

# Altersbedingte Veränderungen der Stimme und Sprechweise von Frauen

Eine perzeptive und akustische Analyse

vorgelegt von  
Markus Brückl M.A.

von der Fakultät I – Geisteswissenschaften  
der Technischen Universität Berlin  
zur Erlangung des akademischen Grades  
Doktor der Philosophie  
– Dr. phil. –  
genehmigte Dissertation

Promotionsausschuss:  
Vorsitzende: Prof. Dr. phil. Heike Rohmann  
Erster Bericht: Prof. Dr. phil. Walter Sendlmeier  
Zweiter Bericht: Prof. Dr.-Ing. Klaus Hobohm

Tag der wissenschaftlichen Aussprache: 27. September 2011

Berlin  
D 83

abzuschätzen, bzw. die linguistische Modulation zu kontrollieren. Um dies zu bewerkstelligen erfolgt die Untersuchung gehaltener Vokale und die Segmentation und Transkription der zusammenhängenden Redebeispiele zum Zweck der Kontrolle linguistischer Modulation.

Sämtliche linguistischen und nicht-linguistischen Informationen modulieren das Sprachsignal. Der Hörer ist im Stande das Signal (bis zu einem gewissen Grade) wieder zu demodulieren, relevante Informationen zu extrahieren und andere Information zu ignorieren. Für die automatische Spracherkennung bedeutet dies, dass es unumgänglich ist, alle Modulationsquellen zu kennen und deren Auswirkungen auf das Sprachsignal, will man gesprochene Sprache ebenso gut oder gar besser automatisch erkennen als der Mensch.

### 1.3.5 Was ist nun Stimme und was Sprechweise?

*Unter Stimme sei im Folgenden, Traunmüller folgend, die Summe aller organischen und expressiven Information verstanden, die in einer lautsprachlichen Äußerung enkodiert ist.*

(1.6)

*Unter Sprechweise sei im Folgenden die Summe der Information über die perspektivische Modulation und die neuronale Steuerung der linguistischen und perspektivischen Modulation verstanden, die in einer lautsprachlichen Äußerung enkodiert ist.*

(1.7)

In einer lautsprachlichen Äußerung sind also generell Stimme, Sprechweise und linguistischer Gehalt (gesprochene Sprache) enkodiert. Da hier die Stimme und Sprechweise untersucht werden soll, ist der linguistische Gehalt konstant zu halten oder zu kontrollieren.

## 1.4 Theorien über das Altern von Stimme und Sprechweise

### 1.4.1 Theorien zur Erklärung der physischen Veränderung des Sprechapparats

#### 1.4.1.1 Das *male-female coalescence model*

Hollien (1987) stellt ein Modell vor, nach dem Unterschiede der altersabhängigen Veränderungen der Stimme zwischen Mann und Frau durch hormonelle Differenzierung während der Pubertät und hormonelle Angleichungen während des rest-

lichen Lebensablaufs erklärt werden. Diese Theorie modelliert also das Zustandekommen altersbedingter Veränderungen der Stimme/Sprechweise durch biologische/physische Veränderungen im spracherzeugenden Mechanismus über deren Verknüpfung mit geschlechtsspezifischen Unterschieden. Durch die Erklärung sowohl geschlechtlicher als auch altersbedingter Unterschiede durch hormonelle Faktoren, ist diese Theorie als Adaption biologischer Programmtheorien des Alters auf die Spracherzeugung zu verstehen.

#### **1.4.1.2 Linvilles Theorie der Versteifung und Abschwächung**

Linville (2001, S. 12 ff.) führt den Großteil altersbedingter Veränderungen der Stimme und Sprechweise auf zwei Faktoren zurück: den Rückgang der Gewebeelastizität und die Schwächung der Muskeln. Die Versteifung des Brustkorbes und die Schwächung der Atemmuskulatur erschweren das Ein- und Ausatmen und führen zu verminderter Leistungsfähigkeit.

Vermehrte Steifheit der Stimmlippen und die Schwächung der Muskeln, die an der Stimmbildung beteiligt sind, führen zu verschlechterter Kontrolle der Position der Stimmlippen und der Bewegungsabläufe, die die Stimme erzeugen, was sich auf die Tonhöhe, die Stimmstabilität, die Lautstärke und den Rauschanteil der Stimme auswirkt.

Die Schwächung der Muskulatur und Versteifung des Gewebes erklärt auch altersabhängige Veränderungen der Artikulation: Die verlangsamte und schlechter koordinierte Interaktion der Artikulatoren führt zu verringerter Sprechgeschwindigkeit und ungenauerer Artikulation.

Ein solches Modell verbindet auch die größere Variabilität altersrelevanter Stimmmerkmale innerhalb der Gruppe der Älteren im Vergleich zur Gruppe der Jungen mit der größeren Variabilität der allgemeinen Fitness unter den Alten: Wenn man in Betracht zieht, dass sich die muskulären und geweblichen Veränderungen nicht nur auf den Sprechapparat reduzieren, sondern den gesamten Körper betreffen, dann ist ein starker Zusammenhang zwischen Merkmalen der Stimme und Sprechweise und Merkmalen des allgemeinen körperlichen Zustandes anzunehmen. Und der allgemeine körperliche Zustand variiert bei älteren Personen stärker als bei jungen.

#### **1.4.1.3 Bakens Theorie der wachsenden Nichtlinearität**

Ausgehend von der Beobachtung von im Alter zunehmender Nichtlinearität in diverser Hinsicht im biomechanischen System, speziell im Larynx und der beispielhaften Implementation der Relation von Nicht-Linearität und Linearität bzgl. der Steifheit der stimmlippen-modellierenden Massen im Glottismodell von Ishizaka

& Flanagan postuliert Baken (2005) sein Modell der wachsenden Nichtlinearität im alternden Larynx: *the nonlinearity model of laryngeal biomechanics*.

Mit der Ausprägung der unabhängigen Variable (Alter) wachsende Nicht-Linearität stellt aber ein Problem für parametrische<sup>12</sup> statistische Auswertungsmethoden dar, die den Standard der inferenzstatistischen Bewertung empirischer Forschung bilden: Durch im Alter zunehmende Nicht-Linearität in multiplen (Teil-) Systemen der Spracherzeugung, die in ihrer Gesamtheit methodisch kaum zu kontrollieren sein dürfte, entsteht mehr (unkontrollierte) Varianz in Äußerungen älterer Sprecher, also, statistisch formuliert Varianzheterogenität. Von parametrischen Verfahren wird aber axiomatisch Varianzhomogenität gefordert.

## 1.4.2 Alter als Information

### 1.4.2.1 Das *physiological aging/communication model*

Die zweite in Hollien (1987) postulierte Theorie modelliert den Informationsfluss bezüglich des Konstrukts (chronologisches) Alter im menschlichen System bis zur motorischen Spracherzeugung. Entscheidende Faktoren/Subsysteme, deren altersbedingten Veränderungen sich auf die Spracherzeugung auswirken, werden darin in Beziehung gesetzt (vgl. Abbildung 1.3).

Zu beachten ist, dass dieses Modell auf den Erzeugungsmechanismus fokussiert ist, bzw. dass die Perzeption in diesem Modell nicht berücksichtigt wird.

### 1.4.2.2 Modulationstheorie

Wie bereits erläutert, behandelt auch Traunmüller (1994) im Rahmen der Modulationstheorie explizit das Sprecheralter als im Sprachsignal zwingend vorhandene Teilinformation. Jedoch ist zu beachten, dass nach dieser Theorie das Sprecheralter zu den sog. *organischen Informationen* zählt, und sich damit nur über die Stimme (nicht über die Sprechweise) manifestiert.

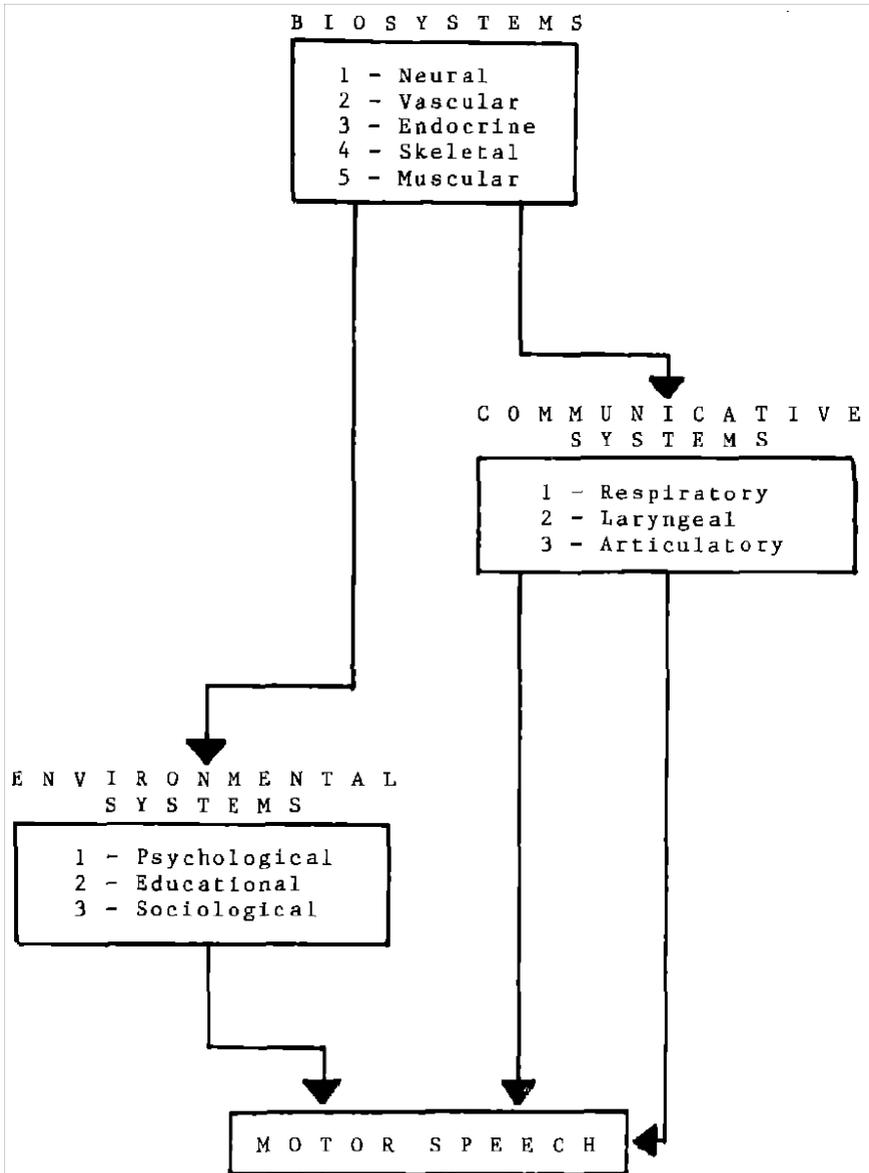
### 1.4.2.3 Zusammenfassung

Altern der lautsprachlichen Äußerung *ist der Prozess der irreversiblen Veränderung der lautsprachlichen Äußerung, die primär auf die irreversible Veränderung des biologischen (Teil-) Systems Sprechapparat zurückgeht. Unter Äußerungsalter soll im Folgenden also die Summe der Informationen in einer lautsprachlichen Äußerung (die als quasi-punktueller Zeitintervall interpretiert wird) über einen Zustand im Verlauf dieses Prozesses verstanden werden.*

(1.8)

---

<sup>12</sup> die auf dem *Allgemeinen Linearen Modell* basieren



**Abbildung 1.3:** Das *physiological aging/communication model* von Harry Hollien, aus Hollien (1987)

Da in dieser Arbeit der linguistische Gehalt konstant gehalten oder kontrolliert werden soll, sollte das Äußerungsalter fast ausschließlich Information über das Alter der Stimme und der Sprechweise beinhalten.

Besteht die Äußerung aus einem (ohne Kontext erzeugten, also nicht koartikulierten) Vokal, so kann davon ausgegangen werden, dass auch kaum Merkmale der Sprechweise in der Äußerung vorhanden sind. Sofern die Äußerung eine Redeäußerung ist, sind sowohl Merkmale der Stimme als auch der Sprechweise enkodiert.

Das biologische (Teil-) System Sprechapparat umfasst (hier) auch die steuernden neuronalen Strukturen. Das Äußerungsalter geht *primär* auf das Altern des biologischen Systems Sprechapparat zurück, weil über die Supervenienzbeziehungen die Produktion lautsprachlicher Äußerungen auch durch das psychische Alter (sekundär) und das soziokulturelle Alter (tertiär) beeinflusst ist.

Das biologische Altern des Sprechapparats ist in erster Linie als Teilaspekt des Alterns des biologischen Systems Mensch zu sehen. Der einflussreichste (Stör-) Faktor des biologischen Alters ist der allgemeine gesundheitliche Zustand (sofern es im Erwachsenenalter überhaupt biologisches Altern ohne gesundheitliche Einbußen gibt). Psychische (Stör-) Faktoren des Äußerungsalters sind der emotionale Zustand, der Grad der Aufmerksamkeit und der Motivation, die kognitive Leistungsfähigkeit sowie Selbstkonzepte der Persönlichkeit bzw. der Identität. Die Selbstkonzepte entstehen durch Wechselwirkungen des Individuums mit seinem soziokulturellen Umfeld. Diese (Stör-) Faktoren können sich auch reversibel kurz- oder mittelfristig ändern.

Erfasst man das Äußerungsalter durch ein Messsystem, so sind also das biologische, das psychische und das soziokulturelle Alter wie auch die Einflüsse der (Stör-) Faktoren und darüber hinaus der Fehler des Messsystems miteinander verwoben.

Erfasst man es durch mehrere Messsysteme (Hörer), die prinzipiell vergleichbare (reliable) Messwerte (Altersurteile) erzielen, so wird allerdings der statistische Fehlerausgleich wirksam: Der „wahre“, zu erhebende Messwert ist allen Messwerten gemein, der individuelle Fehler der einzelnen Messsysteme verteilt sich aber zufällig (mit einer Wahrscheinlichkeitsdichte die durch eine Gauß'sche Glocke modelliert werden kann) um den Wert Null und nivelliert sich dadurch aus, sofern (unendlich) viele Messwerte (verschiedener aber reliabler Systeme) zusammengefasst werden. Bedenkt man, dass verschiedene Hörer u.U. auch die Einflüsse der Störfaktoren anders bewerten, so kommt es durch die Zusammenfassung ihrer Urteile auch zu einer Verringerung der Stör-Einflüsse auf das Äußerungsalter.

Das perzeptiv mithilfe mehrerer Hörer ermittelbare Alter einer Äußerung sollte also der beste empirisch ermittelbare Schätzer des Äußerungsalters sein – zumin-

dest solange die menschliche Perzeptionsleistung der maschinellen Erkennungsleistung überlegen ist.

*Das durch eine Hörergruppe erfassbare perzeptive Alter einer Äußerung ist ein von individuellen, zufälligen Fehlern einzelner Hörer und des Messverfahrens (und eventuell auch einiger Störfaktoren) bereinigtes Abbild des Äußerungsalters. Somit dient es als bester empirisch ermittelbarer Schätzer des Äußerungsalters.*

(1.9)

Perzeptive Alterswerte von (Rede-) Äußerungen eines Sprechers sollten gute Indikatoren des biologischen, psychischen und soziokulturellen Alters dieses Sprechers darstellen.

#### 1.4.2.4 Ein eigenes Modell

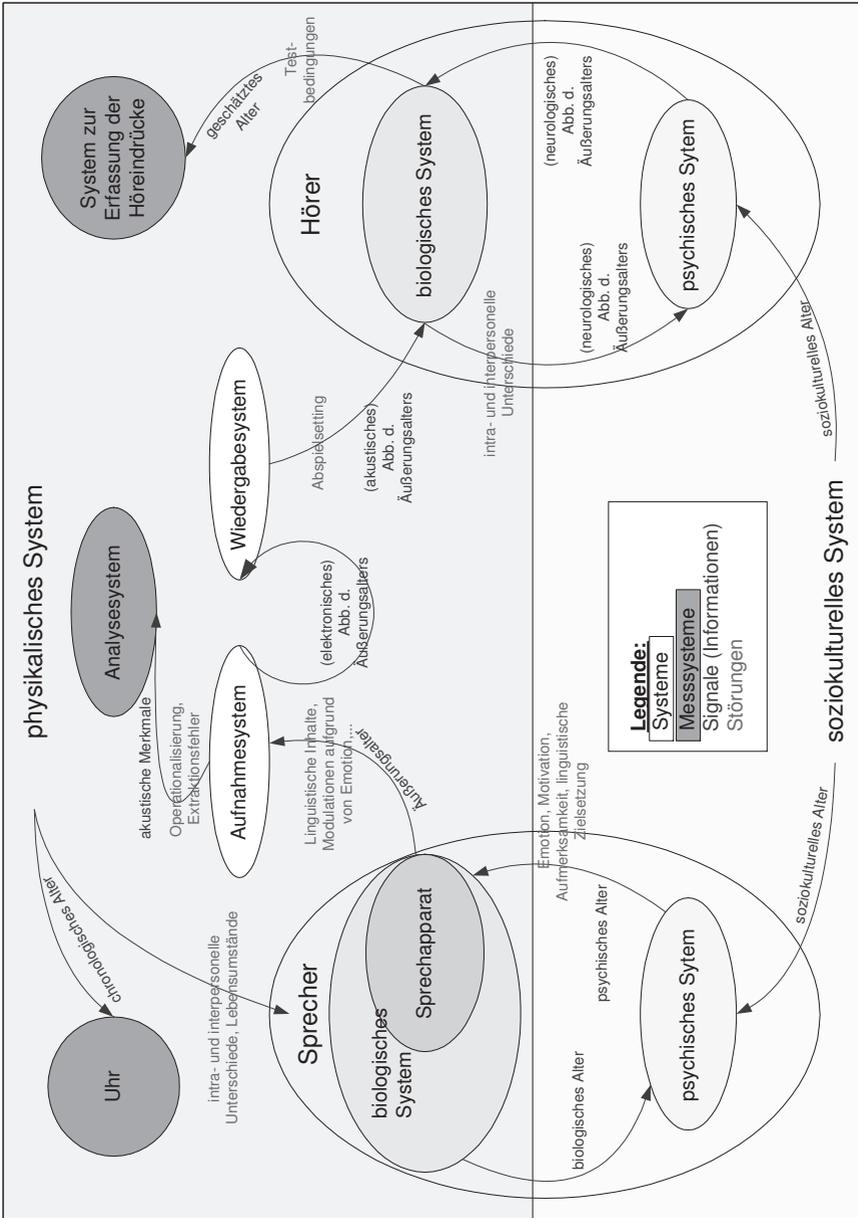
Abbildung 1.4 visualisiert eine Zusammenfassung der theoretischen Überlegungen zu einem Modell darüber, welche Informationen in den folgenden Untersuchungen gemessen werden und welche Störfaktoren zu berücksichtigen sind. Es zeigt die verschiedenen Systeme, die Einfluss auf die Messgrößen dieser Untersuchung haben sowie die altersrelevanten Signale bzw. den Fluss altersrelevanter Information zwischen den Systemen, als auch die Ansatzpunkte einiger Störgrößen.

Über die unterschiedlichen Äußerungstypen, die im empirischen Teil untersucht werden, lässt sich die Gesamtmenge der Information über das Äußerungsalter, wie auch der Anteil, den die Stimme und die Sprechweise transportieren, variieren.

Die Darstellung hat keinen Anspruch auf Vollständigkeit: es können viele weitere Systeme, Untersysteme und Signale postuliert werden. Ferner ist das „soziokulturelle System“ natürlich ein Teil des physikalischen Systems, jedoch einer, der in dieser Betrachtung der monologischen Kommunikationssituation nur über die jeweiligen Vorstellungen des Sprechers und des Hörers zum Tragen kommt.

## 1.5 Zusammenfassung

Die (in den folgenden Untersuchungen) messbaren Merkmale quantifizieren nie Informationen über altersbedingte Veränderungen des Sprechers (als Entität), sondern über die Veränderungen von menschlichen Teilsystemen oder übergeordneten oder autarken Systemen. Will man also die Bedeutung gemessener Merkmale für das Konstrukt Alter einschätzen, muss zunächst der Ursprung und der Weg der Information, die über die jeweiligen Merkmale quantifiziert wird, bestimmt



**Abbildung 1.4:** Informationsflussmodell: Alter als (extra- und paralinguistische) Information in einem allgemeinen Kommunikationsmodell.

werden. Abbildung 1.4 zeigt eine schematische Darstellung eines solchen Informationsflussmodells. Es soll veranschaulichen, in welchen Beziehungen die drei unterschiedlichen Datensorten der folgenden Untersuchung – chronologische Alterswerte der Sprecherinnen, durch Hörer geschätzte Alterswerte und akustische Parameter der Äußerungen – zu einander stehen.

Es ist also nicht zu erwarten, dass allein durch die Analyse lautsprachlicher Äußerungen das chronologische Sprecheralter beliebig genau bestimmbar wäre: Selbst wenn die akustischen Signalanalysemethoden weit genug fortgeschritten wären, um nahezu die ganze altersrelevante Information, die im Sprachsignal enkodiert ist, zu extrahieren, bliebe immer noch das im Gegensatz zum Fortschritt der Zeit von Sprecher zu Sprecher unterschiedlich verlaufende (biologische) Altern als varianzerzeugender (Stör-) Faktor.

Zwar ist die Bestimmung des perceptiven Alters, also des Alters, das dem Sprecher durch auditive Perzeption seiner Äußerungen attribuiert werden kann, frei von biologischen Störungen, doch kann auch sie nie exakt erfolgen, sondern immer nur durch Wahrscheinlichkeitsangaben und basierend auf einer bestimmten Sprachdatengrundlage. Diese Ungenauigkeit ist begründet im zwischen den Hörern unterschiedlichen und nicht vollständig kontrollierbaren Hör- und Beurteilungsverhalten.

Das hier eingeführte theoretische Konstrukt Äußerungsalter ist per Definition durch Signalanalysemethoden theoretisch beliebig genau bestimmbar. Jedoch ist sein Zusammenhang mit dem chronologischen Alter durch unterschiedliche, nicht vollständig erfassbare Alterungsabläufe in den Individuen mit Varianz behaftet, der Zusammenhang mit dem perceptiven durch die unterschiedlichen Hörurteile. Das gebräuchlichere Konstrukt biologisches Alter ist dieser theoretischen Analyse zufolge wohl am besten durch das Äußerungsalter zu approximieren. Im Umkehrschluss bedeutet dies, dass die akustische Analyse hinsichtlich Altersmerkmalen wohl eher den körperlichen Alters-Zustand des Sprechers wiedergeben kann, denn sein chronologisches Alter.

In den Analysen im empirischen Teil dieser Arbeit bleibt also fraglich, welche Varianzanteile bei der Interdependenzanalyse zwischen dem Äußerungsalter und dem chronologischen Alter oder auch dem perceptiven Alter auf eine unzureichende akustische Analyse zurückgehen und welche auf unkontrollierte Faktoren der Alterung der Individuen oder der Ermittlung perceptiver Urteile.

## Kapitel 4

# Methoden

Aus der Zielstellung ergibt sich, dass die Prüfung der Verhältnisse zwischen drei verschiedenen Variablengruppen Gegenstand dieser Arbeit sein soll (vgl. auch Abbildung 1.4):

- Variablen, die Merkmalsausprägungen der Sprecherinnen bezüglich konkret bestimmbarer Konstrukte denotieren: chronologisches Alter, Rauchverhalten, Dialekt (-Selbsteinschätzung), Zugehörigkeit zu einer Berufsgruppe
- Variablen, die Merkmalsausprägungen unterschiedlicher lautsprachlicher Äußerungen der Sprecherinnen bezüglich akustischer Konstrukte denotieren
- Variablen, die Merkmalsausprägungen unterschiedlicher lautsprachlicher Äußerungen der Sprecherinnen bezüglich (durch Hörer) wahrnehmbarer Konstrukte denotieren: perzeptives Alter sowie (perzeptive) Fitness, Erregung und Artikulationsgenauigkeit

Die grundlegende Heuristik, welche die Untersuchung dieser Verhältnisse ermöglicht, soll die (quantitative) Empirie mit einer inferenzstatistischen Prüfung von Hypothesen als letztem Schritt sein. Demzufolge gilt es, den Gütekriterien empirischer Forschung, Objektivität<sup>1</sup>, Reliabilität<sup>2</sup> und Validität<sup>3</sup> zu entsprechen.

Weitere hier verwendete Forschungsstrategien sind (1) der methodische Dreiklang zur Untersuchung mündlicher Kommunikation, (2) die von Salthouse (2001) vorgestellte Strategie zur Untersuchung von Effekten des Alterns und (3) die (Klassische) Testkonstruktion.

Diese werden nachfolgend kurz erläutert. Darauf werden die wichtigsten weiteren Überlegungen zum methodischen Vorgehen dargelegt. Schließlich erfolgt in den Abschnitten 4.2 bis 4.5 eine detailliertere Darstellung der hier verwendeten Methoden.

---

<sup>1</sup> Objektiv sei eine Untersuchung, wenn sie unabhängig vom Untersuchungsleiter zu gleichen Ergebnissen kommt.

<sup>2</sup> Reliabel sei eine Merkmalerhebung, sofern das bzw. die Messinstrument(e) dieselben Objekte (bei wiederholter bzw. vergleichbarer Messung desselben Merkmals) gleich beurteilt/beurteilen, bzw. (bzgl. des Merkmals) gleiche Objekte gleich beurteilt/beurteilen; Reliabilität setzt also Objektivität voraus.

<sup>3</sup> Valide sei eine Merkmalerhebung, wenn wirklich das Merkmal erhoben wird, das vorgeblich erhoben werden soll; Validität setzt also Reliabilität voraus.

## 4.1 Methodischer Rahmen

### 4.1.1 Methodischer Dreiklang zur Untersuchung mündlicher Kommunikation

Die hier unternommenen Analysen folgen der von Sendlmeier (2007) für die Untersuchung lautsprachlicher Äußerungen vorgeschlagenen heuristischen Vorgehensweise: dem methodischen Dreiklang. Dieser Dreiklang besteht aus einer Analyse in drei Schritten, nämlich zunächst einer *auditiv-perzeptiven Analyse*, gefolgt von einer *akustischen* und schließlich einer *statistischen Analyse*, welche die Ergebnisse der auditiv-perzeptiven und akustischen Analyse integriert. Die Gliederung der nachfolgenden Abschnitte und Kapitel folgt der Struktur dieses Dreiklangs.

Die auditiv-perzeptive Analyse kann dabei wiederum in zwei Prozesse unterteilt werden, die in dieser Arbeit als *auditive*<sup>4</sup> und *perzeptive*<sup>5</sup> Analyse bezeichnet werden. Vornehmliches Ziel der auditiven Analyse der Äußerungen ist die Hypothesengenerierung und die Bestimmung von Einheiten der Äußerungen, die über mehrere Realisationen (mehrerer untersuchter Sprecher) hinweg vergleichbar sind. Sie erfolgt in der Regel durch den oder die Untersuchenden. Die Bestimmung vergleichbarer Einheiten beinhaltet in der Regel eine phonetische Transkription der Äußerungen. Die phonetische Segmentation und Transkription wird hier im Abschnitt 4.3 näher erläutert. Der hypothesenfindende Aspekt wurde hauptsächlich in vorangehenden Forschungsarbeiten geleistet (vgl. Abschnitt 2.3.4). Die Konkretisierung dieser Hypothesen findet sich in den Abschnitten zur Motivation und hypothetischen Veränderung des Kapitels 7.

Die perzeptive Analyse ist sinnvoll einsetzbar, wenn die auditive (wenigstens teilweise) erfolgreich abgeschlossen werden konnte: Es muss klar sein, welcherlei Äußerungstypen verglichen werden sollen bzw. vergleichbar sind und wie das Konstrukt von Hörern beurteilt werden kann. Ziel der perzeptiven Analyse, die hier speziell die Wahrnehmung mehrerer (gewöhnlicher Laien-) Hörer erfasst, ist erstens die Bestätigung der Hypothesen der auditiven Analyse und zweitens eventuell die Generierung einer Merkmalskala, die auf der zusammengefassten Perzeption der Hörer beruht. Dies ist vor allem von Vorteil, wenn das zu erfassende Konstrukt (perzeptives Alter) eher latent ist und bisher keine Messgeräte entwickelt werden konnten, die dieses Konstrukt besser erfassen könnten als Hörer (als Gruppe). Die hier vorgenommene perzeptive Analyse findet sich im Kapitel 6.

<sup>4</sup> Auch die auditive Analyse erfolgt auf Grundlage der Wahrnehmung des Untersuchenden und ist somit prinzipiell auch als perzeptiv zu bezeichnen.

<sup>5</sup> Auch die perzeptive Analyse erfolgt natürlich durch Hören und ist somit prinzipiell auch auditiv.

Die akustische Analyse eines lautsprachlichen Signals ist sinnvoll, wenn mehrere Hörer das zu untersuchende Konstrukt reliabel einzuschätzen vermögen. Ziel dieses Analyseschrittes ist es, geeignete akustische Parameter zu finden und zu erheben, die mit dem Konstrukt (Alter) variieren, die also in der Lage sind, die Information in der Äußerung über das Konstrukt zu vermitteln. Die Darstellung der akustischen Analyse erfolgt in den Kapiteln 7 und 8.

Die statistische Zusammenführung der Ergebnisse der auditiv-perzeptiven und der akustischen Analyse findet sich hier hauptsächlich in den Kapitel und 6 und 8.

### 4.1.2 Eine Forschungsstrategie zur Untersuchung von Effekten des Alters

Auch die in Salthouse (2001) vorgestellte Forschungsstrategie dient als Grundlage des eigenen methodischen Vorgehens. Sie soll den Besonderheiten von Alter als zu untersuchendem Merkmal gerecht werden. Sie wurde speziell zur Untersuchung der Auswirkung experimentell nicht manipulierbarer (unabhängiger) Variablen (wie z.B. Alter) auf latente Konstrukte entwickelt. Salthouse bezieht sich dabei zwar auf kognitive Konstrukte, jedoch können diese hier durch akustische oder perzeptive ersetzt werden, ohne das Wesen dieser Strategie zu berühren.

Hauptanliegen dieser Strategie ist es, die (interne Konstrukt-) Validität der (akustischen, perzeptiven oder kognitiven) Maße zur Bestimmung des Alters sicherzustellen – also die Klärung der Frage, wie bestimmte Anteile der Varianz in Merkmalen, die latente Konstrukte erfassen, auf Altersunterschiede zurückgeführt werden können.

Er schlägt dazu die Verknüpfung korrelativer und quasi-experimenteller Verfahren zu einer im weiteren Sinne als Multi-Trait-Multi-Method-Ansatz<sup>6</sup> zu bezeichnenden Forschungsstrategie vor. In der vorliegenden Arbeit werden die perzeptive und die akustische Analyse korrelativ (im Querschnitt) und quasi-experimentell (im Längsschnitt) vorgenommen (vgl. die Kapitel 6 und 8), um besonders valide Korrelate des Alters entdecken zu können. Daneben wird die Erhebung verschiedener Konstrukte durch jeweils mehrere Operationalisierungen (Multi-Trait-Multi-Method) am deutlichsten im Kapitel 7 verwirklicht bzw. dargestellt.

Im Detail besteht diese Strategie aus der Verknüpfung der Erkenntnisse aus fünf postulierten Forschungsphasen, wobei Salthouse (2001) deutlich macht, dass diese Strategie verändert werden kann und auch muss, um sie an konkrete Forschungsvorhaben anzupassen. Zwingend an dieser Stelle ist nur die (theoretische)

---

<sup>6</sup> Erhebung mehrerer Merkmale mit mehreren Methoden an derselben Stichprobe zur Prüfung der Konstrukt-Validität, vgl. hierzu z.B. Bortz & Döring (2006)

1. Phase. Die weiteren (empirischen) Phasen können u.U. auch zusammengefasst (an einem Datensatz) erfolgen:

1. Spezifizierung und Operationalisierung der (Sprachsignal-) Konstrukte: erfolgt hier hauptsächlich im Abschnitt 1.3.5 und im Kapitel 6 für das Konstrukt *perzeptives Alter* und in Kapitel 7 bzgl. der akustischen Konstrukte;
2. Untersuchung von Variablen, die das Konstrukt abbilden sollen, bezüglich der Variation mit Alter: hier in Kapitel 6 und 8;
3. Bestimmung des einzigartigen Beitrags des Konstrukts zur Erklärung des Alters: hier vornehmlich realisiert durch die Selektion der Prädiktoren der multiplen Regressionen (vgl. Kapitel 8);
4. Prüfung der konvergenten und diskriminanten Validität der Konstrukte (zur Indikation von Alter): konvergente Validität (akustischer Maße) kann hier in erster Linie durch die vielfältigen Operationalisierungen derselben Konstrukte überprüft werden, die diskriminante durch den Vergleich zwischen unterschiedlichen Konstrukten (vgl. Kapitel 7); für Letzteres wurden ferner auch die perzeptiven Konstrukte Artikulationsgenauigkeit, Fitness und Erregung erhoben, vgl. Abschnitt 6.4;
5. Evaluierung der Relevanz des Konstruktes für weitere Aspekte mündlicher Kommunikation: dieser Schritt wurde in der vorliegenden Arbeit nicht explizit vorgenommen – davon abgesehen kann angenommen werden, dass die allermeisten hier untersuchten Konstrukte sehr relevant für die mündliche Kommunikation im Allgemeinen sind.

Die ersten beiden Schritte sollten (auch) nach Salthouse (2001) möglichst oft iteriert werden, um möglichst gute Operationalisierungen zu erhalten, bzw. konstruieren und testen zu können. Auch dieser Gedanke wurde hier aufgegriffen, v.a. durch die „wiederholte“, leicht abgeänderte Operationalisierung und Messung ein und desselben Konstruktes (vgl. hierzu Kapitel 7).

### 4.1.3 Die Klassische Testkonstruktion

Die Klassische Testkonstruktion bzw. die ihr zugrundeliegende Klassische Testtheorie (KTT) beschäftigt sich mit der Frage, wie psychometrische Tests zu konstruieren sind. Sie stellt Methoden bereit, durch welche die Erhebung eines (von Menschen erfassbaren) Merkmals hinsichtlich des Erfüllungsgrades der drei Gütekriterien empirischer Forschung beurteilt und gegebenenfalls verbessert wer-

### 4.4.3 Statistische Verfahren zur Ermittlung der Urteilerübereinstimmung

Der erste Schritt der perzeptiven Analyse dieser Arbeit besteht in der Ermittlung der Hörer- (allgemein Urteiler-) Übereinstimmung: Die Hörer dienen hier als Messinstrument, das nur als solches betrachtet werden kann, sofern es reliabel ist, sofern also die Hörer – bis zu einem gewissen Grade – übereinstimmende Urteile abgeben.

Die zur Überprüfung der (Urteiler-) Reliabilität einschlägigen statistischen Verfahren – bzw. ihre in gängigen Softwarepaketen implementierten Algorithmen – benötigen vollständige (Urteilsdaten-) Matrizen: Jeder Urteiler muss jedes Objekt beurteilt haben, um ein Übereinstimmungsmaß berechnen zu können. Fehlt auch nur ein einzelnes Urteil, müssen entweder alle Urteile dieses Urteilers (der nur ein Objekt nicht beurteilt hat) oder aber alle Urteile bezüglich dieses Objekts (das von nur einem Urteiler nicht beurteilt wurde) aus der Reliabilitätsermittlung ausgeschlossen werden.

Wie im Folgenden gezeigt wird, ist ein solcher spalten- oder auch zeilenweiser Ausschluss von Daten zur Übereinstimmungs- bzw. Reliabilitätsbestimmung weder theoretisch noch praktisch (sofern die Algorithmen geringfügig verändert werden) nötig. Dass es theoretisch vorzuziehen ist, alle Daten, die erhoben wurden – und nicht nur „ausgewählte“ –, einer Reliabilitäts- bzw. Übereinstimmungsprüfung zu unterziehen, ist trivial. Dass solche fehlenden Werte während einer Datenerhebung entstehen, kann (und sollte soweit möglich) zwar durch entsprechende Erhebungsmaßnahmen vermieden werden. Dennoch ist es wohl zumindest bei größeren (Urteils-) Datenerhebungen der Regelfall, dass das eine oder andere Objekt vom einen oder anderen Urteiler unbewertet bleibt.

Darüber hinaus erlauben die im Folgenden vorgestellten Verfahrensderivate auch die Ermittlung *einer* Reliabilität bzw. Übereinstimmung *mehrerer* (gepoolter) Datenerhebungen (auch von unterschiedlichen Urteilern und/oder an unterschiedlichen Objekten), sofern alle Urteiler und Objekte jeweils einer Grundgesamtheit entstammen. Mehrere Urteilerhebungen sind generell bei jeder größeren Untersuchung nötig, da die Beurteilungsdauer (pro Urteiler) zwangsläufig limitiert ist.

Im Folgenden werden also Derivate der herkömmlichen Verfahren zur Berechnung von Intra-Klassen-Korrelationen (ICCs) und Kendalls  $W$  vorgestellt, die im Falle der Beurteilung aller Objekte durch alle Urteiler (vollständiger Urteilmatrizen) den herkömmlichen Verfahren äquivalent sind und im Falle „fehlender“ Urteile (unvollständiger Urteilmatrizen) die Information aller vorhandenen Urteile optimal miteinbeziehen.

### 4.4.3.1 Intra-Klassen-Korrelationen

#### 4.4.3.1.1 Allgemeines zur Intra-Klassen-Korrelation

Wie z.B. Wirtz & Caspar (2002) oder auch Bortz & Döring (2006) zu entnehmen ist, ist zur Abschätzung der Reliabilität intervallskalierteter Urteile mehrerer Urteiler (bzgl. mehrerer Objekte) eine Intra-Klassen-Korrelation (ICC) das einschlägige Verfahren. Shrout & Fleiss (1979) definieren:

*[...] the ICC is the correlation between one measurement [...] on a target and another measurement obtained on that target. The ICC is thus a bona fide correlation coefficient that [...] is often but not necessarily identical to the component of variance due to targets divided by the sum of it and other variance components.*

(4.7)

ICCs beruhen also auf dem Verhältnis der „wahren Varianz“<sup>28</sup> zwischen den Objekten zu sich selbst plus verschiedener anderer Varianzkomponenten. Die verschiedenen Zusammensetzungen dieser Komponenten resultieren aus verschiedenen Varianzanalysemodellen, denen unterschiedliche Annahmen zugrundeliegen, weshalb die daraus abgeleiteten ICCs unterschiedlich zu interpretieren sind. Shrout & Fleiss (1979) unterscheiden drei Fälle einer Urteiler-Reliabilitätsanalyse:

1. Fall: Jedes Objekt<sup>29</sup> wird von einer (von Objekt zu Objekt) unterschiedlichen (aber gleich großen) Urteilergruppe, die zufällig aus der Population aller Urteiler gezogen wurde, beurteilt.
2. Fall: Jedes Objekt wird von ein und derselben Urteilergruppe, die zufällig aus der Population aller Urteiler gezogen wurde, beurteilt.
3. Fall: Jedes Objekt wird von ein und derselben Urteilergruppe, die aus allen (für den Untersuchungsgegenstand) relevanten Urteilern besteht, beurteilt.

Die folgenden varianzanalytischen Modelle (der Zusammensetzung eines spezifischen Urteils ( $x_{ij}$ ) aus verschiedenen Komponenten) eignen sich zur Definition von Reliabilität in diesen Fällen:

1. einfaktorielles (Zufallseffekte-) Modell: Modell mit (zufälligen) Objekt-Effekten<sup>30</sup> ( $\alpha_i$ ) und (zufälligen) Residual-Effekten ( $\omega_{ij}$ ):

$$x_{ij} = \mu + \alpha_i + \omega_{ij} \quad (4.8)$$

<sup>28</sup> der Varianz der Objekte in ihrer Population, nicht in der beobachteten Stichprobe

<sup>29</sup> einer Zufalls-Stichprobe aus der Grundgesamtheit der Objekte; gilt auch für die Fälle 2 und 3

<sup>30</sup> Differenzen der „wahren“ Werte der Objekte vom Populationsmittel aller Urteile

2. zweifaktorielles Zufallseffekte-Modell: Modell mit (zufälligen) Objekt-Effekten, zufälligen Urteiler-Effekten<sup>31</sup> ( $\beta_j$ ) und (zufälligen) Residual-Effekten ( $\epsilon_{ij}$ ):

$$x_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \quad (4.9)$$

3. zweifaktorielles Mischeffekte-Modell: die Urteiler-Effekte ( $\theta_j$ ) sind hier „fest“, d.h. jedes einzelne  $\theta_j$  und damit beliebige ihrer Summen seien Null:

$$x_{ij} = \mu + \alpha_i + \theta_j + \epsilon_{ij} \quad (4.10)$$

mit

$$\sum \theta_j = 0 \quad (4.11)$$

Hierbei denotiert  $x_{ij}$  das Urteil des  $j$ -ten Urteilers bezüglich des  $i$ -ten Objekts und  $\mu$  das Populationsmittel aller Urteile.

McGraw & Wong (1996) differenzieren die beiden zweifaktoriellen Modelle ferner in solche mit und in solche ohne Berücksichtigung von Wechselwirkungseffekten, so dass fünf Varianzanalysemodelle zur Entwicklung von Reliabilitätsdefinitionen unterschieden werden können.

Diese Reliabilitätsdefinitionen unterscheiden sich (zumeist) abhängig vom zugrundegelegten Modell und des Weiteren hinsichtlich der Einheit, auf welche die Reliabilitätsanalyse abzielt – auf Einzelmessungen (Reliabilität eines Urteilers) oder gemittelte Messungen (Reliabilität der über eine Urteilergruppe gemittelten Urteile). Und schließlich können die Reliabilitätsdefinitionen, die auf zweifaktoriellen Modellen basieren, entweder die absolute Übereinstimmung oder aber den korrelativen Zusammenhang (die Konsistenz) erfassen. Insgesamt resultieren also 18 verschiedene Reliabilitätsdefinitionen.<sup>32</sup>

Obwohl Wechselwirkungseffekte generell auf (mehrfaktorieller) varianzanalytischer Modellebene zu berücksichtigen sind und auch praktisch auftreten (können), sind sie aber zur Berechnung eines Schätzers der Reliabilität (eines ICC) entweder irrelevant oder aber ihre Berücksichtigung führt dazu, dass die Reliabilität gar nicht abgeschätzt werden kann (wie in den beiden Fällen eines Mischeffekte-Modells zur Schätzung der Reliabilität einer gemittelten Messung).<sup>33</sup> Aus diesem

<sup>31</sup> Differenzen der Urteilmittel (über theoretisch unendlich viele Urteile) vom Populationsmittel aller Urteile

<sup>32</sup> Bei McGraw & Wong (1996) finden sich alle 18 Definitionsgleichungen.

<sup>33</sup> Näheres hierzu findet sich bei McGraw & Wong (1996) und bei Shrout & Fleiss (1979).

Grund resultieren aus den 18 von McGraw & Wong (1996) unterschiedenen Reliabilitätsdefinitionen nur 10 (hinsichtlich ihrer Deutung) verschiedene, tatsächlich berechenbare Reliabilitätskoeffizienten (ICCs).<sup>34</sup>

Darüber hinaus sind aber schon die Reliabilitätsdefinitionen (und dementsprechend auch ihre Schätzungen durch ICC-Koeffizienten) für gemischte und zufällige (zweifaktorielle) Modelle formal äquivalent, so dass die Berechnung und der Wert der ICC-Koeffizienten nur bei sechs der zehn deutungsmäßig zu unterscheidenden Koeffizienten verschieden ist.

Hinzu kommt, dass vier dieser acht „doppelt vorhandenen“ Koeffizienten im Falle der Betrachtung von Urteilern als Messinstrument nicht als Reliabilitätsmaße (oder als Maße der Varianzaufklärung) zu interpretieren sind (sondern nur als deskriptive Zusammenhangsmaße), da sie rechnerisch auf die spezifische Urteilerstichprobe beschränkt sind. Dies betrifft die Koeffizienten (bzw. deren zugrundeliegenden Definitionen), welche entweder absolute Übereinstimmung erfassen („unjustiert“ sind) und auf Mischeffekte-Modellen basieren oder den korrelativen Zusammenhang (die Konsistenz) erfassen („justiert“ sind) und auf Zufallseffekte-Modellen basieren.<sup>35</sup> Zu Letzteren zählt auch das wohl bekannteste Maß der Test-Reliabilität, Cronbachs  $\alpha$ <sup>36</sup>:  $\alpha$  hat immer den gleichen Wert wie der justierte ICC-Koeffizient für gemittelte Messungen, basierend auf dem Mischeffekte-Modell (ICC(3,k), vgl. unten). Es basiert aber definitorisch auf einem Zufallseffekte-Modell (Fall 2)!<sup>37</sup> Zur Abschätzung der Urteiler-Reliabilität ist also „darauf zu achten, dass die Frage nach der Justierung der [Urteiler-] Mittelwerte und die Verallgemeinerbarkeit der Ergebnisse auf andere Rater [Urteiler] unmittelbar ineinander gekoppelt sind“ (Wirtz & Caspar (2002), S. 173).

So verbleibt zur Unterscheidung der ICCs, die Urteiler-Reliabilität abschätzen, zusätzlich zur eingangs gemachten Unterscheidung der drei Fälle die Differenzierung (jedes dieser Fälle) hinsichtlich der Einheit der Reliabilitätsanalyse in

- Koeffizienten der Reliabilität der Urteile eines (1) Urteilers
- Koeffizienten der Reliabilität gemittelter Urteile mehrerer (k) Urteiler

Die Terminologie von Shrout & Fleiss (1979) übernehmend, werden die resultierenden ICCs wie folgt bezeichnet:

<sup>34</sup> Statistische Softwarepakete (z.B. SPSS) ermöglichen generell die Berechnung dieser 10 Koeffizienten an ein und derselben (vollständigen) Datenmatrix.

<sup>35</sup> vgl. Wirtz & Caspar (2002), S. 173 ff.

<sup>36</sup> vgl. McGraw & Wong (1996)

<sup>37</sup> Die Items eines Tests werden „zufällig“ aus der Menge aller (denkbaren) Testitems ausgewählt. Im (generellen) Unterschied zu Urteilern bleiben Items aber (für einen spezifischen Test, dessen Reliabilität bewertet werden soll) dieselben, weshalb  $\alpha$  zwar ein Maß der Testreliabilität ist, nicht aber der Urteilerreliabilität.

- ICC(1,1): Schätzer der Reliabilität *eines* Urteilers für den o.g. 1. Fall
- ICC(1,k): Schätzer der Reliabilität *des Mittels von k* Urteilern für den 1. Fall
- ICC(2,1): Schätzer der Reliabilität *eines* Urteilers für den 2. Fall
- ICC(2,k): Schätzer der Reliabilität *des Mittels von k* Urteilern für den 2. Fall
- ICC(3,1): Schätzer der Reliabilität *eines* Urteilers für den 3. Fall
- ICC(3,k): Schätzer der Reliabilität *des Mittels von k* Urteilern für den 3. Fall

#### 4.4.3.1.2 Ein Verfahren zur Berechnung der einfaktoriellen ICCs an unvollständigen Urteilsmatrizen

Bereits Ebel (1951) beschreibt ein Verfahren zur Ermittlung der Urteiler-Reliabilität über die Berechnung des ICC(1,1) oder des ICC(1,k) für den Fall, dass einzelne Objekte nicht nur durch aus unterschiedlichen Individuen bestehende (aber gleich große) Urteilergruppen beurteilt werden, sondern auch durch unterschiedlich große Urteilergruppen:

Die Reliabilität der Urteile eines einzelnen Urteilers im ersten oben beschriebenen Fall ( $\rho_{1,1}$ ) ist definiert als Anteil der wahren Varianz zwischen den Objekten in der Grundgesamtheit der Objekte ( $\sigma_\alpha^2$ ) an der gesamten beobachteten Varianz ( $\sigma_{tot}^2$ ).<sup>38</sup>  $\sigma_{tot}^2$  wiederum setzt sich additiv aus  $\sigma_\alpha^2$  und der Varianz der Fehler (der einzelnen Urteile) in der Population der Urteile ( $\sigma_\omega^2$ ) zusammen:

$$\rho_{1,1} = \frac{\sigma_\alpha^2}{\sigma_{tot}^2} = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\omega^2} \quad (4.12)$$

Die aus einer einfaktoriellen Varianzanalyse resultierende Schätzung der Fehlervarianz (bzw. der Varianz innerhalb der Objekte  $\hat{\sigma}_{in}^2$ ) ist ein direkter Schätzer von  $\sigma_\omega^2$ . Die varianzanalytische Schätzung der Varianz zwischen den Objekten ( $\hat{\sigma}_{zw}^2$ ) schätzt dagegen den Term  $k\sigma_\alpha^2 + \sigma_\omega^2$ . Somit ergibt sich:

$$\hat{\rho}_{1,1} = \frac{\hat{\sigma}_\alpha^2}{\hat{\sigma}_\alpha^2 + \hat{\sigma}_\omega^2} = \frac{\hat{\sigma}_{zw}^2 - \hat{\sigma}_{in}^2}{\hat{\sigma}_{zw}^2 + (k-1)\hat{\sigma}_{in}^2} = ICC(1,1) \quad (4.13)$$

Mischeffekte- mit

$$\hat{\sigma}_{zw}^2 = \frac{\sum_{i=1}^n k_i (\bar{O}_i^2 - \bar{G}^2)}{n-1} \quad (4.14)$$

<sup>38</sup> Dies trifft auch für  $\rho_{2,1}$  zu; bei  $\rho_{3,1}$  ist es der Anteil an der gesamten Varianz abzüglich der Varianz zwischen den Urteilern.

und

$$\hat{\sigma}_{in}^2 = \frac{\sum_{i=1}^n \sum_{j=1}^k x_{ij}^2 - \sum_{i=1}^n k_i \bar{O}_i^2}{n(k-1)} \quad (4.15)$$

wobei  $n$  die Anzahl der Objekte ist,  $k_i$  die Anzahl der Urteiler, die das  $i$ -te Objekt beurteilt haben,  $\bar{O}_i$  das arithmetische Mittel aller Urteile bzgl. des  $i$ -ten Objekts und  $\bar{G}$  das arithmetische Mittel aller Urteile.

Die Anwendung dieser allgemeinen Formel (4.13) zur Berechnung des ICC(1,1) ist im Falle einer unvollständigen Urteilmatrix problematisch, da es das dort erforderliche *eine*  $k$  nicht gibt – die Anzahl der Urteiler kann sich dann von Objekt zu Objekt unterscheiden ( $k_i$ ). Plausibel ist, anstelle des einen  $k$  einen Mittelwert der Urteiler pro Objekt ( $k_0$ ) zu verwenden. Ebel (1951) greift hierfür auf eine Formel aus (einer früheren Ausgabe von) Snedecor & Cochran (1989) zurück:

$$k_0 = \frac{1}{n-1} \left( \sum_{i=1}^n k_i - \frac{\sum_{i=1}^n k_i^2}{\sum_{i=1}^n k_i} \right) \quad (4.16)$$

$k_0$  entspricht etwa dem harmonischen Mittel der  $k_i$ , ist also immer kleiner als das arithmetische Mittel und im Falle durchweg gleicher  $k_i$  gleich  $k$ .

Ersetzt man also bei der Berechnung des ICC(1,1) das  $k$  in Gleichung 4.13 durch  $k_0$  und das  $k$  im Nenner von Gleichung 4.15 durch das arithmetische Mittel ( $\bar{k}$ ) der  $k_i$ , so erhält man eine adäquate Schätzung von  $\rho_{1,1}$  – auch im Falle unvollständiger Urteilmatrizen.<sup>39</sup>

Zur Definition der Reliabilität der zusammengefassten Urteile im einfaktoriel- len Modell ( $\rho_{1,k}$ ) ist  $\sigma_{\omega}^2$  durch die (Populations-) Varianz der Fehler der Objektmit- tel zu ersetzen, die schon theoretisch gleich  $\sigma_{\omega}^2/k$  ist und dementsprechend durch  $\hat{\sigma}_{in}^2/k$  geschätzt werden kann:

$$\hat{\rho}_{1,k} = \frac{\hat{\sigma}_{\alpha}^2}{\hat{\sigma}_{\alpha}^2 + \hat{\sigma}_{\omega}^2/k} = \frac{\hat{\sigma}_{zw}^2 - \hat{\sigma}_{in}^2}{\hat{\sigma}_{zw}^2} = ICC(1, k) \quad (4.17)$$

Die Schätzung von  $\rho_{1,k}$  anhand unvollständiger Urteilmatrizen ist also beinahe völlig unproblematisch – sofern man die o.g. Ersetzung von  $k$  durch  $\bar{k}$  zur Berechnung von  $\hat{\sigma}_{in}^2$  berücksichtigt.<sup>40</sup>

Ebel (1951) demonstriert auch die Berechnung von Konfidenzintervallen für den Koeffizienten ICC(1,1) (zur Abschätzung des Bereichs, in welchem sich

<sup>39</sup> Die diesbezügliche Abhandlung bei Wirtz & Caspar (2002), S. 177 ist leider etwas kurz geraten und (deshalb) nicht ganz richtig!

<sup>40</sup> Sinnigerweise ist es auch angebracht, diese Reliabilität und den schätzenden ICC dann nicht durch die absolute Anzahl an Urteilern zur bezeichnen, sondern ebenfalls durch ihr Mittel  $\bar{k}$ , also:  $ICC(1, \bar{k})$  sowie  $\rho_{1, \bar{k}}$ .

$\rho_{1,1}$  mit einer bestimmten Wahrscheinlichkeit befindet) für unvollständige Matrizen. Generell ist dazu die Berechnung eines (empirischen) F-Bruches ( $F_{emp} = \hat{\sigma}_{zw}^2 / \hat{\sigma}_{in}^2$ ) nötig. Mit Hilfe von  $F_{emp}$  lassen sich sodann die F-Werte für die untere ( $F_u$ ) und für die obere ( $F_o$ ) Konfidenzintervallgrenze bestimmen:

$$F_u = F_{emp} / F_{tab,u} \quad (4.18)$$

wobei zur Bestimmung von  $F_{tab,u}$  berücksichtigt werden muss, dass  $n - 1$  Zähler-Freiheitsgrade ( $df_Z$ ) und  $n(k - 1)$  Nenner-Freiheitsgrade ( $df_N$ ) bestehen.<sup>41</sup>

$$F_o = F_{emp} \cdot F_{tab,o} \quad (4.19)$$

wobei zur Bestimmung von  $F_{tab,o}$  gilt:  $df_Z = n(k - 1)$  und  $df_N = n - 1$ .

Hierbei sei  $F_{tab,u}$  bzw.  $F_{tab,o}$  der tabellierte F-Wert (der F-Werte-Verteilungsfunktion), der einer gewünschten Wahrscheinlichkeit (Präzision der Aussage) und den durch die Größe der Matrix bestimmten  $df_Z$  und  $df_N$  entspricht. Setzt man also entweder die  $F_o$ - oder die  $F_u$ -Werte in folgende (abgeleitete<sup>42</sup>) ICC(1,1)-Formel ein, so ergibt sich daraus die obere bzw. die untere Konfidenzintervallgrenze des ICC(1,1):

$$ICC(1,1)_{u/o} = \frac{F_{u/o} - 1}{F_{u/o} + k - 1} \quad (4.20)$$

Bei der Anwendung der Gleichungen 4.18, 4.19 und 4.20 auf unvollständige Matrizen besteht das gleiche Problem, wie bei der Berechnung des ICC(1,1): das eine  $k$ , das erfordert wird, gibt es nicht. Zur Berechnung der Freiheitsgrade kann es aber durch  $\bar{k}$  ersetzt werden. In Gleichung 4.20 (die ja eine Ableitung der Gleichung 4.13 ist) wird es nach Ebel (1951) bzw. Snedecor & Cochran (1989) durch  $k_0$  ersetzt.

Es wird zwar nicht explizit bei Ebel (1951) gezeigt, aber es gibt natürlich auch eine Gleichung für die Konfidenzintervallgrenzen des ICC(1,k)<sup>43</sup>, woraus sich auch an unvollständigen Matrizen keine Probleme ergeben:

$$ICC(1,k)_{u/o} = 1 - \frac{1}{F_{u/o}} \quad (4.21)$$

Die Signifikanztests zur Prüfung der Nullhypothese  $\rho = \rho_0$  schließlich ergeben sich generell aus

$$F_{1,1} = F_{emp} \cdot \frac{1 - \rho_0}{1 + (k - 1)\rho_0} \quad (4.22)$$

<sup>41</sup> Bei Ebel (1951) wurden in der Beispielrechnung (Table 3) die Zähler- und Nennerfreiheitsgrade bzw. der tabellierte F-Wert der oberen und unteren Grenze vertauscht!

<sup>42</sup> vgl. Ebel (1951)

<sup>43</sup> vgl. z.B. Shrout & Fleiss (1979) oder McGraw & Wong (1996)

und

$$F_{1,k} = F_{emp} \cdot (1 - \rho_0) \quad (4.23)$$

wobei die Freiheitsgrade beider F-Werte  $df_Z = n - 1$  und  $df_N = n(k - 1)$  sind. Zur Berechnung der Freiheitsgrade an unvollständigen Matrizen ist wiederum  $k$  durch  $\bar{k}$  zu ersetzen; das  $k$  in Gleichung 4.22 (die wiederum auf einer Ableitung von 4.13 beruht) wird durch  $k_0$  ersetzt. Somit sind für einfaktorielle ICCs auch an unvollständigen Urteilmatrizen die (inferenzstatistischen) Abschätzungen von Konfidenzintervallen oder der Fehlerwahrscheinlichkeit beim Signifikanztest bereits gefunden.

Obwohl diese auf unvollständige Matrizen erweiterten Verfahren spätestens seit Ebel (1951) bekannt sind, sind sie jedoch generell nicht in statistischen Softwarepaketen (z.B. SPSS) implementiert. Stattdessen führen diese Programme einen zeilen- oder auch spaltenweisen Ausschluss von fehlenden Werten vor der Berechnung aller ICCs durch.

#### 4.4.3.1.3 Ipsative Urteile

Bortz & Döring (2006) zeigen<sup>44</sup>, wie durch die Korrektur der einzelnen Urteile  $x_{ij}$  eines Urteilers  $j$  um die Differenz seines Urteilmittels ( $\bar{U}_j$ )<sup>45</sup> vom Gesamtmittel ( $\bar{G}$ )<sup>46</sup> sog. *ipsative Urteile*<sup>47</sup> ( $y_{ij}$ ) generiert werden können:

$$y_{ij} = x_{ij} - (\bar{U}_j - \bar{G}) = x_{ij} - \hat{\beta}_j \quad (4.24)$$

Ipsative Urteile haben also die Eigenschaft, dass sie um die (geschätzten) individuellen Tendenzen (also um die Effekte  $\hat{\beta}_j$ ) der Urteiler bereinigt sind, generell höhere oder tiefere Urteile abzugeben als andere Urteiler bzw. als „generalisierte Urteiler“ (dessen Urteile durch arithmetische Mittelung über alle vorhandenen Urteiler geschätzt werden).

An diesen ipsativen Urteilen können natürlich auch – trivialerweise – die o.g. ICC-Algorithmen (1,1 und 1,k) zur Abschätzung der Urteiler-Reliabilität angewendet werden und es resultieren<sup>48</sup> – was weniger trivial ist – die Koeffizienten

<sup>44</sup> auf S. 274 ff.

<sup>45</sup> in seiner Eigenschaft als bester Schätzer der additiven Verknüpfungen des Populationsmittels der Urteile ( $\mu$ ) und des Effekts des j-ten Urteilers ( $\beta_j$ )

<sup>46</sup> in seiner Eigenschaft als bester Schätzer des „Mittels der Urteile des generalisierten Urteilers“, das im Falle vollständiger Urteilmatrizen gleich  $\mu$  ist

<sup>47</sup> vgl. hierzu auch Bortz (2005), S. 335 ff.

<sup>48</sup> Sofern man eine Änderung berücksichtigt: Der Nenner von  $\hat{\sigma}_{in}^2, n(k-1)$ , wird durch  $(n-1)(k-1)$  ersetzt, da die Abschätzung der Populationsvarianz der Urteile (und damit die ihrer Fehler)  $k-1$  Freiheitsgrade weniger aufweist (weil durch die Ipsation  $k$  Urteiler-Effekte auf *einen* mittleren Urteiler-Effekt reduziert werden), wodurch  $\hat{\sigma}_{in}^2$  formal zu  $\hat{\sigma}_{res}^2$  wird.

ICC(3,1) und ICC(3,k) als Schätzer der Reliabilitäten  $\rho_{3,1}$  und  $\rho_{3,k}$  bzgl. der ursprünglichen Urteile:

$$\hat{\rho}_{3,1} = \frac{\hat{\sigma}_\alpha^2}{\hat{\sigma}_\alpha^2 + \hat{\sigma}_\epsilon^2} = \frac{\hat{\sigma}_{zw}^2 - \hat{\sigma}_{res}^2}{\hat{\sigma}_{zw}^2 + (k-1)\hat{\sigma}_{res}^2} = ICC(3,1) \quad (4.25)$$

$$\hat{\rho}_{3,k} = \frac{\hat{\sigma}_\alpha^2}{\hat{\sigma}_\alpha^2 + \hat{\sigma}_\epsilon^2/k} = \frac{\hat{\sigma}_{zw}^2 - \hat{\sigma}_{res}^2}{\hat{\sigma}_{zw}^2} = ICC(3,k) \quad (4.26)$$

mit

$$\hat{\sigma}_{res}^2 = \frac{\sum_{i=1}^n \sum_{j=1}^k y_{ij}^2 - \sum_{i=1}^n k_i \bar{O}_i^2}{(n-1)(k-1)} \quad (4.27)$$

Reliabilität ist im Fall 3,1 definiert als Anteil der Populationsvarianz zwischen den Objekten ( $\sigma_\alpha^2$ ) an der gesamten beobachteten Varianz ( $\sigma_{tot}^2$ ) abzüglich der Populationsvarianz zwischen den Urteilern ( $\sigma_\beta^2$ ). Anders formuliert sei  $\rho_{3,1}$  der Anteil von  $\sigma_\alpha^2$  an der Summe aus  $\sigma_\alpha^2$ , der Populationsvarianz zwischen den „ipsativierten Urteilern“ ( $\sigma_\theta^2 = 0$ ) und der Populationsvarianz der Fehler der ipsativen Urteile ( $\sigma_\epsilon^2$ ):

$$\rho_{3,1} = \frac{\sigma_\alpha^2}{\sigma_{tot}^2 - \sigma_\beta^2} = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\theta^2 + \sigma_\epsilon^2} = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\epsilon^2} \quad (4.28)$$

Die aus der zweifaktoriellen Varianzanalyse resultierende Varianz der Residuen ( $\hat{\sigma}_{res}^2$ ) ist ein direkter Schätzer von  $\sigma_\epsilon^2$ , so dass sich Gleichung 4.25 ableiten lässt. Zur Herleitung von Gleichung 4.26 ist – wie auch zur Definition von ICC(1,k) – ferner zu beachten, dass bereits in der Reliabilitätsdefinition  $\rho_{3,k}$  die Populationsvarianz der Fehler der einzelnen (ipsativen) Urteile  $\sigma_\epsilon^2$  durch die Populationsvarianz der Fehler der (ipsativen) Objektmittel zu ersetzen ist. Jene ist wiederum schon theoretisch<sup>49</sup> gleich  $\sigma_\epsilon^2/k$  und kann somit durch  $\hat{\sigma}_{res}^2/k$  geschätzt werden.

$$\rho_{3,k} = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\epsilon^2/k} \quad (4.29)$$

#### 4.4.3.1.4 Ein Verfahren zur Berechnung der zweifaktoriellen ICCs an unvollständigen Urteilmatrizen

Im Falle unvollständiger Urteilmatrizen ergeben sich durch die Anwendung der Algorithmen 4.25 und 4.26 keine neuen Probleme: die Ersetzung von  $k$  durch  $k_0$  in Gleichung 4.25 und durch  $\bar{k}$  im Nenner der Formel zur Berechnung von  $\hat{\sigma}_{res}^2$  korrigiert die Abschätzungen von  $\rho_{3,1}$  und  $\rho_{3,k}$  ebenso wie diejenigen von  $\rho_{1,1}$  und  $\rho_{1,k}$ .

<sup>49</sup> vgl. McGraw & Wong (1996)

Problematisch bei der Berechnung der Koeffizienten ICC(3,1) und ICC(3,k) an unvollständigen Urteilmatrizen ist die Generierung der ipsativen Urteile: Zwar existiert (nach wie vor) ein arithmetischer Mittelwert aller Urteile ( $\bar{G}$ ), der auch immer noch der beste Schätzer von  $\mu$  ist. Jedoch ist  $\bar{G}$  nicht mehr (generell) der beste ermittelbare Schätzer des Mittels der Urteile des generalisierten Urteilers bzgl. der Objekte, die der Urteiler  $j$  beurteilt hat (sondern nur, sofern der Urteiler  $j$  die gesamte,  $n$  Objekte umfassende Stichprobe beurteilt hat, auf die sich das Mittel aller  $n$  Urteile des generalisierten Urteilers bezieht). Jeder Urteiler kann aber eine unterschiedliche Teilstichprobe der Gesamtstichprobe der Objekte beurteilt haben. Dem entsprechend gibt es für jede Teilstichprobe (potentiell für jeden Urteiler  $j$ ) einen eigenen optimalen Schätzer des Mittels der Urteile des generalisierten Urteilers ( $\bar{G}_j$ ): das arithmetische Mittel der Objektmittel ( $\bar{O}_{i_j}$ ) derjenigen Objekte, welche die jeweilige Teilstichprobe bilden:

$$\bar{G}_j = \frac{1}{n_j} \sum_{i_j=1}^{n_j} \left( \frac{1}{k_{i_j}} \sum_{j=1}^k x_{i_j j} \right) = \frac{1}{n_j} \sum_{i_j=1}^{n_j} \bar{O}_{i_j} \quad (4.30)$$

wobei  $n_j$  die Anzahl der vom  $j$ -ten Urteiler beurteilten Objekte denotiert und  $i_j = 1, \dots, n_j$  der Zähler der Objekte in der Teilstichprobe ist, die der  $j$ -te Urteiler beurteilt hat.

Somit ergibt sich für die Generierung von ipsativen Urteilen an unvollständigen Urteilmatrizen:

$$y_{ij} = x_{ij} - \hat{\beta}_j = x_{ij} - (\bar{U}_j - \bar{G}_j) \quad (4.31)$$

Dieses Ipsations-Derivat verwirklicht dieselbe Grundidee (die Normalisierung um die individuellen Urteiler-Effekte  $\hat{\beta}_j$ ) wie das Ausgangs-Verfahren und es erzielt im Falle vollständiger Matrizen das selbe Ergebnis. Aber auch diese Ipsation für unvollständige Urteilmatrizen führt bei einmaliger Anwendung nur an vollständigen Urteilmatrizen zu einem endgültigen Ergebnis (sodass eine erneute Durchführung der Ipsations-Prozedur an der resultierenden Matrix zu keiner Veränderung mehr führt:  $\sum \theta_j = 0$ ). Führt man die Ipsationsprozedur jedoch iterativ an unvollständigen Matrizen durch, so reduzieren sich die Korrekturbeträge ( $\bar{U}'_j - \bar{G}'_j = \hat{\beta}'_j$ )<sup>50</sup> von Iteration zu Iteration (relativ rasch) und gehen mit steigender Iterationszahl gegen Null (für Iterationszahl gegen Unendlich geht damit auch  $\sum \hat{\theta}_j$  gegen 0).

Dass die Ipsation iteriert werden muss, liegt darin begründet, dass sich durch die (erste) Ipsation der einzelnen Urteile  $x_{ij}$  zu  $y'_{ij}$  (im Falle unvollständiger Matrizen) die auf den  $x_{ij}$  errechneten  $\bar{G}'_j$  nicht den auf den  $y'_{ij}$  errechneten  $\bar{G}'_j$  ent-

<sup>50</sup> wobei „'“ Werte während der Iteration kennzeichnet

### 6.5.3.2 Parametrische Reliabilitätsprüfung

Als parametrische Pendanten werden die Intra-Klassen-Korrelationen berechnet. Hier sind die Koeffizienten der Reliabilität eines Urteilers (ICC(1,1), ICC(2,1), ICC(3,1)) die von der Hörerzahl unabhängigen Maße; die Koeffizienten der Reliabilität der Urteilergruppe (ICC(1,k), ICC(2,k), ICC(3,k)) dagegen sind von der Gruppengröße abhängig.

Diese Gruppenkoeffizienten der Intra-Klassen-Korrelation<sup>26</sup> steigen – ganz im Gegensatz zu dieser Beziehung bei Kendalls W – mit zunehmender Urteilerzahl und gehen gegen 1 für  $k$  gegen  $\infty$  – und das auch bei sehr geringer mittlerer bivariater Korrelation!<sup>27</sup> Die Grundüberlegung hierbei ist, dass viele Urteiler besser geeignet seien, ein (latentes) Konstrukt zu erfassen, als wenige, weil sie über *mehr* Beurteilungskriterien bzw. -Systeme verfügen (können). Sie führt aber leider dazu, dass diese Maße umso schlechter die Inter-Hörer-Reliabilität zu differenzieren vermögen, je mehr Urteiler betrachtet werden. Diese Maße trennen dann also nicht mehr gut unreliable von reliablen Erhebungen, da sie dann generell hoch sind. Um solche Umstände zu erkennen und zu denotieren, oder auch um Reliabilitäten zu vergleichen, die auf unterschiedlicher Urteilerzahl ( $\bar{k}$ ) basieren, sind also die Koeffizienten der Reliabilität eines Urteilers nötig.

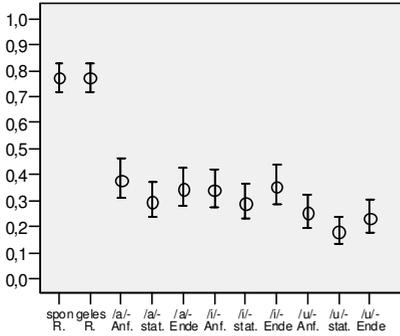
Diese Reliabilitätsmaße, separat für jeden Äußerungstyp berechnet, sind in Tabelle 6.13 verzeichnet und in Abbildung 6.16 graphisch veranschaulicht.

Neben der Unterscheidung der ICCs hinsichtlich ihrer Bezugsgröße in Maße der Reliabilität eines Urteilers oder Maße der Reliabilität einer Urteilergruppe, differieren diese Koeffizienten auch hinsichtlich ihres Anwendungsfalles, vgl. hierzu das Ende des Abschnitts 4.4.3.1. Durch die dort beschriebenen Erweiterungen der ICC-Verfahren lassen sich wohl in den meisten Anwendungsfällen – wie auch hier – der erste und der zweite Fall hinsichtlich ihrer Indikation praktisch nicht mehr unterscheiden. Da die Koeffizienten, die auf Fall 2 beruhen, die Varianz zwischen den Urteilern nicht als Teil der Fehlervarianz betrachtet, sind sie den Koeffizienten, die auf Fall 1 beruhen, aber prinzipiell vorzuziehen. Trotzdem sind auch diese Koeffizienten des ersten Falles hier angegeben und abgebildet, weil sie auf dem (gesicherten) Verfahren von Ebel (1951) beruhen, nicht auf der Erweiterung dieses Verfahrens, die in dieser Arbeit erstmals vorgestellt wird.

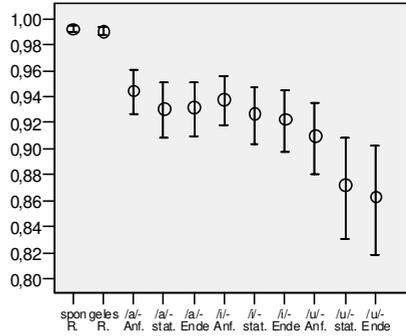
Bezogen auf die hier vorgestellte Anwendung, sind somit die Koeffizienten einschlägig, die auf Fall 2 beruhen, sofern von Interesse ist, wie gut die Urteile eines beliebigen neuen Hörers im Mittel mit den Urteilen der vorhandenen Hörer bivariat zusammenhängen (ICC(2,1)) oder wie gut die (zusammengefassten)

<sup>26</sup> und ihr bekanntester Sonderfall Cronbachs  $\alpha$ , das als parametrisches Maß der internen Konsistenz eines Tests häufig eingesetzt wird

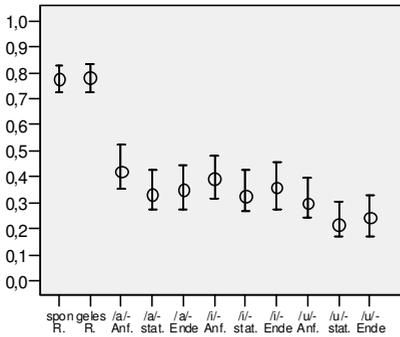
<sup>27</sup> vgl. hierzu auch Gleichung 4.1



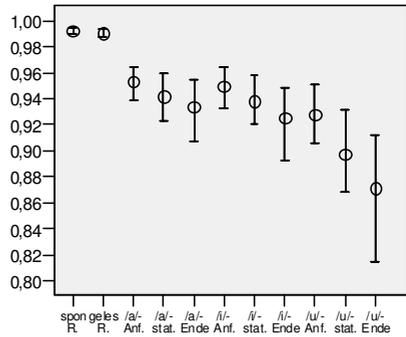
(a) ICC(1,1)



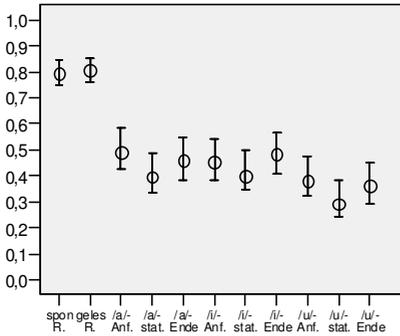
(b) ICC(1,k)



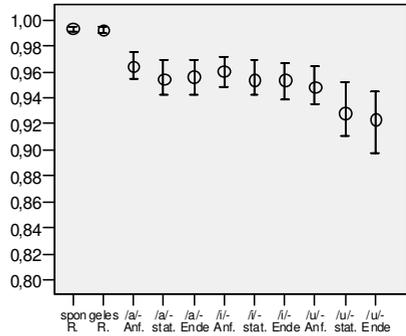
(c) ICC(2,1)



(d) ICC(2,k)



(e) ICC(3,1)



(f) ICC(3,k)

**Abbildung 6.16:** Koeffizienten der Intra-Klassen-Korrelationen der Altersschätzungen aller Querschnitterhebungen. Der Kringel kennzeichnet den wahrscheinlichsten Wert, die Fühler das 95%-tige Konfidenzintervall.

	sp R	gel R	/a/a	/a/s	/a/e	/i/a	/i/s	/i/e	/u/a	/u/s	/u/e
$k_{tot}$	65	43	55	60	26	57	59	22	58	59	21
$\bar{k}$	37,92	29,78	28,11	32,14	25,98	29,31	31,18	22	30,11	31,33	21
$k_0$	37,92	29,78	28,10	32,13	25,98	29,30	31,18	22	30,10	31,32	21
$ICC_{1,1L}$	,719	,719	,309	,236	,280	,275	,231	,285	,197	,135	,177
$ICC_{1,1}$	,772	,773	,378	,296	,346	,340	,290	,353	,253	,179	,232
$ICC_{1,1U}$	,825	,825	,462	,374	,429	,422	,367	,436	,326	,240	,305
$p_{1,1}$	,000	,000	,012	,521	,091	,116	,590	,068	,905	1,00	,967
$ICC_{1,kL}$	,990	,987	,926	,909	,910	,918	,904	,897	,881	,830	,818
$ICC_{1,k}$	,992	,990	,945	,931	,932	,938	,927	,923	,910	,872	,864
$ICC_{1,kU}$	,994	,993	,960	,951	,951	,955	,948	,945	,936	,908	,902
$p_{1,k}$	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000
$ICC_{2,1L}$	,724	,727	,354	,273	,273	,320	,271	,274	,242	,174	,173
$ICC_{2,1}$	,774	,780	,421	,334	,352	,391	,327	,360	,299	,218	,244
$ICC_{2,1U}$	,830	,833	,522	,427	,445	,481	,427	,458	,395	,305	,332
$p_{2,1}$	,000	,000	,000	,152	,059	,003	,204	,036	,492	,992	,945
$ICC_{2,kL}$	,990	,988	,939	,924	,907	,932	,921	,893	,906	,868	,815
$ICC_{2,k}$	,992	,991	,953	,942	,934	,950	,938	,925	,928	,897	,871
$ICC_{2,kU}$	,995	,993	,965	,960	,954	,965	,959	,949	,952	,932	,913
$p_{2,k}$	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000
$ICC_{3,1L}$	,749	,76	,428	,335	,386	,385	,346	,409	,322	,246	,295
$ICC_{3,1}$	,796	,807	,490	,396	,460	,455	,400	,484	,379	,292	,364
$ICC_{3,1U}$	,846	,854	,586	,490	,545	,542	,500	,569	,475	,386	,450
$p_{3,1}$	,000	,000	,000	,003	,000	,000	,002	,000	,011	,567	,036
$ICC_{3,kL}$	,991	,990	,955	,942	,942	,948	,943	,938	,935	,911	,898
$ICC_{3,k}$	,993	,992	,964	,955	,957	,961	,954	,954	,948	,928	,923
$ICC_{3,kU}$	,995	,994	,976	,969	,969	,972	,969	,967	,965	,952	,945
$p_{3,k}$	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000

**Tabelle 6.13:** Intra-Klassen-Korrelationskoeffizienten (ICCs) aller Querschnitts-Wahrnehmungserhebungen. Die  $\alpha$ -Fehler-Wahrscheinlichkeiten ( $p$ ) beziehen sich auf die  $H_0: \rho \leq 0,3$ . Die mit  $L$  gekennzeichneten Koeffizienten denotieren die untere 95%-tige Konfidenzintervallgrenze, die mit  $U$  gekennzeichneten die obere.

Urteile einer neuen (gleich großen) Hörergruppe mit dem Mittel der vorhandenen (den optimalen Schätzern perzeptiven Alters) zusammenhängen ( $ICC(2,k)$ ).

Die Koeffizienten, die auf Fall 3 beruhen, sind einschlägig, sofern von Interesse ist, wie gut neue Urteile eines bereits bekannten Hörers (im Mittel) mit den vorhandenen Urteilen zusammenhängen ( $ICC(3,1)$ ) oder wie gut die (zusammengefassten) neuen Urteile einer gleichen Hörergruppe mit dem Mittel der vor-

	spo. R.	gel. R.	/a/-a	/a/-s	/a/-e	/i/-a	/i/-s	/i/-e	/u/-a	/u/-s	/u/-e
Min( $\hat{\sigma}$ (HS))	3,55	3,68	5,27	8,19	6,60	7,77	8,42	10,18	7,60	8,83	6,67
Max( $\hat{\sigma}$ (HS))	11,29	11,91	19,76	19,01	22,54	20,95	19,32	24,73	20,64	18,78	22,82
Max/Min	3,18	3,24	3,75	2,32	3,41	2,70	2,29	2,43	2,72	2,13	3,42
Min( $\hat{\sigma}$ (HS <sub>ips</sub> ))	4,23	3,22	6,62	6,92	7,13	7,12	6,53	8,02	6,63	6,45	8,01
Max( $\hat{\sigma}$ (HS <sub>ips</sub> ))	10,39	11,41	16,84	16,08	17,22	15,12	14,46	21,01	14,65	14,23	17,47
Max/Min <sub>ips</sub>	2,46	3,55	2,55	2,33	2,41	2,12	2,21	2,62	2,21	2,21	2,18

**Tabelle 6.20:** Werte zur Beurteilung der Homoskedasizität: Minimum und Maximum pro Äußerungstyp der geschätzten Populationsvarianzen der Altersschätzungen pro Äußerung sowie ihr Quotient.

festmachen. Ferner genügt *ein* sehr großer oder sehr kleiner Wert, um diesen Quotienten stark zu verzerren – was, wenn man einen Blick auf Abbildung 6.20(k) wirft, bei gelesener Rede auch der Fall ist. Zudem ist dieser „Ausreißer“ ein hoher Wert bei relativ jungem Alter und widerspricht somit dem Trend des Zusammenhangs zwischen Variabilität und mittlerer Tendenz.

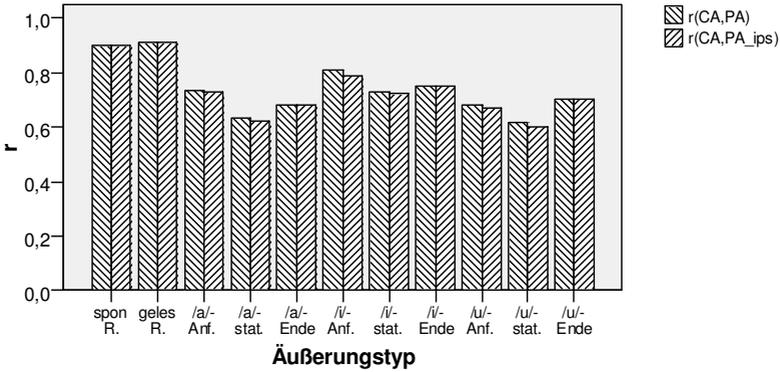
Somit kann bzgl. der ipsativen (also hinsichtlich der Varianz zwischen den Hörern bereinigten) Altersschätzungen anhand aller Äußerungstypen von einer für eine parametrische Analyse hinreichenden Varianzhomogenität ausgegangen werden.

### 6.5.9 Der Zusammenhang zwischen chronologischem und perceptivem Alter

Nachdem im Abschnitt 6.5.3.3 geklärt wurde, dass die Urteile der Hörer bezüglich des Sprecheralters in jedem Fall hinreichend übereinstimmen, im Abschnitt 6.5.3.6, dass verteilungsfreie und parametrische Reliabilitätsanalysen zu vergleichbaren Ergebnissen gelangen, in den Abschnitten 6.5.3.5, 6.5.5 und 6.5.6, dass sich die Hörschätzungen zwischen den Äußerungstypen hinsichtlich ihrer Reliabilität (daher auch ihrer Variabilität) und der mittleren Tendenz unterscheiden, konnte im Abschnitt 6.5.7 gezeigt werden, dass die Variabilität der Hörschätzungen pro Äußerung mit dem Mittelwert pro Äußerung und auch mit dem chronologischen Alter variiert. Im letzten Abschnitt schließlich konnte gezeigt werden, dass diese Varianzheterogenität – zumindest bzgl. der ipsativen Urteile – einer parametrischen Analyse nicht entgegensteht.

Um die Frage zu klären, wie die zusammengefassten (arithmetisch gemittelten) Altersschätzungen mit dem chronologischen Alter zusammenhängen, werden nachfolgend (parametrische) Korrelations- (Pearsons  $r$ ) bzw. lineare Regressi-

onsanalysen durchgeführt und deren Ergebnisse vorgestellt (vgl. die Abbildungen 6.21, 6.22 und 6.23 sowie die Tabelle 6.21).



**Abbildung 6.21:** (Externe) Kriteriums-Validität aller Hörerurteile der Querschnittsuntersuchung: Koeffizient nach Pearson ( $r$ ) der Korrelation zwischen chronologischem und perzeptivem Alter der Sprecherinnen pro Äußerungstyp im Vergleich zwischen originalen und ipsativen Urteilen.

Ein erster Blick auf Abbildung 6.21 zeigt, dass sich die Pearson'schen Koeffizienten  $r$  der Korrelation zwischen dem chronologischen Alter und den Mittelwerten der originalen Hörschätzungen und die entsprechenden Koeffizienten bzgl. der ipsativen Hörschätzungen kaum unterscheiden. (Wobei die ipsativen tendenziell etwas tiefer liegen.) Das ist natürlich darauf zurückzuführen, dass sich schon die Mittelwerte der ipsativen und originalen Schätzungen nur marginal unterscheiden (vgl. auch Abschnitt 6.5.5). Dies kann an den Abbildungen 6.22 und 6.23 für die einzelnen Schätzer des perzeptiven Alters (an der fast identischen Lage der

	spo.R.	gel.R.	/a-a	/a-s	/a-e	/i-a	/i-s	/i-e	/u-a	/u-s	/u-e
$r$	0,900	0,910	0,736	0,632	0,682	0,812	0,731	0,750	0,680	0,619	0,703
$R^2_{korr}$	0,808	0,826	0,536	0,392	0,459	0,656	0,529	0,558	0,456	0,376	0,488
$r_{ips}$	0,901	0,910	0,729	0,623	0,682	0,790	0,723	0,750	0,670	0,602	0,703
$R^2_{korr,ips}$	0,809	0,826	0,525	0,380	0,459	0,619	0,517	0,558	0,442	0,355	0,488

**Tabelle 6.21:** Korrelationskoeffizienten und Determinationskoeffizienten (pro Äußerungstyp) der Korrelationen bzw. Regressionen des perzeptiven Alters durch das chronologische Alter der Sprecherinnen. Alle Regressionen bzw. Korrelationen sind höchst signifikant ( $p < 0,5\%$ ).

einzelnen Datenpunkte) nachvollzogen werden. Da dennoch die Variabilität der einzelnen Hörerurteile bei der Ermittlung dieser Schätzer der Alterswirkung Unterschiede aufweist, die auf vermehrte Varianzheterogenität der originalen Werte schließen lässt, werden die akustischen Analysen zur Bestimmung der Korrelate der Alterswirkung (vgl. Kapitel 8), wie auch die weiteren Analysen in diesem Kapitel, sofern nicht explizit anderes erwähnt ist, mit den Mittelwerten der ipsativen Schätzungen durchgeführt.<sup>44</sup>

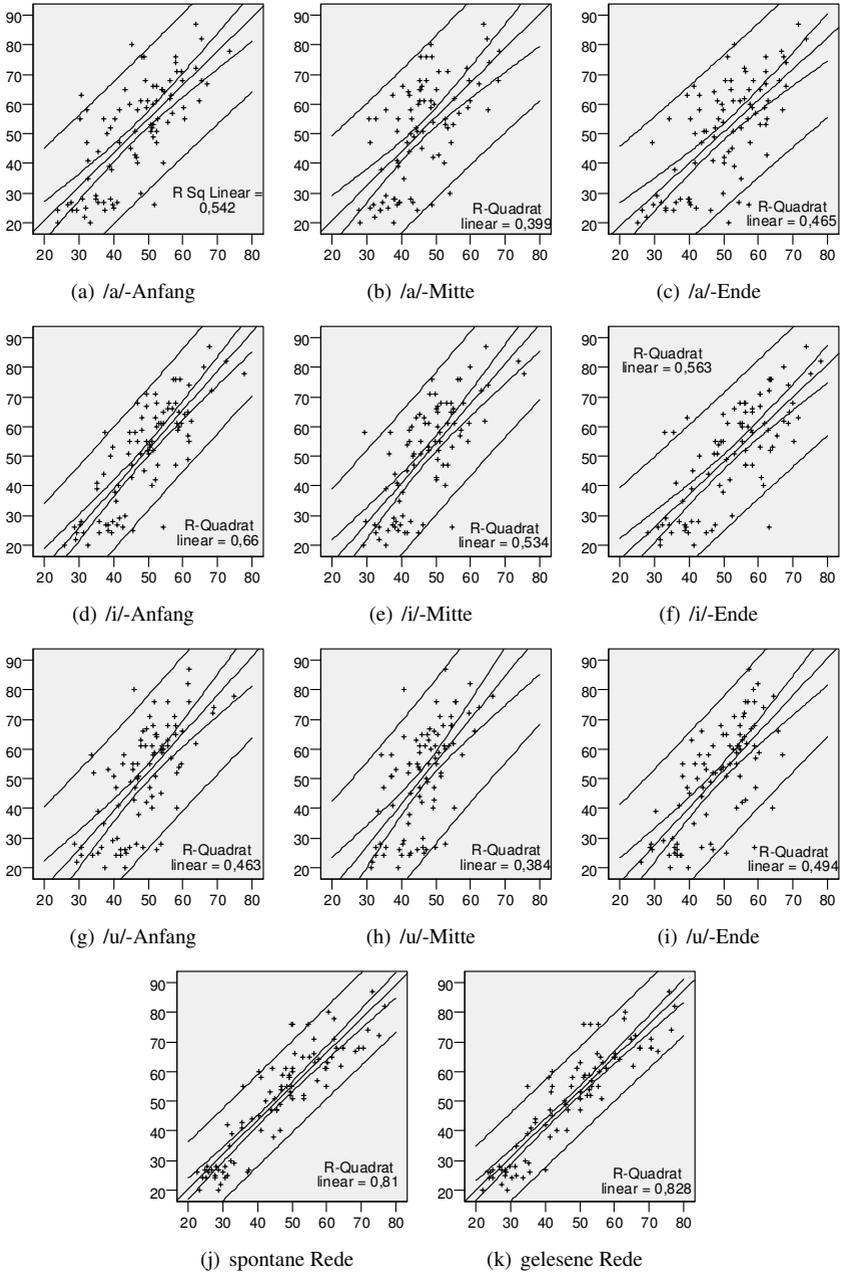
Vergleicht man die einzelnen Äußerungstypen hinsichtlich der externen Kriteriumsvalidität (der Höhe der Korrelationskoeffizienten, also hinsichtlich der Stärke des Zusammenhangs, den Altersschätzungen auf ihrer Grundlage mit dem chronologischen Sprecheralter aufweisen), die sie erzielen, so zeigt sich (fast) das von der vergleichenden Betrachtung der Reliabilitäten bekannte Bild: Redeäußerungen weisen die höchsten Koeffizienten auf. Die Anfänge und Enden der gehaltenen Vokale erzielen höhere Validitäten als die quasi-stationären Mittelteile. Auch dies ist, wie schon bei den Reliabilitäten als Folge des unterschiedlichen Informationsgehaltes über das (jetzt chronologische) Alter zu deuten. Ein Unterschied zu den Relationen bei den Reliabilitäten besteht jedoch: hinsichtlich ihrer Reliabilität (und damit der internen Konstruktvalidität) sind gehaltene /i/s mit den /a/s vergleichbar, die /u/s schlechter. Bezüglich der Fähigkeit das chronologische Alter zu indizieren sind die /i/s valider als die /a/s und /u/s.

Wie genauer der Tabelle 6.21 zu entnehmen ist, liegen die Korrelationskoeffizienten bzgl. der Redeäußerungen höher als 0,9 und denotieren damit sehr starke Zusammenhänge zwischen chronologischem und perzeptiven Alter. Die entsprechenden (korrigierten) Determinationskoeffizienten bescheinigen Varianzaufklärungen von mehr als 80%! Altersschätzungen anhand gehaltener Vokale rangieren diesbezüglich zwischen 35,5% bei den /u/-Mittelstücken und 61,9% anhand der /i/-Anfänge.

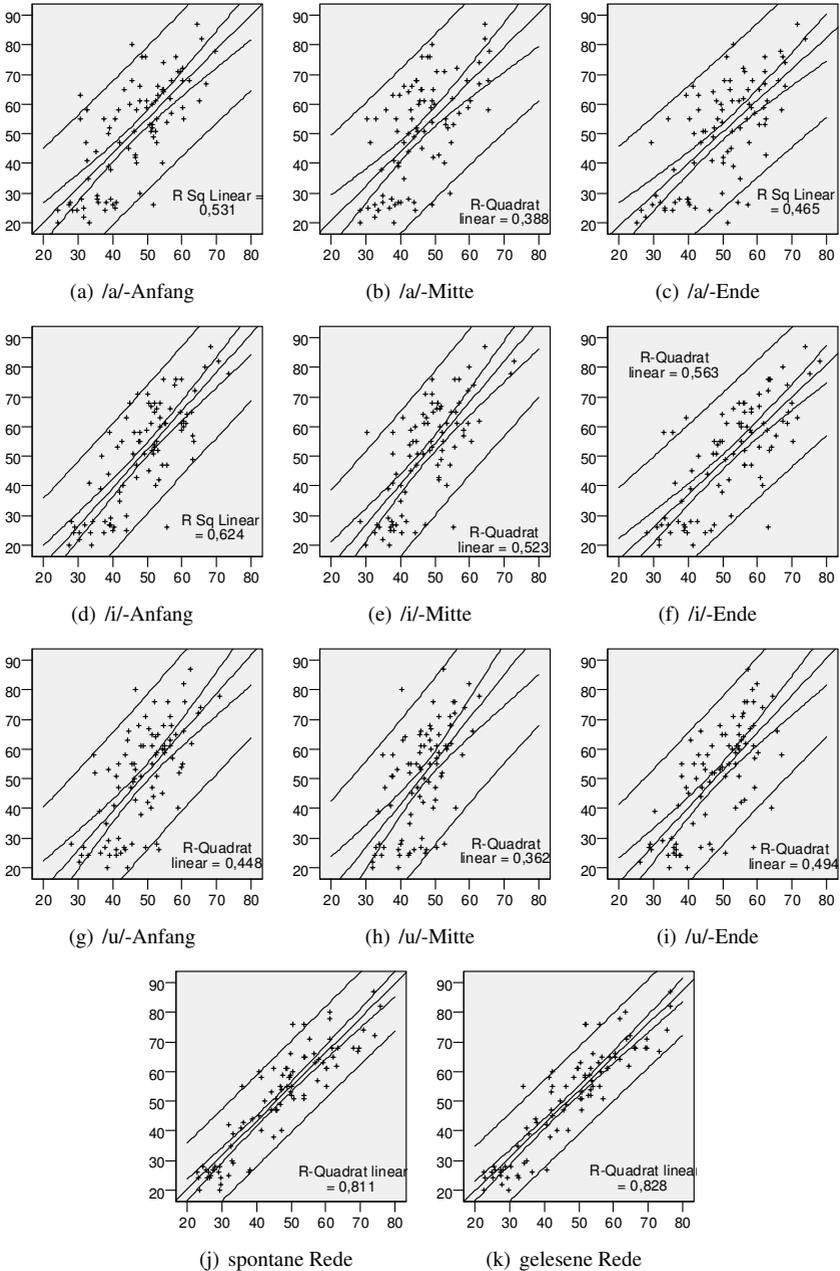
Will man also das chronologische Alter einer Sprecherin aus dem an einer ihrer Äußerungen ermittelten perzeptiven Alter mithilfe eines linearen Regressionsmodells schätzen, so sind die dazu nötigen (am hier vorliegenden Datensatz ermittelten) Kennwerte in Tabelle 6.22 zu finden. Die allgemeine Regressionsgleichung dazu lautet

$$CA = B_1 \cdot PA + B_0, \quad (6.1)$$

<sup>44</sup> Mit gängigen Methoden (Betrachtung der Regressionsresiduen in Abhängigkeit der vorhergesagten Werte) zur Bestimmung von Varianzheterogenität sind beide auf den jeweiligen Mittelwerten gerechneten Regressionen (bei allen Äußerungen) nicht als varianzheterogen zu klassifizieren. Nimmt man allerdings die nicht gemittelten Werte der einzelnen Hörer – was wegen der seriellen Abhängigkeit zwischen den Beurteilungen gleicher Äußerungen zur Unterschätzung des Schätzfehlers der Regression und damit zu zu engen Konfidenzbereichen führen würde – so zeigt sich die im Abschnitt 6.5.6 bereits dargestellte vermehrte Varianzheterogenität der originalen Urteile.



**Abbildung 6.22:** Chronologisches Alter (Ordinate) als Funktion der originalen Werte des perzeptiven Alters aller Hörer der Querschnittsdesigns (Abszisse) pro Äußerungstyp. Die Hyperbeln kennzeichnen die 95%-tigen Konfidenzintervalle.



**Abbildung 6.23:** Chronologisches Alter (Ordinate) als Funktion der ipsativen Werte des perceptiven Alters aller Hörer der Querschnittsdesigns (Abszisse) pro Äußerungstyp. Die Hyperbeln kennzeichnen die 95%-tigen Konfidenzintervalle.

wobei PA für perzeptives Alter, CA für chronologisches Alter und B für den jeweiligen Regressionskoeffizienten steht. Demzufolge ist eine (neue) Sprecherin, die anhand ihrer gelesenen Äußerung auf 50 Jahre geschätzt wird, am wahrscheinlichsten

$$CA_{gel.R.50} = 1,13 \cdot 50 - 2,8 = 53,7 \quad (6.2)$$

Jahre alt. Mit 95%-tiger Wahrscheinlichkeit liegt ihr tatsächliches chronologisches Alter zwischen 38,9 und 68,3 Jahren<sup>45</sup>. Das sind die bestmöglichen Schätzungen chronologischen Alters, die am vorliegenden Datensatz ermittelt werden können. Diese Konfidenzintervalle sind, wie Gleichung 4.55 zu entnehmen ist, ferner (u.a.) davon abhängig, wie weit das geschätzte Alter, das zur Prognose herangezogen wird, vom Mittelwert aller Schätzungen (in der Modellierungstichprobe) entfernt ist. Deshalb ergeben sich die in den Abbildungen 6.22 und 6.23 verzeichneten hyperbolischen Konfidenzintervallgrenzen. Des Weiteren hängt diese Konfidenzintervallabschätzung natürlich auch von der Anzahl der Objekte ab, die zur Erstellung des Modells beigetragen haben: Mit mehr (repräsentativ) ausgewählten Sprechern dürften die Konfidenzintervalle (sehr wahrscheinlich) zunehmend kleiner werden.

	spo.R.	gel.R.	/a/-a	/a/-s	/a/-e	/i/-a	/i/-s	/i/-e	/u/-a	/u/-s	/u/-e
$\hat{\sigma}_{y x}$	7,72	7,36	12,14	13,87	12,96	10,83	12,20	11,67	13,11	14,10	12,56
B <sub>0</sub>	-2,92	-2,80	-3,29	-2,82	-2,42	-12,79	-14,64	-6,83	-13,70	-16,80	-10,33
B <sub>1</sub>	1,16	1,13	1,17	1,17	1,06	1,30	1,38	1,10	1,32	1,46	1,26
$\hat{\sigma}_{E,0}$	2,90	2,73	5,68	7,47	6,35	5,47	6,90	5,64	7,90	9,86	6,86
$\hat{\sigma}_{E,1}$	0,06	0,06	0,12	0,16	0,12	0,11	0,14	0,10	0,16	0,21	0,14
B <sub>0,u</sub>	-8,68	-8,23	-14,59	-17,68	-15,05	-23,67	-28,35	-18,03	-29,41	-36,41	-23,97
B <sub>1,u</sub>	1,04	1,02	0,93	0,85	0,81	1,08	1,10	0,89	1,00	1,04	0,99
B <sub>0,o</sub>	2,84	2,63	8,00	12,05	10,21	-1,90	-0,92	4,38	2,01	2,81	3,31
B <sub>1,o</sub>	1,28	1,24	1,41	1,49	1,31	1,52	1,67	1,31	1,64	1,87	1,54

**Tabelle 6.22:** Standardschätzfehler ( $\hat{\sigma}_{y|x}$ ), Regressionskoeffizienten (B), Standardfehler der Koeffizienten ( $\hat{\sigma}_E$ ) und 95%-tige Konfidenzgrenzen der Regressionen des chronologischen Alters der Sprecherinnen durch das perzeptive Alter pro Äußerungstyp.

Blieben wir bei den vorliegenden Daten, so gelangen die Äußerungen, die am wenigsten Altersinformation tragen, die /u/-Mittelteile, zu einem ähnlichen wahrscheinlichsten Wert, nämlich 56,0 Jahre, sofern die Schätzung (einer vergleichbar

<sup>45</sup> Diese Ober- und Untergrenzen für neue Beobachtungen ergeben sich nicht (!), wenn man die Ober- und Untergrenzen der Regressionskoeffizienten (B<sub>0,u</sub>, B<sub>0,o</sub> usw.) in die Regressionsgleichung einsetzt, sondern mit Hilfe des Standardschätzfehlers ( $\hat{\sigma}_{x|y}$ ) aus Gleichung 4.55.

fischen Vokaltyp<sup>15</sup> – als potentielle Prädiktoren nimmt MFTR[/i/s], Pf0min[/a/e], Pf3[/i/a] und sAPQ[/a/s] als Prädiktoren auf und gelangt damit zu einer korrigierten Varianzaufklärung ( $R_{korr}^2$ ) des chronologischen Alters von 43,1% ( $R = 0,677$ ).

### 8.1.1.10 Regression des chronologischen Alters anhand der Korrelate, die aus allen untersuchten Äußerungen extrahierbar sind

Versucht man, das Regressionsmodell, das bisher die beste Varianzaufklärung des chronologischen Alters erzielte (vgl. Gleichung 8.6), zusätzlich durch die Parameter FTrI\_Ln[/a/a], ATrI\_Ln[/a/a], MFTR[/i/s], FTrI\_Ln[/i/s], FTrI\_Ln[/u/e], Pf0min[/a/e], Pf3[/i/a] und sAPQ[/a/s] aus den Vokaläußerungen zu verbessern, lässt sich die Varianzaufklärung tatsächlich nochmals um 2,3 Prozentpunkte erhöhen. Dabei wird nur das Mittel über die Frequenz-Tremor-Intensitäts-Kontur (MFTR) aus den /i/-Mittelteilen zu den Prädiktoren hinzugefügt.<sup>16</sup> Die Gleichung dieses hinsichtlich des Determinationskoeffizienten optimierten Modells lautet:

$$\begin{aligned}
 CA = & 2,207 \cdot Papq11\_vow[spo.R.] - 3,785 \cdot arPhonKorr[gel.R.] \\
 & - 0,292 \cdot Pf0min\_a[gel.R.] + 0,688 \cdot Pnoiint\_70\_45h\_i[spo.R.] \\
 & + 0,972 \cdot @6000[gel.R.] + 16,173 \cdot Pt0sd\_nas[gel.R.] \\
 & + 19,224 \cdot MFTR[/i/s] + 56,443
 \end{aligned} \tag{8.7}$$

Mit diesem Modell, basierend auf gelesener und spontaner Rede sowie den quasistationären Mittelteilen der gehaltenen /i/s, lassen sich also knapp 75% ( $R_{korr}^2 = 0,747$ ;  $R = 0,883$ ) der Varianz des chronologischen Alters der Sprecherinnen erklären (vgl. Tabelle 8.11). Diese Schätzgenauigkeit ist – wie schon diejenige des Modells 8.6 – durchaus mit der Genauigkeit von Altersschätzungen vergleichbar, die von größeren Hörergruppen an Redeäußerungen ermittelt werden können!

Den größten Erklärungsbeitrag leistet dabei (1) das mit PRAAT an den /a/-Segmenten der spontanen Äußerungen ermittelte Minimum der Grundfrequenz (vgl. die  $\beta$ -Koeffizienten in Tabelle 8.11). Die restlichen Parameter tragen zur Varianzaufklärung etwa gleiche Anteile bei, die ca. zwei Drittel des Anteils des

<sup>15</sup> Die eingegebenen Parameter waren: FTrI\_Ln[/a/a], ATrI\_Ln[/a/a], MFTR[/i/s], FTrI\_Ln[/i/s], FTrI\_Ln[/u/e], sAPQ[/a/s], Pf3[/i/a], Pf0min[/a/e], @3920[/i/a], DSH[/i/s], PFR[/a/e], Pt0sd[/u/s], @80[/i/e], vFo[/a/e], und hnr007\_16[/a/e]. Dabei wurden die fehlenden Werte der Tremorparameter durch den Wert 0 ersetzt.

<sup>16</sup> Bei der Berechnung des (Logarithmus des) Frequenz-Tremor-Intensitäts-Index konnten bei einigen Sprecherinnen keine stärksten Tremorfrequenzen festgestellt und damit kein Wert für die Intensität dieser Tremorfrequenz ermittelt werden. Die so entstandenen fehlenden Werte wurden auch in dieser Analyse durch Nullen ersetzt, wodurch er nicht mehr die hohen Korrelationswerte aufweisen konnte.

Prädiktoren	Reg.-Koeff.	SE	$\beta$ -Koeff.	p	Reg.-Koeff. <sub>u</sub>	Reg.-Koeff. <sub>o</sub>
(Konstante)	56,443	23,002		0,018	10,169	102,718
Papq11_vow[spo.R.]	2,207	0,809	0,216	0,009	0,579	3,835
arPhonKorr[gel.R.]	-3,785	1,083	-0,258	0,001	-5,963	-1,607
Pf0min_a[gel.R.]	-0,292	0,074	-0,314	0,000	-0,441	-0,144
Pnoiint_70_45h_i[spo.R.]	0,688	0,235	0,223	0,005	0,215	1,161
@6000[gel.R.]	0,972	0,338	0,222	0,006	0,293	1,651
Pt0sd_nas[gel.R.]	16,173	6,118	0,206	0,011	3,867	28,480
MFTR[i/i-s]	19,224	8,306	0,173	0,025	2,515	35,933

**Tabelle 8.11:** Kennwerte der Regression zur Abschätzung des chronologischen Alters ( $R^2_{\text{korrr}} = 0,747$ ) durch Parameter aus allen Äußerungen.

eben genannten Parameters ausmachen. Den größten dieser Anteile erklärt (2) die an gelesener Rede gemessene, bzgl. Elaborationen und Reduktionen korrigierte (Phon-) Artikulationsrate. Weitere sehr vergleichbare Anteile liefern (3) die mit PRAAT an Vokal-Segmenten spontaner Rede erhobene, über elf Perioden geglättete Amplituden-Perturbation, (4) das unter PRAAT selbstentwickelte Maß der Intensität aperiodischer Energie im Frequenzbereich zwischen 70 und 4500 Hertz, gemessen an spontaner Rede, (5) die über das Langzeitspektrum an gelesener Rede ermittelte Intensität im Frequenzbereich von 5920 bis 6080 Hertz und (6) die mit PRAAT an Nasal-Segmenten gelesener Rede erhobene Standardabweichung der Periodendauer. Den geringsten Anteil (der jedoch immerhin mehr als die Hälfte des größten beträgt) erklärt (7) das mit MDVP an quasi-stationären Teilen gehaltener /i/s ermittelbare Integral über die spektrale Kontur im Bereich zwischen zwei und zwölf Hertz der Frequenzremor-Intensität.

Auf der Beschreibungsebene der Sprachsignal-Konstrukte bedeutet das, dass die Stimmen bzw. Sprechweisen chronologisch älterer Frauen sich dadurch auszeichnen, dass (1) (wenigstens) das untere Ende ihres (gewöhnlichen) Sprechtonlagenbereichs während der Produktion von /a/s in spontaner Rede tiefer liegt, (2) sie langsamer und/oder weniger präzise artikuliert laut vorlesen (3) ihre Stimme während spontaner Rede rauer klingt, (4) die Behauchung (bzw. der Geräuschanteil) im gesamten Frequenzbereich, in dem lautsprachliche Information hauptsächlich übertragen wird, zunimmt, (5) das Sprachsignal in einem Frequenzbereich lauter wird, der über dem Haupt-Sprachbereich liegt, (6) (wenigstens während der Produktion von Nasalen beim Lesen) die Stimme heiserer klingt und (7) die Stimme (wenigstens in gehaltenen /i/s) zittriger wird.

Und diese sieben Merkmale sind ausreichend, um die Veränderungen durch die Alterung so genau zu erfassen wie eine größere Gruppe von Hörern, also viel genauer als die meisten Hörer alleine dazu in der Lage sind.

## 8.1.2 Zusammenhänge mit perzeptivem Alter

Wie theoretisch zu erwarten ist, wird sich in diesem Abschnitt zeigen, dass die Zusammenhänge akustischer Parameter mit dem perzeptiven Alter generell stärker bzw. vielfältiger sind als mit dem chronologischen Alter. Die Alterswirkung lautsprachlicher Äußerungen lässt sich also noch genauer akustisch erfassen als das chronologische Alter der Sprecherinnen.

### 8.1.2.1 Korrelate, die aus allen Äußerungstypen extrahierbar sind

Die Standardabweichung der Periodendauern ( $Pt0sd$ ) ist auch hier, wie schon bei der obigen Betrachtung der Korrelate des chronologischen Alters, der beste Altersindikator (vgl. Tabelle 8.12). Ihre empirische Signifikanz der Korrelation bei den /i/-Anfängen ( $r = 0,295$ ) beträgt immerhin  $p = 0,003$ ; bei den Mittelteilen und Endstücken gar  $p = 0,002$ . Auch diese  $\alpha$ -Fehler-Risiken sind also schwer mit dem Zufall vereinbar. Dennoch ist die Höhe der Zusammenhänge mit ca.  $r = 0,3$  als gering einzustufen. Unter Berücksichtigung dieser Einschränkungen bzgl. der gehaltenen /i/-Vokale kann also  $Pt0sd$  als Korrelat des perzeptiven Alters verstanden werden, das aus allen hier betrachteten Äußerungstypen extrahiert werden kann. Leider ist sie der einzige solche Parameter.

Festzuhalten ist auch, dass  $Pt0sd$  bei gehaltenen Vokalen etwas besser als Korrelat des perzeptiven Alters fungiert als bei Redeäußerungen. Dies kann darauf zurückgeführt werden, dass Sprechtonlagenänderungen bei Redeäußerungen auch durch die linguistische Modulation induziert werden, die bei gehaltenen Vokalen nicht besteht. Sogar (für einzelne akustische Parameter) hoch-altersindizierende Werte finden sich bei den Korrelationen von  $Pt0sd$  mit dem perzeptiven Alter bei den /a/-Enden ( $r = 0,610$ ) und den (quasi-stationären) /u/-Mittelteilen ( $r = 0,644$ ).

Sieht man von den /i/-Abschnitten ab, so erweisen sich die Parameter, welche die zentrale Tendenz der Sprechtonlage erfassen ( $F_0$ ,  $MF_0$ ,  $T_0$ ,  $Pf_0am$ ,  $Pf_0md$  und  $Pt_0am$ ), als ebenbürtig zu  $Pt0sd$  – besonders das arithmetische Mittel der  $F_0$ -Werte pro Analysefenster, extrahiert durch MDVP ( $MF_0$ ) und der Median der von PRAAT extrahierten  $F_0$ -Werte pro Analysefenster ( $Pf_0md$ ).

Neben diesen Sprechtonlagen- und Sprechtonverteilungsparametern erzielen nur der Harmonics-to-Noise Ratio extrahiert mit PRAAT aus dem Frequenzbereich 70–4500 Hz ( $hnr007\_45$ ) sowie zwei benachbarte hochfrequente LTAS-Bänder (mit Mittenfrequenzen bei 5200 und 5360 Hz) bei beiden Redeäußerungen und

## Kapitel 9

# Diskussion und Zusammenfassung

## 9.1 Diskussion der Ergebnisse

### 9.1.1 Diskussion der Ergebnisse der perceptiven Analyse

Will man möglichst genaue Abschätzungen des chronologischen Alters von Frauen anhand ihrer Stimme bzw. Sprechweise erhalten, so empfiehlt es sich, eine möglichst große Hörergruppe zu suchen, deren Schätzungen zu mitteln und diese Mittelwerte mithilfe der Regressionsgleichungen in Tabelle 6.22 in Schätzwerte des chronologischen Alters zu überführen.

Kann man sich dabei die Äußerungen aussuchen, so wäre für diese Abschätzung eine gelesene Redeäußerung zu empfehlen. Diese Redeäußerung sollte möglichst lange dauern. Allerdings ist zu erwarten, dass der Zuwachs an Genauigkeit der Schätzung durch längere Dauern rasch kleiner wird. Vermutlich reichen nicht wesentlich längere Äußerungen als die hier untersuchten (also ca. 20 Sekunden dauernde), um nur noch unwesentlich vom Maximum der erreichbaren Genauigkeit entfernt zu sein.

Ferner erscheint klar, dass das einfach nur gemittelte Alter der Schätzungen einer Hörergruppe das reelle chronologische Alter erwachsener Sprecherinnen eher unterschätzt und das umso mehr, je höher die Schätzung (bzw. das unbekannte chronologische Alter) ist. So gibt es wohl mehr jung klingende Alte als alt klingende Junge. In der vorliegenden Arbeit wurde ferner deutlich, dass einige Hörer sich scheuen, höhere Schätzungen abzugeben.

Die Art der Höreerauswahl bzw. auch die persönlichen Merkmale der Hörer sind dabei (wahrscheinlich) vergleichsweise unbedeutend. Nach Linville (2001) sind die Schätzungen jüngerer erwachsener Hörer genauer. Vielleicht liegt der eben beschriebene Effekt aber auch daran, dass sich diese jüngeren Hörer eher scheuen, höhere Schätzungen abzugeben. Das muss an anderer Stelle näher empirisch untersucht werden.

Jedenfalls ist es allen Hörern, sofern sie motiviert sind, möglichst gute Altersschätzungen abzugeben, ähnlich gut (konkordant, nicht absolut) möglich, diese Aufgabe zu erfüllen.

Die hier erreichten Varianzaufklärungen des chronologischen Alters durch die Hörschätzungen sind zwar geringfügig höher als das anderen Veröffentlichungen zu entnehmen ist (sofern sie vergleichbar sind), jedoch scheint ein (Pearson'scher)

Korrelationskoeffizient von ca.  $r = 0,9$  (ca. 80% Varianzaufklärung) an Stichproben, welche die Grundgesamtheit erwachsener Sprecher(-innen) abbilden, zu erwarten zu sein.

Auch die Unterschiedlichkeit der hier eingesetzten Wahrnehmungserhebungsverfahren wirkt sich offenbar nicht wesentlich auf die Schätzungen aus. Leicht optimieren ließe sich ein Schätzverfahren durch Hörer, wenn die „Vorlieben“ der Hörer bekannt wären, wenn also die Hörergruppe zuvor an einer repräsentativen Äußerungs-Stichprobe „geeicht“ worden wäre.

Das wichtigste Ergebnis der Wahrnehmungserhebung an den Längsschnittdaten ist, dass es Hörern (als Gruppe) bei der Beurteilung von Redeäußerungen gelingt, den Unterschied von fünf Jahren korrekt und signifikant zu erkennen. An gehaltenen Vokalen ist das nicht mehr möglich. Somit kann gefolgert werden, dass bei ca. fünf Jahren – abhängig von der vorliegenden Äußerungsart – die eben merkbliche Differenz der Wahrnehmung von Altersunterschieden an Stimme und Sprechweise liegt.

## 9.1.2 Diskussion der Ergebnisse der akustischen Analyse

### 9.1.2.1 Ergebnisse der Replikationen der Messungen aus anderen Untersuchungen

#### 9.1.2.1.1 Grundfrequenz

Die generelle Absenkung der Sprechtonlage bzw. der Grundfrequenz der Rede von Frauen mit zunehmendem Alter konnte auch hier bestätigt werden. Als besonders gutes Korrelat des Alters erwies sich dabei der Median der mit PRAAT aus vokalischen Segmenten spontaner Rede extrahierten Grundfrequenz (Pf0md\_vow). Der Median gewichtet extreme Werte nicht so stark wie das arithmetische Mittel dies tut, woraus zu schließen ist, dass die Extremwerte der (automatischen) Grundfrequenzextraktion zumindest bzgl. der Alterserkennung vermehrt Fehlervarianz erzeugen.

An gelesener Rede ist allerdings kein Mittelwert das beste Sprechtonlagenmaß des Alters, sondern das Minimum, erhoben an Nasalen und /a/-Vokalen (Pf0min\_nas und Pf0min\_a): Vor allem die tiefste beim Lesen produzierte Tonlage verweist auf das Alter der Sprecherinnen.

Die Standardabweichung der Periodendauern (Pt0sd) ist zwar kein Maße der Tonlage, sondern ihrer Verteilung (-sbreite) und damit je nach Äußerungstyp vielleicht auch ein Maß der (intonatorisch bedingten) Grundfrequenzvariation oder auch der Perturbation. Trotz oder auch wegen dieses Facettenreichtums ist sie derjenige Parameter, der (unter den hier gemessenen) aus den meisten Äußerungstyp-

pen extrahiert als Korrelat des Alters fungiert. Allerdings immer nur in moderatem Ausmaß.

Auch andere Operationalisierungen der Sprechtonlage variieren nicht wesentlich geringer mit dem Sprecheralter, dennoch ist die Sprechtonlage als alleiniges oder auch Haupt-Korrelat des Alters (in einer automatischen Erkennung) wohl ungeeignet: Schließlich variiert die Sprechtonlage in Abhängigkeit der (linguistisch motivierten) Prosodie, des Geschlechts, der Persönlichkeit, des Gesundheitszustandes, des Tagesverlaufs, etc. Im Einzelfall eine Tonhöhe (-änderung) auf ein(en) Alter (-sunterschied) zurückzuführen, ist also kaum möglich.

### 9.1.2.1.2 Formanten

Formanten sind wohl nur als Korrelate der Alterswirkung, nicht aber des chronologischen Sprecheralters von Relevanz. Extrahiert aus spontaner Rede weisen der zweite und der vierte Formant von vokalischen Abschnitten bzw. /a/-Segmenten (Pf4\_vow, Pf2\_a) Zusammenhänge mit dem perceptiven Alter auf. An gehaltenen Vokalen sind es der zweite und dritte, an gelesener Rede die Standardabweichung des vierten. Nasalformanten finden sich gar nicht unter den Korrelaten des Alters. Auch kein erster Formant sinkt mit zunehmendem Alter der Sprecherinnen.

Eigentlich lassen die physiologischen Befunde eine Verlängerung des alternenden Ansatzrohrs erwarten und dadurch deutlichere Absenkungen aller Formanten. Vor allem aber der erste Formant sollte bei längeren Ansatzrohren tiefer liegen. Diesem Effekt wirkt aber (bei Frauen) wohl der Effekt der Vokalzentralisation entgegen, welcher ebenfalls mit dem Alter zunehmen dürfte. Der einzige Vokal, bei dem (theoretisch) beide Effekte am gleichen Strang ziehen, ist das [æ], das hier nicht explizit untersucht wurde, da es nicht Teil des dt. Phonemsystems ist (vgl. dazu aber Linville & Fisher (1985)).

Hinzu kommt, dass (die ersten beiden Vokal-) Formanten das Hauptunterscheidungskriterium verschiedener Vokalqualitäten sind, also extrem der linguistischen Modulation ausgesetzt sein müssen. Diese erzeugt (sehr wahrscheinlich) so viel Varianz bzgl. der Formantlagen, dass beide zuvor genannten Effekte vor diesem Hintergrund schwer zu entdecken sind.

Dies mag die von Untersuchung zu Untersuchung wechselnden Befunde erklären, die sich auch in den verschiedenen hier erzielten Ergebnissen widerspiegeln. Mit großer Sicherheit ist aber davon auszugehen, dass nicht alle vier Formanten im Verlauf von fünf Jahren in sämtlichen Phonen absinken (vgl. Endres et al. (1971)). Dies widerlegt deutlich die eigene Fünf-Jahres-Untersuchung, in welcher der Point of Formant Concentration hochsignifikant mit dem Alter steigt.

Trotz einer Verlängerung des Ansatzrohres mit zunehmendem Alter sind die Formanten als akustische Ableitung der dadurch veränderten Resonanzverhältnis-

se zur Alterserkennung ungeeignet. Es bedarf wohl langer (und linguistisch hinsichtlich der verwendeten Vokalqualitäten vergleichbarer) Redepassagen um diesen Effekt überhaupt nachweisen zu können.

#### 9.1.2.1.3 LTAS-Intensitäten

Die Veränderung der Intensitäten in bestimmten Frequenzbändern, die mit dem Long-term Average Spectrum (LTAS) bestimmbar sind, auf das Alter der Sprecherinnen zurückzuführen, erweist sich ebenfalls als schwierig.

Zunächst ist festzustellen, dass die LTAS-Parameter offenbar kaum als Korrelate der Alterswirkung in Erscheinung treten. Und da Hörer (als Gruppe) noch immer die besten „Alterserkenner“ an Stimme und Sprechweise sind, muss das wenigstens als Hinweis zur Vorsicht verstanden werden.

Dennoch erwies sich die Intensität um 6000 Hertz (@6000) sowohl bei Linville (2002) als auch hier als Korrelat des chronologischen Alters von Sprecherinnen – zumindest wenn (längere) Redeäußerungen als Grundlage der Parameterextraktion dienen. Auch diese Zusammenhänge sind hier eher moderat ausgefallen, jedoch erklären sie Varianzanteile des chronologischen Alters, die andere Parameter nicht erklären können. Auf welche zugrundeliegende Veränderung dieser Intensitätsanstieg allerdings zurückgehen könnte, ist bislang unklar.

Die von Linville (2002) gefundenen Effekte bzgl. tieferer Frequenzen können so nicht bestätigt werden.

#### 9.1.2.1.4 Intensitäten der Harmonischen

Der von Decoster & Debruyne (1997a) gefundene Effekt von mit zunehmendem Alter kleiner werdenden Differenzen zwischen den Intensitäten der Grundfrequenz und der ersten Oberschwingung kann nicht bestätigt werden.

#### 9.1.2.1.5 Maße des spektralen Rauschens

Die erstmals von Ramig (1983) durch visuelle Begutachtung von Spektrogrammen entdeckte Zunahme spektralen Rauschens mit zunehmendem Alter konnte hier durch die Implementation neuer Rauschmaße unter PRAAT objektiv bestätigt werden. Die enorm sensitive Extraktion des Harmonics-to-Noise Ratio unter PRAAT ist dabei wohl entscheidend. Das beste Korrelat des Alters unter den hier erhobenen Maßen des spektralen Rauschens ist die Intensität „aperiodischer Energie“ im Frequenzbereich zwischen 70 und 4500 Hertz, extrahiert aus /i/-Segmenten spontaner Rede (pnoiint70\_45h\_i). /i/s sind zur Extraktion dieses Parameters wohl besonders geeignet, weil sie in großen Teilen dieses Frequenzbereichs prinzipiell eher wenig Energie aufweisen.

Auch die von Linville (2001) festgestellte Forschungslücke, die spektrales Rauschen als Korrelat der Alterswirkung von Frauenstimmen betrifft, kann als gefüllt betrachtet werden: auch bei Frauenstimmen korrelieren Geräuschmaße mit dem perceptiven Alter.

Allerdings scheint spektrales Rauschen (wenigstens in den hier realisierten Operationalisierungen) konzeptuell ähnlich zu den (Amplituden-) Perturbationsmaßen zu sein. An quasi-periodischen Signalen, bei denen die Perturbationsmaße höhere Zusammenhänge aufweisen, erklären diese offenbar auch den Anteil der Varianz des Alters, den die Rauschmaße erklären könnten. Dies ist aber nicht weiter verwunderlich. Schließlich „erzeugt“ jede Stimm-perturbation per Definition vermehrt unmodulierten also geräuschartigen Schall.

#### 9.1.2.1.6 Sprechgeschwindigkeit

Eine Verringerung der Artikulationsgeschwindigkeit mit zunehmendem (perzeptivem wie auch chronologischem) Alter zeigt sich in der vorliegenden Untersuchung zwar an vielerlei einzelnen (Dauer- und Geschwindigkeits-) Parametern, jedoch nur sofern diese aus gelesener Rede extrahiert wurden. An spontaner Rede konnte ein solcher Effekt nicht nachgewiesen werden.

Dreierlei Möglichkeiten bestehen zur Deutung dieses Befundes: Entweder die Reduktion der Sprechgeschwindigkeit tritt tatsächlich nur (in messbarem Ausmaß) auf, wenn Sprecher(-innen) laut vorlesen oder die untersuchten spontanen Redeäußerungen sind zu kurz und (hinsichtlich ihres linguistischen Gehalts) zu variabel, als dass der Effekt erfassbar wäre. Oder beides trifft zu. Sofern das Lesen den (Haupt-) Effekt verursacht, besteht die Möglichkeit, dass eine Verlangsamung visuelle und/oder kognitive und/oder motorische Ursachen hat – vielleicht aber auch nur konventionelle, derart, dass ältere Menschen langsamer lesen, weil sie das so gelernt haben, oder jüngere schneller, weil sie meinen, das wäre besser. Dieser Befund lässt also Fragen offen, die an anderer Stelle geklärt werden müssen.

#### 9.1.2.1.7 Perturbationen

Zwei von Linville (2001, S. 196) geäußerte Befunde bzw. Vermutungen konnten hier bestätigt werden: Erstens sind Amplituden-Perturbationen im Vergleich zu Grundfrequenz-Perturbationen auch hier die besseren Korrelate des (perzeptiven wie auch chronologischen) Alters. Zweitens sind „vermehrt geglättete“ Maße, die Unterschiede zwischen zeitlich größeren Abschnitten erfassen, bessere Indikatoren des Alters als solche, die Unterschiede in zeitlich kleineren Abschnitten (z.B. von Periode zu Periode) fokussieren.

So sind der über elf Perioden geglättete Amplituden-Perturbations-Quotient (Papq11), extrahiert aus spontaner Rede, der über 55 Perioden geglättete

Amplituden-Perturbations-Quotient (sAPQ) bei gehaltenen Vokalen und die Standardabweichung der Periodendauern (pt0sd), also jeweils Parameter mit der größten zeitlichen Auflösung (wenn man von  $vF_0$  und  $vAm$  bei den gehaltenen Vokalen absieht) generell diejenigen, die am stärksten mit Alter korrelieren.

Linville vermutet, dass das daran liegen könnte, dass gerade diese Parameter nicht nur unregelmäßige Fluktuationen erfassen, sondern auch periodische bestimmter tiefer Frequenzen – also Zittern. Auch das kann bestätigt werden.

Dass  $F_0$ -Perturbationen weniger als Korrelate des Alters fungieren, wird generell dem Umstand zugeschrieben, dass diese Maße mehr den Gesundheitszustand abbilden als das Alter. Auch dieser Vermutung widersprechen die hier gewonnenen Daten nicht.

Nachteilig für eine Anwendung der Perturbationsmaße zur automatischen Alterserkennung ist aber nach wie vor, dass auch sie relativ stark durch (viele) weitere Konstrukte bzw. Begebenheiten beeinflusst sind. Vor allem der Einfluss der zwischen verschiedenen Analysesystemen sehr unterschiedlichen Grundfrequenz- und Glottisimpulsextraktion führt im Einzelnen zu Werten, die unreliabler erscheinen als sie es sind. Hier besteht also Bedarf an einer Explizierung der Extraktionsverfahren und ihrer Parameter, um beispielsweise deutlich zu machen, dass mit PRAAT erhobener Shimmer nicht dasselbe ist wie mit MDVP erhobener.

### 9.1.2.2 Ergebnisse der neu und weiter entwickelten Maße

#### 9.1.2.2.1 Korrigierte Sprechrates

Die um Elaborations- und Reduktionseffekte korrigierte Phon-Sprechrates (arPhon-Korr) erwies sich als zuverlässigster Altersindikator der Sprechtempo- bzw. Artikulationsgenauigkeitsmaße unter den hier erhobenen Parametern. Dieses hier zum ersten Mal vorgestellte Maß erfasst in einem ersten Schritt die Koartikulationseffekte für jedes realisierte Phon, bewertet diese, und verrechnet die Summe dieser Werte mit der Anzahl der Phone. Die so korrigierte Phonzahl wird dann wie gewöhnlich zur Berechnung der Sprechrates verwendet.

Nachteilig ist daran, dass die Bestimmung der Elaborationen und Reduktionen sehr aufwändig ist und nur mit (relativ viel) Übung im Interpretieren von Sonagrammen und einer daraus entstehenden engen phonetischen Transkription erfolgen kann. Eine solche Transkription ist derzeit (und wohl auch auf absehbare Zeit) nicht automatisch (durch programmierbare Algorithmen) zu leisten.

#### 9.1.2.2.2 Maße der prosodischen Veränderung von Silben und Phonemen

Die neu entwickelten Maße der Artikulationsmonotonie erwiesen sich als hilfreich zur Deutung von Perturbationseffekten in Redeäußerungen und teilweise auch als

eigenständige Korrelate des Alters. Davon abgesehen erscheint es für ihre zukünftige Verwendung sinnvoller, wenn nicht alle Silben bzw. Phone (gleichzeitig) in ein Maß einfließen, sondern eher vergleichbare; also beispielsweise nur betonte Silben, oder nur unbetonte, usw. Das setzt aber andererseits eine verlässliche Betonungsklassifizierung voraus (die es als programmierten Algorithmus bisher nicht gibt).

#### 9.1.2.2.3 Segmentweise Extraktion stimmqualitativer Parameter

Dagegen als sehr erfolgreich kann die phon-(klassen-)weise Extraktion stimmqualitativer Maße aus Redeäußerungen betrachtet werden. Die besten Korrelate aus beiden Redeäußerungstypen, die in dieser Arbeit extrahiert werden konnten, entstammen dieser Grundidee: Pf0md\_vow, Pf0min\_nas, Papq11\_vow, Pnoint\_70\_45h\_i, Pnhr\_all, Pf4\_vow, Pjita\_nas,...

Ohne diese Anpassung von Maßen, die gewöhnlich nur an quasistationären Abschnitten gemessen werden, an die Redeäußerungen hätte die Varianz des (chronologischen wie perzeptiven) Alters nicht im erfolgten Ausmaß erklärt werden können.

#### 9.1.2.2.4 Maße des spektralen Rauschens

Auch ohne die neu entwickelten Geräuschindizes würden die Varianzaufklärungskoeffizienten nicht die vorzufindende Höhe erreichen. Diese Entwicklungen haben zwei Ursprünge: Die prinzipielle Idee ist den MDVP-Geräuschindizes entnommen, die (Güte der) tatsächliche(n) Realisierung basiert auf der sehr sensitiven Extraktion des Harmonics-to-Noise Ratio unter PRAAT. So konnten Geräuschindizes entwickelt werden, die zur Alterserkennung besser geeignet sind als ihre beiden Ursprünge.

#### 9.1.2.2.5 Tremormaße

Die ebenfalls unter Verwendung von PRAAT-Prozeduren neu entwickelten Tremormaße sind die besten Maße des chronologischen und perzeptiven Alters, die in dieser Arbeit an gehaltenen Vokalen ermittelt werden konnten. Das gilt sowohl hinsichtlich der Ausprägung des Zusammenhangs als auch hinsichtlich der Anwendbarkeit auf die meisten Äußerungstypen.

Sie sind damit (an gehaltenen Vokalen) bessere Korrelate des Alters als (sämtliche) Grundfrequenzparameter, Perturbationsparameter oder Geräuschmaße. Ihr Erfolg ist auch problemlos zu deuten: schließlich sollen diese Maße die Zittrigkeit der Stimme erfassen, die einerseits in der Medizin seit Langem bekannt ist – eben

auch als „Symptom“ des Älterwerdens. Andererseits wurde über signaltechnische Umwege (vgl. Abschnitt 9.1.2.1.7) ihre Relevanz von Linville prognostiziert.

Sie stellen auch bessere Korrelate des Sprecheralters dar als die MDVP-Maße, die wiederum ihre (konzeptuellen) Vorfahren sind. Davon abgesehen sind sie aber mit diesen korreliert genug, um diese und damit auch sich selbst als Maße des Stimmtremors zu validieren.

Der Nachteil hierbei ist, dass zur Extraktion von Tremor zwingend gehaltene Vokale erforderlich sind und dass daran wohl nichts geändert werden kann. Somit können Tremormaße (per Definition) nicht (sinnvoll) zur Alterserkennung an Redeäußerungen eingesetzt werden.

### **9.1.2.3 Ergebnisse der Verknüpfung mehrerer Parameter zur Erklärung des Alters**

Die in dieser Arbeit durchgeführten Regressionsanalysen mit akustischen Parametern als Prädiktoren des chronologischen Alters erreichen eine Varianzaufklärung des Alters von knapp 75%. Damit liegen sie nur etwa 5% unter derjenigen, die eine (relativ große) Hörergruppe erzielt.

Mit anderen derartigen Ergebnissen ist ein Vergleich zwar prinzipiell schwierig, weil in anderen Untersuchungen andere Stichproben vorlagen. Dennoch dürfte eine solch hohe Aufklärungsrate (eine solch hohe Präzision der Vorhersage) des chronologischen Alters bisher nicht erreicht worden sein.

Die Parameter, die zu diesem Ergebnis beitragen, sind größtenteils selbst entwickelte bzw. verfeinerte Operationalisierungen der Konstrukte Sprechtonlage, Rauigkeit, Behauchung, Zittrigkeit und Artikulationsgeschwindigkeit bzw. Artikulationsgenauigkeit.

Auch die Alterswirkung der Äußerungen auf Hörer (als Gruppe) ließ sich in dieser Arbeit besser durch akustische Merkmale vorhersagen als bisher.

## **9.1.3 Diskussion der beide Analysen übergreifenden Ergebnisse**

Die Relationen des Informationsgehaltes über das perzeptive wie auch das chronologische Alter sind über die verschiedenen Wahrnehmungserhebungsverfahren hinweg sehr konstant. Größere Unterschiede bestehen diesbezüglich zwischen perzeptivem und chronologischem Alter.

Noch größere Unterschiede des Informationsgehaltes finden sich zwischen den verschiedenen Äußerungstypen. Der größte davon zwischen Rede- und Vokaläußerungen. Das kann nur so interpretiert werden, dass nicht nur stimmliche Merk-

male Information über das Sprecheralter bereitstellen, sondern auch Merkmale der Sprechweise.

Der perceptiven Analyse ist zu entnehmen, dass innerhalb der Vokale die Faktoren Vokalqualität und Vokalabschnitt hinsichtlich der Informationsmenge in etwa gleich große Effekte erzeugen. Die meiste Information über die Alterswirkung ist in /a/s enkodiert, die meiste Information über das chronologische Alter in /i/s. /u/s sind bzgl. beider Altersskalen die informationsärmste Äußerungsart. Vergleicht man Anfangs- Mittel- und Endabschnitte gehaltener Vokale, so sind die quasistationären Mittelstücke informationsärmer als Anfänge und Enden. Dies kann nur damit begründet werden, dass im Einschwing- bzw. Ausschwingvorgang jeweils eigene Informationsanteile enkodiert sind. Zwischen den Sprechstilen (spontan vs. gelesen) besteht kaum ein Unterschied hinsichtlich der Informationsmenge.

Die akustische Analyse offenbart jedoch, dass sich auch die Redeäußerungen hinsichtlich der Art und Weise *wie* sie nahezu gleich viel Information bereitstellen, erheblich unterscheiden. Bei den Vokaläußerungen hingegen scheint das Bild der wie-Frage etwas einfacher als das der wieviel-Frage: nur die Qualität scheint hier ein Faktor der Informationsart zu sein.

## 9.2 Anwendungsgebiete, Anwendungsmöglichkeiten

Wie schon eingangs erwähnt, ist diese Dissertation vornehmlich ausgelegt, die Grundlagen der Alterserkennung zu erforschen. Dennoch ergeben sich direkte Anwendungsmöglichkeiten der hier gewonnenen Erkenntnisse. Die wichtigsten werden im Folgenden kurz erläutert.

### 9.2.1 Reliabilitätsstudien

Für zukünftige Reliabilitätsstudien dürften die hier vorgestellten, um die Möglichkeit zur Berechnung an Matrizen mit fehlenden Werten erweiterten Reliabilitätsanalyseverfahren (der Intra-Klassen-Korrelation und der Konkordanzanalyse von Kendall) eine Erleichterung sein. Diese Verfahren können natürlich auf Daten aller Art angewendet werden. Das Problem der Analyse, das durch fehlende Werte entsteht, ist aber wohl vermehrt in den Humanwissenschaften (Psychologie, Medizin, Soziologie) aber auch in der Biologie anzutreffen.

Die hier vorgestellten Verfahren sind ferner auch – leicht abgewandelt – zur urteilerinternen Konsistenzanalyse geeignet.

Für die Berechnung der beiden Verfahren ist es unerheblich, wie viele Werte fehlen – dennoch ist es immer, auch mit diesen Verfahren, zur korrekten Interpretation der Ergebnisse günstiger, wenn es weniger sind. Eigentlich ist, und das gilt für alle inferenz-statistischen Verfahren, zu fordern, dass die Stichproben probabilistisch sind. Bei Anwendung dieser Verfahren wäre demzufolge ferner zu fordern, dass auch die Teilstichproben an Objekten, die einzelne Urteiler beurteilen, probabilistischer Natur sind.

Da die Intra-Klassen-Korrelation auf der (ein- bzw. zweifaktoriellen) Varianzanalyse basiert und auch Kendalls W mit der Rangvarianzanalyse nach Friedman sehr verwandt ist, besteht wahrscheinlich die Möglichkeit auch diese Verfahren mit Hilfe der hier eingesetzten neuen Teil-Algorithmen auf die Anwendung an Matrizen mit fehlenden Werten zu erweitern.

### 9.2.2 Forensische Sprechererkennung

Für den forensischen Sprechervergleich ist es nötig, auf sogenannte Hintergrundstatistiken zurückgreifen zu können, vor denen Tat- und Vergleichsmaterial evaluiert werden können. Weibliche Täter sind zwar wohl eher die Ausnahme, besonders in der Praxis der Sprechererkennung, doch falls weibliche Stimmen zu beurteilen sind, können die Daten dieser Arbeit zu einer besseren Beurteilung des Alters beitragen.

Darüber hinaus lassen sich wohl auch die hier erwähnten Signalanalysemethoden auf männliche Sprecher anwenden. Dazu müssen aber zuerst Modelle (wie hier für Frauen geschehen) erstellt werden. Problematisch dürften dabei aber die (teilweise nicht nachvollziehbaren) Aufnahmebedingungen der forensisch zu bewertenden Materialien sein.

In jedem Fall verwertbar ist das Ergebnis der Perzeptionsanalyse, demzufolge eine Hörergruppe die besten (derzeit) erhältlichen Altersschätzungen erzielt.

### 9.2.3 Automatische Sprechererkennung bzw. -Klassifizierung

Sofern im Telekommunikationsbereich eine Verbesserung der „Kundenerkennung“, beispielsweise zur Identifizierung eines Bankkunden über Telefon, erwünscht ist, kann auf die hier gewonnen Erkenntnisse zurückgegriffen werden. Alter ist einerseits ein relativ schwer absichtlich zu änderndes Merkmal. Andererseits sollten solche Systeme aber auch auf die zwangsläufige allmähliche Veränderung der Stimme eingestellt sein.

Dass generell seitens der Industrie die Klassifikation des Alters anhand von Stimme und Sprechweise als Desiderat der Forschung gesehen wird, mag beispielsweise der „Innovationspreis der Voice-Days 2007“ belegen, den ein System

gewann, das Alterserkennung in vier Kategorien über die gesamte Lebensspanne hinweg leistet. Die Erkennungsleistung dieses Systems war weit geringer als die hier erzielte.

### 9.2.4 Automatische Spracherkennung

Zur automatischen Spracherkennung sind die hier gewonnenen Erkenntnisse dann einsetzbar, wenn nicht (wie bei momentanen „Diktiersystemen“ üblich) die Rede eines einzigen Sprechers erfasst werden soll, sondern die mehrerer unterschiedlich alter Sprecher(-innen), bzgl. deren Stimmen bzw. Sprechweisen das System vorab nicht trainiert werden konnte.

### 9.2.5 Sprechwissenschaftliche, logopädische und medizinische Anwendungsbereiche

Durch die hier entwickelten Algorithmen zur Extraktion stimmqualitativer Maße ist mit PRAAT eine eingehendere bzw. verbesserte Stimmanalyse möglich.

Hauptsächlich die Tremormäße und die Maße der spektralen Energieverteilung, aber auch die Algorithmen zur Perturbationsmessung an Redeäußerungen erweitern die Funktion von PRAAT.

Zur medizinischen Diagnose einsetzbar ist vor allem der Tremor-Algorithmus. (Stimm-) Tremor kann viele Ursachen haben (beispielsweise Parkinson, Alzheimer, Depressionen, Angsterkrankungen, usw.). Diese weisen unterschiedliche Tremorfrequenzen auf, die mit dem vorgestellten Algorithmus relativ genau bestimmt werden können.

## 9.3 Ausblick auf weitere Arbeiten

Will man die hier erarbeiteten akustischen Parameter bzw. die Regressionsmodelle, die auf diesen Parametern gründen, nutzen, um verlässliche Vorhersagen zu erlangen, so ist eine Normierung dieser Modelle an einer großen und möglichst repräsentativen Stichprobe nötig.

Davon abgesehen hat die Forschung nach akustischen Merkmalen des Sprecheralters bislang aber noch nicht das Stadium erreicht, in dem es möglich wäre, hinreichend verlässliche Maße des Alters für die (meisten profitorientierten) technologischen Anwendungen bereitzustellen.

Obwohl es eine Menge akustischer Parameter gibt, von denen man weiß, dass sie durch den Prozess des Alterns beeinflusst sind, konnte bislang kein einziger Parameter gefunden werden, der nur durch das Alter beeinflusst wäre, oder wenigstens dessen hauptsächlichster Varianzfaktor das Altern wäre – tatsächlich aber gibt

es sehr wahrscheinlich nicht ein einziges noch so kleines Subsystem des Sprechapparates oder ein anderes Subsystem des menschlichen Organismus, das nicht durch das Altern beeinflusst wäre.

Denn es scheint so zu sein, dass alle akustischen Parameter, die mit dem Sprecheralter variieren, auch durch *viele* andere Faktoren beeinflusst werden oder wenigstens beeinflussbar sind – vor allem durch linguistische Modulation, das Geschlecht, die Persönlichkeit, den Gesundheitszustand, die Sprechsituation, den emotionalen Zustand, ... Also scheint es schwierig, wenn nicht gar unmöglich, ausschließlich das Sprecheralter akustisch zu bestimmen, ohne einen Großteil der anderen zwingend vorhandenen Einflussfaktoren konstant zu lassen oder zu kontrollieren. Um wesentlich über dieses momentan erreichte Stadium der Forschung hinaus zu gelangen, erscheint es also notwendig, diese anderen Varianzfaktoren mündlicher Kommunikation besser erfassen zu können. Der wichtigste Schritt dazu erscheint eine sprecher-, mikrofon-, raum- sowie dialekt- sprach- (...) *unabhängige* Erkennung des linguistischen Gehalts lautsprachlicher Äußerungen zu sein. Also ein Algorithmus der (alle) Phone erkennen, abgrenzen und benennen kann – ohne weiteren Input als das Sprachsignal zu benötigen. Ob dies möglich ist, ohne die Erfassung der Sprecherpersönlichkeit, der Sprache, des Dialekts, des Raums, der Mikrofoncharakteristiken, und vor allem des linguistischen Gehalts bleibt zukünftiger Forschung vorbehalten, erscheint aber unwahrscheinlich.

Wahrscheinlicher ist, da all diese Information in gesprochener Sprache enkodiert ist, dass sie auch (größtenteils) dekodiert werden *muss*, auch wenn nur ein Teil der Information erfasst werden soll.

## 9.4 Zusammenfassung

Diese Arbeit beschäftigt sich damit, wie sich die Stimme und Sprechweise von Frauen im Erwachsenenalter verändert.

Dazu wurden phonetische und akustische Analysen sowie Wahrnehmungsexperimente mit verschiedenen Äußerungstypen an zwei sich ergänzenden methodischen Settings (Längs- und Querschnitt) durchgeführt und durch teilweise neu entwickelte statistische Verfahren ausgewertet.

Die wichtigsten Ergebnisse dieser Arbeit sind:

- Die Stimme älterer Sprecherinnen ist zittriger, tiefer, rauer und behauchter. Sie sprechen (beim Vorlesen) langsamer und/oder weniger genau artikuliert und machen mehr Sprechpausen.
- Unterschiedliche Äußerungstypen vermitteln ein unterschiedliches Ausmaß an Information über das Sprecheralter und diese Information wird in un-

terschiedlichen Äußerungstypen durch andere akustische Merkmale bzw. durch unterschiedliche Merkmalsausprägungen enkodiert.

- Die (bislang) bestmögliche Schätzung des chronologischen Alters einer Sprecherin erhält man über die Zusammenfassung der Schätzungen einer Hörergruppe.
- Dabei sollte beachtet werden, dass einige Hörer dazu neigen, keine hohen Schätzungen abzugeben, dass ältere Sprecherinnen unter Umständen (viel) jünger klingen als sie sind und dass beides dazu führt, dass das chronologische Alter generell unterschätzt wird.
- Der an Stimme und Sprechweise eben merkbare Unterschied zwischen verschiedenen alten Sprecherinnen im Erwachsenenalter liegt bei ca. 5 Jahren.
- Die Genauigkeit der Abschätzung des chronologischen und perceptiven Sprecheralters durch akustische Messungen konnte wesentlich erhöht werden und erreicht annähernd die Genauigkeit zusammengefasster Schätzungen einer größeren Hörergruppe.
- Dies wurde durch den Einsatz von akustischen Maßen möglich, die bisher (so) nicht erhoben wurden und bisher bekannten vergleichbaren Algorithmen (wenigstens bzgl. der Altersindikation) überlegen sind. Hervorzuheben sind dabei
  - die Tremormaße,
  - die Maße der spektralen Energieverteilung,
  - die gezielte Anwendung stimmqualitativer Maße an phonetisch motivierten Segmenten von Redeäußerungen und
  - die um Koartikulationseffekte korrigierte Sprechrate als Maß der Sprechgeschwindigkeit bzw. Artikulationsgenauigkeit.

Diese Algorithmen sind auf der Basis von PRAAT implementiert und kostenfrei zugänglich.

- Zwei statistische Verfahren zur Bestimmung der (Urteiler-) Reliabilität konnten so umformuliert werden, dass sie auch auf unvollständige Urteilmatrizen anwendbar sind.