerdem soll das Problem der Interpretierbarkeit der Kompetitionseffekte in der Sprecherbedingung durch ein neues Design erhöht werden. Durch die Vermeidung identischer Sprecher und Gegenstandswörter sollten sich die Sprecher- und Worteffekte eindeutiger zuordnen lassen.

5.3. Experiment 2

5.3.1. Hypothesen und Ziele

In dem vorangegangenen Experiment sprachen die Ergebnisse dafür, dass Sprecherinformation etwas schneller aufgenommen und verarbeitetet wurde als phonetische. Allerdings könnte die schnellere Entscheidung für den Sprecher auf einer strategischen Überlegung der Versuchspersonen beruht haben. In der Gegenstandsbedingung sahen die Versuchspersonen stets vier Bilder, die zweimal den identischen Sprecher enthielten, aber vier verschiedene Objekte. Daher bestand eine größere Ähnlichkeit zwischen den beiden Bildern eines Sprechers als zwischen den Objekten mit dem selben Anfangskonsonanten. Eventuell haben die Versuchspersonen diese Tatsache erkannt und eine Entscheidungsstrategie entwickelt, bei der sie zuerst immer den Sprecher identifiziert haben, da dieser sich (visuell) leichter erkennen ließ. Die phonetisch ähnlichen Objekte waren sich visuell hingegen unähnlich, sodass ihre Gemeinsamkeit erst durch die sprachliche Repräsentation zum Tragen kam. Da diese Gemeinsamkeit nicht so unmittelbar zum Tragen kommt, wie die visuelle Gleichheit der Sprecherbilder, wurden die phonetischen Kompetitoren erst später erkannt.
Um den Einfluss des Sprechers im Vergleich zum ersten Experiment und zu anderen vorangegangenen Experimenten zu verringern und somit die Sprecheridentifikation zu erschweren, wurden verschiedene Maßnahmen getroffen. (1) Zunächst wurde die Anzahl der Sprecher von zwei auf vier erhöht, indem zwei weibliche Sprecher hinzugefügt wurden. Da es für die Hörer schwieriger sein sollte, sich vier Stimmen mit den entsprechenden Bildreferenten zu merken und wiederzuerkennen, sollte die Sprecheridentifikation erschwert werden. (2) Während im ersten Experiment am Bildschirm zwei identische Sprecherbilder mit unterschiedlichen Gegenständen präsentiert wurden, waren die Sprecherbilder diesmal nicht identisch, sondern stimmten nur im Geschlecht überein. Der Sprecherkompetitor hatte somit zwar das gleiche Geschlecht wie der Target-Sprecher, war aber nicht mehr (visuell) identisch. Daher sollte die Sprecheridentifikation diesmal schwieriger sein. (3) Außerdem reichte in diesem Experiment ein Informationstyp (entweder phonetische oder Sprecherinformationen) zur Identifikation des Targets aus. Da zwar die Wörter den Hörern bereits vor dem Experiment bekannt waren, die Sprecher jedoch nicht, wurde vermutet, dass die Targetidentifikation anhand phonetischer Informationen einfacher als anhand von Sprecherinformationen wäre. (4) Alle Wörter waren echte Wörter, welche ebenfalls vor dem Experiment noch einmal präsentiert wurden. Daher sollte die Identifikation der Wörter einfacher gewesen sein als die Identifikation der Sprecher. [48] zeigte, dass der Einfluss von Sprecherinformationen

schwächer war, wenn „echte“ Wörter statt Unsinnwörtern verwendet wurden. (5) Um dem Kritikpunkt der unterschiedlichen Größe von Sprecher- und Gegenstandsbild zu begegnen, wurden die Sprecherbilder auf die gleiche Größe wie die Gegenstandsbilder gebracht.
Neben der zeitlichen Koordination von Sprecher- und Worterkennung sollte in diesem Experiment der Einfluss der phonetischen Überlappung am Wortanfang untersucht werden. Deshalb wurden die Wörter so kombiniert, dass die phonetische Überlappung in manchen Fällen die erste Konsonant-Vokal-Sequenz (CV) umfasste und manchmal nur (wie in Experiment 1) den ersten Konsonanten (C). Im Falle der größeren phonetischen Überlappung wurde auch eine stärkere phonetische Kompetition zwischen den Objekten erwartet. Außerdem sollte die Erstellung von (möglichst vielen) CV-Overlap Wörtern auch dafür sorgen, dass die phonetische Kompetition insgesamt stärker wird.
Ein weiteres Anliegen dieses Experiments war es, Gender-Kompetition nachzuweisen. Das bedeutet, dass bei einem männlichen Targetsprecher die Hörer auch stärker auf den anderen männlichen Sprecher schauen würden als auf die weiblichen Sprecher. Da sich die Stimmen von Sprechern innerhalb eines Geschlechts ähnlicher sind (ähnlichere Grundfrequenz, Formantwerte, ect.) wurde davon ausgegangen, dass Gender-Kompetition auftreten würde. Bisher wurde sie aber noch nicht empirisch belegt. Da die meisten Untersuchungen bisher nur immer einen männlichen und einen weiblichen Sprecher verwendeten (z.B. [48] [50] [56] [157]), ließ sich die Sprecherkompetition nicht von der Gender-Kompetition trennen. Dies war auch ein Grund dafür, weshalb in dieser Studie zwei Sprecher von beiden Geschlechtern untersucht wurden.
In vielen Studien bestand zwischen den Wörtern eine stärkere phonetische Überlappung als in diesem Experiment (z.B. [48]). Außerdem wurden zumeist auch nur zwei Sprecher verschiedenen Geschlechts verwendet. Dadurch war eine Unterscheidung des Sprechers sehr einfach, wohingegen die Unterscheidung der Wörter durch ihre größere phonetische Ähnlichkeit schwieriger war. In diesem Experiment wurden mehr Sprecher beider Geschlechter verwendet, sodass die Sprecheridentifikation schwieriger war. Durch die kürzere phonetische Überlappung in den Wörtern wurde die Erkennung der Wörter erleichtert. Außerdem wurden nur echte Wörter verwendet, die allen Sprechern vorher bekannt waren, wohingegen in anderen Experimenten Unsinnwörter verwendet wurden, die zuvor erst erlernt werden mussten ([48] [50]). Der einzige Faktor, der die Worterkennung im Vergleich zur Sprechererkennung erschwerte, war die größere Anzahl an Wörtern im Vergleich zur Anzahl der Sprecher. Durch diese Faktoren müsste eine Sprechererkennung relativ schwieriger gewesen sein als eine Worterkennung. Wird dennoch eine frühere Sprecheridentifikation als Wortidentifikation

nachgewiesen, spräche dies stark für eine schnellere Wahrnehmung und Verarbeitung von Sprecherinformationen gegenüber phonetischen Informationen.

5.3.2. Methode

Versuchspersonen

45 Studenten der Ludwig-Maximilians-Universität München nahmen gegen ein kleines Entgelt an der Studie teil. Alle Teilnehmer waren deutsche Muttersprachler und zwischen 19 und 38 Jahren alt. Bis auf eine Teilnehmerin gab niemand eine Sprach- oder Hörstörung an. Eine Teilnehmerin berichtete von einem Tinitus, der aber aufgrund ihrer Ergebnisse die Hörfähigkeiten nicht eingeschränkt zu haben schien. Es wurden nur die Daten von Teilnehmern verwendet, die in dem Feedbackteil des Experiments die Sprecher zu mindestens 68,75 % korrekt identifizieren konnten. Daher wurden nur die Daten von 32 Hörern in die Analyse einbezogen. Keiner der Teilnehmer hatte zuvor an dem ersten Visual-World Eye-Tracking Experiment teilgenommen.

Sprachmaterial

Für dieses Experiment wurden die gleichen 128 Gegenstandsbilder verwendet, wie für das erste Experiment. Bei den Sprecherbildern wurden zusätzlich zu den Bildern der zwei männlichen Sprecher auch zwei Bilder von zwei jungen Frauen ausgewählt, welche die weiblichen Sprecher repräsentieren. Die Auswahlkriterien waren die gleichen wie in Experiment 1. In diesem Experiment hatten die Sprecher- und die Gegenstandsbilder die gleiche Größe von 200 x 200 Pixeln.

Akustische Merkmale der Sprecher

Für das zweite Experiment wurden insgesamt 4 Sprecher ausgewählt, wovon zwei männlich und zwei weiblich waren. Dadurch sollte der Einfluss des Faktors Sprecher verringert werden. Um zu überprüfen, ob und wie stark sich die Sprecher in ihrer Grundfrequenz und den mittleren Formantfrequenzen unterschieden, wurde wieder eine akustische Analyse mit Praat ([35]) durchgeführt. Weder die männlichen, noch die weiblichen Sprecher unterschieden sich signifikant in ihren akustischen Merkmalen (siehe Tabelle 5.4). Dennoch klangen die weiblichen Sprecher auditiv ähnlicher als die männlichen. Außerdem existieren außer den Grundfrequenzmerkmalen und den mittleren Formantfrequenzen noch andere Parameter, durch die sich die Stimmen der Sprecher unterscheiden können. Aufgrund der perzeptiv

höheren Ähnlichkeit der weiblichen Stimmen wäre anzunehmen, dass sie schwieriger zu unterscheiden sein werden als die männlichen Stimmen.

Sprecher	F0	F0-SD	min. F0	max. F0	F0-Bereich	F1	F2	F3	F4
m1	134 Hz	30 Hz	111 Hz	204 Hz	93 Hz	532 Hz	1501 Hz	2693 Hz	3890 Hz
m2	115 Hz	40 Hz	85 Hz	222 Hz	137 Hz	600 Hz	1656 Hz	2628 Hz	3494 Hz
f1	218 Hz	38 Hz	183 Hz	315 Hz	132 Hz	682 Hz	1920 Hz	3045 Hz	4396 Hz
f2	195 Hz	51 Hz	139 Hz	300 Hz	161 Hz	721 Hz	1935 Hz	3029 Hz	4064 Hz

Tabelle 5.4.: Mittlere akustische Merkmale der männlichen (m1, m2) und weiblichen Sprecher (f1, f2)

Design

Die Bilder der Sprecher und Gegenstände wurden zu zwei verschiedenen Typen von Displays kombiniert: In der *Complex-Bedingung* wurden den Sprechern vom gleichen Geschlecht die beiden phonetisch ähnlichen Wörter zugeordnet, sodass die Gender-/Sprecherkompetition und die phonetische Kompetition zusammenfielen. Im weiteren Verlauf wird dieser Kompetitor daher als *komplexer Kompetitor* bezeichnet. Da in vorherigen Studien trotz des Geschlechterunterschieds immer vom Faktor *Sprecher* die Rede war ([48] [50]), wird auch bei dieser Untersuchung so verfahren, obwohl der Sprecherkompetitor genau genommen ein Gender-Kompetitor war. In den Beispieldisplays (siehe Abbildung 5.6) haben die beiden Männer die phonetisch ähnlichen Wörter *Ball* und *Band* bzw. *Sessel* und *Soße*. Die beiden anderen Bilder unterscheiden sich sowohl im Sprecher/Geschlecht als auch in der Phonetik der Wörter (*Sessel* und *Soße*). In den Abbildungen ist das Target immer links-oben, der komplexe Kompetitor links-unten und die beiden Distraktoren auf der rechten Seite. Im linken Beispieldisplay überlappen das Target und der phonetische Kompetitor in der initialen CV-Seqenz, während im rechten Beispieldisplay die beiden Wörter nur im initialen Konsonanten übereinstimmen. Aufgrund der phonetischen Struktur der Wörter entstanden mehr C-Trials als CV-Trials im Verhältnis 5:3.
In der *Simplex-Bedingung* wurden den Sprechern vom anderen Geschlecht die beiden phonetisch ähnlichen Wörter zugeordnet, sodass die Sprecherkompetition und die phonetische Kompetition nicht mehr zusammenfielen. In den Beispieldisplays (siehe Abbildung 5.7) haben jeweils immer ein Mann und eine Frau die phonetischen Kompetitoren *Ball* und *Band* bzw. *Sessel* und *Soße*, während sich die Wörter innerhalb der Geschlechter völlig unähnlich sind. Das linke Beispieldisplay zeigt wieder die in CV überlappenden phonetischen Kompeti-

toren und das rechte die nur in C überlappenden. Die Paarungen der Konsonanten wurden so gestaltet, dass immer die maximal unterschiedlichen Laute in einem Trial zusammen gezeigt wurden (/b/ - /s/, /d/ - /f/, /m/ - /t/, /n/ - /p/).

(a) Target und phonetischer Kompetitor überlappen initial in CV. Dargestellt werden die Gegenstände *Ball*, *Band*, *Sessel* und *Soße*.

(b) Target und phonetischer Kompetitor überlappen initial in C. Dargestellt werden die Gegenstände *Sessel*, *Soße*, *Ball* und *Band*.

Abbildung 5.6.: Beispieldisplays für die Complex-Bedingung

(a) Target und phonetischer Kompetitor überlappen initial in CV. Dargestellt werden die Gegenstände *Ball*, *Band*, *Sessel* und *Soße*.

(b) Target und phonetischer Kompetitor überlappen initial in C. Dargestellt werden die Gegenstände *Sessel*, *Soße*, *Ball* und *Band*.

Abbildung 5.7.: Beispieldisplays für die Simplex-Bedingung

Die Aufgabe der Versuchspersonen war es, sich die aufgenommenen Wörter anzuhören

und den korrekten Gegenstand und den Sprecher, der ihn genannt hat, zu identifizieren. Im Gegensatz zu Experiment 1 war es hier in keiner der Bedingungen notwendig, den Sprecher zu identifizieren. Da alle Bilder (Sprecher und Gegenstände) immer einzigartig waren, konnte man sich ausschließlich an der phonetischen Information in den genannten Wörtern orientieren oder aber ausschließlich an den Sprechermerkmalen.

Durchführung

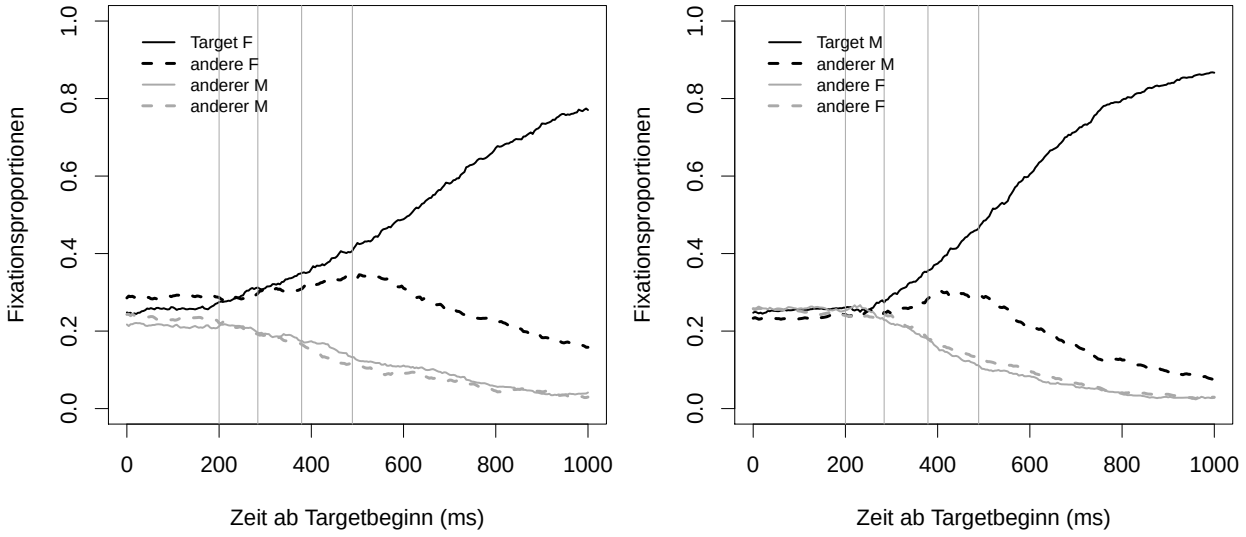
Die Durchführung des Experiments war der des ersten sehr ähnlich. Allein durch die neue Unterteilung der Bedingungen ergaben sich einige Unterschiede im Ablauf. Die Aufgabe der Hörer war es wieder, mit der Maus auf die passende Sprecher-Gegenstand Bildkombination zu klicken. Jeder Teilnehmer bearbeitete 128 randomisierte Trials, wobei jeweils ein Viertel von einem der Sprecher gesprochen wurde; eine Hälfte kam aus der Complex-Bedingung, die andere aus der Simplex-Bedingung. Die CV- und C-Trials waren gleichmäßig im Verhältnis 3:5 in beiden Bedingungen verteilt. Da jedes der 128 Wörter von vier Sprechern aufgenommen wurde, gab es insgesamt 512 Wörter, die sich entweder in ihrer Phonetik oder ihren Sprechermerkmalen unterschieden. Jedem Hörer wurden alle 128 phonetisch unterschiedlichen Wörter als Targetwörter präsentiert. Dabei wurde immer über 8 Hörer der Sprecher des Targetwortes und die Gender-Bedingung variiert. Auf diese Weise blieb das Target immer unvorhersagbar für die Teilnehmer, auch wenn Bilder wiederholt wurden, um Antworten für andere Gegenstände, Sprecher oder Bedingungen zu erhalten. Die restliche Durchführung entsprach der von Experiment 1.

5.3.3. Ergebnisse - Sprechertraining

Die Ausgabedateien des Eye-Trackers wurden mit Hilfe eines angepassten Perl-Skripts [244] in eine Datentabelle mit Informationen (Versuchspersonennummer, Targetbild, ect.) und eine Tabelle mit den zeitlich synchronisierten und kategorisierten Fixationspositionen umgewandelt. Beide Datentabellen wurden dann zur Auswertung in R Version 3.2.0 [241] eingelesen. Zur statistischen Analyse wurden die Pakete „lme4“ ([26]) und „lmerTest“ ([161]) verwendet. Die Effekte der Kompetitoren werden immer im Vergleich zu Distraktor 2 angegeben, wobei sich Distraktor 1 und 2 in keiner Bedingung signifikant voneinander unterschieden.

Die Versuchspersonen konnten in durchschnittlich 91,9 % aller Fälle den Targetsprecher richtig identifizieren. Die Genauigkeit unterschied sich allerdings erheblich zwischen den Trials

mit weiblichem und denen mit männlichem Target. Während die Männer durchschnittlich zu 96,9 % richtig erkannt wurden, identifizierten die Hörer die Frauen nur in 87,0 % aller Fälle richtig. Dies ist ein Indiz für die höhere perzeptive Ähnlichkeit der weiblichen Stimmen in Relation zu den männlichen, die auch von den Versuchsteilnehmern berichtet wurde. Dennoch wurden alle Sprecher über dem Zufallsniveau von 25 % bzw. von 50 % (getrennt nach Geschlechtern) richtig erkannt.
Die Auswertung der Feedbacktrails zeigte, dass tatsächlich der andere Sprecher gleichen Geschlechts wie der Targetsprecher signifikant häufiger/stärker betrachtet wurde als die beiden Sprecher des anderen Geschlechts (siehe Abbildung 5.8). Zur statistischen Analyse wurden die Daten in zwei Zeitintervalle aufgeteilt: T1 von 200 - 500 ms (Targetanfang bis Wortende) und T2 von 500 - 900 ms. Zur Analyse wurden die Fixationsproportionen logistisch transformiert (siehe [23] [128]). Anschließend wurden mit diesen Werten Linear Mixed Effect Models berechnet. Das Modell enthielt das Target als abhängige Variable, das Targetgeschlecht (männlich vs. weiblich) als zweistufigen Fixed Factor, wobei das männliche Target auf das Intercept gesetzt wurde. Außerdem wurden zwei Random Intercepts hinzugefügt, nämlich *Versuchsperson* und *Target*.



(a) Frau als Target

(b) Mann als Target

Abbildung 5.8.: Gender-Kompetition für weiblichen und männlichen Target-Sprecher

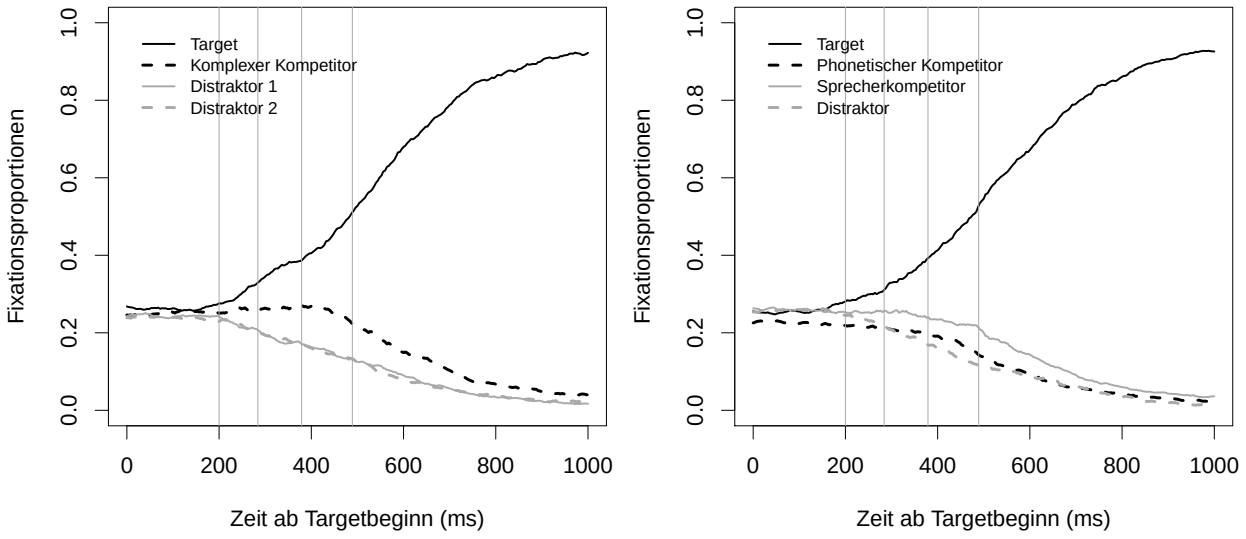
In T1 spielte es keine Rolle, ob das Target männlich oder weiblich war. Beide Targets wurden gleich stark/schnell fixiert. In T2 allerdings wurde das männliche Target stärker/schneller fixiert als das weibliche ($b = 0{,}85$; $t = 4{,}66$; $p < 0{,}001$). Dies spricht dafür, dass

es im Falle eines männlichen Targets leichter war, den Sprecher richtig zu identifizieren. Bei beiden Geschlechtern ist zu beobachten, dass die Gender-Kompetitoren selbst nach dem Ende des Wortes nicht wieder vollständig absanken. Das spricht dafür, dass die Versuchspersonen bis zum Ende des Trials nicht nur das Target, sondern auch immer den anderen Sprecher des gleichen Geschlechts betrachtet haben.
Nach der Targetanalyse wurde der Gender-Kompetitor untersucht. Zunächst für männliche und weibliche Targets zusammen. Die Differenz zwischen dem Gender-Kompetitor und dem andersgeschlechtlichen Sprecher wurde in einem Mixed Model mit der Differenz der Kompetitoren als abhängige Variable, einem Intercept Term als unabhängige Variable und *Versuchsperson* und *Sound-File* als Random Intercepts sowohl für T1 ($b = 0{,}79$; $t = 4{,}78$; $p < 0{,}001$) als auch für T2 ($b = 1{,}18$; $t = 10{,}37$; $p < 0{,}001$) signifikant.
Anschließend wurde der Einfluss des Targetgeschlechts auf die Kompetitordifferenzen untersucht. Dafür wurde in dem eben beschriebenen Model der Intercept Term als unabhängige Variable durch den Faktor *Targetgeschlecht* ersetzt. Während dieser Faktor in T1 noch keinen Einfluss hatte, zeigte sich in T2, dass der Gender-Kompetitor bei einem weiblichen Target signifikant höher war als bei einem männlichen Target ($b = -0{,}7$; $t = -4{,}02$; $p < 0{,}001$). Dies zeigt, dass die perzeptiv höhere Ähnlichkeit der weiblichen Stimmen zu einer stärkeren Gender-Kompetition führte als die männlichen Stimmen. So konnte gezeigt werden, dass Hörer nicht nur den Sprecher des gleichen Geschlechts wie der Targetsprecher stärker fixieren, sondern dass dieser Effekt auch von der (perzeptiven) Ähnlichkeit der Stimmen innerhalb eines Geschlechts abhängt.

5.3.4. Ergebnisse - Hauptteil

Die Ausgabedateien des Eye-Trackers wurden auf die gleiche Weise verarbeitet wie die Feedback-Dateien. Zusätzlich wurde aber im Perl-Skript ([244]) eine Zeitnormalisierung der wortinitialen Konsonant-Vokal-Folge (CV- Folge) durchgeführt. Ein Nachteil dieser Normierung ist, dass kürzere Segmente mehr Gewicht bekommen, da gleiche Messpunkte wiederholt werden. Der Vorteil dieser Methode besteht aber darin, dass verschiedene Segmente (mit unterschiedlich langen Dauern) besser miteinander verglichen werden können. Da in diesem Experiment ebenfalls der Einfluss des initialen Konsonanten auf die Sprecherkompetition untersucht werden sollte, wurden hier ausschließlich zeitnormierte Daten analysiert.
Die Effekte der komplexen Kompetitoren (Complex-Bedingung) werden immer im Vergleich zum zweiten Distraktor angegeben, wobei sich der erste und zweite in keiner Bedingung signifikant voneinander unterschieden.

Alle Teilnehmer wählten durchschnittlich zu 99,1 % der Trials das richtige Bild aus. Für die nachfolgende Analyse wurden nur diese richtigen Trials verwendet. Abbildung 5.9 zeigt die Fixationsproportionen über die Zeit der vier Sprecher-Gegenstand Kombinationen. In der Complex-Bedingung fielen Sprecherkompetitor und phonetischer Kompetitor zusammen. Daher markieren schwarze volle Linien das Target, schwarze gestrichelte Linien den komplexen Kompetitor (den Sprecher vom gleichen Geschlecht kombiniert mit dem phonetisch ähnlichen Gegenstand); graue Linien markieren die Distraktoren (volle Linie: Distraktor 1; gestrichelte Linie: Distraktor 2).
In der Simplex-Bedingung markieren die schwarzen vollen Linien ebenfalls das Target, schwarze gestrichelte Linien nur den phonetischen Kompetitor (den Sprecher des anderen Geschlechts gepaart mit dem phonetisch ähnlichen Gegenstand); graue volle Linien den Sprecherkompetitor (den Sprecher des gleichen Geschlechts kombiniert mit dem phonetisch unähnlichen Gegenstand) und graue gestrichelte den Distraktor.
Die vertikalen Linien repräsentieren den Wort-/Konsonantenbeginn, Konsonantenende, Vokalende und Wortende. Dabei sind alle Werte um 200 ms verschoben. Es wird angenommen, dass zwischen dem akustischen Signal und der Augenbewegung aufgrund dieses Signals 200 ms verstreichen ([4]).



(a) Complex-Bedingung (b) Simplex-Bedingung

Abbildung 5.9.: Fixationsproportionen in Abhängigkeit von der Zeit

Es wurden zwei separate Zeitfenster analysiert: T1 von 200 ms bis 500 ms und T2 von 500 ms bis 900 ms. T1 wurde gewählt, um die Fixationen während der Wortverarbeitung

widerzuspiegeln (durchschnittliche Wortdauer: 480 ms). T2 reichte vom Wortende bis zu dem Punkt, wo die Targetfixationen aufhörten zu steigen und die Kompetitorfixationen auf ein Minimum fielen.
Zur Analyse wurden die Fixationsproportionen logistisch transformiert (siehe [23] [128]). Anschließend wurden mit diesen Werten Linear Mixed Effect Models berechnet. Zunächst wurde ein globales Mixed Model berechnet, welches das Target als abhängige Variable enthielt, die *Bedingung*, den *Overlap* und die Interaktion zwischen beiden als Fixed Factors und ein Random Intercept für *Versuchspersonen* und *Targetbild*. Ein zusätzliches Random Intercept über *Iteration* bewirkte keinen signifikanten Unterschied, was bedeutet, dass es keinen Lern- oder Ermüdungseffekt gab. Ein Random Slope über *Targetbild pro Versuchsperson* war nicht notwendig, da jede Versuchsperson jedes Targetbild gleich häufig gesehen hat, sodass sich eventuelle individuelle Präferenzen einer Person für ein bestimmtes Bild ausgeglichen haben. In T1 hatte keiner der Faktoren einen signifikanten Einfluss auf das Target, in T2 jedoch zeigte sich, dass die Targetwörter mit initialem CV-Overlap schneller/stärker anstiegen als Wörter mit C-Overlap ($b = -0{,}43$; $t = -2{,}84$; $p = 0{,}005$). In einem späteren Abschnitt werden daher die CV- und C-Overlap-Wörter separat analysiert und besprochen.
Zur Analyse der Kompetitoren wurden beide Bedingungen (Simplex und Complex) getrennt voneinander betrachtet und separate Linear Mixed-Effect Models berechnet. Die abhängige Variable war die Fixationspräferenz zwischen dem jeweiligen Kompetitor und dem Distraktor. Als Fixed Factor enthielten die Modelle nur einen Intercept Term. Wenn sich dieser signifikant von Null unterschied, wurde einer der Kompetitoren stärker fixiert als der andere. Für *Versuchsperson* wurde ein Random Intercept einbezogen.
Während T1 wurde das Target in beiden Bedingungen sehr früh gegenüber den Kompetitoren bevorzugt. In der Complex-Bedingung wurde erwartungsgemäß der komplexe Kompetitor stärker fixiert als die Distraktoren (siehe Tabelle 5.5). Da dieser Kompetitor dem Target sowohl in Geschlecht als auch im phonetischen Muster ähnlich war, löste er vor allem in T1 eine starke Kompetition aus. Die beiden unähnlichen Bilder (Distraktoren) wurden so gut wie gar nicht betrachtet.
In der Simplex-Bedingung wurde der Sprecherkompetitor stärker betrachtet als der phonetische Kompetitor und der Distraktor, welche sich nicht unterschieden (siehe Tabelle 5.6). Das heißt, die Hörer beachteten den Gegenstand, der mit dem Sprecher des gleichen Geschlechts wie der Targetsprecher kombiniert war, trotz des phonetischen Mismatches am Wortanfang.

Für T2 wurden die gleichen Signifikanzen erzielt wie für T1 mit leichten Abweichungen in den Regressionsgewichten (b) und den *t*-Werten.

	T1 (200 - 500 ms)			**T2 (500 - 900 ms)**		
Vergleich	b	t	p	b	t	p
Komplexer Komp. - Distraktor 1	0,58	5,69 *	$< 0{,}001$	0,41	6,3 *	$< 0{,}001$
Komplexer Komp. - Distraktor 2	0,58	5,31 *	$< 0{,}001$	0,43	7,22 *	$< 0{,}001$

Tabelle 5.5.: Ergebnisse der Complex-Bedingung in T1 (200 - 500 ms) und T2 (500 - 900 ms): Fixationspräferenzen zwischen den verschiedenen Kompetitoren (* markiert signifikante Werte)

	T1 (200 - 500 ms)			**T2 (500 - 900 ms)**		
Vergleich	b	t	p	b	t	p
phonetischer Komp. - Distraktor	0,11	1,26	0,21	0,06	1,24	0,21
Sprecherkomp. - Distraktor	0,43	4,7 *	$< 0{,}001$	0,32	4,06 *	$< 0{,}001$
Sprecherkomp. - phonetischer Komp.	-0,32	-3,15 *	0,004	-0,26	-4,06 *	$< 0{,}001$

Tabelle 5.6.: Ergebnisse der Simplex-Bedingung in T1 (200 - 500 ms) und T2 (500 - 900 ms): Fixationspräferenzen zwischen den verschiedenen Kompetitoren (* markiert signifikante Werte)

Vergleich der Complex- und der Simplex-Bedingung

Die Complex- und die Simplex-Bedingungen wurden beide in T1 und T2 miteinander verglichen. Dabei wurden die Effekte der Kompetitoren immer in Bezug auf die Distraktoren der entsprechenden Bedingung bestimmt. Der Verlauf der Distraktoren unterschied sich weder in T1 noch in T2 zwischen beiden Bedingungen. In den Abbildungen scheint der komplexe Kompetitor (Complex-Bedingung) etwas stärker zu sein als der Sprecherkompetitor (Simplex-Bedingung) (siehe Abbildung 5.10).
Mit Hilfe eines *t*-tests wurde die Differenz zwischen dem komplexen Kompetitor und dem Distraktor 2 (Complex-Bedingung) und die Differenz zwischen dem Sprecherkompetitor und dem Distraktor (Simplex-Bedingung) miteinander verglichen. Beide Differenzen unterschieden sich weder in T1 ($t = 1{,}03$; $p = 0{,}31$) noch in T2 ($t = 1{,}25$; $p = 0{,}22$) voneinander. Scheinbar konnte die zusätzliche phonetische Ähnlichkeit der Gegenstände der Ähnlichkeit

des Sprechergeschlechts nicht viel hinzufügen. Dieses Ergebnis ist in Anbetracht des nicht signifikanten Effekts des phonetischen Kompetitors (Simplex-Bedingung) nicht überraschend. Das Verhältnis von komplexem Kompetitor und Distraktor 2 und von phonetischem Kompetitor und Distraktor unterschied sich dagegen sowohl in T1 ($t = 3{,}49$; $p < 0{,}001$) als auch in T2 ($t = 5{,}19$; $p < 0{,}001$) stark voneinander. Die zusätzliche Sprecherinformation im komplexen Kompetitor hatte also im Vergleich zur phonetischen Information einen großen Einfluss auf die Stärke der Kompetition.

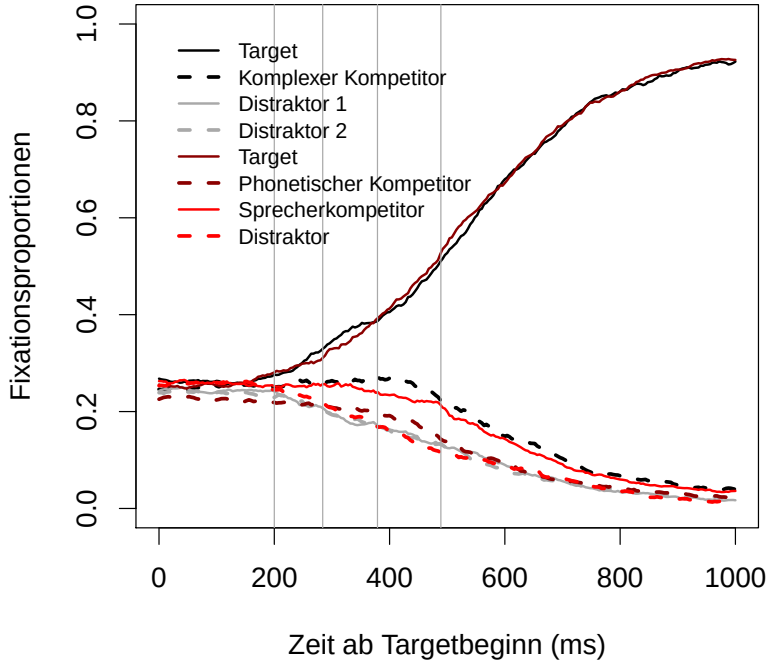
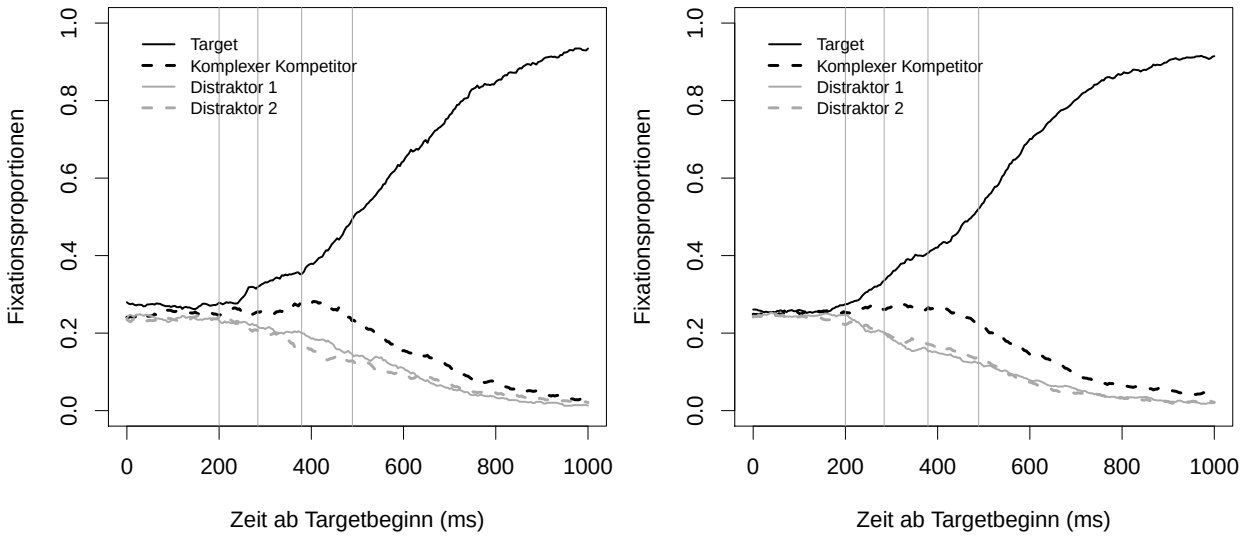


Abbildung 5.10.: Simplex- und Complex-Bedingung: Fixationsproportionen in Abhängigkeit von der Zeit

Länge des phonetischen Overlaps: CV vs. C

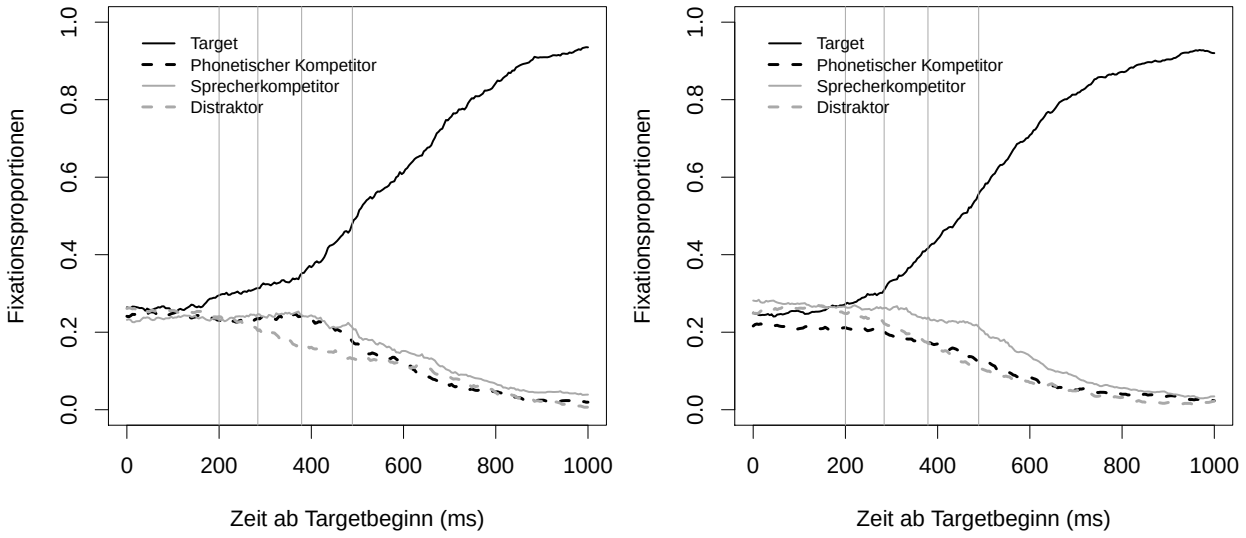
Abgesehen von der Absicht, die phonetische Ähnlichkeit der Wörter zu maximieren, war es auch unser Ziel, den (möglichen) Einfluss der Länge des phonetischen Overlaps zu untersuchen. In der Complex-Bedingung (siehe Abbildung 5.11) schien der komplexe Kompetitor durch den längeren CV-Overlap nur leicht verstärkt zu werden. In der Simplex-Bedingung hingegen ist ein sehr deutlicher Unterschied zwischen dem phonetischen Kompetitor mit CV-Overlap im Vergleich zu dem mit C-Overlap zu erkennen (siehe Abbildung 5.12). Während sich bei C-Overlap der phonetische Kompetitor (fast) auf der gleichen Höhe wie der Distraktor befindet, steigt er im Falle von CV-Overlap auf die gleiche Höhe wie der

Sprecherkompetitor an. Da der Effekt von dem initialen phonetischen Overlap ca. 200 ms nach Target-Onset erwartet wurde, sind vor allem in T1 Unterschiede zu vermuten. In T2 ist dieser Effekt wahrscheinlich bereits wieder abgeklungen.



(a) CV (b) C

Abbildung 5.11.: Complex-Bedingung: Fixationsproportionen in Abhängigkeit von der Zeit



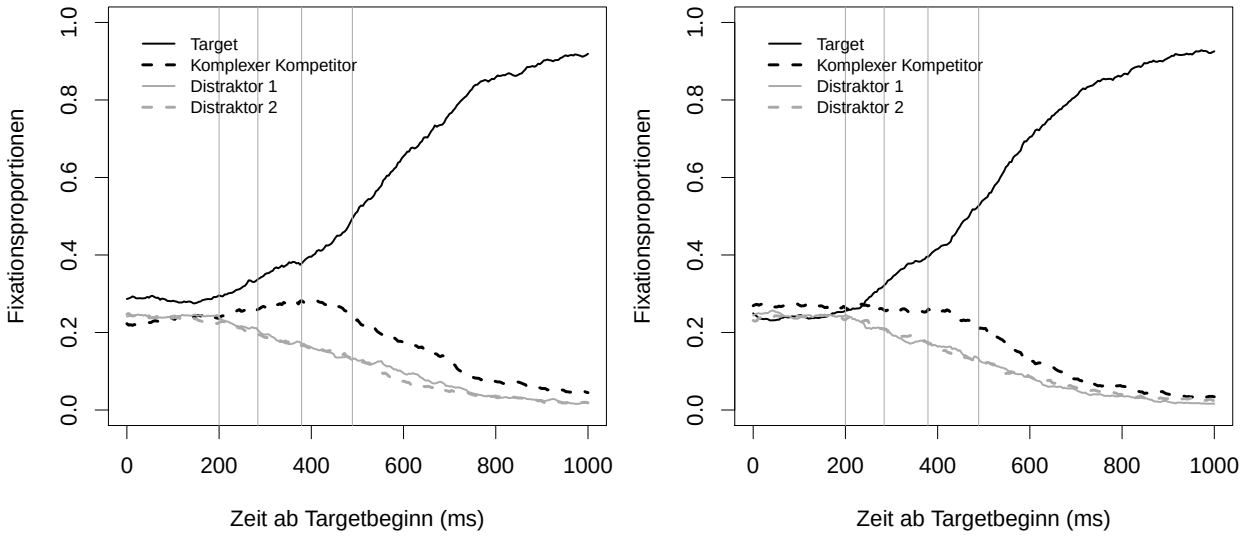
(a) CV (b) C

Abbildung 5.12.: Simplex-Bedingung: Fixationsproportionen in Abhängigkeit von der Zeit

Zur statistischen Analyse wurde ein Mixed Model berechnet mit der Differenz des phonetischen Kompetitors zum Distraktor als abhängige Variable, *Overlap* als Fixed Factor und einem Random Intercept für *Versuchspersonen*. Dabei wurde die C-Overlap Bedingung auf dem Intercept abgebildet und die CV-Overlap Bedingung im Vergleich dazu bestimmt. Das Ergebnis der statistischen Auswertung bestätigte das Bild der deskriptiven Analyse. In der Complex-Bedingung war der Einfluss des Faktors *Overlap* auf den phonetischen Kompetitor nicht signifikant, weder in T1 ($b = 0{,}06$; $t = 0{,}3$; $p = 0{,}76$), noch in T2 ($b = -0{,}05$; $t = -0{,}43$; $p = 0{,}67$). Da sich in dieser Bedingung Sprecher- und phonetische Informationen überlagerten, scheint sich der kleine Unterschied (CV vs. C) zwischen den phonetischen Kompetitoren nicht so stark auf die gesamte Kompetition auszuwirken. In der Simplex-Bedingung hingegen war der phonetische Kompetitor mit CV-Overlap signifikant stärker als mit C-Overlap ($b = 0{,}36$; $t = 2{,}05$; $p = 0{,}04$).
Nachdem der signifikante Einfluss des Faktors Overlap gezeigt wurde, wurden zur weiteren Analyse die Daten nach Overlap-Bedingung aufgesplittet und separate Mixed Models berechnet. In diesen Modellen unterschied sich der CV-Overlap Kompetitor in T1 nicht vom (signifikanten) Sprecherkompetitor ($b = -0{,}06$; $t = -0{,}35$; $p = 0{,}73$) während der C-Overlap Kompetitor signifikant schwächer war ($b = -0{,}48$; $t = -4{,}15$; $p < 0{,}001$) und sich nicht signifikant vom Distraktor unterschied ($b = -0{,}03$; $t = -0{,}27$; $p = 0{,}79$). Daraus folgt, dass der längere phonetische Overlap zu einer längeren Ambiguität der Wörter führt und deshalb eine stärkere phonetische Kompetition hervorruft. Da aber im Verhältnis zu den C-Overlap Trials nur wenige CV-Overlap Trials vorhanden waren (5:3), hatten diese einen geringeren Einfluss auf den gesamten (CV- + C-Overlap) phonetischen Kompetitor, sodass dieser in der Gesamtanalyse nicht signifikant wurde. Es wird jedoch klar, dass der phonetische Overlap einen Einfluss auf die Stärke der phonetischen Kompetition hat und dass bei einer höheren Anzahl an CV-Overlap Trials der Effekt des phonetischen Kompetitors wahrscheinlich signifikant geworden wäre.

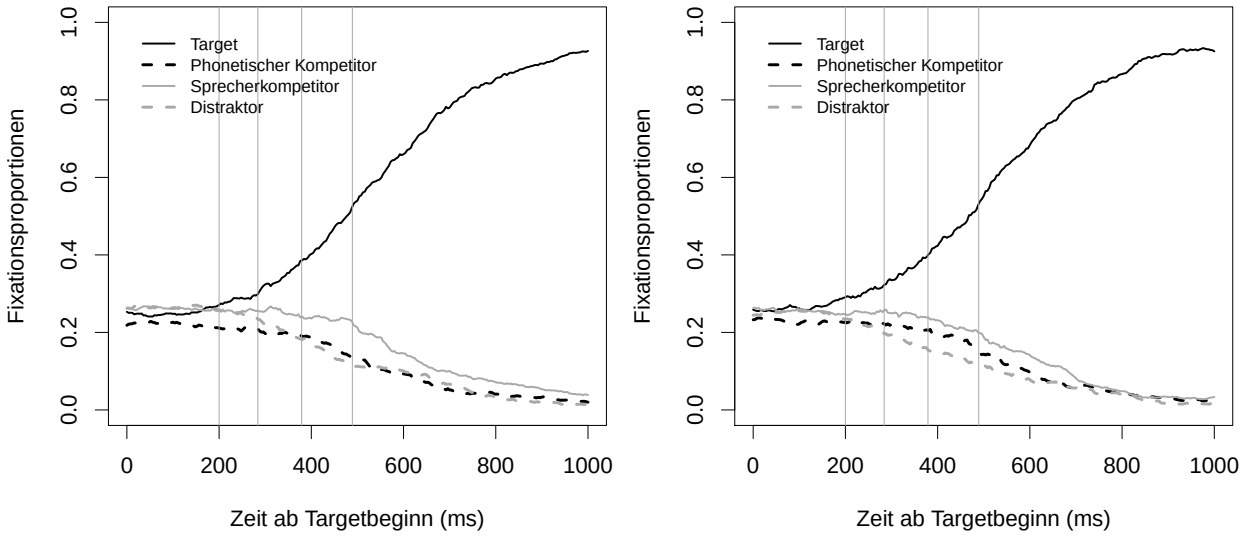
Geschlecht des Targetsprechers: Frau vs. Mann

Da viele der Versuchspersonen berichteten, dass sich die beiden weiblichen Sprecher ähnlicher anhörten als die männlichen Sprecher, wurde auch der Einfluss des Geschlechts des Targetsprechers auf die Fixationsproportionen untersucht. Dieser Eindruck der Versuchspersonen schien sich auch in den Ergebnissen widerzuspiegeln: Bei einer Frau als Targetsprecher entstand in beiden Zeitintervallen eine stärkere komplexe Kompetition (siehe Abbildung 5.13).



(a) Frau als Target (b) Mann als Target

Abbildung 5.13.: Complex-Bedingung: Einfluss des Targetgeschlechts auf die Fixationsproportionen in Abhängigkeit von der Zeit



(a) Frau als Target (b) Mann als Target

Abbildung 5.14.: Simplex-Bedingung: Einfluss des Targetgeschlechts auf die Fixationsproportionen in Abhängigkeit von der Zeit

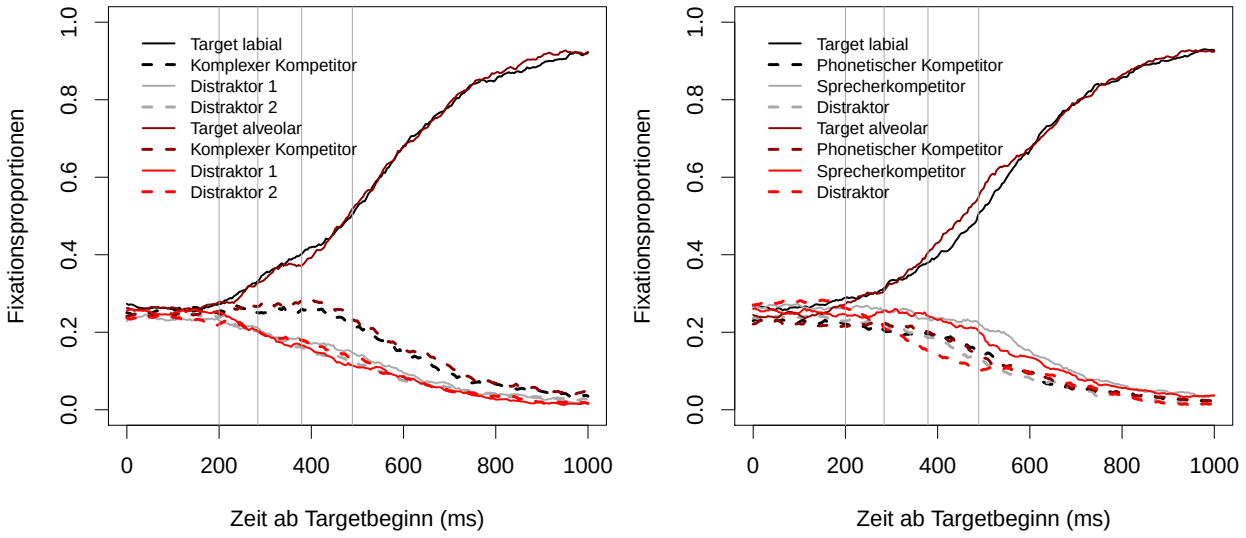
Für die statistische Analyse wurde jeweils ein Mixed Model mit der Differenz zwischen

komplexem Kompetitor und Distraktor 1 bzw. 2 als abhängige Variable und *Target-Gender* als zweistufigem Fixed Factor (männlich vs. weiblich) erstellt, wobei der weibliche Targetsprecher auf das Intercept gesetzt wurde. Der Faktor *Versuchsperson* wurde als Random Intercept eingefügt. Der kleine Unterschied zwischen weiblichem und männlichem komplexen Kompetitor wurde in der statistischen Analyse für T1 nicht signifikant. In T2 trat der Unterschied aber deutlicher hervor, sodass der Einfluss des Geschlechts des Targetsprechers in der Differenz von komplexem Kompetitor und Distraktor 2 signifikant wurde (b = -0,31; t = -2,91; p = 0,004).
In der Simplex-Bedingung schien sich das Geschlecht des Targetsprechers auf die Stärke des phonetischen Kompetitors auszuwirken und weniger auf den Sprecherkompetitor (siehe Abbildung 5.14). Ein männlicher Sprecher als Target ruft scheinbar eine stärkere phonetische Kompetition hervor als ein weiblicher Sprecher. Trotz des Eindrucks der Abbildung konnte dieser Effekt in der statistischen Analyse nicht nachgewiesen werden. Lediglich ein fast signifikantes Ergebnis in T2 (b = 0,19; t = 1,83; p = 0,067) zeigte, dass der phonetische und der Sprecherkompetitor bei einem männlichen Targetsprecher näher zusammenliegen. Da der Sprecherkompetitor seine Lage nicht verändert, muss also der phonetische Kompetitor in T2 bei einem männlichen Sprecher höher liegen als bei einem weiblichen Targetsprecher. Es scheint also, dass die Hörer die Wörter von einem männlichen Sprecher etwas (aber nicht signifikant) schlechter verstehen als von einem weiblichen und es dadurch zu einer erhöhten phonetischen Kompetition kommt.

Artikulationsstelle des Initialkonsonanten

Eine Frage der Untersuchung bestand darin, ob sich die Artikulationsstelle (labial vs. alveolar) des Initialkonsonanten des gesprochenen Wortes auf die phonetische und/oder die Sprecherkompetition auswirken würde. In der Complex-Bedingung schien sich die Artikulationsstelle nicht auf den komplexen Kompetitor ausgewirkt zu haben. Beide Kompetitoren und Distraktoren liegen sehr nah beieinander (siehe Abbildung 5.15). In der Simplex-Bedingung wirkte der Sprecherkompetitor mit labialer Artikulationsstelle leicht stärker als der mit alveolarer.
Zur Überprüfung dieser deskriptiven Ergebnisse wurde ein Mixed Model berechnet. Abhängige Variable war die Differenz des jeweiligen Kompetitors zum Distraktor, Fixed Factor die *Artikulationsstelle* und Random Intercept der Faktor *Versuchsperson*. Der zweistufige Faktor *Artikulationsstelle* unterteilte sich nach labialer und alveolarer Artikulationsstelle, wobei labial auf dem Intercept lag. In der Complex-Bedingung zeigten sich in keinem der

Zeitintervalle T1 und T2 signifikante Effekte der Artikulationsstelle. Nur in T2 zeigte sich ein tendenzieller Effekt auf den Kompetitor im Verhältnis zum Distraktor 1 (b = -0,18; t = -1,69; p = 0,09), bei dem der komplexe Kompetitor bei einem alveolar-initialen Targetwort minimal stärker war als als bei einem labial-initialen. In der Simplex-Bedingung war das Ergebnis ähnlich mit einem tendenziell stärkeren Sprecherkompetitor bei einem Initialkonsonanten mit labialer Artikulationsstelle in T2 (b = 0,17; t = 1,66; p = 0,098). Nachdem aber keiner dieser Effekte signifikant wurde, ist anzunehmen, dass die Artikulationsstelle dieses (kurzen) ersten Konsonanten keinen Einfluss auf die Stärke der Kompetitoren hat.



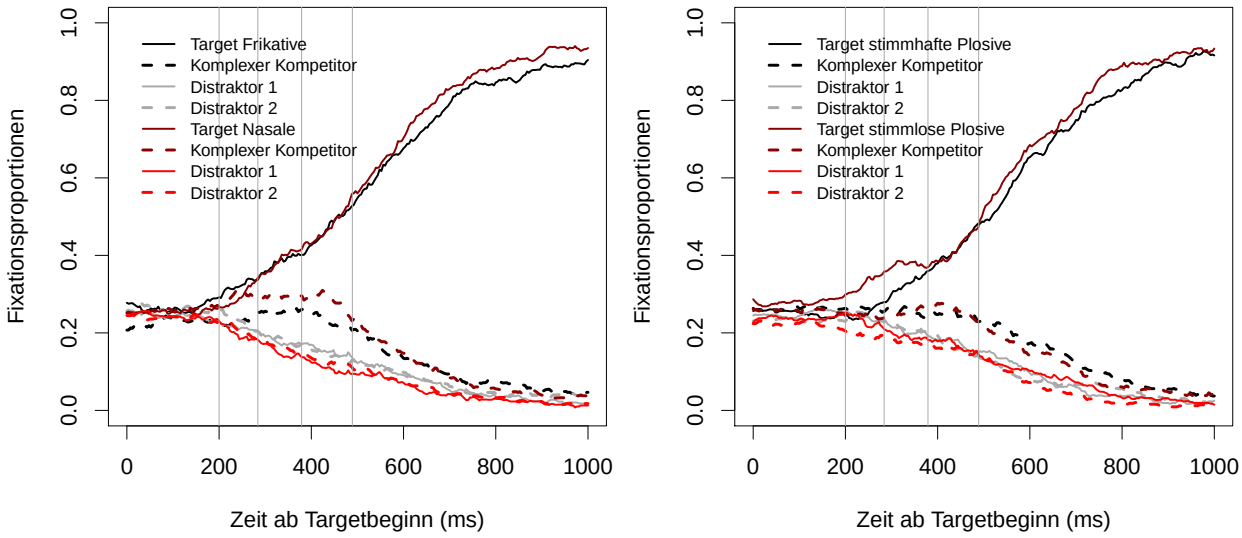
(a) Complex-Bedingung (b) Simplex-Bedingung

Abbildung 5.15.: Einfluss der Artikulationsstelle auf die Fixationsproportionen in Abhängigkeit von der Zeit

Artikulationsmodus des Initialkonsonanten

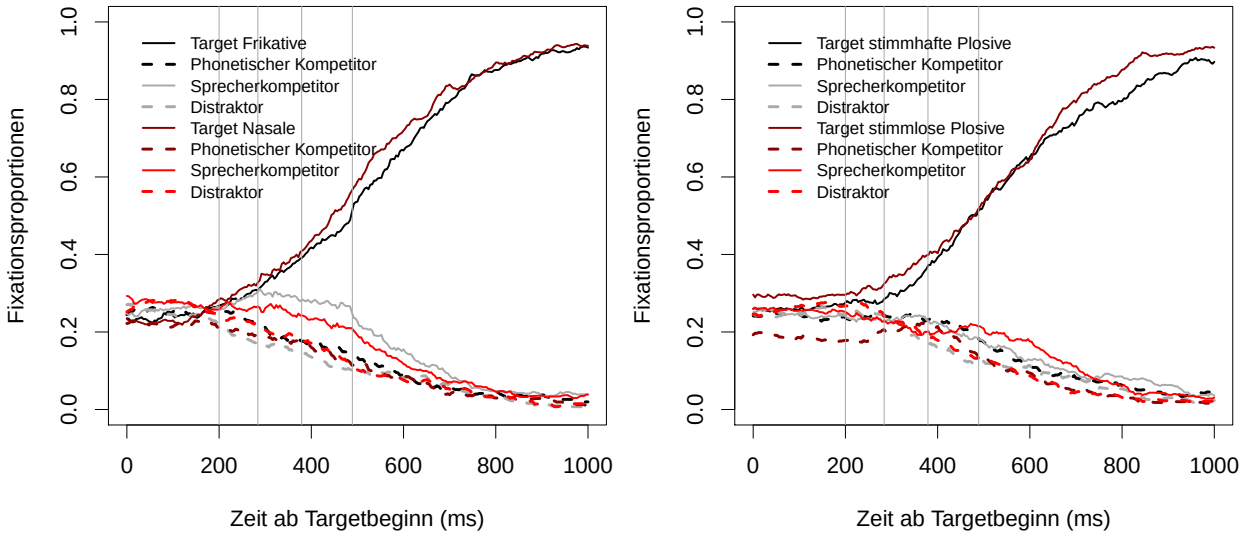
Um heraus zu finden, ob sich der Artikulationsmodus des Initialkonsonanten auf die Fixationsproportionen des Targets und der Kompetitoren auswirkt, wurde der Artikulationsmodus detaillierter analysiert. In der Complex-Bedingung schienen die Nasale als Initialkonsonanten im Vergleich zu den Frikativen sowohl einen stärkeren/schnelleren Anstieg des Targets als auch des komplexen Kompetitors zu bewirken. Bei den Plosiven schienen die stimmlosen Konsonanten das Target schneller/stärker steigen zu lassen als die stimmhaften, während der Verlauf der komplexen Kompetitoren ähnlich war (siehe Abbildung 5.16). In der Simplex-Bedingung bewirkten die Frikative einen stärkeren Anstieg des Sprecher-

kompetitors als die Nasale, hatten aber keine sichtbare Wirkung auf den phonetischen Kompetitor. Die stimmlosen Plosive ließen gegenüber den stimmhaften das Target und den Sprecherkompetitor stärker ansteigen (siehe Abbildung 5.17).



(a) Frikative und Nasale (b) Plosive

Abbildung 5.16.: Complex-Bedingung: Fixationsproportionen in Abhängigkeit von der Zeit



(a) Frikative und Nasale (b) Plosive

Abbildung 5.17.: Simplex-Bedingung: Fixationsproportionen in Abhängigkeit von der Zeit

Für die statistische Überprüfung dieser Effekte wurden Mixed Models berechnet mit dem Target als abhängige Variable und *Artikulationsmodus* als vierstufigen Fixed Factor (Nasale, Frikative, stimmhafte Plosive, stimmlose Plosive), wobei die Frikative stets auf dem Intercept lagen. Ein weiterer Fixed Factor war der zweistufige Faktor *Bedingung* (Complex vs. Simplex). Hier lag die Complex-Bedingung auf dem Intercept. Der Faktor *Versuchsperson* wurde als Random Intercept hinzugefügt. Beim Einfluss auf das Target unterschied sich keine der Lautgruppen von den Frikativen, weder in T1 noch in T2.
In der Complex-Bedingung bewirkten in T1 die nasal-initialen Wörter einen signifikant stärkeren Anstieg des komplexen Kompetitors als die Wörter mit anderen Lauten (b = 0,55; t = 2,13; p = 0,03). Die Nasale haben somit entweder die phonetische oder die Sprecherkompetition (oder beide) signifikant erhöht. In T2 wurde dieser Effekt nicht mehr signifikant.
In der Simplex-Bedingung wirkte sich kein Artikulationsmodus signifikant auf den Verlauf des phonetischen Kompetitors aus, während sich alle Lautgruppen signifikant von den Frikativen in ihrem Einfluss auf den Sprecherkompetitor unterschieden (Nasale: b = -0,50; t = -1,97; p = 0,048; stimmhafte Plosive: b = -0,81; t = -3,17; p = 0,002; stimmlose Plosive: b = -0,87; t = -3,41; p < 0,001). Dabei lösten die Frikative die stärkste Sprecherkompetition aus und alle anderen Laute weniger. Anschließende *t*-Test zeigten, dass sich die Nasale und die Plosive untereinander nicht unterschieden. Dieses Ergebnis zeigt, dass die verschiedenen Lautgruppen unterschiedlich stark zur Identifikation des Sprechers beitragen. Beim Einfluss auf die Differenz von phonetischem und Sprecherkompetitor unterschieden sich nur die stimmhaften Plosive von den Frikativen (b = 0,69; t = 2,66; p = 0,008), da bei diesen Lauten der phonetische und der Sprecherkompetitor auf gleicher Höhe lagen.
Alle diese Ergebnisse, wenn auch signifikant, sind mit Vorsicht zu betrachten, da durch die starke Aufteilung der Daten die Aussagekraft der statistischen Ergebnisse sinkt.

Jackknife-Analyse

Um die zeitliche Verarbeitung von phonetischer und Sprecherinformation genauer zu untersuchen, wurde eine Maximaleffektanalyse mittels der Jackknife-Methode durchgeführt. Für dieses Verfahren wurden beide phonetischen Kompetitoren (Targetwort und phonetisch ähnliches Wort) als auch beide Sprecherkompetitoren (Targetsprecher und der andere Sprecher des gleichen Geschlechts) gemittelt. Dadurch erhält man einen Graphen für den Faktor Sprecher und einen für den Faktor Phonetik/Wort. Anschließend wurden Zeitpunkte berechnet, an denen die Fixation der Hörer auf Sprecher oder Wort einen bestimmten Prozentwert des Maximums erreichte. Die Prozentpunkte lagen bei 10, 15, 20, 30, 40, 50,

60 und 70 %, um einen möglichst großen Bereich abzudecken.

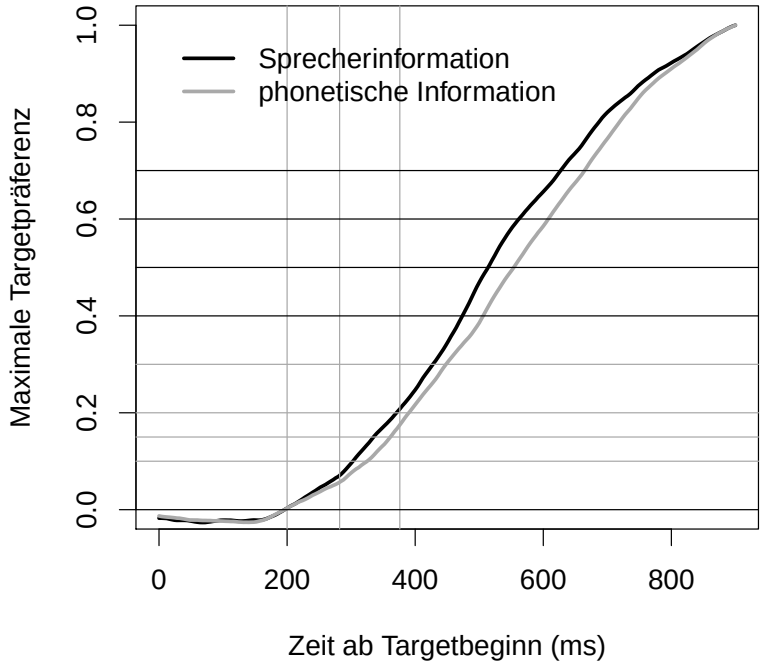
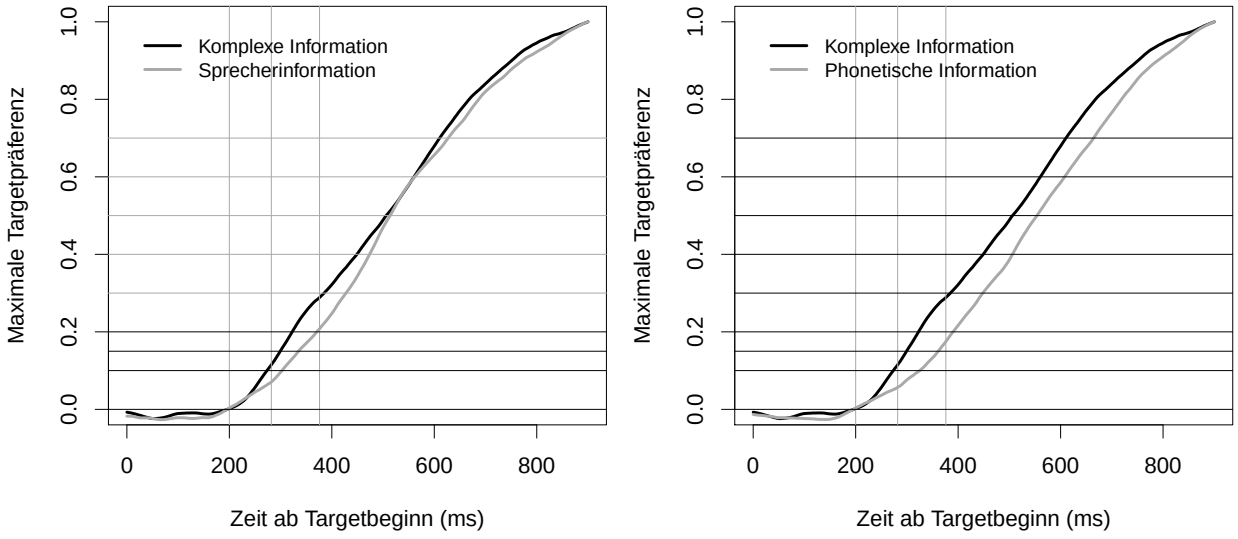


Abbildung 5.18.: Simplex-Bedingung: Proportionen der maximalen Targetpräferenz in Abhängigkeit von der Zeit zwischen 200 und 900 ms.

Durch diese Analyse wird nicht vorrangig die Größe des Kompetitionseffekts betrachtet, sondern vor allem dessen Anstiegsgeschwindigkeit. Steigt ein Effekt schneller als der andere, lässt sich vermuten, dass die Informationen, die zu diesem Effekt führen, von den Hörern eher wahrgenommen und verarbeitet wurden. Die Zeitpunkte zu denen die Prozentwerte des Maximumeffekts für Sprecher und Phonetik/Wort erreicht wurden, wurden anschließend mit *t*-tests verglichen. Die *t*-Werte wurden wie in Experiment 1 beschrieben angepasst. Ab einem *t*-Wert von $t < 2$ oder $t > 2$ ist ein Ergebnis signifikant.
In der Complex-Bedingung fielen der phonetische und der Sprecherkompetitor zusammen, sodass kein Vergleich stattfinden kann. In der Simplex-Bedingung wurde die Differenz zwischen den phonetischen und den Sprecherinformationen ab einem Maximaleffekt von 40 % signifikant (siehe Abbildung 5.18). In den Abbildungen sind nicht-signifikante Differenzen durch graue horizontale Linien gekennzeichnet und signifikante durch schwarze. Dabei stieg der Sprechereffekt stets schneller als der phonetische Effekt (40 %: $t = -2{,}22$; 50 %: $t = -2{,}46$; 60 %: $t = -2{,}38$; 70 %: $t = -2{,}44$). Der Sprechereffekt war vor allem im späteren Verlauf dem phonetischen Effekt voraus. Das deutet darauf hin, dass am Anfang sowohl phonetische

als auch Sprecherinformationen verarbeitet werden, während sich die Hörer im späteren Verlauf hauptsächlich auf die Sprecherinformationen konzentrieren. Außerdem erreichte der Sprechereffekt sein Maximum schneller als der phonetische, was darauf hindeutet, dass Sprecherinformationen schneller verarbeitet werden als phonetische Informationen.
Anschließend wurde der komplexe Kompetitor aus der Complex-Bedingung mit den zwei Kompetitoren aus der Simplex-Bedingung verglichen. Der komplexe Kompetitor stieg am Anfang schneller als der Sprecherkompetitor, später näherten sich beide aber an, sodass der Unterschied nicht mehr signifikant wurde (siehe Abbildung 5.19). Die genauen Werte sind in Tabelle 5.7 aufgeschlüsselt. Diese Ergebnisse geben Grund zur Annahme, dass die Wichtigkeit der Sprecherinformationen gegen Ende des Wortes (und nach Wortende) zunimmt.



(a) Einfluss der Phonetik

(b) Einfluss des Sprechers

Abbildung 5.19.: Proportionen der maximalen Targetpräferenz in Abhängigkeit von der Zeit zwischen 200 und 900 ms.

Im Vergleich von komplexem und phonetischem Effekt zeigte sich, dass die Differenz zwischen beiden zu jedem Prozentpunkt signifikant wurde, wobei der phonetische Effekt langsamer stieg als der komplexe (siehe Abbildung 5.19). Dies zeigt deutlich, dass die phonetische Kompetition nicht nur geringer war als die komplexe Kompetition, sondern auch langsamer. Der Sprechereffekt scheint einen deutlich höheren Anteil an dem schnellen Anstieg des komplexen Effekt zu haben als die phonetischen Informationen. Mit anderen Worten, die Hörer verarbeiten Sprecherinformationen schneller als phonetische, wobei zu

Beginn des Wortes beide Informationstypen gleich schnell Eingang zu finden scheinen. Gegen Wortende (und nach Wortende) verschiebt sich dieses anfangs ausgeglichene Verhältnis aber signifikant zu Gunsten der Sprecherinformationen.

	***t*-Werte zu Prozentpunkten**							
Vergleich	10 %	15 %	20 %	30 %	40 %	50 %	60 %	70 %
Sprecher-Phonetik	-0,81	-1,20	-0,82	-0,93	-2,22 *	-2,46 *	-2,38 *	-2,44 *
Komplex - Sprecher	-2,26 *	-2,67 *	-2,99 *	-1,77	-1,56	-1,56	-0,07	-1,01
Komplex - Phonetik	-2,17 *	-3,47 *	-3,56 *	-2,55 *	-3,76	-3,38 *	-3,80 *	-3,36 *

Tabelle 5.7.: Vergleich der Kompetitoren aus beiden Bedingungen zu verschiedenen Prozentpunkten des Maximaleffekts (* markiert signifikante Werte)

5.3.5. Diskussion

In diesem Nachfolgeexperiment wurden den Hörern 4 Sprecher präsentiert, zwei Männer und zwei Frauen. Da sich die Stimmen innerhalb der Geschlechter ähnlicher waren, wurde vermutet, dass bei einer Sprecheridentifikationsaufgabe Gender-Kompetition auftreten würde. Das bedeutet, dass die Hörer beim Klang einer weiblichen Stimme auch stärker auf das Bild der anderen Frau geschaut haben als auf die Bilder der männlichen Sprecher (und anders herum). Das heißt, sie haben das Geschlecht des Sprechers (und die damit verbundene höhere Ähnlichkeit) erkannt und nur noch die Sprecher des richtigen Geschlechts betrachtet und deren Bilder auf mögliche Identität mit der gehörten Stimme überprüft. Darüber hinaus riefen auch Ähnlichkeitsunterschiede innerhalb des Geschlechts Unterschiede in der Sprecherkompetition hervor. Da sich die weiblichen Stimmen (perzeptiv) ähnlicher waren, lösten sie eine stärkere Sprecherkompetition aus als die männlichen Sprecher. Die schwierigere Unterscheidbarkeit der weiblichen Stimmen zeigte sich auch in der durchschnittlichen Identifikationsrate von 87,0 % (bei den Männern waren es 96,9 %).

Die Ergebnisse des Hauptteils bestätigten die Ergebnisse des ersten Eye-Tracking Experiments. Die Versuchsteilnehmer waren in durchschnittlich 99,1 % der Fälle in der Lage, die richtige Sprecher-Gegenstands-Kombination zu identifizieren. Erwartungsgemäß rief der

komplexe Kompetitor (phonetisch ähnliches Wort + Sprecher gleichen Geschlechts) die stärkste Kompetition zum Target hervor. Der reine Sprecherkompetitor (Simplex-Bedingung) war jedoch nicht signifikant schwächer als der komplexe Kompetitor. Die Hörer schauten also stärker auf den anderen Sprecher des gleichen Geschlechts (gepaart mit dem phonetisch unähnlichen Gegenstand) als auf die Sprecher des anderen Geschlechts (gepaart mit dem phonetisch ähnlichen Gegenstand). Der phonetische Kompetitor hingegen wurde kaum betrachtet, genauso wenig wie der Distraktor.
Die starke Sprecherkompetition bedeutet einerseits, dass es schwierig war, die Sprecher zu unterscheiden und andererseits, dass die Hörer auf Sprecherinformationen geachtet haben. Im Gegensatz dazu kann die geringe phonetische Kompetition darauf hinweisen, dass die Wörter so leicht zu unterscheiden waren, dass keine Kompetition auftrat oder, dass die Hörer nicht auf die phonetischen Merkmale achteten. Letzteres ist allerdings unwahrscheinlich, da gezeigt wurde, dass die Wörter mit CV-Overlap durchaus eine signifikante phonetische Kompetition hervorriefen. Das lässt vermuten, dass die Hörer schon auf die phonetischen Merkmale im Sprachsignal achteten, aber der alleinige C-Overlap nicht ausreichend lang war, um phonetische Kompetition zu erzeugen. Sprich, die C-Overlap-Wörter waren phonetisch zu einfach zu unterscheiden. Da entweder die phonetischen oder die Sprechermerkmale alleine ausreichend waren zur Targetidentifikation und die phonetischen Merkmale noch dazu einfacher zuzuordnen waren als die Sprecherinformationen, stellt sich die Frage, warum die Hörer den Sprecherinformationen überhaupt Aufmerksamkeit schenkten. Es gibt mehrere Gründe, die dazu geführt haben könnten: (1) Es könnte einerseits darin liegen, dass Menschen von Natur aus eher auf Gesichter schauen als auf Gegenstände und dass die Hörer deshalb stärker die Sprecherbilder fokussierten. (2) Außerdem gab es wesentlich weniger Sprecher als Wörter (4:128), sodass die Kategorie Sprecher eine geringere Anzahl an zu unterscheidenden Instanzen aufwies. Die Hörer könnten darin einen (strategischen) Vorteil gesehen haben und deshalb ihre Aufmerksamkeit auf die Sprecherbilder konzentriert haben. (3) Eine letzte Möglichkeit wäre, dass die Hörer in dem Experimentteil davor, darauf trainiert wurden, Sprecher zu identifizieren. Vielleicht haben sie in dem darauffolgenden Teil einfach diese Aufgabe fortgesetzt ohne eine neue Herangehensweise in Betracht zu ziehen. Dennoch bleibt die Frage, weshalb die Hörer Sprecherinformationen, die nur optional genutzt werden konnten, so stark fokussierten und verwendeten. Diese Tatsache weist auf die wichtige Bedeutung von Sprecherinformationen in der gesprochenen Sprache hin. Sobald sie vorhanden sind, auch wenn nicht entscheidend, werden sie von den Hörern wahrgenommen und verarbeitet.

Zur Überprüfung des Einflusses des Anfangskonsonanten auf die Fixationspräferenzen, wurden die Gegenstandswörter einmal nach Artikulationsstelle (labial vs. alveolar) und einmal nach Artikulationsmodus (Nasale, Frikative, stimmhafte und stimmlose Plosive) aufgeteilt analysiert. Der Einfluss der Artikulationstelle wurde zu keinem Zeitpunkt in keiner Bedingung signifikant, es zeigte sich aber eine leichte Tendenz, dass der Sprecherkompetitor mit labialem Anfangskonsonanten stärker war als mit alveolerem Konsonanten. Dies könnte darauf hindeuten, dass sich Sprecher durch alveolare Konsonanten leichter identifizieren lassen als durch labiale und das deshalb die Kompetition schwächer war. Auch wenn einige Studien bereits den größeren Sprecherinformationsgehalt von alveolaren Lauten nachwiesen (z.B. [6]), konnte diese Hypothese in diesem Experiment nicht bestätigt werden.
Für die verschiedenen Artikulationsmodi ergaben sich jedoch einige signifikante Differenzen. In der Complex-Bedingung lösten Wörter mit Nasalen als Anfangskonsonant eine signifikant stärkere komplexe Kompetition aus als Wörter mit Frikativen oder Plosiven. Das bedeutet, dass wenn das gesprochene Wort mit einem Nasal begann, der (komplexe) Kompetitor stärker betrachtet wurde. Ob diese verstärkte Fixierung auf phonetische oder Sprecherinformationen zurückzuführen ist, kann aufgrund der Daten dieser Bedingung nicht beantwortet werden. In der Simplex-Bedingung hatte der Artikulationsmodus keinen signifikanten Einfluss auf den phonetischen Kompetitor, wohl aber auf den Sprecherkompetitor. Die Wörter mit initialen Frikativen lösten die stärkste Sprecherkompetition aus. Alle anderen Laute waren signifikant schwächer, wobei sich die Plosive nicht von den Nasalen unterschieden. Außerdem war die Differenz zwischen phonetischem und Sprecherkompetitor bei den stimmhaften Plosiven signifikant kleiner als bei allen anderen Lauten. Begann das Wort mit einem stimmhaften Plosiv, so schauten die Hörer gleichermaßen auf das phonetisch-ähnliche als auch auf das sprecher-ähnliche Wort. Daraus könnte man schließen, dass stimmhafte Plosive genauso viele phonetische wie Sprecherinformationen enthalten, sodass sie für beide Identifikationen gleichermaßen relevant sind. Alle anderen Laute lösten hingegen eine starke Sprecherkompetition und fast keine phonetische Kompetition aus. Entweder sind diese Laute phonetisch einfacher zu unterscheiden oder sie enthalten mehr Sprecherinformationen, sodass durch ihr Auftreten eher die Sprecher als die Wörter unterschieden werden können. Auf die gleiche Weise kann die stärkere Sprecherkompetition bei den Frikativen interpretiert werden: Entweder sind Sprecher anhand von Frikativen schwerer zu unterscheiden oder sie liefern viele Sprechermerkmale, weshalb sich die Hörer eher auf eine Identifikation des Sprechers verlegen, wenn frikativische Informationen vorhanden sind.

Die Ergebnisse vorangegangener Studien lassen letztere Erklärung wahrscheinlicher wirken ([143] [12]).
Da durch die starke Zersplitterung der Daten die Aussagekraft der statistischen Ergebnisse sinkt, sollten diese Ergebnisse - wenn auch signifikant - mit Vorsicht betrachtet werden. Besonders da in dem ersten Experiment keinerlei signifikante Einflüsse der Anfangskonsonanten gefunden wurden.

Die Jackknife-Analyse der Daten offenbart detaillierter die zeitliche Koordination von phonetischen und Sprechereffekten. Hier zählt nicht mehr der Maximalwert der Kompetitionseffekte, sondern es wird hauptsächlich die Geschwindigkeit, mit der das Maximum erreicht wird, untersucht. Dabei ist davon auszugehen, dass ein schneller steigender Effekt auch eine frühere Verarbeitung der Informationen, die zu diesem Effekt führen, voraussetzt. In der Simplex-Bedingung zeigt sich, dass der Sprechereffekt - vor allem im späteren Verlauf - dem phonetischen Effekt voraus geht. Das deutet darauf hin, dass am Anfang sowohl phonetische als auch Sprecherinformationen verarbeitet werden, während sich die Hörer im späteren Verlauf hauptsächlich auf die Sprecherinformationen konzentrieren. Außerdem erreicht der Sprechereffekt sein Maximum schneller als der phonetische, was darauf hinweist, dass Sprecherinformationen schneller verarbeitet werden als phonetische Informationen. Dies bestätigt auch der Vergleich des komplexen Kompetitors aus der Complex-Bedingung mit dem phonetischen und dem Sprecherkompetitor aus der Simplex-Bedingung. Die phonetische Kompetition ist nicht nur deutlich geringer, sondern auch langsamer als die komplexe Kompetition. Der Sprechereffekt hingegen unterscheidet sich nur anfangs von dem komplexen Effekt und ist in der Nähe des Maximums (ab 30 %) dann genauso schnell. Der Sprechereffekt scheint also einen deutlich höheren Anteil an dem schnellen Anstieg des komplexen Effekt zu haben als die phonetischen Informationen. Das heißt, die Hörer verarbeiten Sprecherinformationen schneller als phonetische, wobei zu Beginn des Wortes beide Informationstypen gleich schnell Eingang zu finden scheinen. Gegen Wortende (und nach Wortende) verschiebt sich dieses anfangs ausgeglichene Verhältnis aber signifikant zu Gunsten der Sprecherinformationen.
Die relativ gleichgroße Differenz zwischen phonetischem vs. komplexen Kompetitor und Sprecherkompetitor vs. komplexem Kompetitor zeigt, dass phonetische und Sprecherinformationen sich in ihrer Verarbeitungsgeschwindigkeit anscheinend nicht additiv verhalten. Vielmehr scheint eine Interaktion zwischen beiden Informationstypen vorzuliegen, welche nur im Zusammenwirken zu der schnellen Aufnahme und Verarbeitung der komplexen Infor-

mationen führt. Durch das Wissen über Sprechermerkmale konnten phonetische Merkmale besser aufgenommen und genutzt werden und umgekehrt. Bereits in vorherigen Studien wurde die Interaktion zwischen phonetischen/lexikalischen und sprecherspezifischen Merkmalen nachgewiesen ([14] [259]). Somit fügen sich die hier gefundenen Ergebnisse sehr gut in den bisherigen Wissenstand ein.

5.4. Generelle Diskussion

Jedes Sprachsignal enthält immer Informationen über die Wörter (lexikalische bzw. phonetische Informationen) und über den Sprecher (stimmliche Merkmale). Im vorhergehenden Kapitel 4 wurde bereits gezeigt, dass sich der phonetische Inhalt (verschiedene Konsonanten) auf den Sprecherinformationsgehalt das Segments auswirkt. Je nach Art der vorhandenen Informationen („statische + dynamische“ vs. „nur statische“) lieferten dabei eher die Nasale und Plosive oder die Frikative mehr Sprecherinformationen. Die Faktoren Wort und Sprecher interagieren miteinander. Da aber im Gegensatz zur Worterkennung noch wenig über den Prozess der Sprechererkennung bekannt ist, sollte dieser Vorgang in den vorgestellten Experimenten untersucht werden. Bisher war unklar, ob sich die Sprechererkennung ähnlich verhält wie die Worterkennung (lexikalische/phonetische Kompetition bis zur Disambiguierung des Wortes) und ob sich Sprecherkompetition finden ließe. Neben der Sprecheridentifikation an sich, war auch wenig über die zeitliche Koordination von Sprecher- und Worterkennung bekannt. Nutzen die Hörer zuerst die phonetischen/lexikalischen Merkmale des Wortes, um dann den Sprecher erkennen zu können oder verwenden sie die Sprechermerkmale in der Stimme, um dann das Wort besser verstehen zu können? In diesem Kapitel sollte einer Antwort auf diese Fragen näher gekommen werden.
Dazu wurde ein Visual-World Eye-Tracking Experiment durchgeführt, bei dem Hörer Sprecher-Gegenstand-Kombinationen identifizieren mussten. Die Ergebnisse zeigen, dass die Hörer sehr gut in der Lage sind, die entsprechenden Sprecher-Wort-Targets zu identifizieren. Sollte im ersten Experiment der Faktor Sprecher maximiert werden, um einen Effekt dieses Faktors auf das Fixationsverhalten der Versuchsteilnehmer messbar zu machen, so wurde die Prominenz des Sprechers im zweiten Experiment reduziert. Trotzdem wird in beiden Experimenten in allen Bedingungen eine signifikante Sprecherkompetition gefunden. Diese Kompetition zeigt, dass es schwierig ist, die Sprecher zu identifizieren, sowohl bei 2 als auch bei 4 Sprechern. Außerdem demonstriert sie, dass die Hörer auf die Sprechermerkmale achten und sie verarbeiten. Im zweiten Experiment wird außer der Sprecherkompetition

auch Gender-Kompetition nachgewiesen. Das heißt, dass die Hörer bei einem männlichen Target den anderen männlichen Sprecher auch stärker betrachten als die weiblichen Sprecher und umgekehrt. Dabei zeigt sich, dass bei größerer Ähnlichkeit der Stimmen innerhalb eines Geschlechts die Gender-Kompetition stärker ist; d.h., die Sprecher sind schwerer zu unterscheiden.
Phonetische Kompetition kann nicht in allen Bedingungen gefunden werden. Ist der phonetische Overlap am Wortanfang zu gering (nur C-Overlap), dann ist die Worterkennung so einfach, sodass keine phonetische Kompetition entsteht. Wird der Overlap allerdings länger, verstärkt sich die phonetische Kompetition. Im Falle von CV-Overlap ist sie genauso stark wie die Sprecherkompetition und bei vollständiger Überlappung des ganzen Wortes wird sie signifikant stärker als die Sprecherkompetition.
Interessanterweise wird in allen Bedingungen in beiden Experimenten Sprecherkompetition gefunden, selbst wenn die Sprecherinformationen zur Targetidentifizierung nicht nötig (Experiment 2) oder ausreichend (Experiment 1 Gegenstandsbedingung) sind. In der Sprecherbedingung von Experiment 1 ist es nicht verwunderlich, dass Sprecherinformationen verwendet werden. Da das Wort vollständig ambig ist, kann man nur durch die zusätzliche Verwendung von Sprecherinformationen das Target finden. Aber auch in der Gegenstandsbedingung wird offenbar auf den Sprecher geachtet, obwohl es (a) nicht nötig und (b) nicht ausreichend ist für die Targeterkennung. Und auch wenn sich (durch eine höhere Anzahl von Sprechern) die Sprechererkennung erschwert, bringt es die Hörer nicht dazu, die Sprecherinformationen zu ignorieren, denn in diesem Fall hätten wir in Experiment 2 keine Sprecherkompetition finden dürfen. Dieses Verhalten der Versuchspersonen spricht für den integralen Anteil den Sprecherinformationen als nicht-lexikalische Informationen auf die Wahrnehmung und Verarbeitung von gesprochener Sprache haben.
Neben der absoluten Stärke der phonetischen und Sprecherkompetition wurde mittels eine Jackknife-Analyse auch die Geschwindigkeit des Anstiegs der Kompetitionseffekte bestimmt und verglichen. Dabei zeigt sich in beiden Experimenten, dass die Sprecherkompetition leicht schneller steigt als die phonetische Kompetition. Dieser Unterschied wird meist erst ab der Hälfte (ca. 40 %) des Maximaleffekts signifikant, ist tendenziell aber meist auch vorher schon vorhanden. Dieser zeitliche Vorsprung deutet darauf hin, dass die Hörer die Sprecherinformationen früher aufnehmen und verarbeiten als die phonetischen Informationen.
Die Analyse des Einflusses von Artikulationsstelle und -modus brachte keine eindeutigen Ergebnisse. Im ersten Experiment bewirkt keiner der Faktoren signifikante Unterschiede in

einem der Kompetitoren. Im zweiten Experiment hat die Artikulationsstelle ebenfalls keinen Einfluss, während der Artikulationsmodus einige signifikante Differenzen hervorrief. Es zeigt sich, dass frikativ-initiale Wörter eine signifikant stärkere Sprecherkompetition auslösen als Wörter mit Nasalen oder Plosiven als Anfangskonsonanten. Mit Bezug auf vorherige Studien ([143] [12]) erscheint es am wahrscheinlichsten, dass Frikative mehr Informationen über den Sprecher enthalten als andere Laute und die Hörer bei diesen Lauten deshalb stärker auf die Sprechermerkmale achten. Dieser erhöhte Fokus führt dann zu der stärkeren Betrachtung der Sprecher. Außerdem bewirken die nasal-initialen Wörter in der Complex-Bedingung (Experiment 2) eine stärkere komplexe Kompetition. In dieser Bedingung lässt sich aber leider nicht zwischen dem Anteil der phonetischen und der Sprecherinformationen an der Kompetition unterscheiden. In der Literatur wurde Nasalen häufig auch eine besondere Eignung zur Sprecheridentifikation zugeschrieben ([295] [68] [12]). Allerdings bewirken sie in der Simplex-Bedingung (Experiment 2) genauso wenig Sprecherkompetition wie die Plosive, weshalb aufgrund der Ergebnisse dieses Experiments nicht davon ausgegangen werden kann, dass Nasale besonders viele sprecherspezifische Informationen enthalten.
Die Ergebnisse lassen eine wichtige Frage offen: Warum wurden so stark (immer signifikant) und konstant (in allen Bedingungen) Sprecherinformationen verwendet, selbst wenn sie gegenüber den phonetischen Informationen keinen Vorteil boten? Eventuell ist dieses Ergebnis auf unterschiedliche Strategien der Hörer zurückzuführen. Da in der Analyse nur über Versuchspersonen gemittelte Fixationsdaten verwendet werden (konnten), lässt sich nicht sagen, ob manche Teilnehmer vielleicht eher auf die Sprecher und andere eher auf die Gegenstände bzw. Wörter geachtet haben. Eine solche Aufteilung der Hörer könnte zumindest die Ergebnisse von Experiment 2 erklären. Es würde aber noch nicht (vollständig) erklären, weshalb auch in der Gegenstandsbedingung von Experiment 1 Sprecherinformationen verwendet wurden, obwohl sie für die Targetidentifzierung nicht ausreichend waren.
Die gefundene frühe und konstante Wahrnehmung und Verarbeitung von Sprecherinformationen spricht stark für deren Repräsentation im menschlichen mentalen Lexikon im Sinne eines exemplarischen Modells der Sprachperzeption ([133] [135] [231] [232]). Dort werden neben den lexikalischen, inhaltlichen Anteilen der Wörter auch sprecherspezifische, nicht-lexikalische Informationen gespeichert. Würden Sprecherinformationen bei der Sprachwahrnehmung ignoriert oder herausgefiltert werden, so wäre es nicht plausibel, in allen Bedingungen dieser Experimente Sprecherkompetition zu finden. Daraus folgt, dass diese Informationen aufgenommen und verwendet werden. Da bereits in anderen Studien gezeigt wurde, dass Sprecherinformationen die Worterkennung beeinflussen, erscheinen diese

Annahme und die neuen Ergebnisse plausibel und überzeugend.
Allerdings müsste das Modell so spezifiziert werden, dass es eine gewisse Abstraktion zulässt. Die Hörer waren offenbar in der Lage, von den sprecherindividuellen Merkmalen zu abstrahieren, sodass sie die beiden verschiedenen Geschlechter erkannten und unterscheiden konnten. Die gefundene Gender-Kompetition zeigt, dass die Hörer zuerst das Geschlecht des Sprechers erkannten und erst danach den konkreten Sprecher innerhalb des Geschlechts identifizierten. Das deutet darauf hin, dass die Hörer über die Merkmale, die allen männlichen bzw. allen weiblichen Stimmen gemeinsam sind, generalisierten und so eine Geschlechterunterscheidung vornahmen. Zwar ist aus der Literatur bekannt, dass eine Geschlechteridentifikation nicht zwingend einer Sprecheridentifikation vorausgehen muss [83], aber in dieser Studie konnte diese Reihenfolge beobachtet werden. Hätten die Hörer direkt den Sprecher identifiziert, ohne vorher das Geschlecht zu erfassen, hätte keine Gender-Kompetition auftreten dürfen. Somit sprechen diese Ergebnisse sowohl für eine Repräsentation von sprecherspezifischen Merkmalen im mentalen Lexikon, die aber aufgrund bestimmter Gemeinsamkeiten zu Gruppen (wie z.B. Männer und Frauen) zusammengefasst werden können. Diese Struktur ermöglicht es den Hörern, sehr schnell die relevanten Informationen in einer Sprachäußerung zu erfassen und zu verarbeiten, wobei neben den inhaltlich wichtigen phonetischen (lexikalischen) Merkmalen auch nicht-lexikalische Merkmale wie Sprechermerkmale Beachtung finden.

5.5. Zusammenfassung

Es wurden zwei Experimente unter Verwendung des Visual-World Eye-Tracking Paradigmas zur Verarbeitung von Sprecherinformationen und deren zeitliche Koordination zur Verarbeitung von phonetischen Informationen durchgeführt. Es zeigte sich, dass mit zunehmender Anzahl an Sprechern die Schwierigkeit der Sprecheridentifikation stieg. Außerdem wurde gefunden, dass die Hörer zuerst das Geschlecht der Sprecher erkennen, bevor sie den konkreten Sprecher identifizieren. Je ähnlicher sich die Stimmen innerhalb eines Geschlechts sind, desto schwieriger ist die Identifikation des Sprechers.
In beiden Experimenten identifizierten die Teilnehmer die visuellen Referenten der Sprecher-/Gegenstand-Kombinationen mit hoher Genauigkeit. Dabei schien in Experiment 1 die Sprecherbedingung (zwei Gegenstände und zwei Sprecher am Bildschirm) schwieriger gewesen zu sein, wahrscheinlich weil die Sprecher-Gegenstand-Kombinationen in dieser Bedingung ambiger und somit schwieriger zu unterscheiden waren.

Im ersten Experiment wurden nur zwei männliche Sprecher verwendet, was die Unterscheidbarkeit zwar etwas schwieriger als die Unterscheidung zwischen männlicher und weiblicher Stimme machen sollte, aber immer noch einfach genug, dass die Hörer Sprecherinformationen verwenden würden. Im zweiten Experiment wurde die Unterscheidbarkeit der Sprecher etwas erschwert, indem jeweils zwei Sprecher beider Geschlechter verwendet wurden. Dadurch wurde die Sprecheridentifikation schwieriger und die Wahrscheinlichkeit, dass Hörer diese Informationen zur Targetidentifikation nutzen würden, verringert. Dennoch zeigte sich für beide Experimente eine (relativ) starke Sprecherkompetition, die dafür spricht, dass die Hörer in beiden Experimenten Sprecherinformationen verwendeten. Dabei war die Sprecherkompetition in Experiment 1 (2 Sprecher) leicht schwächer als in Experiment 2 (4 Sprecher), was wahrscheinlich an der einfacheren Unterscheidbarkeit von 2 im Gegensatz zu 4 Sprechern lag. Die phonetische Kompetition hing stark vom phonetischen Overlap am Wortanfang ab: War der Overlap zu gering (nur der Konsonant), trat keine phonetische Kompetition auf, wurde der Overlap länger (Konsonant und Vokal) stieg die phonetische Kompetition auf die gleiche Höhe wie die Sprecherkompetition und wurde signifikant. Überlappten die Wörter phonetisch vollständig (identische Wörter), so war sie sogar signifikant stärker als die Sprecherkompetition und blieb bis zum Trialende erhalten.
Diese Resultate werfen die Frage auf, weshalb die Hörer so stark und konstant Sprecherinformationen verwendeten, selbst wenn sie gegenüber den phonetischen Informationen keinen Vorteil (oder sogar einen Nachteil) boten. Möglicherweise verursachten unterschiedliche Entscheidungsstrategien zwischen den Hörern diesen Effekt, was aber die Ergebnisse auch nicht vollständig erklärt.
Die Jackknife-Analyse der Daten zeigte, dass die Sprechereffekte schneller/früher stiegen als die phonetischen Effekte. Diese Tatsache spricht dafür, dass Sprecherinformationen tatsächlich etwas schneller aufgenommen und verarbeitet werden als phonetische Informationen. Dabei waren beide Informationstypen am Anfang gleich schnell, aber ab der knappen Hälfte des Maximaleffekts überholten dann die Sprecherinformationen die phonetischen.
Anschließend wurde diskutiert, inwieweit die Ergebnisse ein exemplarbasiertes Modell der Sprachwahrnehmung unterstützen und welche Spezifikationen für ein solches Modell noch nötig und sinnvoll wären.
Nachdem gezeigt wurde, dass Visual-World Eye-Tracking nicht nur zur Verfolgung der Verarbeitung phonetischer (bzw. lexikalischer) Information verwendet werden kann, könnte es ein nützliches Werkzeug sein, um in Zukunft viele neue Einblicke in den Prozess der Sprechererkennung im Verhältnis zur Worterkennung zu gewinnen.

6. Zusammenfassung

Die gesprochene Sprache enthält neben den linguistischen (inhaltlichen) Informationen auch extra-linguistische oder indexikalische Informationen. Zu letzteren gehören zum Beispiel Informationen über die Identität des Sprechers einer Äußerung. Diese Sprecherinformationen werden bei der Wahrnehmung von Sprache ebenfalls verarbeitet. Da jeder Mensch aufgrund seiner speziellen anatomisch-physiologischen und verhaltensbasierten Faktoren eine einzigartige Stimme hat, können wir Personen mittels dieser Merkmale identifizieren und unterscheiden. Dies kann sowohl anhand von akustischen als auch von perzeptiven Stimmmerkmalen geschehen. In der automatischen Sprechererkennung werden ausschließlich akustisch messbare Sprechermerkmale verwendet, die maschinell extrahiert und verglichen werden können [149]. Im Bereich der forensischen Phonetik werden sowohl akustische als auch perzeptive Sprecherspezifika genutzt. So werden z.B. sowohl die Vokalformanten gemessen (zunehmend auch automatisch) als auch der auditive Höreindruck der Stimmqualität bestimmt [98]. Des weiteren können die Sprechermerkmale auch einen Beitrag zum Verständnis der menschlichen Sprachwahrnehmung leisten. Durch das Wissen, wie wir Sprechermerkmale im Prozess der Sprachwahrnehmung und -verarbeitung nutzen, lernen wir, auf welche Weise lexikalische Einheiten (z.B. Wörter) im mentalen Lexikon gespeichert werden (z.B. [48]).

Um herauszufinden, welche Parameter sich generell eignen, um Sprecher voneinander zu unterscheiden, wurde in **Kapitel 3** eine akustische Analyse einiger konsonantischer Merkmale des Deutschen durchgeführt. Dabei lag der Fokus vor allem auf den Merkmalen von Nasalen und Frikativen, wobei aber auch Plosive und Vokale mit betrachtet wurden. Es wurde außerdem der Einfluss von Artikulationsmodus und -stelle bestimmt sowie auch der Einfluss der Sprachaufnahmequalität. Es zeigt sich, dass der Artikulationsmodus der Nasale und Frikative sprecherspezifischer ist als der der Plosive und alveolare Konsonanten sprecherspezifischer als labiale. Eine Verminderung der Qualität des Sprachsignals durch die Übertragung per Telefon wirkt sich auf alle untersuchten Artikulationsmodi gleichermaßen schlecht aus. Die Sprecherspezifität der Laute wird durchschnittlich ca. auf die Hälfte

reduziert.
In **Kapitel 4** wurde die Sprecherspezifität der Konsonanten dann perzeptiv untersucht. Dafür wurde ein Sprecherdiskriminationstest durchgeführt, bei dem Hörer mehrere Sprecher anhand verschiedener Stimuli unterscheiden mussten. Es zeigt sich nicht nur, dass manche Konsonanten mehr Sprecherinformationen enthalten als andere, sondern dass diese sich auch in unterschiedlichen zeitlichen Bereichen des Segments befinden. So enthalten die Nasale und Plosive einen Großteil ihrer Sprecherinformationen in den Transitionen zu den umliegenden Vokalen, während die Frikative die Sprecherinformationen eher im Konsonanten selbst enthalten.
Anschließend wurde in **Kapitel 5** die zeitliche Koordination bei der Verarbeitung von phonetischen (lexikalischen) und Sprecherinformationen untersucht. Dafür wurde mit Hilfe eines Visual World Eye-Tracking Experiments analysiert, wie der Entscheidungsprozess der Versuchspersonen bei der Sprecher- und Wortidentifikation abläuft. Dabei spielen verschiedene Faktoren eine Rolle, wie die Anzahl der Sprecher und ihre Ähnlichkeit als auch die phonetische Überlappung der Stimuli. In dieser Untersuchung scheinen die Sprecherinformationen einen leichten zeitlichen Vorsprung vor den phonetischen Informationen in der Sprachwahrnehmung zu haben.
Die neu gewonnenen Erkenntnisse sollen in diesem Kapitel mit Hinblick auf die Literatur kritisch diskutiert und zusammengefasst werden. Daraus folgende Implikationen und weitere Forschungsfragen sollen kurz umrissen und ein Ausblick gegeben werden.

6.1. Interaktion von phonetischen und Sprecherinformationen

Die verschiedenen Informationstypen (linguistische und extra-linguistische) im Sprachsignal interagieren auf unterschiedliche Weise miteinander. So tragen beispielsweise bestimmte phonetische Segmente mehr Informationen über die Identität eines Sprechers als andere [14]. Klassischerweise waren das hauptsächlich Vokale, da sie aufgrund ihrer Stimmhaftigkeit mehr Sprecherinformationen (z.B. die Grundfrequenz eines Sprechers) transportieren können ([177] [187] [289] [268]). Einige Studien zeigten aber bereits, dass Konsonanten ebenfalls relevante Sprecherinformationen beinhalten und somit einen wichtigen Beitrag zur Sprecheridentifikation und -diskrimination leisten können ([101] [295] [68] [179] [16] [12] [14] [15] [142] [143] [72]).
Die durchgeführten Studien hatten als Grundlage aber fast immer englisches (oder japa-

nisches) Sprachmaterial. Für die deutsche Sprache gab es bisher allerdings nur wenige Untersuchungen (z.B. [76] [179]). Da die Artikulation der Konsonanten zwischen den Sprachen zwar ähnlich, aber nicht identisch ist (bzw. das Lautinventar sich unterscheidet), lohnt sich ein genauerer Blick auf die Sprecherspezifität deutscher Konsonanten. Dafür wurden ausgewählte Konsonanten unterschiedlicher Artikulationsmodi und -stellen akustisch und perzeptiv auf ihre Sprecherspezifität hin untersucht. Die Ergebnisse dieser Untersuchungen sollen im Folgenden zusammenfassend dargestellt und diskutiert werden.

6.1.1. Einfluss des Artikulationsmodus

In der Literatur hatten sich besonders die Artikulationsmodi der Nasale und Frikative als sprecherspezifisch erwiesen ([101] [295] [68] [16] [12] [14] [15] [142] [72]), während die Plosive häufig als wenig sprecherspezifisch beschrieben wurden ([141] [41] [114] [68] [76] [140] [15]); mit wenigen Ausnahmen (z.B. [82]). Daher konzentrierte sich die Untersuchung der deutschsprachigen Konsonanten hauptsächlich auf Nasale und Frikative. Diese wurden zunächst akustisch und anschließend auditiv-perzeptiv analysiert.

Akustische Analyse

Für die akustische Analyse der Konsonanten wurden die spektralen Momente als akustische Merkmale ausgewählt. Diese werden üblicherweise zur phonetischen Klassifikation von Frikativen verwendet [93]. Es stellt sich die Frage, ob Parameter, die Informationen über die phonetische Identität eines Lautes enthalten, auch über Informationen über die Sprecheridentität verfügen. Es ist anzunehmen, dass in Bereichen mit einer hohen Dichte an akustischen Informationen sowohl phonetische als auch Sprecherinformationen vorhanden sind. Daher ist es nicht abwegig, die gleichen Parameter zur Analyse der Sprechermerkmale zu verwenden wie für phonetische Merkmale. In der Tat wurde im Bereich der automatischen Sprechererkennung bereits gezeigt, dass die Mel-Frequency Cepstral Coefficients (MFCC), die schon lange in der Spracherkennung Anwendung finden, sich auch zur automatischen Sprechererkennung sehr gut eignen ([148] [149]). In der Untersuchung von [143] wurde bereits demonstriert, dass die spektralen Momente von /s/ sich zwischen Sprechern relativ stark unterscheiden.
Für Nasale wurden bisher eher Formanten und Antiformanten als spektrale Momente verwendet ([142] [72]). Allerdings gibt es keinen Grund, warum sich die spektralen Unterschiede zwischen den verschiedenen Nasalen nicht auch in ihrem parametrisierten Spektrum zeigen

sollten. Daher werden in dieser Arbeit für die Nasale ebenfalls die spektralen Momente zur Messung der Sprecherspezifität verwendet.
Als Maß für die Sprecherspezifität wurde die F-ratio verwendet, welche das Verhältnis von Inter- zu Intra-Sprecher-Variation angibt. Je höher der F-ratio-Wert eines Merkmals, desto stärker variiert es zwischen verschiedenen Sprechern und desto weniger innerhalb eines Sprechers. Somit spricht ein hoher (signifikanter) F-ratio-Wert für eine hohe Sprecherspezifität eines Merkmals.
Die Ergebnisse der akustischen Analyse zeigen, dass vor allem die Nasale /m, n/ und die stimmlosen Frikative /f, s/ sehr sprecherspezifisch sind. Die hohe Sprecherspezifität der Nasale lässt sich artikulatorisch durch die Anatomie und Physiologie des Nasenraumes erklären. Dieser ist in seiner Form fest, wodurch keine große Intra-Sprecher-Variation entstehen kann und zwischen verschiedenen Sprechern unterschiedlich, sodass die Inter-Sprecher-Variation groß ist.
Die guten Ergebnisse von /s/ lassen sich dadurch erklären, dass Sibilanten viel Energie im Spektrum aufweisen. Nicht nur weil sie mit einem großen vorderen Hohlraum produziert werden, sondern vor allem weil der Luftstrom auf die Zähne trifft, was eine hochfrequente und energiereiche Turbulenz erzeugt [230].
Auffällig ist in den Ergebnissen, dass die stimmlosen Frikative durchgehend sprecherspezifischer waren als die stimmhaften Frikative. Dies liegt möglicherweise an der schwächeren Intensität stimmhafter Frikative gegenüber stimmlosen. Da die Stimmlippen geschlossen gehalten werden müssen, um ein Schwingen zu ermöglichen, kann die Luft nicht so schnell durch die Glottis strömen. Dies führt zu einem langsameren Luftstrom und somit zu einem schwächeren Friktionsgeräusch [230]. Das könnte erklären, weshalb die stimmlosen Frikative /f, s/ sprecherspezifischer sind als die stimmhaften Frikative /v, z/.
Obwohl auch die Plosive ein signifikantes Maß an Sprecherspezifität erreichen, erweisen sie sich erwartungsgemäß als der am wenigsten sprecherspezifische Artikulationmodus. Begründet liegt das möglicherweise in ihrer sehr kurzen Dauer, der Verschlussphase, während der keine akustischen Informationen vorhanden sind und dem sehr kurzen Burst, der wenig Raum für sprecherindividuelle Merkmale bietet.

Perzeptive Analyse

An die akustische Analyse schloss sich die perzeptive Untersuchung der Konsonanten an. Dazu wurden Stimuli mit verschiedenen Konsonanten in einem /a/-Kontext erstellt. Anhand dieser Stimuli mussten Hörer Sprecher in zwei AX-Diskriminationstests unterscheiden. Im

ersten Test waren sowohl statische als auch dynamische sprecherspezifische Informationen vorhanden, im zweiten nur noch statische. Diese Informationsverringerung führte, wie erwartet, zu einer geringeren Sprecherdiskriminationsrate im zweiten Experiment (62 %) im Vergleich zum ersten (83 %). Dennoch lag die Erkennungsrate in beiden Experimenten über dem Zufallsniveau. Daraus lässt sich schlussfolgern, dass die Hörer (bis zu einem bestimmten Grad) in der Lage sind, die Sprecher auch ausschließlich anhand der statischen Konsonantinformationen zu unterscheiden. Dieser Unterschied in der Diskriminationsleistung verdeutlicht aber auch, dass ein beträchtliches Maß an Sprecherinformationen in den dynamischen Bereichen der Vokalen bzw. der Vokaltransitionen liegt.

Im ersten Experiment, in Anwesenheit der transitionalen und vokalischen Sprecherinformationen, erreichten die Nasale und Plosive eine höhere Sprecherdiskriminationsfähigkeit als die Frikative. Scheinbar profitieren die Nasale und Plosive stärker von den dynamischen Transitionsinformationen als die Frikative. Dies mag daran liegen, dass Nasale und Plosive sehr schnelle und abrupte Transitionen haben, was mit einer starken Energieveränderung einhergeht. Dynamische Merkmale wie diese Energieveränderungen enthalten typischerweise viele akustische Informationen ([147] [95] [151]), die bei der Spracherkennung und vielleicht auch bei der Sprechererkennung hilfreich sind. Die Frikative hingegen haben eher langsame Transitionen, da die Artikulation eine hohe Präzision verlangt und daher stärker kontrolliert werden muss. Aufgrund dieser langsameren Transitionen enthalten die Frikative vermutlich nicht so viele akustische Informationen in diesen Bereichen und somit auch weniger Sprecherinformationen.

Im zweiten Experiment hingegen, in dem die Vokale und Transitionen keine Sprecherinformationen mehr enthalten, zeigen die Frikative eine höhere Sprecherdiskriminationsfähigkeit als die Nasale und Plosive. Durch die Neutralisation der Sprecherinformationen in den Transitionen wurde die Sprecherspezifität aller Konsonanten reduziert, die der Nasale und Plosive aber deutlich stärker als die der Frikative. Offenbar enthalten die Frikative in ihrem statischen konsonantischen Teil mehr Sprecherinformationen (und weniger in den Transitionen) als es bei den Nasalen und Plosiven der Fall ist. Diese Erklärung erscheint plausibel, wenn man annimmt, dass phonetische und Sprecherinformationen beide in Bereichen hoher akustischer Merkmalsdichte liegen. Denn da die Frikative zum Beispiel in ihrem Frequenzschwerpunkt viele phonetische Informationen (über die Artikulationsstellen) einschließen, beinhalten diese Merkmale wahrscheinlich auch einige Sprecherinformationen. Daher enthalten sie auch ohne ihre Transitionen noch relativ viele sprecherspezifische Merkmale.

Unerwartet ist, dass die Nasale in keinem der Experimente sprecherspezifischer sind als die Plosive, obwohl dies nicht den Berichten der Literatur entspricht ([12] [68]). Allein die hohe Sprecherspezifität von /t/ wurde bereits belegt ([82]). Andererseits ist auch bekannt, dass Nasale und Plosive aufgrund ihrer Artikulation gleichermaßen stark von ihren Transitionen abhängen [230]. Dahingehend wäre es also nicht verwunderlich, wenn sich die Neutralisation der dynamischen Transitionsinformationen gleichermaßen stark auf die Sprecherspezifität beider Artikulationsmodi auswirken würde. Eine weitere Ursache für den starken Einfluss der Transitionen könnte das offene /a/ als Kontextvokal gewesen sein, da besonders offene und hintere Vokale in den Transitionen sehr viele Informationen über die Artikulationsstelle enthalten, während das bei vorderen Vokalen wie /i/ nur in geringem Maß der Fall ist [247].

Vergleich von Akustik und Perzeption

Wenn man die Ergebnisse aus der akustischen und perzeptiven Analyse vergleicht, zeigen sich einige Unterschiede. Während die akustischen Untersuchungsergebnisse eindeutig den Berichten aus der Literatur entsprechen (Nasale und Frikativ /s/ am sprecherspezifischsten; Plosive wesentlich sprecherunspezifischer), zeigte das Perzeptionsexperiment überraschendere Resultate. Die Plosive sind im ersten Perzeptionsexperiment sehr viel sprecherspezifischer als erwartet; möglicherweise wegen der zusätzlichen Sprecherinformationen in den Transitionen. Die Nasale erweisen sich im ersten Perzeptionsexperiment als sehr sprecherspezifisch; genauso, wie es die akustischen Messungen vermuten ließen. Die Frikative sind zwar signifikant sprecherunspezifischer als die Nasale, weisen aber dennoch relativ hohe Werte auf, wie auch in den akustischen Untersuchungen.
Im zweiten Teil des Perzeptionsexperiments ergibt sich für die Nasale nur eine sehr geringe Sprecherspezifität, was vermutlich an den fehlenden dynamischen Informationen der Transitionen und möglicherweise einem Mismatch der „Nasenrauminformationen“ im Konsonanten und in den Vokalen liegt. Daraus lässt sich schlussfolgern, dass entweder bei der akustischen Analyse aufgrund der (automatischen) Segmentierung ebenfalls Transitionsinformationen in den Nasalen vorhanden sind oder dass Hörer auf andere auditive Merkmale zur Sprecherdiskrimination achten als in der akustischen Analyse untersucht wurden. In letzterem Fall resultiert dann die hohe Sprecherspezifität der Nasalstimuli im ersten Experimentteil wahrscheinlich aus den zusätzlichen Vokalinformationen. Betrachtet man die Literatur, fällt auf, dass einzelne Nasale fast nur in akustischen Untersuchungen zur Sprecheridentifikation verwendet wurden ([101] [295] [142] [72]), während für perzeptive Sprecheridentifikation immer Nasal-Vokal-Silben verwendet wurden ([12] [14]). Beide Untersuchungsvarianten

kamen zu ähnlichen Ergebnissen wie die akustische Analyse in dieser Arbeit und das erste Perzeptionsexperiment (mit Nasal- und Vokalinformationen). Das deutet darauf hin, dass akustische Messverfahren die statischen konsonantischen Sprechermerkmale der Nasale besser erfassen und verwenden können als menschliche Hörer dazu in der Lage sind. Letztere benötigen zur Sprecheridentifikation (-diskrimination) auch die Sprecherinformationen aus den vokalischen Transitionen. Diese Erklärungshypothese müsste in weiterführenden Experimenten bestätigt werden, steht aber mit den bisherigen Resultaten in Einklang.
Die Plosive zeigen ohne die Transitionsinformationen ähnlich niedrige Werte in ihrer Sprecherspezifität wie die Nasale. Nimmt man die Tatsache hinzu, dass die Plosive mit den Transitionsinformationen genauso hohe Sprecherspezifitätswerte erreichen wie die Nasale, so scheinen beide Artikulationsmodi in gleichem Maße von den Sprecherinformationen in ihren Transitionen abhängig zu sein. Allerdings zeigen Plosive - im Gegensatz zu Nasalen - bisher weder in akustischen noch in perzeptiven Untersuchungen mit Plosiv-Vokal-Stimuli sprecherspezifische Eigenschaften ([141] [41] [114] [68] [76] [140] [15]). Die einzige Ausnahme bildet die Studie von [82], welche für den alveolaren Plosiv /t/ eine hohe Sprecherspezifität ermittelte. Berücksichtigt man hauptsächlich letzteres Ergebnis und die aktuelle Perzeptionsstudie, so würde sich für die unterschiedliche Sprecherspezifität der Plosive folgende Erklärung anbieten: Plosive in Isolation bieten in ihren akustischen Eigenschaften wenig sprecherspezifische Informationen, können aber gemeinsam mit dynamischen Transitionsinformationen in perzeptiven Untersuchungen relevante Sprecherinformationen liefern. Jedenfalls lässt sich die Behauptung, dass Plosive in keinem Kontext (ob mit oder ohne Vokale) signifikant zur Sprecheridentifikation (-diskrimination) beitragen können, in Anbetracht der neuen Ergebnisse (aus [82] und der aktuellen Studie) nicht halten.
Die Frikative zeigen die stärkste Übereinstimmung in ihrer Sprecherspezifität zwischen der akustischen und der perzeptiven Untersuchung. Allerdings erzielen die stimmhaften Frikative in der perzeptiven Analyse sprecherspezifischere Ergebnisse als in den akustischen Messungen. Möglicherweise profitieren die stimmhaften Frikative stärker von den Sprecherinformationen in den Transitionen als die stimmlosen, was sie in den perzeptiven Studien mit Frikativ-Vokal-Silben informativer macht als in den akustischen Analysen der statischen konsonantischen Sprechermerkmale. Insgesamt enthalten die Frikative im Vergleich zu den anderen Artikulationsmodi offenbar die meisten Sprecherinformationen direkt im statischen Bereich des Konsonanten (statt in den Transitionen), sodass sie auch im zweiten Perzeptionsexperiment immer noch relativ sprecherspezifisch sind. Somit stimmen die akustischen und perzeptiven Resultate im Artikulationsmodus der Frikative am besten überein. Vielleicht

liegt diese höhere Übereinstimmung in der Wahl der akustischen Merkmale, die analysiert wurden, begründet. Dafür wurden nämlich die spektralen Momente verwendet, welche typische Merkmale zur Bestimmung der Artikulationsstelle bei Frikativen sind. Für alle anderen Artikulationsmodi wurden sie bisher nicht angewendet. Unter Umständen zeigt sich darin, die bessere Eignung dieser spektralen Merkmale für die Frikative im Vergleich zu den Nasalen und Plosiven. Um diese Annahme zu bestätigen, wären weitere Studien mit verschiedenen akustischen Parametern von Nöten. Den zu untersuchenden Faktor könnte man als „Eignung des akustischen Parameters für den Konsonantentyp“ bezeichnen. Es ist fraglich, ob man akustische Merkmale finden kann, die für alle Konsonanten gleichermaßen gut geeignet sind. Andernfalls müsste man für jeden Artikulationsmodus andere Parameter verwenden, was deren Vergleichbarkeit beeinträchtigen würde.
Die Ergebnisse zeigen deutlich die unterschiedliche Eignung der statischen konsonantischen Sprechermerkmale in Abhängigkeit vom Analyseintervall (Konsonant vs. Konsonant und Vokal), der Untersuchungsmethode (akustisch vs. perzeptiv) und dem gewählten akustischen Merkmal (und seiner Eignung für den Konsonanten).

6.1.2. Einfluss der Artikulationsstelle

In einigen perzeptiven Studien hatte sich abgezeichnet, dass alveolare Konsonanten sprecherspezifischer sind als labiale (z.B. [7]). Mit Fokus auf diese Differenz sollen die akustische und perzeptive Analyse vergleichend gegenüber gestellt werden.

Akustische Analyse

In der durchgeführten akustischen Analyse bestätigte sich dieser Fakt für die Nasale und Frikative. Um zu überprüfen, ob dieser Unterschied in der Sprecherspezifität zwischen den Artikulationsstellen systematisch ist, wurden zusätzlich auch die Plosive als weiterer Artikulationsmodus analysiert. Die höhere Sprecherspezifität der alveolaren Laute bestätigt sich für alle Artikulationsmodi (Nasale, Frikative und Plosive). Für die Frikative /s, f, z, v/ lassen sich die Unterschiede in der Sprecherspezifität hauptsächlich mit der höheren Energie von alveolaren Frikativen gegenüber labialen erklären.
Bei Nasalen ist die Ursache für diesen Unterschied nicht so eindeutig. Nasale enthalten viele Informationen über die Artikulationsstelle in ihren Transitionen. Daher wäre anzunehmen, dass diese Unterschiede ebenfalls Differenzen in der Sprecherspezifität der Artikulationsstellen verursachen. Allerdings ist unklar, inwieweit die Transitionen der Nasale in dieser

Untersuchung mitgemessen wurden. Da die akustischen Merkmale immer zum Segmentmittelpunkt bestimmt wurden, ist es eher unwahrscheinlich, dass Transitionsinformationen in den Messwerten enthalten sind. Demzufolge wären die Gründe für die Differenzen eher im statischen Bereich des Konsonanten zu suchen. Ein möglicher Grund könnte die unterschiedliche Lage der Antiresonanzen in den verschiedenen Nasalen sein. Je länger der orale Vokaltrakt ist, desto tiefer ist die Antiresonanzfrequenz. Aufgrund der schwachen Energie von Nasalen oberhalb von ca. 500 Hz sind diese Unterschiede aber nur schwer wahrnehmbar. Möglicherweise werden sie aber bei akustischen Messungen mit erfasst. Außerdem hängt es auch vom Vokalkontext ab, wie viele Informationen im (statischen Bereich des) Konsonanten und in den (dynamischen) Transitionen stecken [247]. Da bei der akustischen Untersuchung der Vokalkontext nicht beschränkt war, haben möglicherweise die Nasale eines bestimmten Kontextes (z.B. aus dem /i/-Kontext) mehr zur durchschnittlichen Sprecherspezifität der Nasale beigetragen als Nasale aus anderen Kontexten (z.B. aus dem /a/-Kontext).
Auch die stimmhaften und stimmlosen Plosive zeigen einen deutlichen Unterschied in ihrer Sprecherspezifität zwischen der alveolaren und labialen Artikulationsstelle. Die Plosive enthalten, genau wie die Nasale, den Großteil ihrer Ortsinformationen in ihren Transitionen. Da es aber, wie bereits erwähnt, unwahrscheinlich ist, dass Transitionsinformationen in die Messwerte eingeflossen sind, wurden diese Differenzen wahrscheinlich eher von statischen konsonantischen Merkmalen verursacht. Neben den Transitionen enthält auch der Burst des Plosives Informationen zur Artikulationsstelle [269], allerdings unterschiedlich viele in Abhängigkeit vom jeweiligen Plosiv und vom Vokalkontext. Beispielsweise ist der Burst von labialen Plosiven schwächer als der von alveolaren wegen der geringeren Länge des vorderen Vokaltrakts [220]. Da in dieser Untersuchung bereits gezeigt wurde, dass energiereichere Laute mehr Sprecherinformationen enthalten, könnte dieser Unterschied in der Energie des Burst die höhere Sprecherspezifität der alveolaren Plosive erklären.

Perzeptive Analyse

Die perzeptive Analyse zeigt tendenziell ein ähnliches Bild wie die akustischen Messungen: Die alveolaren Konsonanten sind oft signifikant sprecherspezifischer als die labialen. Dieser Unterschied offenbart sich für die Nasale und Plosive in der Bedingung, die sowohl statische als auch dynamische Merkmale enthält und für die Frikative in der Bedingung, die ausschließlich statische und keine dynamischen Merkmale beinhaltet. Vermutlich liegt diesem Phänomen die Tatsache zu Grunde, dass Nasale und Plosive die Informationen über ihre Artikulationsstelle hauptsächlich in den Vokaltransitionen enthalten, wohingegen die

Frikative einen größeren Informationsanteil im statischen Bereich des Konsonanten besitzen. Im Fall der Nasale scheint der Informationsanteil in den Transitionen besonders hoch zu sein, da sich nach der Neutralisierung der dynamischen Sprecherinformationen kein signifikanter Unterschied in der Sprecherdiskriminationsfähigkeit der alveolaren und labialen Nasale nachweisen lässt. Die (statischen) innerkonsonantischen Ortsinformationen sind möglicherweise zu schwach, um von den Hörern wahrgenommen zu werden oder sie sind für beide Nasale gleichermaßen sprecherspezifisch.
Die Plosive hingegen weisen auch ohne die Transitionsinformationen noch signifikante Unterschiede auf. Offenbar verursacht die unterschiedliche Energieintensität des labialen und alveolaren Bursts [220] eine signifikante Differenz in der Sprecherspezifität der Plosive. Die Frikative, welche bei der Kombination von statischen und dynamischen Informationen keine Differenzen in der Sprecherspezifität aufweisen, zeigen unter der Bedingung der reinen statischen Konsonantinformationen einen signifikanten Unterschied zwischen alveolarer und labialer Artikulationsstelle. Auch wenn die Neutralisation der Transitionsinformationen die Sprecherspezifität der Frikative senkt, so bringt sie doch deren Unterschiede zwischen den Artikulationsstellen stärker zum Vorschein. Somit scheinen bei den Frikativen eindeutig die statischen Merkmale des Konsonanten für den Sprecherspezifitätsunterschied zwischen den alveolaren und labialen Frikativen verantwortlich zu sein.
In der Tatsache unter welchen Bedingungen (statische + dynamische Informationen vs. nur statische Informationen) sich die Unterschiede in der Sprecherspezifität der Laute zeigen, offenbart sich auch deren (zeitliche) Verteilung der akustischen Informationen, welche sowohl phonetische als auch Sprecherinformationen beinhalten.

Vergleich von Akustik und Perzeption

Beim Vergleich der Ergebnisse der akustischen und der perzeptiven Untersuchung muss bedacht werden, dass die Messung der akustischen Merkmale immer zum Segmentmittelpunkt stattfand, sodass vermutlich keine dynamischen Transitionsinformationen in den Messungen vorhanden sind. Daher entspricht die akustische Analyse eher der zweiten Bedingung des Perzeptionsexperiments, in der nur konsonantische Informationen vorhanden waren. Vergleicht man diese beiden Untersuchungen, so zeigt sich auch eine fast vollständige Übereinstimmung der Ergebnisse. Für die Plosive und Frikative ergibt sich eine signifikant höhere Sprecherspezifität für die alveolaren Laute gegenüber den labialen. Nur die Resultate der Nasale unterscheiden sich erheblich. Während in der akustischen Messung der Sprecherspezifitätsunterschied zwischen den alveolaren und labialen Nasalen deutlich

ist, kann in der perzeptiven Analyse keine Differenz gemessen werden. Es existiert sogar eine leichte Tendenz zu höherer Sprecherspezifität der labialen Nasale. Offenbar können die statischen Konsonantinformationen der Nasale durch die akustischen Messungen besser erfasst werden als von den menschlichen Hörern.
In der ersten Bedingung des Perzeptionsexperiments (mit statischen und dynamischen Informationen) stimmen die Ergebnisse für die Nasale eher mit den akustischen Ergebnissen überein. Die hier vorhandenen zusätzlichen akustischen Informationen in den Transitionen scheinen die Ursache für die signifikant höhere Sprecherspezifität der alveolaren Nasale zu sein. Die Plosive zeigen die gleiche Differenz wie auch ohne die Transitionsinformationen; sind aber insgesamt sprecherspezifischer. Dies zeigt, dass die zusätzlichen Transitionsinformationen vielleicht nicht die Differenz zwischen alveolaren und labialen Plosiven erhöhen, aber doch stark zur absoluten Sprecherspezifität beitragen. Bei den Frikativen erhöhen die Informationen in den Vokaltransitionen ebenfalls die Sprecherspezifität, überlagern aber auch die Differenzen zwischen den Artikulationsstellen, sodass alveolare und labiale Frikative annähernd gleich sprecherspezifisch sind.
Die menschlichen Hörer scheinen die Sprecher besser diskriminieren zu können, wenn zusätzlich zu den statischen konsonantischen Informationen auch die dynamischen Sprecherinformationen aus den Vokaltransitionen zur Verfügung stehen, wohingegen die akustischen Messungen auch allein mit statischen Merkmalen eine hohe Sprecherspezifität erzielen.

6.2. Einfluss der Qualität des Sprachsignals

In dieser Arbeit wurde auch der Einfluss der Qualität des Sprachsignals auf die Sprecherspezifität der Konsonanten untersucht. Besonders im Bereich der forensischen Phonetik liegt häufig per Telefon aufgenommenes Sprachmaterial vor [159]. Daher ist es von Bedeutung, Sprechermerkmale zu finden, welche trotz der geringeren Qualität des Sprachsignals noch ausreichend Informationen zur Sprecheridentifikation (-diskrimination) liefern. Deshalb sollte untersucht werden, wie stark sich die Telefonübertragung auf die Sprecherspezifität des jeweiligen Konsonants auswirkt. Es wurde vermutet, dass der eingeschränkte Frequenzdurchgang (ca. 300 bis 3400 Hz) hauptsächlich die Sprecherinformationen der Frikative verringern würde, da diese ihren Energieschwerpunkt oberhalb dieser Grenze haben [230]. Für Nasale, welche eher im unteren Spektralbereich die meiste Energie haben [230], wurde eine geringere Verminderung der Sprecherspezifität angenommen. Allerdings zeigt sich in der akustischen Analyse für alle Artikulationsmodi eine ähnlich starke Reduktion der Sprecherinformationen.

Vermutlich liegt der Energieschwerpunkt der Nasale um oder unterhalb der 300 Hz, sodass diese Informationen durch die Übertragung ebenfalls herausgefiltert werden. Alle Nasale und Frikative werden in ihrer Sprecherspezifität um ca. die Hälfte reduziert.
Für die perzeptive Untersuchung fehlt bisher ein solcher Vergleich. Die Stimuli wurden aber, in Vorbereitung auf eine solche Untersuchung, bei der Aufnahme gleichzeitig über Mikrofon und Handy aufgenommen. Auch wenn der zeitliche Rahmen dieser Arbeit keine Analyse der Telefon-Stimuli mehr zuließ, so wäre es eine relevante und gewinnbringende Studie, die sich für die Zukunft anbieten würde.

6.3. Sprecherinformationen in der Sprachwahrnehmung

Aus Studien über die menschliche Sprachwahrnehmung und -verarbeitung ist bekannt, dass lexikalische bzw. phonetische und Sprecherinformationen miteinander interagieren ([219] [48] [49] [54]). So können Sprecherinformationen beispielsweise die Worterkennung erleichtern [259] oder behindern [103]. In diesem Zusammenhang stellte sich die Frage, welche Informationen Hörer zuerst wahrnehmen und verarbeiten. Erkennen sie erst das Wort und dann den Sprecher oder andersherum? Wie sieht die genaue zeitliche Koordination zwischen diesen beiden Informationstypen aus und haben die phonetischen Informationen (verschiedene Anfangskonsonanten) einen Einfluss auf die Geschwindigkeit der Sprechererkennung?
Um diesen Fragen auf den Grund zu gehen, wurde ein Visual-World Eye-Tracking Experiment durchgeführt, bei dem die Versuchspersonen eine Kombination aus einem Wort (Bild eines Gegenstands) und eines Sprechers (Bild einer Person) identifizieren mussten. Dafür sahen sie am Bildschirm vier verschiedene Sprecher-Gegenstand-Kombinationen, hörten dann einen der Sprecher ein Wort (Name eines Gegenstands) sagen und mussten den richtigen Gegenstand und Sprecher erkennen. Bei der Untersuchung wurden verschiedene Einflussfaktoren berücksichtigt: die Anzahl der Sprecher (zwei vs. vier), das Geschlecht der Sprecher (Männer vs. Frauen), die Ähnlichkeit der Sprecherstimmen (ähnlich vs. unähnlich), die Ambiguität der Bedingungen (ambig (Sprecherbedingung) vs. nicht-ambig (Gegenstandsbedingung)), die phonetische Überlappung (gering vs. stark) als auch die Identität des Anfangskonsonants (/m, n, f, s, p, t, b, d/).
In beiden Experimenten identifizierten die Teilnehmer die visuellen Referenten der Sprecher-Gegenstand-Kombinationen mit hoher Genauigkeit. Mit zunehmender Anzahl an Sprechern stieg aber die Schwierigkeit der Sprecheridentifikation. Im Identifikationsprozess des Spre-

chers entschieden sich die Hörer zuerst für das Geschlecht des Sprechers, bevor sie die genaue Identität wählten. Dabei erhöhte sich die Schwierigkeit der Sprecheridentifikation mit zunehmender Ähnlichkeit der Sprecher. Die Frauenstimmen waren sich perzeptiv ähnlicher als die Männerstimmen und daher schwerer zu unterscheiden. Bei höherer Ambiguität der Sprecher-Gegenstand-Kombinationen (wie z.B. in der Sprecherbedingung von Experiment 1) war die Identifikation schwieriger als bei nicht-ambigen Sprecher-Gegenstand-Kombinationen. Der phonetische Overlap am Wortanfang (gering vs. stark) wirkte sich stark auf die phonetische Kompetition aus: War der Overlap zu gering (nur der Konsonant) trat keine phonetische Kompetition auf, wurde der Overlap länger (Konsonant und Vokal) stieg die phonetische Kompetition auf die gleiche Höhe wie die Sprecherkompetition und wurde signifikant. Überlappten die Wörter phonetisch vollständig (identische Wörter), so war sie sogar signifikant stärker als die Sprecherkompetition und blieb bis zum Trialende erhalten. Der Anfangskonsonant zeigte wenig Einfluss auf die Sprecherkompetition. Einzig bei Wörtern mit einem Frikative am Anfang zeigte sich eine signifikant stärkere Sprecherkompetition im Vergleich mit Wörtern anderer Anfangskonsonanten. Dies könnte ein Indiz dafür sein, dass Frikative viele Sprecherinformationen enthalten und daher die Hörer während dieser Laute stärker Sprecherinformationen nutzen. Die Artikulationsstelle (labial vs. alveolar) rief keinen signifikanten Effekt hervor. Vielleicht würde man Differenzen sehen, wenn man die Targetidentifizierung schwieriger machen würde (so, wie es im Perzeptionsexperiment gemacht wurde). Aber so traten bei der hohen Identifikationsrate keine Einflüsse der Konsonanten auf.

In beiden Experimenten zeigt sich eine (relativ) starke Sprecherkompetition, die dafür spricht, dass die Hörer Sprecherinformationen verwenden. Mit zunehmender Anzahl an Sprechern steigt die Sprecherkompetition, was wahrscheinlich an der einfacheren Unterscheidbarkeit von zwei im Gegensatz zu vier Sprechern liegt. Diese Resultate werfen die Frage auf, weshalb die Hörer so stark und konstant Sprecherinformationen nutzen, selbst wenn sie gegenüber den phonetischen Informationen keinen Vorteil (oder sogar einen Nachteil) bieten. Möglicherweise verursachen unterschiedliche Entscheidungsstrategien zwischen den Hörern diesen Effekt, was aber die Ergebnisse auch nicht vollständig erklären kann.

Die Jackknife-Analyse der Daten zeigt, dass die Sprechereffekte früher steigen als die phonetischen Effekte. Diese Tatsache spricht dafür, dass Sprecherinformationen tatsächlich etwas schneller aufgenommen und verarbeitet werden als phonetische Informationen. Dabei sind beide Informationstypen am Anfang gleich schnell, aber ab der knappen Hälfte des Maximaleffekts überholen dann die Sprecherinformationen die phonetischen.

Die Ergebnisse sprechen für die Wichtigkeit von Sprecherinformationen in der Sprachwahrnehmung. Erkennt ein Hörer einen Sprecher, so kann er die phonetischen Informationen besser verarbeiten und z.B. ein Wort schneller erkennen. Andersherum funktioniert der Effekt aber auch: Hat ein Hörer ein Wort erkannt, ist er in der Lage auch den Sprecher schneller zu erkennen. Welche der Informationen - phonetisch oder Sprecherinformationen - zuerst verarbeitet werden, liegt möglicherweise an ihrer disambiguierenden Wirkung. Sind sich die Sprecherstimmen sehr ähnlich, liefern sie wenig disambiguierende Informationen für den Hörer und er wird eher auf die phonetischen Informationen achten; so diese informativer sind. Bei sehr ähnlichen Wörtern hingegen, liegt die Wahl der (informativeren) Sprecherinformationen näher. So ist es im Fall von identischen Wörtern unumgänglich Sprecherinformationen zu verwenden. Die zwei Sprecher waren aber sehr einfach zu unterscheiden, sodass die Sprecherkompetition zwar signifikant, aber nicht so stark war. Erstaunlicherweise war aber auch im Fall, dass die phonetischen Informationen sehr disambiguierend waren, die Sprecherkompetition recht hoch. Das zeigt, dass die Hörer offenbar trotz der informativen Phonetik immer noch auf die Sprecherinformationen achteten und sie nutzten. Dieses Phänomen spricht für die Relevanz von Sprecherinformationen in der Sprachwahrnehmung. Mittels der Jackknife-Analyse konnte dann, neben der Wichtigkeit der Sprecherinformationen, auch ihre frühe Wahrnehmung und Verarbeitung gegenüber den phonetischen Informationen gezeigt werden.

6.4. Abstrakte oder episodische Sprachwahrnehmung

Nachdem die Ergebnisse der einzelnen Untersuchungen erklärt und diskutiert wurden, soll betrachtet werden, inwieweit sich die neu gewonnen Erkenntnisse in bestehende Modelle der Sprachwahrnehmung einfügen. Auf der einen Seite stehen die abstraktionistischen Modelle ([184] [212] [214] [107]), die von einer Normierung des Sprachsignals auf einem prälexikalischen Level ausgehen, sodass nur „reine“ lexikalische Informationen übrig bleiben, die dann einer passenden Phonem- und Wortkategorie im mentalen Lexikon zugeordnet werden. Dem gegenüber stehen die episodischen oder Examplar-theoretischen Modelle ([260] [97] [69] [133] [103] [231] [135]), die von dem Erhalt aller akustischer Details des Sprachsignals ausgehen, sodass Wörter (oder andere Entitäten) zusammen mit Sprecherinformationen im mentalen Lexikon abgelegt werden.
Die gefundene frühe Wahrnehmung von Sprecherinformationen spricht stark für deren

Repräsentation im menschlichen mentalen Lexikon im Sinne eines exemplarischen Modells der Sprachperzeption. Die Ergebnisse der Jackknife-Analyse zeigten, dass gerade im späteren Verlauf (oberhalb der 40 % des Maximaleffekts) die Sprecherinformationen früher als die phonetischen verarbeitet wurden. Wären sie auf einem frühen, prälexikalischen Level herausgefiltert worden, könnten sie kaum diesen Effekt bewirken. Auch die Tatsache, dass eine Erhöhung der Anzahl der Sprecher eine Erhöhung der Schwierigkeit der Targetidentifikation bedeutete, weist darauf hin, dass die Sprecherinformationen mitverarbeitet werden. Andernfalls sollte es keinen Unterschied machen, ob zwei oder vier Sprecher zu unterscheiden sind, da deren spezifische Merkmale vor der Identifikation des Wortes normiert würden. Der stärkere Fokus auf Sprecherinformationen im Fall von ambigeren phonetischen Informationen spricht ebenfalls für eine Mitverarbeitung sprecherindividueller Merkmale während der Sprachwahrnehmung. Ohne ein Vorhandensein dieser Informationen im (vorverarbeiteten) Sprachsignal wäre eine stärkere Nutzung dieser Informationen wohl nicht möglich. Die einzige Tatsache, die möglicherweise eine Sprachwahrnehmung nach abstraktionistischem Modell unterstützen würde, wäre die Fähigkeit der Hörer, von sprecherindividuellen Merkmalen auf das Geschlecht des Sprechers zu abstrahieren. Aber auch das würde sich mit einem exemplarischen Modell erklären lassen, wenn man davon ausgeht, dass die im mentalen Lexikon gespeicherten Entitäten (z.B. Wörter) in Clustern zusammengefasst werden, so würden sich Exemplare von weiblichen Sprechern untereinander stärker ähneln und sich von männlichen Sprechern stärker unterscheiden. Die ähnlicheren Exemplare würden durch eine Clusterbildung näher beieinander liegen, und so könnte über die Zugehörigkeit eines Exemplars zu einem Cluster auf ein gemeinsames Merkmal aller dieser Exemplare geschlossen werden. Diese Struktur ermöglicht es den Hörern, sehr schnell die relevanten Informationen in einer Sprachäußerung zu erfassen und zu verarbeiten, wobei neben den inhaltlich wichtigen, phonetischen (lexikalischen) Merkmalen auch nicht-lexikalische Merkmale, wie Sprechermerkmale, Beachtung finden. Somit sprechen die gefundenen Effekte eher für ein episodisches, Examplar-theoretisches Sprachwahrnehmungsmodell als für ein abstraktionistisches.

6.5. Zusammenfassung, Schlussfolgerungen und Ausblick

Diese Arbeit widmete sich der Frage nach der Bedeutung von Sprecherinformationen in verschiedenen Forschungs- und Anwendungsbereichen der Phonetik. Menschen (und inzwischen auch Maschinen) sind in der Lage, die sprecherspezifischen Merkmale einer Person in ihrer

Sprachäußerung wahrzunehmen und sie anhand derer zu erkennen bzw. von anderen zu unterscheiden. Die forensische Phonetik macht sich die Existenz dieser sprecherspezifischen Merkmale zu nutze, um anhand von Stimmvergleichen Straftäter zu ermitteln. Die automatische Sprecherverifikation nutzt sie, um die Berechtigung einer Person für den Zugang physischer und virtueller Räume zu überprüfen. Und die Sprachwahrnehmungsforschung untersucht ihre Verarbeitung und Repräsentation im mentalen Lexikon des Menschen.
Als erstes sollte ermittelt werden, in welchen Dimensionen bzw. Parametern sich Sprecher unterscheiden, um möglichst aussagekräftige Merkmale zu finden. Dazu wurde eine akustische Analyse der spektralen Momente von ausgewählten Konsonanten des Deutschen durchgeführt. Alle konsonantischen Merkmale zeigen eine signifikant höhere Inter-Sprecher-Variation als Intra-Sprecher-Variation, was sie potenziell zu geeigneten Parametern der Sprecherdiskrimination und -identifikation macht. Die höchste Sprecherspezifität weisen dabei die Nasale /m, n/ und die Frikative /f, s/ auf. Insgesamt sind diese beiden Artikulationsmodi sprecherspezifischer als die Plosive, und die Konsonanten der alveolaren Artikulationsstelle sprecherspezifischer als die der labialen. Im Vergleich zu den Vokalen zeigen sich die Konsonanten in dieser Untersuchung als deutlich sprecherspezifischer, egal ob als akustische Merkmale für die Vokale die üblichen Formanten oder ebenfalls die spektralen Momente gemessen werden. Kritisch ist anzumerken, dass die Segmentierung und die Messung der Vokalformanten automatisch erfolgte, was immer ein gewisses Fehlerpotenzial birgt. Allerdings sollten sich diese möglichen Ungenauigkeiten in einem engen Rahmen bewegen [263]. Aufgrund der Verwendung von Spontansprache und der unterschiedlichen Häufigkeit der Laute im Deutschen, war die Anzahl der Laute pro Konsonant unterschiedlich. Möchte man dies vermeiden, muss man stark kontrolliertes Sprachmaterial und kein spontansprachliches verwenden. Dies wiederum würde aber die Natürlichkeit des Sprachmaterials beeinträchtigen. In dieser Arbeit wurden akustisch nur statische Merkmale (zum zeitlichen Mittelpunkt des Segments) berechnet. Es wäre aber auch interessant, dynamische Merkmale (die zeitliche Veränderung von Merkmalen) zu betrachten, da sich in Bereichen starker Veränderung im Signal auch immer viele akustische Informationen befinden ([147] [95] [151]).
Anschließend wurde in zwei Perzeptionsexperimenten die menschliche Sprecherdiskriminationsfähigkeit mittels unterschiedlich informativer /aKa/-Stimuli untersucht. In der Kombination von statischen und dynamischen Informationen sind alle Laute sprecherspezifisch, wobei die Nasale und Plosive etwas besser abschneiden als die Frikative. Im Fall, dass nur statische Konsonantinformationen zur Verfügung stehen, sind die Frikative deutlich sprecherspezifischer als die anderen beiden Artikulationsmodi. Beim Vergleich

der Sprecherspezifität der alveolaren und labialen Artikulationsstelle fällt der Einfluss der Lokalisation der akustischen Informationen auf. Nasale und Plosive, die ihre Ortsinformationen eher in den Transititionen zu den benachbarten Vokalen enthalten, zeigen in Anwesenheit dieser Informationen einen signifikanten Unterschied in der Sprecherspezifität zwischen diesen beiden Artikulationsstellen. Hingegen demonstrieren die Frikative erst in Abwesenheit der Transitionsinformationen einen signifikanten Einfluss der Artikulationsstelle auf die Sprecherspezifität. Die Auswahl der zeitlichen Bereiche eines Konsonanten hat somit einen starken Effekt auf seine Sprecherspezifität. Außerdem stimmen die Ergebnisse der akustischen und der perzeptiven Analyse am stärksten für die Frikative und Plosive überein, während sie bei den Nasalen differieren. Vermutlich liegen diese Unterschiede an der Eignung der gewählten akustischen Merkmale für den jeweiligen Konsonanten. Bei der perzeptiven Untersuchung ist kritisch anzumerken, dass die Amplitudennormierung nicht getrennt nach Konsonant, sondern einheitlich für alle Konsonanten durchgeführt wurde. Das könnte konsonantspezifische Lautstärkenunterschiede, die sich ebenfalls auf die Sprecherspezifität auswirken, beseitigt haben. Bei einer nachfolgenden Analyse sollten diese spezifischen Lautstärken der Konsonanten berücksichtigt werden.
In der akustischen Analyse wurde auch der starke Einfluss von der Qualität des Sprachsignals auf die Sprecherspezifität der Konsonanten belegt. Unabhängig von der Lage ihres Energieschwerpunktes wurde die Sprecherspezifität aller Nasale und Frikative durch die Übertragung über den Telefonkanal (Frequenzdurchgang von ca. 300 bis 3400 Hz) um ca. die Hälfte reduziert. Dennoch enthielten sie weiterhin ein signifikantes Maß an Sprecherinformationen.
Abschließend wurde die Rolle der Sprechermerkmale in der menschlichen Sprachwahrnehmung und deren zeitliche Koordination zur Verarbeitung phonetischer Informationen beleuchtet. Dafür mussten Hörer in einem Visual-World Eye-Tracking Experiment Sprecher-Gegenstand-Kombinationen identifizieren. Die untersuchten Einflussfaktoren Anzahl der Sprecher (zwei vs. vier), Geschlecht des Sprechers (Mann vs. Frau), Ähnlichkeit der Sprecherstimmen (ähnlich vs. unähnlich), Ambiguität der phonetischen Informationen (ambig (Sprecherbedingung) vs. nicht-ambig (Gegenstandsbedingung)) als auch die phonetische Überlappung der Gegenstandswörter (gering vs. stark) haben einen signifikanten Einfluss auf die Geschwindigkeit der Targetidentifikation. Einzig die Identität des Anfangskonsonants zeigt keinen bzw. nur einen schwachen Effekt. Das interessanteste Ergebnis ist aber der Fakt, dass auch im Falle von vollständig disambiguierenden phonetischen Informationen, Sprecherinformationen beachtet und verwendet werden, um das Target zu identifizieren.

Weiterhin fällt auf, dass besonders im späteren Entscheidungsprozess die Sprecherinformationen leicht früher verarbeitet werden als die phonetischen. Diese beiden Erkenntnisse sprechen stark für die Relevanz und Informativität von sprecherindividuellen Informationen in der Sprachwahrnehmung.
Es wurde diskutiert, für welches Sprachwahrnehmungsmodell (abstraktionistisch vs. exemplartheoretisch) die Ergebnisse sprechen. Die Tatsache, dass Faktoren wie die Sprecheranzahl und die phonetische Ambiguität einen Einfluss auf die Geschwindigkeit der Targetidentifizierung haben, deutet in Richtung eines Exemplar-Modells. Auch dass in jeder Bedingung Sprecherinformationen verwendet wurden sowie der frühe Anstieg des Sprechereffekts (in der Maximalwertanalyse) sprechen für eine Mitverarbeitung der sprecherindividuellen Merkmale in der Worterkennung. Für ein abstraktionistisches Modell würde höchstens die Fähigkeit der Hörer, von sprecherindividuellen Merkmalen auf das Geschlecht des Sprechers zu abstrahieren, sprechen. Allerdings wäre auch diese Abstraktion in einem Exemplar-Modell mit Hilfe von Cluster-Bildung im mentalen Lexikon möglich. Somit unterstützen die gewonnenen Erkenntnisse über den Prozess der Sprachwahrnehmung den Ansatz der exemplartheoretischen Sprachwahrnehmungsmodelle.

Zum Abschluss soll ein kurzer Ausblick mit möglichen weiteren Forschungsfragen gegeben werden. In dieser Arbeit wurden die meisten Experimente (nur) mit Männerstimmen durchgeführt. Diese Entscheidung lag in der höheren Relevanz von männlichen Sprechern in der forensischen Phonetik begründet. Allerdings wäre es für zukünftige Studien auch interessant und sinnvoll Frauenstimmen zu untersuchen, sowohl akustisch als auch perzeptiv. In der akustischen Analyse wurden annähernd spontansprachliche Dialoge verwendet, sodass die Anzahl der einzelnen Konsonanten nicht identisch war. Dies könnte sich auf die Stärke der Sprecherspezifität ausgewirkt haben. Daher wäre es vielleicht sinnvoll, eine Analyse mit gleich vielen Exemplaren jedes Konsonanten durchzuführen. In diesem Fall wäre aber die Verwendung von spontansprachlichen Äußerungen nicht möglich, da sich die Laute in ihrer Häufigkeit unterscheiden. Eine andere Möglichkeit wäre die Verwendung eines anderen Maßes als der F-ratio. Beispielsweise könnte die in der forensischen Phonetik verbreitete log-likelihood-Ratio ([2]) verwendet werden, die für eine gegebene Sprachprobe angibt, um wie viel wahrscheinlicher es ist, dass sie vom gleichen Sprecher stammt als von zwei verschiedenen ([199]). Außerdem wurden die spektralen Momente immer (nur) zum zeitlichen Mittelpunkt jedes Segments gemessen. Eine breitete Messung von beispielsweise 30 - 70 % des Segments und eine anschließende Durchschnittsbildung, könnten womöglich

mehr Sprecherinformationen des Segments einfangen. Es wurden bereits erste Analysen in dieser Richtung durchgeführt, die diese Vermutung bestätigen. Weiterhin könnten natürlich auch andere akustische Parameter untersucht werden als die hier verwendeten spektralen Momente. Dabei wäre zu überlegen, ob sich Parameter finden ließen, die für alle Konsonanten (aller Artikulationsmodi) gleichermaßen gut geeignet wären, ihre akustischen Merkmale einzufangen. Oder ob man für jeden Artikulationsmodus die bestgeeignetsten Parameter ermitteln und dann diese verwenden sollte.
In der Perzeptionsstudie wurde die Sprecherdiskriminationsfähigkeit der Konsonanten bisher nur im /a/-Kontext untersucht. Da sich der Vokalkontext auf den Informationsgehalt des Konsonanten und der Transitionen auswirken kann [247], wäre ein weiteres Experiment mit einem /i/-Kontext sinnvoll und gewinnbringend. Entsprechende Stimuli wurden bereits aufgenommen und bearbeitet und stünden für ein solches Experiment zur Verfügung. Des Weiteren stehen auch Telefon-Stimuli zur Verfügung, die zeitgleich mit den Mikrofonaufnahmen erzeugt wurden. Mit diesen könnte ein weiteres Perzeptionsexperiment durchgeführt und so die auditive Sprecherdiskriminationsfähigkeit von telefongefilterten Stimuli untersucht werden. Nachdem sich die Transitionen als wichtige Träger von Sprecherinformationen herausgestellt haben, wäre es interessant diese akustisch (oder artikulatorisch) detaillierter zu untersuchen, um herauszufinden, weshalb manche Transitionsinformationen sprecherspezifischer sind als andere. Allgemein wäre es interessant eine akustische Analyse der Stimuli analog zur Analyse der Sprachdaten aus dem Verbmobil-Korpus durchzuführen, um zusehen, inwieweit die Sprecherspezifität der akustischen Merkmale übereinstimmen würde. Des Weiteren bietet es sich an, die perzeptive Diskriminationsfähigkeit der Hörer für die einzelnen Sprecher genauer zu betrachten. Zu diesem Zweck könnte eine multidimensionale Skalierung durchgeführt werden. Dies würde einen detaillierteren Einblick geben, in welchem Maß sich die einzelnen Sprecher auditiv voneinander unterscheiden. Dabei wäre ein Vergleich der Sprecherdiskriminationsfähigkeit anhand von akustischen Merkmalen und anhand von auditiven Merkmalen lohnend.
Im Visual-World Eye-Tracking Experiment zeigte sich die nennenswerte Rolle der Sprecherinformationen in der Sprachwahrnehmung und -verarbeitung. Eine lohnende Frage wäre, ob Hörer verschiedene Entscheidungsstrategien bei der Targetidentifizierung verwenden. Das heißt, ob manche Menschen eher auf die Sprecherinformationen und andere auf die phonetischen Informationen achten. Nach dem Experiment wurden die Hörer darüber befragt. Zusätzlich könnte man sich anschauen, ob die Hörer tatsächlich eher auf das Bild des Sprechers oder das des Gegenstands schauten und diese Ergebnisse mit der subjekti-

ven Einschätzung der Versuchspersonen vergleichen. Möglicherweise verhalten sich nicht alle Menschen nach dem gleichen Schema, sondern fokussieren sich auf unterschiedliche Informationen im Sprachsignal. Ein zusätzliches Experiment könnte darin bestehen, die gleiche Anzahl an Sprechern und Wörtern zu verwenden. Bisher waren es nur zwei bzw. vier Sprecher und 128 Wörter. Man könnte nun argumentieren, dass die Hörer deshalb die Sprecherinformationen nutzten, weil sie aufgrund der geringeren Anzahl einfacher zu unterscheiden waren. Diesem Kritikpunkt könnte man durch ein solches Experiment nachgehen.

In dieser Arbeit wurde die Rolle von phonetischer Information in der Sprechererkennung untersucht. Dabei zeigt sich die starke Interaktion dieser beiden Informationstypen. Je nach Konsonant enthalten die Laute unterschiedlich sprecherspezifische akustische bzw. perzeptive Merkmale. Dabei werden die Ergebnisse vorangegangener Studien auch für die deutsche Sprache belegt. Die gewonnenen Erkenntnisse können dabei helfen, geeignete Parameter zur Sprecheridentifikation bzw. -diskrimination zu finden. Weiterhin leisten sie einen Beitrag zum Wissen über die Verarbeitung von Sprecherinformationen bei der menschlichen Sprachwahrnehmung und somit auch zu unserem allgemeinen Verständnis der Sprachwahrnehmungsprozesse. In dieser Hinsicht sprechen die Ergebnisse dieser Arbeit eher für ein episodisches, exemplartheoretisches Sprachwahrnehmungsmodell als für ein abstraktionistisches. Neben diesen Erkenntnissen, wurde auch erstmals gezeigt, dass sich das Visual-World Eye-Tracking Paradigma nicht nur zur Verfolgung der Verarbeitung phonetischer (bzw. lexikalischer) Information eignet, sondern auch zum Erkenntnisgewinn über den Prozess der Sprechererkennung. Somit kann diese Methode in Zukunft ein nützliches Werkzeug sein für weitere Einblicke in die Sprechererkennung und ihr Verhältnis zur Worterkennung.

A. Sprachmaterial

D-Wörter	B-Wörter	T-Wörter	P-Wörter	N-Wörter	M-Wörter	S-Wörter	F-Wörter
Dach	Bach	Tacho	Pass	Nacht	Mars	Sack	Fass
Darm	Bank	Tanne	Panther	Narr	Mantel	Sarg	Farn
Dachs	Ball	Taxi	Pastor	Napf	Maske	Salz	Falke
Damm	Band	Tasche	Panda	Nase	Mandel	Sand	Falte
Dampfer	Bambus	Tanker	Panzer	Nachthemd	Mammut	Salbe	Fallschirm
Dattel	Balken	Taste	Paddel	Natter	Masche	Sattel	Falter
Dackel	Backblech	Tasse	Palme	Nacken	Mappe	Saft	Fackel
Diskus	Birne	Tinte	Pinsel	Nixe	Mistel	Sichel	Finger
Dieb	Biene	Tiger	Pisa	Niete	Mine	Siegel	Fiedel
Dichtung	Bild	Tisch	Pilz	Nikolaus	Milch	Silber	Filzhut
Decke	Becher	Teller	Pendel	Netz	Messer	Sekt	Ferse
Deckel	Becken	Tempel	Pelz	Nest	Messstab	Sessel	Fernrohr
Degen	Besen	Teesieb	Pegel	Nebel	Mehl	Segel	Feder
Duft	Butter	Tulpe	Pulver	Nuss	Mund	Suppe	Fuchs
Dose	Boot	Tor	Polo	Note	Motor	Soße	Vogel
Dorn	Bonbon	Torte	Pony	Nonne	Motte	Socken	Formel

Tabelle A.1.: Sprachmaterial für das Eye-Tracking Experiment

Alle /s/-initialen Wörter werden im Standarddeutschen stimmhaft (phonemisch /z/) ausgesprochen. Im süddeutschen Raum gibt es aber keinen Stimmhaftigkeitskontrast zwischen den beiden alveolaren Frikativen ([155]). Außerdem wurden die Sprecher zusätzlich gebeten, die /s/-initialen Wörter tatsächlich mit einem stimmlosen Frikativ zu sprechen.

B. Publikationen

Der Autor dieser Dissertation war an folgenden Veröffentlichungen beteiligt:

Eva Reinisch & Carola Schindler, *Tracking the time-course of speaker recognition relative to phonetic processing*, 21st Annual Conference on Architectures and Mechanisms for Language Processing, Valletta, Malta, 2015.

Carola Schindler & Eva Reinisch, *Tracking the temporal relation between speaker recognition and processing of phonetic information*, In Proc. of the 18th International Congress of Phonetic Sciences, Glasgow, UK, 2015.

Carola Schindler, Eva Reinisch & Jonathan Harrington, *Perceptual speaker discrimination based on German consonants*, 23rd Annual Conference of the International Association for Forensic Phonetics and Acoustics, Zurich, 2014.

Carola Schindler & Christoph Draxler, *The influence of the place of articulation on the speaker specificity of German phonemes*, 4th International Summer School on "Speech Production and Perception: Speaker-Specific Behavior", Aix-en-Provence, 2013.

Carola Schindler & Christoph Draxler, *Using Spectral Moments as a Speaker Specific Feature in Nasals and Fricatives*, In Proc. 14th Annual Conference of the International Speech Communication Association, Lyon, 2013.

Carola Schindler & Christoph Draxler, *The influence of bandwidth limitation on the speaker discriminating potential of nasals and fricatives*, 22nd Annual Conference of the International Association for Forensic Phonetics and Acoustics, Tampa, 2013.

Carola Mook & Christoph Draxler, *A study on the speaker discriminating power of vowels, nasals and fricatives*, 21st Annual Conference of the International Association for Forensic Phonetics and Acoustics, Santander, 2012.

C. Literaturverzeichnis

[1] P. Adank und J.M. McQueen: *The effect of an unfamiliar regional accent on spoken word comprehension. The effect of an unfamiliar regional accent on spoken word comprehension*, In *Proceedings of the 16th International Congress of Phonetic Sciences*. (2007).

[2] C. Aitken und D. Lucy, *Applied Statistics* **53** (2004), 109.

[3] M.J. Albalá, E. Battaner, J. Gil, J. Llisterri, M. Machuca und V. Marrero: *Vowel formant structure and speaker identification - A perceptual study. Vowel formant structure and speaker identification - A perceptual study*, In *Conferência Ibérica de Percepção (CIP)*. Guimarães, Portugal (2009).

[4] P.D. Allopenna, J.S. Magnuson und M.K. Tanenhaus, *Journal of Memory and Language* **38** (1998), 419.

[5] G.T.M. Altmann und Y. Kamide, *Cognition* **73** (1999), 247.

[6] K. Amino und T. Arai: *Contribution of the consonants and vowels to the perception of speaker identity. Contribution of the consonants and vowels to the perception of speaker identity*, In *The Japan-China Joint Conference of Acoustics (JCA)*. (2007).

[7] K. Amino und T. Arai, *The Acoustical Society of Japan* **28** (2007), 128.

[8] K. Amino und T. Arai: *Perceptual Speaker Identification Using Monosyllabic Stimuli - Effects of the Nucleus Vowel and Speaker Characteristics Contained in Nasals. Perceptual Speaker Identification Using Monosyllabic Stimuli - Effects of the Nucleus Vowel and Speaker Characteristics Contained in Nasals*, In *Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Brisbane (2008) Seiten 1917–1920.

[9] K. Amino und T. Osanai, *The Acoustical Society of Japan* **33** (2012), 96.

[10] K. Amino, T. Sugawara und T. Arai: *The Correspondences between the Perception of the Speaker Individualities Contained in Speech Sounds and Their Acoustic Properties. The Correspondences between the Perception of the Speaker Individualities Contained in Speech Sounds and Their Acoustic Properties*, In *Annual Conference of the International Speech Communication Association (INTERSPEECH)*. (2005) Seiten 2015–2028.

[11] K. Amino, T. Sugawara und T. Arai. *Effects of the Syllable Structure on Perceptual Speaker Identification.* Technischer Bericht, The Institute of Electronics, Information and Communication Engeneers, 2006.

[12] K. Amino, T. Sugawara und T. Arai, *Acoustical Science and Technology* **27** (2006), 233.

[13] K. Amino, T. Sugawara und T. Arai: *Speaker Similarities in Human Perception and their Spectral Properties. Speaker Similarities in Human Perception and their Spectral Properties*, In *WESPAC IX 2006*. (2006).

[14] A. Andics, J.M. McQueen und M. Van Turennout: *Phonetic Content Influences Voice Discriminability. Phonetic Content Influences Voice Discriminability*, In *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS)*. Saarbrücken (2007) Seiten 1829–1832.

[15] M. Antal: *Phonetic Speaker Recognition. Phonetic Speaker Recognition*, In *Proceedings of the 7th International Conference COMMUNICATIONS*. Bucharest, Romania (June 2008) Seiten 73–76.

[16] M. Antal und G. Toderean: *Speaker Recognition and Broad Phonetic Group. Speaker Recognition and Broad Phonetic Group*, In *Proceedings of the IASTED International Conference on Signal Processing, Pattern Recognition, and Applications*. SPPRA (2006).

[17] J.E. Atkinson, *Journal of the Acoustic Society of America* **60** (1976), 440.

[18] R.H. Baayen: *Analyzing Linguistic Data - A practical introduction to statistics using R.* Cambridge University Press, 2008.

[19] J.A. Bachorowski und M.J. Owren. *Handbook of Emotions.* The Guilford Press, 2008. Kapitel Vocal Expressions of Emotions, Seiten 196–210.

[20] R. Bakeman, *Behavior Research Methods* **37** (2005), 379.

[21] S. Barfüsser und F. Schiel: *Disfluencies in alcoholized speech. Disfluencies in alcoholized speech*, In *Annual Conference of International Association for Forensic Phonetics and Acoustics*. (2010).

[22] B.A. Barkera und R.S. Newman, *Cognition* **94** (2004), 45.

[23] D.J. Barr, *Journal of Memory and Language* **59** (2008), 457.

[24] D.J. Barr, R. Levy, C. Scheerpers und H.J. Tily, *Journal of Memory and Language* **68** (2013), 255.

[25] N. Bastug: *A Contrastive Analysis of the English and the German Sound System*. GRIN Verlag, Munich, 2011.

[26] D. Bates, M. Maechler, B. Bolker und S. Walker. *lme4: Linear mixed-effects models using Eigen and S4* (2015). R package version 1.1-8.

[27] E. Battaner, J. Gil, V. Marrero, J. Llisterri, C. Caró, M.J. Machuca, C. de la Mota und A. Ríos: *VILE: Acoustic Study of Inter and Intra Speaker Variation in Spanish. VILE: Acoustic Study of Inter and Intra Speaker Variation in Spanish*, In *Actas del II Congreso de la Sociedad Española de Acústica Forense (SEAF)*. (April 2003) Seiten 59–70.

[28] D. Bauer, J. Kannampuzha, P. Hoole und B.J. Kröger. In *Development of Multimodal Interfaces: Active Listening and Synchrony*. Springer (2010), Seiten 346–353.

[29] B. Baumeister und F. Schiel: *On the effects of alcoholisation on fundamental frequency. On the effects of alcoholisation on fundamental frequency*, In *Annual Conference International Association for Forensic Phonetics and Acoustics*. (2010).

[30] T. Becker: *The Influence of intra-speaker variability in automatic speaker identification. The Influence of intra-speaker variability in automatic speaker identification*, In *Proceedings of the International Association of Forensic Phonetics and Acoustics 2007 Annual Conference (IAFPA)*. (2007).

[31] T. Becker, M. Jessen und C. Grigoras: *Forensic speaker verification using formant features and Gaussian mixture models. Forensic speaker verification using formant features and Gaussian mixture models.*, In *Interspeech*. (2008) Seiten 1505–1508.

[32] A. Belotel-Grenié und M. Grenié: *The Creaky Voice Phonation And The Organisation Of Chinese Discourse. The Creaky Voice Phonation And The Organisation Of Chinese Discourse*, In *Proceedings of the International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages*. (2004).

[33] J.J.A. van Berkum, D. van den Brink, C.M.J.Y. Tesink, M. Kos und P. Hagoort, *Journal of Cognitive Neuroscience* **20** (2008), 580.

[34] R. van Bezooijen: *Characteristics and Recognizability of Vocal Expressions of Emotion*. Foris Publications, Dordrecht, 1984.

[35] P. Boersma, *Glot International* **5** (2001), 341.

[36] A.R. Bradlow, D.B. Pisoni, R. Akahane-Yamada und Y. Tohkura, *Journal of the Acoustic Society of America* **101** (1997), 2299.

[37] A. Braun. *Studies in Forensic Phonetics*. Wissenschaftlicher Verlag Trier, 1995. Kapitel Fundamental frequency - how speaker-specific is it?, Seiten 9–23.

[38] A. Braun und A. Rosin: *On the speaker-specificity of hesitation markers. On the speaker-specificity of hesitation markers*, In *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow, Scotland, UK (August 2015).

[39] P.D. Bricker und S. Pruzansky, *Journal of the Acoustic Society of America* **40** (1966), 1441.

[40] F. Brock: *Phonetische Untersuchungen der Stimmverstellung von professionellen Sprechern*. Ludwig-Maximilians-Universität München, Diplomarbeit, 2011.

[41] T. Broeders und T.C.M. Rientveld: *Segmental marking as a cue in auditory voice identification of telephone speech. Segmental marking as a cue in auditory voice identification of telephone speech*, In *Eurospeech*. ISCA (1989) Seiten 1071–1074.

[42] C. Byrne und P. Foulkes, *The International Journal of Speech, Language and the Law* **11** (2004), 83.

[43] J.P. Campbell: *Speaker Recognition: A Tutorial. Speaker Recognition: A Tutorial*, In *Proceedings of the IEEE*, Band 85. (1997).

[44] S. Cassidy und J. Harrington, *Speech Communication* **33** (2001), 61.

[45] M. Clayards, M.K. Tanenhaus, R.N. Aslin und R.A. Jacobs, *Cognition* **108** (2008), 802.

[46] O. Cooney: *Acoustic analysis of the effects of alcohol on the human voice.* Dublin City University, Diplomarbeit, 1998.

[47] R.M. Cooper, *Cognitive Psychology* **6** (1974), 84.

[48] S.C. Creel, R.N. Aslin und M.K. Tanenhaus, *Cognition* **106** (2008), 633.

[49] S.C. Creel und M.R. Bregman, *Language and Linguistics Compass* **5** (2011), 190.

[50] S.C. Creel und M.A. Tumlin, *Journal of Memory and Language* **65** (2011), 264.

[51] R.G. Crowder, *Perception & Psychophysics* **31** (1982), 477.

[52] D. Crystal: *Prosodic Systems and Intonation in English.* London: Cambridge University Press, 1969.

[53] A. Cutler. *The Handbook of Speech Perception.* Blackwell, Oxford, 2005. Kapitel Lexical Stress, Seiten 264–289.

[54] A. Cutler, A. Andics und Z. Fang: *Inter-Dependent Catagorization of Voices and Segments. Inter-Dependent Catagorization of Voices and Segments*, In *Proceedings of the 17th International Congress of Phontic Sciences (ICPhS).* Hong Kong (2011) Seiten 552–555.

[55] A. Cutler und D. Norris, *Journal of Experimantal Psychology, Human Perception and Performance* **14** (1988), 113.

[56] D. Dahan, S.J. Drucker und R.A. Scarborough, *Cognition* **108** (2008), 710.

[57] D. Dahan, J.S. Magnuson und M.K. Tanenhaus, *Cognitive Psychology* **42** (2001), 317.

[58] D. Dahan, J.S. Magnuson, M.K. Tanenhaus und E.M. Hogan, *Language and Cognitive Processes* **16** (2001), 507.

[59] D. Dahan und M.K. Tanenhaus, *Journal of Experimantal Psychology: Learning, Memory and Cognition* **30** (2004), 498.

[60] M.H. Davis und W.D. Marslen-Wilson, *Journal of Experimantal Psychology, Human Perception and Performance* **28** (2002), 218.

[61] P.C. Delattre, A.M. Liberman und F.S. Cooper, *Journal of the Acoustical Society of America* **27** (1955), 769.

[62] V. Dellwo, A. Leemann und M.J. Kolly: *Speaker idiosyncratic rhythmic features in the speech signal. Speaker idiosyncratic rhythmic features in the speech signal*, In *13th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Portland, Oregon, USA (2012).

[63] V. Dellwo, A. Leemann, M.J. Kolly und M. Meyer: *Auditory speaker identification based on suprasegmental temporal characteristics. Auditory speaker identification based on suprasegmental temporal characteristics*, In *22th Annual Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA)*. Tampa, Florida (2013).

[64] C. Draxler und K. Jänsch: *SpeechRecorder - a Universal Platform Independent Multi-Channel Audio Recording Software. SpeechRecorder - a Universal Platform Independent Multi-Channel Audio Recording Software*, In *Proc. of the IV. International Conference on Language Resources and Evaluation*. Lisbon, Portugal (2004) Seiten 559–562.

[65] C. Draxler, F. Schiel und T. Ellbogen: *F0 of Adolescent speakers - First Results for the German Ph@tt Sessionz Database. F0 of Adolescent speakers - First Results for the German Ph@tt Sessionz Database*, In *Proceedings of LREC*. (2008).

[66] A.T. Duchowski: *Eye Tracking Methodology: Theory and Practice*. Springer, 2003.

[67] M. Duckworth und K. McDougall: *Assessing the consistency of disfluency measures in characterising speakers. Assessing the consistency of disfluency measures in characterising speakers*, In *23th Annual Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA)*. Zurich, Switzerland (2014).

[68] J.P. Eatock und J.S.D. Mason: *A Quantitative assessment of the relative speaker discriminating properties of phonemes. A Quantitative assessment of the relative speaker discriminating properties of phonemes*, In *IEEE*. (1994).

[69] G.M. Edelman: *Neural Darwinism: The Theory of Neuronal Group Selection*. New York: Basic Books, 1987.

[70] B. Efron und C. Stein, *The Annals of Statistics* **9** (1981), 586.

[71] F. Eisner und J.M. McQueen, *Perception & Psychophysics* **67** (2005), 224.

[72] E. Enzinger und P. Balazs: *Speaker Verification using Pole/Zero Estimates of Nasals. Speaker Verification using Pole/Zero Estimates of Nasals*, In *Proceedings of the Multi-Conference on Systems & Structures.* (2011).

[73] E.J. Eriksson: *That voice sounds familiar: factors in speaker recognition.* Umea University, Doctoral thesis, 2007.

[74] M. Ernestus, I. Hanique und E. Verboom, *Journal of Phonetics* **48** (2014), 60.

[75] G. Fairbanks und W. Pronovost, *Speech Monographs* **6** (1939), 87.

[76] R. Faltlhauser und G. Ruske: *Improving speaker recognition using phonetically structured Gaussian mixture models. Improving speaker recognition using phonetically structured Gaussian mixture models*, In *7th European Conference on Speech Communication and Technology (Eurospeech).* Aalborg, Denmark (2001).

[77] G. Fant: *Acoustic Theory of Speech Production.* The Hague: Mouton, 1960.

[78] G. Fant: *Speech Sounds and Features.* MA: MIT Press, Cambridge, 1973.

[79] F. Farahani, P.G. Georgiou und S.S. Narayanan: *Speaker identification using supra-segmental pitch pattern dynamics. Speaker identification using supra-segmental pitch pattern dynamics*, In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP).* Montreal, Canada (2004) Seiten 89–92.

[80] M. Farrùs, J. Hernando und P. Ejarque: *Jitter and Shimmer measurements for Speaker Recognition. Jitter and Shimmer measurements for Speaker Recognition*, In *Annual Conference of the International Speech Communication Association (INTERSPEECH).* (2007).

[81] N. Fecher und D. Watt: *Speaking under cover: The impact of face-concealing garments on the acoustics of fricatives. Speaking under cover: The impact of face-concealing garments on the acoustics of fricatives*, In *The 17th International Congress of Phonetic Sciences.* Hong Kong, China (August 2011).

[82] N. Fecher und D. Watt: *Speaker discrimination based on facewear-speech. Speaker discrimination based on facewear-speech*, In *23th Annual Conference of the Internatio-*

nal Association for Forensic Phonetics and Acoustics (IAFPA). Zurich, Switzerland (2014).

[83] J.M. Fellowes und R.E. Remez, *Perception & Psychophysics* **59** (1997), 839.

[84] E. Fischer-Jørgensen, *Miscellenea Phonetica* **2** (1954), 42.

[85] J. Fischer-Weppler, M. Jessen und F. Schiel: *The Effect of the 'Telephone Situation' on Formant Frequencies. The Effect of the 'Telephone Situation' on Formant Frequencies*, In *IAFPA 2010 Annual Conference*. International Association for Forensic Phonetics and Acoustics (July 2010).

[86] W.T. Fitch, *The Journal of the Acoustical Society of America* **102** (1997), 1213.

[87] W.T. Fitch und J. Giedd, *J Acoust Soc Am* **106** (1999), 1511.

[88] J.L. Flanagan: *Speech Synthesis, Analysis and Perception*. New York: Springer-Verlag, 1972.

[89] H.J. Foley. *Introduction to the Analysis of Variance*, 2013.

[90] I. Fónagy und K. Magdics, *Phonetica* **16** (1963), 293.

[91] K. Forrest, G. Weismer, P. Milenkovic und R.N. Dougall, *Journal of the Acoustic Society of America* **84** (1988), 115.

[92] R.A. Fox und S.L. Nissen, *Journal of Speech, Language, and Hearing Research* **48** (2005), 753.

[93] H. Fu, R.D. Rodman, D.F. McAllister, D.L. Bitzer und B. Xu: *Classification of Voiceless Fricatives through Spectral Moments. Classification of Voiceless Fricatives through Spectral Moments*, In *SCI and ISAS International Conference*. International Institute of Informatics and Systemics (1999).

[94] O. Fujimura, *Journal of the Acoustic Society of America* **34** (1962), 18.

[95] S. Furui, *Journal of the Acoustic Society of America* **80** (1986), 1016.

[96] L.F. Gallardo: *Human and Automatic Speaker Recognition over Telecommunication Channels*. Springer, 2015.

[97] F. Galton: *Inquiries into human faculty and its development.* London: Macmillan, 1883.

[98] S.G. Gfrörer: *Auditory-instrumental forensic speaker recognition. Auditory-instrumental forensic speaker recognition*, In *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech).* Geneva, Switzerland (2003) Seiten 705–708.

[99] H.R. Gilbert und G.G. Weismer, *Journal of Psycholinguistic Research* **3** (1974), 225.

[100] Glen. *Effect of sample size in F-test*, March 2015.

[101] J.W. Glenn und N. Kleiner, *Journal of the Acoustic Society of America* **43** (1968), 368.

[102] S.D. Goldinger, *Journal of Experimantal Psychology: Learning, Memory and Cognition* **22** (1996), 1166.

[103] S.D. Goldinger, *Psychological Review* **105** (1998), 251.

[104] J. González und A. Carpi, *Medical Science Monitor* **10** (2004), 649.

[105] M. Gordon, P. Barthmaier und K. Sands, *Journal of the International Phonetic Association* **32** (2002), 141.

[106] D. Graddol und J. Swann, *Language and Speech* **26** (1983), 351.

[107] T. Hannagan, J.S. Magnuson und J. Grainger, *Frontiers in Psychology* **4** (2013), 79.

[108] J. Harrington. *A Handbook of Phonetics.* Wiley-Blackwell: Oxford, 2010. Kapitel Acoustic Phonetics, Seiten 81–129.

[109] J. Harrington und S. Cassidy: *Techniques in Speech Acoustics.* Kluwer Academic Publishers, Dordrecht, Boston, London, 1999.

[110] P. Harrison: *Variability of Formant Measurements.* University of York, Diplomarbeit, September 2004.

[111] M.R. Hasan, M. Jamil und M.G.R.M.S. Rahman: *Speaker identification using mel frequency cepstral coefficients. Speaker identification using mel frequency cepstral coefficients*, In *The 3rd International Conference on Eletrical & Computer Engineering*, Band 1. (2004) Seite 4.

[112] S. Hawkins. *The Acoustics of Speech Communication - Fundamentals, Speech Perception Theory, and Thechnology*. Allyn and Bacon, 1998. Kapitel Looking for the invariant correlates of linguistic units: Two classical theories of speech perception, Seiten 198–231.

[113] W.J. Hess, K.J. Kohler und H.G. Tillmann: *The Phondat-Verbmobil Speech Corpus. The Phondat-Verbmobil Speech Corpus*, In *Fourth European Conference on Speech Communication and Technology*. EUROSPEECH '95, Madrid, Spain (September 1995).

[114] H.v.d. Heuvel und T. Rietveld: *Speaker related variability in cepstral representations of dutch speech segments. Speaker related variability in cepstral representations of dutch speech segments*, In *Second International Conference on Spoken Language Processing*. (1992).

[115] J. Hillenbrand, L.A. Getty, M.J. Clark und K. Wheeler, *Journal of the Acoustic Society of America* **97** (1995), 3099.

[116] J.M. Hillenbrand und M.J. Clark, *Attention, Perception & Psychophysics* **71** (2009), 1150.

[117] D.L. Hintzman, *Psychological Review* **93** (1986), 411.

[118] A. Hirson und M. Duckworth, *Beiträge zur Phonetik und Linguistik* **64** (1995), 67.

[119] H. Hollien, R. Green und K. Massey, *Journal of the Acoustic Society of America* **96** (1994), 2646.

[120] H.F. Hollien und T. Ship, *Journal of Speech and Hearing Research* **15** (1972), 155.

[121] D.M. Houston und P.W. Jusczyk, *Journal of Experimantal Psychology, Human Perception and Performance* **29** (2003), 1143.

[122] F. Huettig und G.T.M. Altmann, *Cognition* **96** (2005), B23.

[123] F. Huettig und J.M. McQueen, *Journal of Memory and Language* **57** (2007), 460.

[124] F. Huettig, R.K. Mishra und C.N.L. Olivers, *Frontiers in Psychology* **2** (2012), 1.

[125] F. Huettig, J. Rommers und A.S. Meyer, *Acta Psychologica* **137** (2011), 151.

[126] U. Ilg und P. Thier. *Neuropsychologie*. Springer-Lehrbuch. Springer Berlin Heidelberg, 2006. Kapitel Zielgerichtete Augenbewegungen, Seiten 296–307.

[127] A.E.F. J. M. Chambers und R.M. Heiberger. *Statistical Models in S*. Wadsworth & Brooks/Cole, California, 1992. Kapitel Analysis of Variance: Designed Experiments, Seiten 145–193.

[128] T.F. Jaeger, *Journal of Memory and Language* **59** (2008), 434.

[129] A. Jesse und J.M. McQueen: *Prelexical Adjustments to Speaker Idiosyncrasies: Are they Position-specific? Prelexical Adjustments to Speaker Idiosyncrasies: Are they Position-specific?*, In *Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Antwerp, Belgium (2007).

[130] A. Jesse, J.M. McQueen und M. Page: *The Locus of Talker-Specific Effects in Spoken-Word Recognition. The Locus of Talker-Specific Effects in Spoken-Word Recognition*, In *The 16th International Congress of Phonetic Sciences*. Saarbrücken (6-10 August 2007).

[131] M. Jessen: *Einfluss von stress auf Sprache und Stimme - Unter besonderer Beruecksichtigung polizeilicher Anforderungen*. Schulz-Kirchner Verlag, 2006.

[132] Q. Jin, A.R. Toth, A.W. Black und T. Schultz: *Is Voice Transformation a Threat to Speaker Identification? Is Voice Transformation a Threat to Speaker Identification?*, In *IEEE International Conference on Acoustic, Speech, and Signal Processing*. Las Vegas (March 2008) Seiten 4845 – 4848.

[133] K. Johnson, *OSU Working Papers in Linguistics* **50** (1997), 101.

[134] K. Johnson: *Acoustic and Auditory Phonetics*. Oxford: Blackwell, 2004.

[135] K. Johnson, *Journal of Phonetics* **34** (2006), 485.

[136] T. Johnstone und K.R. Scherer: *The effects of emotions on voice quality. The effects of emotions on voice quality*, In *Proceedings of the XIVth International Congress of Phonetic Sciences*. San Francisco, USA (27/06/2005 1999).

[137] A. Jongman, R. Wayland und S. Wong, *The Journal of the Acoustical Society of America* **108** (2000), 1252.

[138] M. Joos, M. Rötting und B.M. Velichkovsky: *Die Bewegungen des menschlichen Auges: Fakten, Methoden, innovative Anwendungen.* Berlin & NY: de Gruyter, 2003.

[139] M.A. Just und P.A. Carpenter, *Psychological Review* **87** (1980), 329.

[140] S.S. Kajarekar und H. Hermansky: *Speaker verification based on broad phonetic categories. Speaker verification based on broad phonetic categories*, In *2001: A Speaker Odyssey-The Speaker Recognition Workshop.* (2001).

[141] R. Kashyap, *Acoustics, Speech and Signal Processing* **24** (1976), 481.

[142] C. Kavanagh: *Speaker discrimination using English nasal durations and formants. Speaker discrimination using English nasal durations and formants*, In *The International Association for Forensic Phonetics and Acoustics.* (2010).

[143] C. Kavanagh: *Inter- and intra-speaker variability in acoustic properties in English /s/. Inter- and intra-speaker variability in acoustic properties in English /s/*, In *The International Association for Forensic Phonetics and Acoustics.* (2011).

[144] P. Keating und R. Buhr, *Journal of the Acoustic Society of America* **63** (1978), 567.

[145] P. Keating und G. Kuo, *Journal of the Acoustic Society of America* **132** (2012), 1050.

[146] D. Kewley-Port, *The Journal of the Acoustical Society of America* **73** (1983), 322.

[147] D. Kewley-Port, D.B. Pisoni und M. Studdert-Kennedy, *Journal of the Acoustic Society of America* **73** (1983), 1779.

[148] T. Kinnunen. *Spectral Features for Automatic Text-Independent Speaker Recognition.* Licentiate's thesis, University of Joensuu, Joensuu, Finland, December 2003.

[149] T. Kinnunen und H. Li: *An Overview of Text - Independent Speaker Recognition: from Features to Supervectors. An Overview of Text - Independent Speaker Recognition: from Features to Supervectors*, In *Proceedings of Speech Communication.* (2009).

[150] D.F. Kleinschmidt und T.F. Jaeger, *Psychological Review* **122** (2015), 148.

[151] K.R. Kluender, J.A. Coady und M. Kiefte, *Speech Communication* **41** (2003), 59.

[152] T.R. Knösche, S. Lattner, B. Maess, M. Schauer und A.D. Fiederici, *NeuroImage* **17** (2002), 1493.

[153] C. Koch: *Rolle visueller Referenzen bei der Objektlokalisation.* München, Ludwig-Maximilians-Universität, Dissertation, 2005.

[154] M.J. Kolly, A. Leemann, P.B. de Mareüil und V. Dellwo: *Speaker-idiosyncrasy in pausing behavior: Evidence from a cross-linguistic study. Speaker-idiosyncrasy in pausing behavior: Evidence from a cross-linguistic study*, In *Proceedings of the 18th International Congress of Phonetic Sciences.* Glasgow, Scotland, UK (August 2015).

[155] W. König: *Dtv-Atlas zur deutschen Sprache: Tafeln und Texte;[mit Mundart-Karten]*, Band 3025. Deutscher Taschenbuch Verlag, 1978.

[156] J. Koreman: *The role of articulation rate in distinguishing fast and slow speakers. The role of articulation rate in distinguishing fast and slow speakers*, In *Proceedings of the 3rd International Conference on Speech Prosody.* Dresden, Germany (May 2006).

[157] T. Kraljic und A.G. Samuel, *Journal of Memory and Language* **56** (2007), 1.

[158] J.K. Kruschke, *Psychological Review* **99** (1992), 22.

[159] H.J. Künzel, *The International Journal of Speech, Language and the Law* **8** (2001), 80.

[160] K. Kurowski und S.E. Blumenstein, *Journal of the Acoustic Society of America* **81** (1987), 1917.

[161] A. Kuznetsova, P. Bruun Brockhoff und R. Haubo Bojesen Christensen. *lmerTest: Tests in Linear Mixed Effects Models* (2015). R package version 2.0-25.

[162] P. Ladefoged und D.E. Broadbent, *The Journal of the Acoustical Society of America* **29** (1957), 98.

[163] P. Ladefoged und J. Ladefoged, *UCLA Working Papers in Phonetics* **49** (1980), 43.

[164] E.J.C. Laing, R. Liu, A.J. Lotto und L.L. Holt, *Frontiers in Psychology* **3** (2012), 1.

[165] D. van Lancker und J. Kreiman, *Neuropsychologia* **25** (1987), 829 .

[166] D.V. Lancker und J. Kreiman, *Journal of Phonetics* **13** (1985), 39.

[167] D.V. Lancker, J. Kreiman und K. Emmorey, *Journal of Phonetics* **13** (1985), 19.

[168] C.L. LaRiviere: *Some acoustic and perceptual correlates of speaker identification. Some acoustic and perceptual correlates of speaker identification*, In *Proceedings of the 7th International Congress of Phonetic Sciences.* (1972).

[169] A.M. Laukkanen, E. Vilkman, P. Alku und H. Oksanen, *Journal of Phonetics* **24** (1996), 313.

[170] J. Laver: *Phonetic Description of voice quality.* Cambridge: University Press, 1980.

[171] Y. Lavner, I. Gath und J. Rosenhouse, *International Journal of Speech Technology* **4** (2001), 63.

[172] N. Li und P.C. Loizou, *Journal of the Acoustic Society of America* **124** (2008), 3947.

[173] A.M. Liberman, P.C. Delattre, F.S. Cooper und L.J. Gerstman, *Psychological Monographs: General and Applied* **68** (1954), 1.

[174] A.M. Liberman und I.G. Mattingly, *Cognition* **21** (1985), 1.

[175] J. Lindh: *Preliminary Descriptive F0-staistics for Young Male Speakers. Preliminary Descriptive F0-staistics for Young Male Speakers*, In *Conference of the International Assiociation for Forensic Phonetics and Acoustics.* Gothenburg (2006) Seiten 89–92.

[176] L. Lisker und A.S. Abramson, *WORD* **20** (1964), 384.

[177] D. Loakes: *Front Vowels as Speaker-Specific: Some Evidence from Australian English. Front Vowels as Speaker-Specific: Some Evidence from Australian English*, In *Proceedings of the 10th Australian International Conference on Speech Science and Technology.* (2004).

[178] D. Loakes: *A forensic phonetic investigation into the speech patterns of identical and non-identical twins.* The University of Melbourne, Dissertation, 2006.

[179] R. Lorenzen: *Eine akustisch-phonetische Untersuchung zur Stimmverstellung.* Christian-Albrechts-Universität, Diplomarbeit, 2004.

[180] L. Loveday, *Language and Speech* **24** (1981), 71.

[181] J.S. Magnuson, T. Strauss und H.D. Harris: *Interaction in spoken word recognition models: Feedback helps. Interaction in spoken word recognition models: Feedback helps*,

In *Proceedings of the 27th Annual Meeting of the Cognitive Science Society*. (2005) Seiten 1379–1394.

[182] A. Malécot, *Language* **32** (1956), 274.

[183] V. Marrero, J. Gil und E. Battaner: *Inter-speaker variation in Spanish. an experimental and acoustic preliminary approach. Inter-speaker variation in Spanish. an experimental and acoustic preliminary approach*, In *Proceedings of the 15th International Congress of Phonetic Sciences*. (2003) Seiten 703–706.

[184] J.L. McClelland und J.L. Elman, *Cognitive Psychology* **18** (1986), 1.

[185] J.L. McClelland, D. Mirman und L.L. Holt, *TRENDS in Cognitive Sciences* **10** (2006), 363.

[186] K. McDougall, *International Journal of Speech Language and the Law* **11** (2004), 103.

[187] K. McDougall und F. Nolan: *Discrimination of Speakers using the Formant Dynamics of /u:/ in British English. Discrimination of Speakers using the Formant Dynamics of /u:/ in British English*, In *ICPhS XVI*. Saarbrücken (6-10 August 2007 2007).

[188] C.T. McLennan und P.A. Luce, *Journal of Experimantal Psychology: Learning, Memory and Cognition* **31** (2005), 306.

[189] B. McMurray, M.A. Clayards, M.K. Tanenhaus und R.N. Aslin, *Psychonomic Bulletin & Review* **15** (2008), 1064.

[190] J.M. McQueen. *The Handbook of Cognition*. London: Stage Publications, 2005. Kapitel Speech perception, Seiten 255–275.

[191] J.M. McQueen. *The Oxford Handbook of Psycholinguistics*. Nummer 3. Oxford University Press, 2007. Kapitel Eight questions about spoken word recognition, Seiten 38–53.

[192] J.M. McQueen, D. Norris und A. Cutler, *TRENDS in Cognitive Sciences* **10** (2006), 533.

[193] J.M. McQueen und M.C. Viebahn, *The Quarterly Journal of Experimental Psychology* **60** (2007), 661.

[194] H. Meier: *Deutsche Sprachstatistik*. Olms Verlag, Hildesheim, 1967.

[195] L. Ménard, J.L. Schwartz und L..J. Boë, *Journal of Speech, Language and Hearing Research* **47** (2004), 1059.

[196] D. Michaelis, M. Fröhlich und H.W. Strube, *Journal of the Acoustic Society of America* **103** (1998), 1628.

[197] J. Miller, T. Patterson und R. Ulrich, *Psychophysiology* **32** (1998), 99.

[198] R.G. Miller, *Biometrika* **61** (1974), 1.

[199] G.S. Morrison, *Science and Justice* **49** (2009), 298.

[200] C. Müller: *Zweistufige kontextsensitive Sprecherklassifikation am Beispiel von Alter und Geschlecht*. Naturwissenschaftlich-Technische Fakultät, Universität des Saarlandes, Dissertation, 2005.

[201] C.H. Murphy und P.C. Doyle, *Otolaryngology - Head and Neck Surgery* **97** (1987), 376.

[202] D.G. Myers, C. Grosser und S. Hoppe-Graff: *Psychologie*. Springer, 2004.

[203] S.S. Narayanan, A.A. Alwan und K. Haker, *Journal of the Acoustic Society of America* **98** (1995), 1325.

[204] S. Neuhauser, *Journal of Speech, Language and the Law* **15** (2008), 131.

[205] K.M. Newell und P.A. Hancock, *Journal of Motor Behavior* **16** (1984), 320.

[206] R.S. Newman, S.A. Clouse und J.L. Burnham, *Journal of the Acoustic Society of America* **109** (2001), 1181.

[207] S. Nittrouer, M. Studdert-Kennedy und R.S. McGowan, *Journal of Speech and Hearing Research* **32** (1989), 120.

[208] F. Nolan: *The phonetic bases of speaker recognition*. Cambrige University Press, 1983.

[209] F. Nolan, *Forensic Linguistics* **9** (2002), 1.

[210] F. Nolan, K. McDougall und T. Hudson. *Research Report on 'Voice similarity and the effect of the telephone: a study of the implications for earwitness evidence (VoiceSim)'*. Technischer Bericht, Department of Linguistics, University of Cambridge, 2009.

[211] F. Nolan, K. McDougall, G. de Jong und T. Hudson: *A Forensic Study of "Dynamic" Sources of Variability in Speech: The DyViS Project. A Forensic Study of "Dynamic" Sources of Variability in Speech: The DyViS Project*, In *Proceedings of the 11th Australasian International Conference on Speech Science and Technology.* (December 2006) Seiten 13–18.

[212] D. Norris, *Cognition* **52** (1994), 189.

[213] D. Norris und J.M. McQueen, *Psychological Review* **115** (2008), 357.

[214] D. Norris, J.M. McQueen und A. Cutler, *Behavioral and Brain Sciences* **23** (2000), 299.

[215] D. Norris, J.M. McQueen und A. Cutler, *The Behavioral and Brain Sciences* **23** (2000), 299.

[216] D. Norris, J.M. McQueen und A. Cutler, *Cognitive Psychology* **47** (2003), 204.

[217] R.M. Nosofsky, *Journal of Experimantal Psychology: General* **115** (1986), 39.

[218] L.C. Nygaard und D.B. Pisoni, *Perception & Psychophysics* **60** (1998), 355.

[219] L.C. Nygaard, M.S. Sommers und D.B. Pisoni, *Psychological Sciences* **5** (1994), 42.

[220] R.N. Ohde und K.N. Stevens, *Journal of the Acoustic Society of America* **74** (1983), 706.

[221] S.E. Öhman, *The Journal of the Acoustical Society of America* **39** (1966), 151.

[222] M.J. Owren und G.C. Cardillo, *Journal of the Acoustic Society of America* **119** (2006), 1727.

[223] C. Pallier. *Computing discriminability and bias with the R software*, 2002.

[224] T.J. Palmeri, S.D. Goldinger und D.B. Pisoni, *Journal of Experimantal Psychology: Learning, Memory and Cognition* **19** (1993), 309.

[225] G. Parikh und P.C. Loizou, *Journal of the Acoustic Society of America* **118** (2005), 3874.

[226] M.I. Pegoraro-Krook, *Folia Phoniatrica et Logopaedica* **40** (1988), 82.

[227] C.R. Pemet, P. Belin und A. Jones, *Frontiers in Psychology* **4** (2014), 1.

[228] E. Pépiot: *Male and female speech: a study of mean f0, f0 range, phonation type and speech rate in Parisian French and American English speakers. Male and female speech: a study of mean f0, f0 range, phonation type and speech rate in Parisian French and American English speakers*, In *Speech Prosody 7*. (2014).

[229] G.E. Peterson und H.L. Barney, *Journal of the Acoustic Society of America* **24** (1952), 175.

[230] J.M. Pickett: *The Acoustics of Speech Communication - Fundamentals, Speech Perception Theory, and Technology*. Allyn and Bacon, 1998.

[231] J.B. Pierrehumbert. *Laboratory Phonology 7*. De Gruyter Publishing, 2002. Kapitel Word-specific phonetics, Seiten 1–24.

[232] J.B. Pierrehumbert, *Language and Speech* **46** (2003), 115.

[233] O.d. Pinto und H. Hollien, *Journal of Phonetics* **10** (1982), 367.

[234] D.B. Pisoni und C.S. Martin, *Alcoholism: Clinical and Experimental Research* **13** (1989), 577.

[235] B. Pompino-Marschall: *Einführung in die Phonetik*. Gruyter, 1995.

[236] R.K. Potter, G.A. Kopp und H.G. Kopp: *Visible Speech*. David Van Nostrand, 1947.

[237] T. Pruthi und C.Y. Espy-Wilson: *An MRI Based Study of the Acoustic Effects of Sinus Cavities and Its Application to Speaker Recognition. An MRI Based Study of the Acoustic Effects of Sinus Cavities and Its Application to Speaker Recognition*, In *Ninth International Conference on Spoken Language Processing*. INTERSPEECH 2006 - ICSLP, Pittsburgh, PA, USA (September 2006).

[238] I. Psychology Software Tools. *E-Prime 2.0*, 2012.

[239] M. Pützer: *Stimmqualität und Artikulation bei Dysarthrophonien in der individuellen, tendenziellen und referentiellen Bewertung*. Institut für Phonetik, Universität des Saarlandes, Dissertation, 2008.

[240] M.H. Quenouille, *Biometrika* **43** (1956), 353.

[241] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2015).

[242] W. Rappaport, *Acustica* **8** (1958), 220.

[243] E. Reinisch. *Perl-Skript zur Randomisierung der Target- und Kompetitorpostitionen als Input-Datei für den Eye-Tracker*.

[244] E. Reinisch. *Perl-Skript zur Umwandlung von Eye-Tracking Ausgabedateien in Datentabellen*.

[245] E. Reinisch und M.J. Sjerps, *Journal of Phonetics* **41** (2013), 101.

[246] R.E. Remez, J.M. Fellowes und P.E. Rubin, *Journal of Experimantal Psychology, Human Perception and Performance* **23** (1997), 651 .

[247] B.H. Repp, *Journal of the Acoustic Society of America* **79** (1986), 1987.

[248] U. Reubold, J.M. Harrington und F. Kleber, *Speech Communication* **52** (2010), 638.

[249] D.A. Reynolds, *Speech Communication* **17** (1995), 91.

[250] D.A. Reynolds, M. Zissman, T.F. Quatieri, G.C. O'Leary, B.A. Carlson et al. : *The effects of telephone transmission degradations on speaker recognition performance. The effects of telephone transmission degradations on speaker recognition performance*, In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Band 1. IEEE (1995) Seiten 329–332.

[251] D.C. Richardson und M.J. Spivey. *Encyclopedia of Biomaterials and Biomedical Engineering*. Marcel Dekker, 2004. Kapitel Eye Tracking: Characteristics And Methods, Seiten 1028–1032.

[252] D.C. Richardson und M.J. Spivey. *Encyclopedia of Biomaterials and Biomedical Engineering*. Marcel Dekker, 2004. Kapitel Eye-Tracking: Research Areas and Applications, Seiten 1033–1042.

[253] R.D. Rodman: *Speaker recognition of Disguised Voices: A Program for research. Speaker recognition of Disguised Voices: A Program for research*, In *Proceedings of the Consortium on Speech Technology in Conjunction with the Conference on Speaker Recognition by Man and Machine: Directions for Forensic Applications, Ankara, Turkey, COST250 Publishing Arm.* (1998) Seiten 9–22.

[254] P. Rose: *Forensic Speaker Identification*. Taylor & Francis, London, New York, 2002.

[255] P. Rose: *Forensic Speaker Discrimination With Australian English Vowel Acoustics. Forensic Speaker Discrimination With Australian English Vowel Acoustics*, In *The 16th International Congress of Phonetic Sciences*. Saarbrücken (6-10 August 2007 2007).

[256] R.C. Rose, E.M. Hofstetter und D.A. Reynolds, *Transactions on Speech and Audio Processing (IEEE)* **2** (1994), 245.

[257] A.P. Salverda, D. Dahan und J.M. McQueen, *Cognition* **90** (2003), 51.

[258] A.P. Salverda, D. Kleinschmidt und M.K. Tanenhaus, *Journal of Memory and Language* **71** (2014), 145.

[259] A.G. Samuel und T. Kraljic, *Attention, Perception & Psychophysics* **71** (2009), 1207.

[260] D.L. Schacter, J.E. Eich und E. Tulving, *Journal of Verbal Learning and Verbal Behavior* **17** (1978), 721.

[261] K.R. Scherer, *Journal of Psycholinguistic Research* **3** (1974), 281.

[262] F. Schiel: *MAUS goes iterative. MAUS goes iterative*, In *4th International conference on Language resources and evaluation (LREC)*. Lisbon, Portugal (2004) Seiten 1015–1018.

[263] F. Schiel, C. Draxler und J. Harrington: *Phonemic Segmentation and Labelling using the MAUS Technique. Phonemic Segmentation and Labelling using the MAUS Technique*, In *Proceedings of the Workshop 'New Tools and Methods for Very-Large-Scale Phonetics Research'*. University of Pennsylvania, USA (2011).

[264] L. Sichelschmidt, *ZiF-Mitteilungen* **1995** (1995), 1.

[265] M. Sigmund. *Frontiers in Robotics, Automation and Control*. InTech, October 2008. Kapitel Automatic Speaker Recognition by Speech Signal, Seiten 41–54.

[266] N. Silbert und K. de Jong, *Journal of the Acoustic Society of America* **123** (2008), 2769.

[267] M.J. Sjerps, H. Mitterer und J.M. McQueen, *Neuropsychologia* **49** (2011), 3831.

[268] R. Skarnitzl, J. Vaňková und L. Weingartová: *Speaker discrimination using short- and long-term segmental information in vowels. Speaker discrimination using short- and long-term segmental information in vowels*, In *Annual Conference of the International Association for Forensic Phonetics and Acoustics*. (2012).

[269] R. Smits, L.T. Bosch und R. Colliert, *Journal of the Acoustic Society of America* **100** (1996), 3852.

[270] D. Sorensen und Y. Horii, *Journal of Communication Disorders* **15** (1982), 135.

[271] L. SR Research. *Experiment Builder*, 2013.

[272] K.N. Stevens, *The Journal of the Acoustical Society of America* **50** (1971), 1180.

[273] K.N. Stevens, *Journal of the Acoustic Society of America* **111** (2002), 1872.

[274] K.N. Stevens und S.E. Blumenstein, *Journal of the Acoustic Society of America* **64** (1978), 1358.

[275] K.N. Stevens und A.S. House, *Journal of the Acoustic Society of America* **28** (1956), 578.

[276] M.L. Stoicheff, *Journal of Speech and Hearing Research* **24** (1981), 437.

[277] W. von Suchodoletz. *Diagnostik sonderpädagogischen Förderbedarfs*, Band 5 von *Jahrbuch der pädagogisch-psychologischen Diagnostik. Tests und Trends.* Hogrefe Verlag GmbH & Co, 2006. Kapitel Diagnostik bei Artikulationsstörungen, Seiten 187–210.

[278] H. Suzuki, T. Nakai und H. Sakakibara: *Analysis of Acoustic Properties of the Nasal Tract Using 3-D FEM. Analysis of Acoustic Properties of the Nasal Tract Using 3-D FEM*, In *4th International Conference on Spoken Language Processing*. Philadelphia, PA, USA (October 1996).

[279] M.K. Tanenhaus und M.J. Spivey-Knowlton, *Language and Cognitive Processes* **11** (1996), 583.

[280] M.K. Tanenhaus, M.J. Spivey-Knowlton, K.M. Eberhard und J.C. Sedivy, *Science* **268** (1995), 1632.

[281] K. Tanner, N. Roy, A. Ash und E.H. Buder, *Journal of Voice* **19** (2005), 211.

[282] M. Thompson. *Colored noise*, 1989.

[283] H.G. Tillmann und P. Mansell: *Phonetik. Lautsprachliche Zeichen Sprachsignale und lautsprachlicher Kommunikationsprozess.* Klett, Stuttgart, 1980.

[284] V. Tiwari, *International Journal on Emerging Technologies* **1** (2010), 19.

[285] H. Traunmüller und A. Eriksson. *The frequency range of the voice fundamental in the speech of male and female adults* (December 1995). Retrieved from <http://www.ling.su.se/staff/hartmut/aktupub.htm>.

[286] A.M. Trude und S. Brown-Schmidt, *Language and Cognitive Processes* **27** (2011), 979.

[287] J.W. Tukey, *The Annals of Mathematical Statistics* **29** (1958), 614.

[288] M. Vasilakis und Y. Stylianou, *Folia Phoniatrica et Logopaedica* **61** (2009), 153.

[289] M. Weirich, *ZASPiL* **52** (2010), 19.

[290] C.E. Williams: *Aural speaker recognition and speech communication system evaluation. Aural speaker recognition and speech communication system evaluation*, In *Proceedings oth the 7th International Congress on Acoustics.* Budapest (1971).

[291] C.E. Williams und K.N. Stevens, *Journal of the Acoustic Society of America* **52** (1972), 1238.

[292] B.R. Witkin, *International Journal of Listening* **4** (1990), 7.

[293] M.J. Witteman, N.P. Bardhan, A. Weber und J.M. McQueen, *Language and Speech* **May 06** (2014), 1.

[294] M.J. Witteman, A. Weber und J.M. McQueen, *Attention, Perception & Psychophysics* **75** (2013), 537.

[295] J.J. Wolf, *Journal of the Acoustic Society of America* **51** (1972), 2044.

[296] R. Wright: *Intra-speaker variation and units in human speech perception and ASR. Intra-speaker variation and units in human speech perception and ASR*, In *Speech Recognition and Intrinsic Variation Workshop.* (2006).

[297] I.P. Yuasa, *American Speech* **85** (2010), 315.