Zweistufige kontextsensitive Sprecherklassifikation am Beispiel von Alter und Geschlecht

Dissertation
zur Erlangung des Grades
Doktor der Ingenieurwissenschaften (Dr.-Ing.)
der Naturwissenschaftlich-Technischen Fakultät I der Universität des Saarlandes

vorgelegt von

Christian Müller

Saarbrücken,
den 19. Dezember 2005

Datum des Kolloquiums:
13. Januar 2006
Dekan:
Prof. Dr. Jörg Eschmeier
Vorsitzender:
Prof. Dr. Philipp Slusallek
Berichterstatter:
1. Prof. Dr. Dr. h.c. mult. Wolfgang Wahlster
2. Prof. Dr. William Barry
Promovierter akademischer Mitarbeiter der Fakultät:
Dr. Jörg Baus

Erklärung

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet.

Diese Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form in anderen Prüfungsverfahren vorgelegt.

Saarbrücken, den 19. Dezember 2005

Danksagung

Mein besonderer Dank gilt Prof. Wahlster, der mir von Beginn an Freiraum gewährte, wann immer es möglich war, und mich mit Vorgaben unterstützte, wann immer dies nötig war. Ich danke meinen Kollegen für ihre Unterstützung, insbesondere Frank Wittig, von dem ich in der Anfangsphase viel lernte, und Michael Feld, für dessen wertvolle Hilfe besonders in der Schlussphase.

Ganz herzlich danke ich meiner Familie, ohne die ich es sicherlich nicht geschafft hätte, dieses Dokument zu erstellen: Jakob, der mich zum Frühaufsteher machte und meinem Tagesablauf deutlich mehr Struktur verliehen hat. Anne, für ihre moralische Unterstützung, die Diskussionen, ihren Blick von außen und den ständigen Griff nach dem Duden. Meinen Eltern, für die vielen Tage, an denen sie Jakob betreut haben, und so für mich eine ideale Umgebung zum Schreiben schufen.

Christian Müller im Dezember 2005

Kurzzusammenfassung

In der vorliegenden Dissertation wird ein zweistufiger Ansatz zur Sprecherklassifikation am Beispiel Alter und Geschlecht vorgestellt. Dazu werden zunächst die Ergebnisse umfangreicher Korpusanalysen präsentiert, die als Referenzbasis humanwissenschaftlicher Studien geeignet sind. Es wird gezeigt, dass die Modelle, die mithilfe dieser Daten trainiert wurden, in der Lage sind, die genannten Sprechereigenschaften mit einer Genauigkeit zu erkennen, die teilweise das Fünffache des jeweiligen Zufallsniveaus beträgt. Darüber hinaus zeichnet sich der vorgestellte Ansatz vor allen Dingen durch die so genannte Zweite Ebene aus, auf der mithilfe von Dynamischen Bayes'schen Netzen eine Fusion multipler Klassifikationsergebnisse unter Berücksichtigung des auditiven Kontextes erfolgt. In der Arbeit wird außerdem ein konkretes Sprecherklassifikationssystem beschrieben, welches für das Anwendungsszenario von mobilen, sprachbasierten Dialogsystemen entwickelt worden ist.

Short Abstract

This dissertation describes a two-layered speaker classification approach on the example of age and gender. First of all, the results of comprehensive corpus analyses are presented that are suitable to serve as a reference basis for further studies in human sciences. It is showed, that the models which are trained using these data are able to recognize the above mentioned characteristics with an accuracy that is up to five times better than the respective chance level. In addition, the presented approach distinguishes itself by the so called Second Layer, on which a context sensitive fusion of multiple classification results is accomplished using Dynamic Bayesian Networks. The dissertation also describes a concrete speaker classification system which was developed for the application scenario of mobile spoken dialog systems.

Zusammenfassung

Der Hauptuntersuchungsgegenstand der vorliegenden Dissertation ist die Entwicklung eines Verfahrens zur Nutzung der in der Sprache enthaltenen paralinguistischen Informationen, um das Sprecheralter und -geschlecht einzuschätzen. Die technischen Rahmenbedingungen werden durch ein mobiles natürlichsprachliches Dialogsystem gegeben, welches durch die Integration eines solchen Verfahrens zu einem nicht-intrusiven Aufbau eines Benutzermodells befähigt wird. Aufgrund der im Vordergrund stehenden Sprechercharakteristika wurde das Verfahren unter der Projektbezeichnung AGENDER (abgeleitet von den englischen Begriffen *age* für Alter und *gender* für Geschlecht) entwickelt. Es handelt sich um einen Teil des Projektes *m3i* (Mobile Multi-Modal Interaction), welches wiederum zu dem vom Bundesministerium für Bildung und Forschung geförderten Projekt COLLATE (Computational Linguistics and Language Technology for Real Life Applications) gehört, das an der Universität des Saarlandes und dem Deutschen Forschungszentrum für Künstliche Intelligenz (DFKI) ausgeführt wurde.

Zu den Ziel-Applikationen von AGENDER gehören mobile Shopping-Assistenten und Fußgängernavigationssysteme. Ein besonderer Aspekt dieser Applikation ist die mobile und multi-modale Interaktion, die in Form von Gesten, Sprache, Handschrift und einer Kombination aus diesen bestehen kann. Auf Basis des Benutzermodells, das von AGENDER zur Verfügung gestellt wird, kann der Einkaufsassistent eine spezifische Auswahl von Produkten treffen – im Fall von Digitalkameras kann er beispielsweise, wenn ein Sprecher als weiblich erkannt worden ist, zunächst ein Modell präsentieren, das vom Hersteller speziell für Frauen entwickelt worden ist. Das Navigationssystem kann die Auswahl von alternativen Routen anpassen: Wenn z. B. erkannt worden ist, dass es sich bei dem Sprecher um ein Kind handelt, kann im Fall eines Touristenführers eine Tour durch die Innenstadt mit Sehenswürdigkeiten speziell für Kinder ausgewählt werden.

Aufgrund des regen Interesses an der AGENDER-Technologie von Seiten der Telekommunikationsindustrie ist ein weiterer Anwendungsbereich in den Fokus gerückt, nämlich telefonbasierte Dienste. Ein Callcenter, wie z. B. eine Bestell- oder Servicehotline, ist für den Betreiber mit hohen Kosten verbunden – entsprechend groß ist das Interesse der Telekommunikationsindustrie als Plattformanbieterin an Lösungen zur Effizienzsteigerung. Analog zu Emotionserkennern, kann die AGENDER-Technologie in die so genannte *Automatic Call Distribution* (ACD) integriert werden. Zu den telefonbasierten Diensten gehören auch die so genannten *Spoken Dialog Systems* (SDS), die

sich von den Callcenter-Diensten dadurch unterscheiden, dass sie nicht von menschlichen Agenten, sondern von sprachverstehenden Computersystemen geleistet werden. Neben dem bekannten Beispiel der Fahrplanauskunft der Bahn finden weitere Produktinformations- und Einkaufssysteme zunehmend Verbreitung. Die Dimension, auf die sich in diesem Fall die Verbesserungsbestrebungen richten, ist weniger die Kostenverringerung als eher die Steigerung der Kundenzufriedenheit. Auf Basis des Sprechermodells soll eine kundengruppengerichtete Produktauswahl getroffen und gleichzeitig das Dialogverhalten des Systems angepasst werden.

Der AGENDER-Ansatz zur Sprecherklassifikation stellt eine Kombination aus datengesteuerten Aspekten und wissensbasierten Aspekten dar. Die Modelle wurden auf der Grundlage von Daten erzeugt, die aus umfangreichen Korpusanalysen stammen. Zu den untersuchten Merkmalen gehören Charakteristika der Stimme wie mittlere Grundfrequenz, Jitter und Shimmer und Charakteristika des Sprechverhaltens wie Artikulationsgeschwindigkeit und Anzahl und Dauer der Sprechpausen. Die Definition der Altersklassen ist wie folgt: Als KINDER werden Sprecher bis einschließlich 12 Jahren bezeichnet. Die Klasse JUGENDLICHE umfasst Sprecher von 13 bis einschließlich 19 Jahren. Sprecher zwischen 20 und einschließlich 64 Jahren werden als (jüngere) ERWACHSENE bezeichnet. Ab 65 Jahren gehören die Sprecher der Klasse SENIOREN an. Zusammen mit der zusätzlichen Differenzierung nach dem Geschlecht besteht das Klassifikationsproblem demnach aus acht Klassen.

Diejenigen Phasen der Mustererkennung, welche die Merkmalsextraktion und die Klassifikation betreffen, werden in AGENDER *Erste Ebene* genannt. Bezüglich der Klassifikation wurden die folgenden bekannten Methoden des maschinellen Lernens untersucht: 1. Naive Bayes (NB), 2. Gaussian-Mixture-Models (GMM), 3. k-Nearest-Neighbor (KNN), 4. C 4.5 Entscheidungsbäume (C 4.5), 5. Support-Vector-Machines (SVM) und 6. Künstliche Neuronale Netze (Artificial Neural Networks ANN). Die Ergebnisse, die erzielt werden konnten, sind insgesamt vielversprechend: Die Klassifikationsgenauigkeiten sämtlicher getesteter Verfahren liegen deutlich über dem Zufallsniveau. Mithilfe der Methode ANN konnte beispielsweise für das Acht-Klassen-Problem eine Gesamtgenauigkeit von 63.5 % erreicht werden, was dem Fünffachen des Zufallsniveaus entspricht. Bei der reinen Geschlechtsklassifikation wurde mit demselben Verfahren eine Genauigkeit von 93.14 % erreicht (1.8-faches des Zufallsniveaus).

Abgesehen von diesem positiven Resultat zeichnet sich der AGENDER-Sprecherklassifikationsansatz durch die so genannte *Zweite Ebene* aus, die auf Dynamischen Bayes'schen Netzen (DBNs) basiert. Anhand von Beispielen wurde gezeigt, wie diese genutzt werden können, um erstens die klassifikationsinhärente Unsicherheit explizit zu modellieren, zweitens um Top-Down-Wissen in den Entscheidungsprozess einfließen zu lassen – wie z. B. die Tatsache, dass je nach Kontext die Resultate bestimmter Klassifizierer als zuverlässiger eingeschätzt werden sollten als die anderer – und drittens um eine Fusion multipler Klassifikationsergebnisse bezüglich einer Äußerung (statische Fusion) oder mehrerer aufeinander folgender Äußerungen (dynamische Fusion) erreichen zu können.

Abstract

This dissertation describes an approach on how to exploit the so called paralinguistic information contained in the speech for estimating the speakers' age and gender. The technical constraints are given by mobile speech-based dialog systems, which shall acquire user models in a non-intrusive fashion. According to the main speaker characteristics *age* and *gender*, the approach was developed under name AGENDER as a part of the project *m3i* (Mobile Multi-Modal Interaction). It belongs to the project COLLATE (Computational Linguistics and Language Technology for Real Life Applications), which is sponsored by the German Federal Ministry for Research and Education (BmBf) and is being carried out by the German Research Center for Artificial Intelligence (DFKI) and the University of the Saarland.

The target applications of AGENDER involve mobile shopping as well as pedestrian navigation systems. A central issue of these applications is the interaction style, which is multimodal i.e. consists of gestures, speech, handwriting, and a combination of these. On the basis of the user models that are provided by AGENDER, the shopping assistant is able to make a specific selection of products (e.g. digital cameras): When the speaker is recognized as being a female person, the system can e.g. choose a camera that is especially designed for women. Analogously, the navigation system can adapt the selection of alternative routes: When the user is a child, a tourist guide could choose sights especially interesting for kids.

The vital interest in the AGENDER technology from telecommunications industry yielded another application domain, namely telephone-based services. Since call centers like that used for shopping or service hotlines cause high costs for the provider, the telecommunications industry is interested in solutions to improve the efficiency. Analogously to emotion recognizers, the AGENDER technology can be integrated into *Automatic Call Distribution* (ACD) systems. Besides this, telephone-based services involve spoken dialog systems (SDSs). Unlike call centers, SDSs are provided by speech-enabled computers instead of human agents. Railroad information systems currently belong to the best-known examples of such services but more and more product information and shopping systems are emerging in the market. The ambitions for improving SDSs through the AGENDER technology are related to the increase of consumer satisfaction: On the basis of the speaker model, the product shall be selected to meet the requirements of the respective customer group. At the same time, the dialog shall be adapted accordingly.

The AGENDER approach on speaker classification represents a combination of data-driven and knowledge-based aspects. The models are built on the basis of data stemming from extensive corpus analyses. The set of features for estimating the speakers' age and gender comprise characteristics of the voice like pitch, jitter, and shimmer as well as characteristics of the speaking behavior like articulation rate and number and duration of speech pauses. The age classes are defined as follows: The class CHILDREN represents speakers up to and including an age of 12 years. The class TEENAGER encompasses speakers between 13 and 19 years. Speakers between 20 and 64 years belong to the class (younger) ADULTS. The class of SENIORS begins with 65 years. In conjunction with the gender, the classification task consists of a total of eight classes.

In AGENDER, the phases of pattern recognition, which concern the feature extraction and classification, are called the *first layer*. With respect to the classification, the following well known machine learning methods have been investigated: 1. Naive Bayes (NB), 2. Gaussian-Mixture-Models (GMM), 3. k-Nearest-Neighbor (KNN), 4. C 4.5 Decision Trees (C45), 5. Support-Vector-Machines (SVM) and 6. Artificial Neural Networks, (ANN). The results are very promising: The classification accuracy of all methods in the test were significantly above the chance level. Regarding the method ANN e.g. an overall accuracy of 64.5 % for the eight-class problem was obtained. This is five times better than the chance level. With respect to a pure gender estimation an accuracy of 93.14 % (1.8 times chance level) was achieved using the same method.

Despite of these positive results the AGENDER speaker classification approach distinguishes itself by means of a special post processing technique, the so called *second layer*: Multiple post processing problems are solved with one single mechanism, namely Dynamic Bayesian Networks (DBNs). Examples are provided on how DBNs can be used for: 1. explicitely modeling the classification-inherent uncertainty; 2. incorporating top down knowledge into the decision making process, like e.g. the fact that depending on the context, certain classifiers are more reliable than others; 3. fusing the results of multiple classifiers with respect to one utterance (static fusion) as well as several consecutive utterances (dynamic fusion).

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

1.1 Motivation

Die maschinelle Verarbeitung natürlicher Sprache geht zurück auf die 1950er Jahre. Zunächst in den USA, ab den 80er Jahren verstärkt auch in Deutschland, konnte sich die aus der Sprachwissenschaft und der Informatik entstandene Computerlinguistik als eigenständiger Fachbereich etablieren (vgl. Carstensen et al., 2004). Innerhalb des praxisorientierten Bereichs dieser Wissenschaft werden Formalismen entwickelt, die dazu genutzt werden können, natürliche Sprache auf unterschiedlichen Ebenen der Beschreibung zu modellieren. Auf der lautsprachlichen Ebene beispielsweise überführen automatische Spracherkenner (*automatic speech recognizer*, ASR) den vom Mikrofon aufgezeichneten Sprachschall in eine Folge von Phonemen, die dann durch Abgleich mit einem Lexikon auf Wörter abgebildet wird. In den heutigen Systemen kommen in beiden Schritten üblicherweise stochastische Verfahren zum Einsatz. Ein prototypisches Beispiel für die Anwendung tiefer linguistischer Analyse ist *Verbmobil* (Wahlster, 2000b), ein System zur Übersetzung spontan gesprochener Äußerungen in drei Sprachen (Deutsch, Englisch und Japanisch) aus dem Bereich der Terminvereinbarung. Dabei gilt es zunächst, die gesprochene Sprache zu erkennen (ASR), und die vom System generierten Übersetzungen auszugeben (*text-to-speech*, TTS). Die Eingabeäußerungen müssen jedoch darüber hinaus auf syntaktischer, semantischer und pragmatischer Ebene analysiert werden, d. h. ihre Struktur muss geparst[1] und die damit verbundene Bedeutung in der Quellsprache ermittelt und einer Funktion im Dialogkontext zugeordnet werden. Zur Realisierung der äquivalenten Funktion in der Zielsprache werden sämtliche Ebenen erneut von unten nach oben durchlaufen.

In der zwischenmenschlichen Kommunikation transportiert der Sprachschall nicht ausschließlich die Bedeutung einer Äußerung, sondern darüber hinaus auch so genannte *paralinguistische* Informationen, die z. B. Rückschlüsse auf Sprechercharakteristika zulassen. Es entspricht unserer alltäglichen Erfahrung, dass wir fremde Menschen, mit denen wir telefonieren, anhand ihrer Stimme charakterisieren können: Wir können in den allermeisten Fällen das Geschlecht einschätzen, das ungefähre Alter und, ob der Anrufer gut gelaunt oder genervt, aufgeregt oder gelassen, ängstlich oder sicher ist – und das unabhängig davon, was gesagt wird. Die

[1] von engl. *to parse* = analysieren. Der Begriff *parsing* bezieht sich im Kontext der Sprachverarbeitung auf die Analyse der syntaktischen Struktur einer komplexen sprachlichen Einheit, wie einem Satz oder einer Phrase.

Mitarbeiter von Callcentern werden häufig speziell ausgebildet, um den Gemütszustand des Kunden richtig einzuschätzen und angemessen darauf zu reagieren. Genauso sollten die Mitarbeiter einer Notrufzentrale anhand der Stimme und der Sprechweise eines Anrufers bemerken, ob er panisch ist oder gar unter Schock steht, und dies bei der Aufnahme des Vorfalls berücksichtigen.

Systeme zu entwickeln, die ihr (Dialog-)Verhalten an die Bedürfnisse des Benutzers anpassen, ist Gegenstand der *Benutzermodellierung*, die dem Bereich der Künstlichen Intelligenz angehört und deren Tradition bis in die 70er Jahre zurückreicht (vgl. z. B. Hahn, Henskes, Hoeppner und Wahlster, 1975). Die Problemfelder der heutigen Benutzermodellierung sind im Wesentlichen die folgenden: 1) Wie kann das Benutzermodell auf möglichst *nicht-intrusive* Art und Weise *akquiriert* werden? 2) Welches sind die geeigneten *Strategien* für eine angemessene *Adaption* des Systems auf Basis des Benutzermodells? Beide Probleme sind größtenteils domänenabhängig, auch wenn einige allgemeingültige Prinzipien zur Akquisition und Adaption formuliert werden können (vgl. Kobsa und Wahlster, 1989). Unter nicht-intrusiven Akquisitionsmethoden versteht man diejenigen, die keine zusätzlichen Dialogschritte verursachen und die den Benutzer auch ansonsten nicht bei der Interaktion mit dem System stören.

Die Benutzermodellierung gewinnt immer mehr an Bedeutung, da die Anwendung von Computersystemen sich mit fortschreitender Entwicklung der Geräte längst vom Schreibtisch gelöst und Einzug in viele Lebensbereiche gefunden hat. Der Begriff „*mobile and ubiquitous*“ bezeichnet diesen Paradigmawechsel, der sich dahingehend auswirkt, dass immer mehr Menschen Gebrauch von mobilen Geräten wie PDAs (Personal Digital Assistants) oder *Smartphones* machen. Die besonderen Anforderungen an die Systeme ändern sich mit den unterschiedlichen Situationen, in denen sie benutzt werden. Ein mobiles Fußgängernavigationssystem z. B. sollte den Umstand berücksichtigen, dass der Benutzer sich möglicherweise an einer lauten Straßenkreuzung in der Innenstadt befindet und einen Großteil seiner Aufmerksamkeit auf seine Umgebung richten muss, während er ein anderes Mal auf einer ruhigen Parkbank sitzt und sich voll dem Dialog mit dem System widmen kann (vgl. Müller, 2002).

Wie am Beispiel des Autonavigationssystems deutlich wird, ist in Situationen, bei denen das System möglichst wenig Aufmerksamkeit konsumieren sollte, und darüber hinaus die Hände des Benutzers nicht immer frei sind, die Sprache eine geeignete Interaktionsmodalität. Tatsächlich ist die Weiterentwicklung der Sprachtechnologie auf dem ASR-Sektor soweit gediehen, dass Spracherkenner auf mobilen Geräten genutzt werden können. Wasinger, Stahl und Krüger (2003) beschreiben z. B. ein Fußgängernavigationssystem, bei dem der Benutzer gesprochene Anfragen stellen kann in der Form: „Wie komme ich von hier zur Mensa?“[2]. Die Spracherkennung basiert bei diesem System auf dem von IBM entwickelten *Embedded Viavoice*.

Es liegt nahe zu untersuchen, ob die Sprache als Informationsquelle genutzt werden kann, um Rückschlüsse über Charakteristika des Sprechers ziehen zu können. Müller et al. (2001) beschreiben beispielsweise ein Experiment zur Identifikation von Merkmalen der Sprache, auf deren Basis die kognitive Belastung des Sprechers eingeschätzt werden kann (vgl. auch Jameson et al., 2005).

[2] Die prototypische Implementierung basiert auf einer Modellierung des Campus der Universität des Saarlandes.

Das zugrunde liegende Szenario ist das folgende: Ein mobiles Assistenzsystem soll einen Reisenden auf seinem Weg durch einen Großflughafen begleiten. Es ist zu erwarten, dass der Reisende in erhöhtem Maße kognitiv belastet ist, weil er sich die Gate-Nummer und die Boarding-Zeit merken muss und weil eine Fülle an Informationen am Flughafen auf ihn einwirkt. Außerdem steht er möglicherweise unter Zeitdruck, weil er in der kurzen Zeit bis zum Abflug nicht nur das Gate finden, sondern auf dem Weg dahin noch ein Geschenk kaufen möchte. Das System soll diese Belastung erkennen und bei der Erzeugung von Weghinweisen berücksichtigen.

Das Experiment, das Müller et al. beschreiben, simuliert die Situation am Flughafen durch ein komplexes Doppelaufgabenexperiment, bei dem die Probanden künstlich unter kognitive Belastung und Zeitdruck gesetzt werden und dabei Anfragen an ein Assistenzsystem stellen, wie: „Ich muss noch mein Baby wickeln – wie komme ich zum nächsten Wickelraum?“ Als Ergebnis konnte ein Liste von Merkmalen herausgestellt werden, anhand derer sich die Belastung in der Sprache manifestiert: langsamere Artikulationsgeschwindigkeit, mehr und längere Sprechpausen und insbesondere ein erhöhtes Vorkommen so genannter *Disfluenzen*. Darunter versteht man Elemente, die den flüssigen Ablauf einer Äußerung stören, wie z. B. Selbstkorrekturen, Wiederholungen, Satzabbrüche oder Fehlansätze.

Obwohl die Ergebnisse statistisch signifikant sind, konnten sie aufgrund der vergleichsweise geringen Menge von zugrunde liegenden Daten nicht unmittelbar zur Bildung von Modellen zur Erkennung von kognitiver Belastung und Zeitdruck herangezogen werden. Als problematisch hat sich auch das Vorhaben erwiesen, sprachliche Disfluenzen automatisch zu erkennen, da hierfür in jedem Fall eine Interpretation der Äußerung notwendig ist, und gerade diese Disfluenzen zum Scheitern der Sprachverarbeitung führen.

Der Hauptuntersuchungsgegenstand der vorliegenden Arbeit ist daher die Entwicklung eines Verfahrens zur Nutzung der in der Sprache enthaltenen paralinguistischen Informationen, um den Sprecher zu charakterisieren. Der Vorteil gegenüber der Betrachtung linguistischer Merkmale ist der, dass die automatische Extraktion wesentlich weniger Kosten – im Sinne von komputationellen Ressourcen und Zeit – in Anspruch nimmt, vor allem deshalb, weil keine Interpretation der Äußerung notwendig ist. Als Sprechercharakteristika werden jedoch nicht kognitive Belastung und Zeitdruck betrachtet, sondern Alter und Geschlecht. Die technischen Rahmenbedingungen werden durch ein mobiles natürlichsprachliches Dialogsystem gegeben, welches durch die Integration eines solchen Verfahrens zu einem nicht-intrusiven Aufbau eines Benutzermodells befähigt wird. Da bei einem solchen System die (Nutzungs-)Umgebung eine wichtige Rolle spielt, wird bei der Entwicklung des Verfahrens der so genannte *situative Kontext* ebenfalls berücksichtigt. Im Folgenden wird daher von *Sprecher-* und *Kontextklassifikation* gesprochen.

Aufgrund der im Vordergrund stehenden Sprechercharakteristika wurde das Verfahren unter der Projektbezeichnung AGENDER[3] entwickelt. Es handelt sich um einen Teil des Projektes *m3i* (Mobile Multi-Modal Interaction), welches wiederum zu dem vom Bundesministerium für Bildung und Forschung geförderten Projekt COLLATE (Computational Linguistics and Language Technology for Real Life Applications) gehört, das an der Universität des Saarlandes und dem Deutschen Forschungszentrum für Künstliche Intelligenz (DFKI) ausgeführt wurde.

[3] Abgeleitet von *age* (Alter) und *gender* (Geschlecht).

1.2 Alter und Geschlecht als Charakteristika für die Sprecherklassifikation

Im Kontext mobiler Dialogsysteme ist eine nicht-intrusive Akquisition des Benutzermodells insbesondere dann von Bedeutung, wenn es sich um Merkmale handelt, die sich innerhalb eines kurzen Zeitraums verändern, wie z. B. der kognitive oder emotionale Zustand. In Anbetracht der zunehmenden Einbettung intelligenter, sprachbasierter Systeme in Gebäude, Infrastrukturkomponenten und Nutzgegenstände, die für eine Vielzahl von Personen zugänglich sind, ist auch im Fall von statischen Variablen wie Alter und Geschlecht die automatische Erkennung einer manuellen Eingabe durch den Benutzer vorzuziehen. Die Besucher eines Museums werden beispielsweise nicht gewillt sein, vor Beginn ihres Rundgangs zunächst Profile auf einem elektronischen Musemsführer anzulegen – zumal dessen Nutzen in der Regel erst im Laufe der Interaktion deutlich wird. Genauso wenig werden die Fahrer von Mietwagen, die mit adaptiven *Driver Assistance Systemen* ausgestattet sind, Zeit in die Eingabe ihrer Daten investieren wollen, auch wenn dies nur einmalig zu Beginn der Fahrt erforderlich ist.

Das Alter eines Benutzers ist auch deshalb eine wichtige Variable, weil speziell älteren Menschen häufig der Zugang zu technischen Informationsdiensten verwehrt bleibt. Da jedoch die elektronische Interaktion zunehmend zu einem wesentlichen Bestandteil des täglichen Lebens wird, können den Bevölkerungskreisen, die nicht in der Lage sind, diese Technologien für sich zu nutzen, erhebliche Nachteile entstehen (vgl. Wahlster, 2000a). In diesem Zusammenhang wurde das von HEWLETT-PACKARD geförderte „*Voice Web Philantropic Programme*“ initiiert (vgl. Hickey, 2003), dessen Ziel unter anderem ist, mithilfe der Sprachtechnologie älteren und behinderten Menschen die Benutzung elektronischer Systemen zu erleichtern (vgl. Newell, 2003). Die automatische Einschätzung des Sprecheralters sollte einen Beitrag dazu leisten, dass Systeme entwickelt werden können, die sich an die besonderen Bedürfnisse älterer Menschen anpassen können (vgl. Müller et al., 2003). Das Geschlecht wiederum ist eine Variable, die unmittelbar damit verbunden ist: Aufgrund der Tatsache, dass die Stimmen von Männern und Frauen auf sehr unterschiedliche Art und Weise altern, kann das Sprecheralter nur dann sinnvoll analysiert werden, wenn das Geschlecht ebenfalls berücksichtigt wird (vgl. Kapitel 3). Im weiteren Verlauf des Projektes haben sich jedoch weitere Anwendungsbereiche herauskristallisiert, die im folgenden Abschnitt zusammengefasst werden.

1.3 Anwendungen von AGENDER

1.3.1 Adaptive Mobile Systeme

Wie eingangs bereits erwähnt wurde, bilden mobile Systeme, die gesprochene natürlichsprachliche Eingaben erlauben, den technischen Rahmen für die in der vorliegenden Arbeit vorgestellte Version von AGENDER. Bei den beiden Ziel-Applikationen handelt es sich um die im Rahmen des Projektes *m3i* entwickelten Systeme *m3i Mobile ShopAssist* und *m3i Personal Navigator*.

Der Mobile ShopAssist (Wasinger, Krüger und Jacobs, 2005) ist eine Pocket-PC Anwendung,

die dazu dient, die Nutzung von natürlicher Sprache in einer typischen Einkaufsumgebung zu demonstrieren. Ein zentrales Thema dieser Applikation ist die mobile und multi-modale Interaktion, die in Form von Gesten, Sprache, Handschrift und einer Kombination aus diesen bestehen kann (vgl. Beispiel 1.1).

(1.1)

Benutzer hält in einer Hand einen Pocket-PC, auf dessen Display eine digitale Kamera abgebildet wird. Er steht in einem Geschäft vor einem Regal mit Kameras anderer Marken.	
Benutzer:	„Vergleiche diese Kamera <tippt auf das Display> mit dieser Kamera <nimmt eine Schachtel aus dem Regal>."
System:	„Die Canon Powershot hat eine Auflösung von 4 Megapixeln. Die Olympus Camedia hat eine Auflösung von 5 Megapixeln."

Der Personal Navigator (Wasinger et al., 2003) ist bezüglich der Interaktionstechnik eine sehr ähnliche Anwendung: Die Benutzer können durch eine Kombination von Sprache und Gesten Weganfragen stellen oder sich Informationen über nahe liegende Gebäude geben lassen (vgl. Beispiel 1.2).

(1.2)

Benutzer hält in einer Hand den Pocket-PC. Auf dem Display ist eine Karte des Campus der Universität des Saarlandes abgebildet.	
Benutzer:	„Beschreibe dieses <tippt auf das Display> Gebäude."
System:	„In dem Gebäude mit der Nummer 43 befindet sich das Deutsche Forschungszentrum für Künstliche Intelligenz D-F-K-I."

Der *m3i Client*, eine Pocket-PC Version des AGENDER Sprecherklassifikationssystems, wurde für eine Integration in diese beiden Applikationen entwickelt (vgl. Abbildung 1.1). Auf die technischen Details der Kommunikation wird in Kapitel 11.2 genauer eingegangen. Auf Basis des Benutzermodells, das von AGENDER zur Verfügung gestellt wird, kann der Einkaufsassistent eine spezifische Auswahl von Produkten treffen – im Fall von Digitalkameras kann er beispielsweise, wenn ein Sprecher als weiblich erkannt worden ist, zunächst ein Modell präsentieren, dass vom Hersteller speziell für Frauen entwickelt worden ist. Das Navigationssystem kann die Auswahl von alternativen Routen entsprechend anpassen: Wenn z. B. erkannt worden ist, dass es sich bei dem Sprecher um ein Kind handelt, kann im Fall eines Touristenführers eine Tour durch die Innenstadt mit Sehenswürdigkeiten speziell für Kinder ausgewählt werden, die darüber hinaus möglichst wenige gefährliche Kreuzungen enthält.

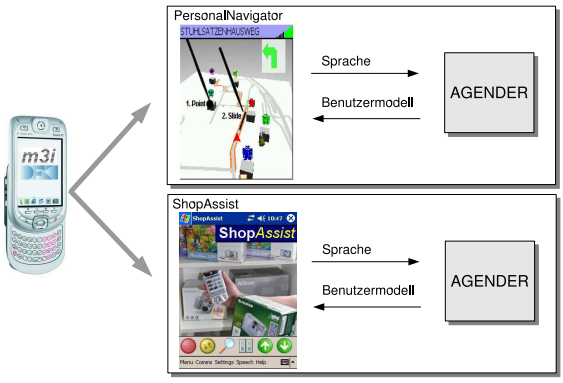


Abbildung 1.1: AGENDER im Anwendungsszenario „Adaptive mobile Systeme“ am Beispiel von *m3i PersonalNavigator* und *m3i ShopAssist*.

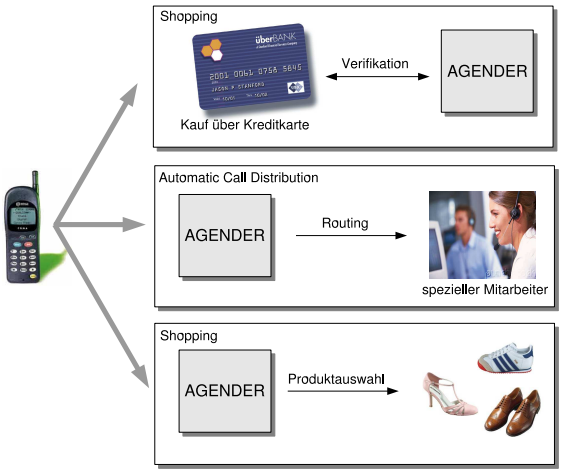


Abbildung 1.2: AGENDER im Anwendungsszenario „Callcenter“ am Beispiel von Kreditkartenverifikation, Service-Hotline und Shopping System.

1.3.2 Telefonbasierte Dienste

Spätestens seit der erfolgreichen Präsentation des Prototypen auf der *CeBIT 2005* erfährt die AGENDER-Technologie ein reges Interesse von Seiten der Telekommunikationsindustrie, wodurch ein völlig anderer Anwendungsbereich in den Fokus gerückt ist, nämlich die telefonbasierten Dienste.

Ein Callcenter, wie z. B. eine Bestell- oder Servicehotline, ist für den Betreiber mit hohen Kosten verbunden – entsprechend groß ist das Interesse der Telekommunikationsindustrie als Plattformanbieterin an Lösungen zur Effizienzsteigerung. Ein zentraler Bestandteil der Callcenter-Technik ist die *Automatic Call Distribution* (ACD), ein computergestütztes System, das Anrufe entgegennimmt und an einzelne Mitarbeiter oder Mitarbeitergruppen verteilt (vgl. Greff und Fojut, 2003). Dabei werden bei einem Verfahren, das *Conditional Routing* genannt wird, Anrufe auf Basis von zuvor festgelegten Regeln geschaltet. Bei den heute üblichen ACD-Systemen betreffen diese Regeln zumeist das Verhältnis von Anrufen und Auslastung. In jüngster Zeit sind jedoch auch Ansätze entwickelt worden, Anrufe aufgrund bestimmter von einem System erkannter Emotionen zu *routen*. Die Grundlage hierfür bilden Emotionserkenner, wie sie z. B. im Rahmen des bereits erwähnten *Verbmobil*-Projektes entwickelt worden sind (vgl. Batliner et al., 2000). Die Möglichkeit der Integration in ein ACD-System besteht auch für die AGENDER-Technologie (vgl. Abbildung 1.2 mitte).

Zu den telefonbasierten Diensten gehören auch die so genannten *Spoken Dialog Systems*, die sich von den Callcenter-Diensten dadurch unterscheiden, dass sie nicht von menschlichen Agenten, sondern von sprachverstehenden Computersystemen geleistet werden. Neben dem bekannten

Beispiel der Fahrplanauskunft der Bahn finden weitere Produktinformations- und Einkaufssysteme zunehmend Verbreitung. Die Dimension, auf die sich in diesem Fall die Verbesserungsbestrebungen richten, ist weniger die Kostenverringerung als eher die Steigerung der Kundenzufriedenheit. Die Anwendung für AGENDER, die sich daraus ergibt, ist der ähnlich, die für die mobilen Dialogsysteme beschrieben wurde: Auf Basis des Sprechermodells wird eine kundengruppengerichtete Produktauswahl getroffen und gleichzeitig das Dialogverhalten des Systems angepasst (vgl. Abbildung 1.2 unten sowie Beispiele 1.3 und 1.4).

(1.3)

Anrufer:	„Welche Handytarife gibt es?“
AGENDER:	Erkennt einen jungen, männlichen Sprecher und gibt diese Information an das System weiter.
System:	„Der XY-Tarif ist genau der richtige für dich. Damit kannst du im Monat 150 Frei-SMS versenden“.

(1.4)

Anrufer:	„Welche Handytarife gibt es?“
AGENDER:	Erkennt einen älteren, männlichen Sprecher und gibt diese Information an das System weiter.
System:	„Wir empfehlen Ihnen den ABC-Tarif. Neben einer geringen Grundgebühr bietet er den Vorteil einer vollen Kostenkontrolle auch im Ausland.“

In Abbildung 1.2 oben wird ein Dienst skizziert, der sowohl automatische als auch von menschlichen Agenten geleistete Anteile aufweist. Es wird der Fall betrachtet, bei dem ein Produkt über ein elektronisches Bestellsystem gekauft und mit Kreditkarte bezahlt wird. Über das Alter und Geschlecht des Kunden ist das System also informiert, da es Zugriff auf die Kreditkartendaten hat. Das Profil, das von AGENDER ermittelt wird, könnte jedoch genutzt werden, um einen möglichen Widerspruch aufzudecken: Wenn z. B. ein weiblicher Sprecher erkannt wurde, die Kreditkarte jedoch einem Mann gehört, könnte das System veranlassen, dass der Vorgang an einen menschlichen Agenten weitergeleitet wird, der die Korrektheit der Angaben überprüft. Dasselbe gilt für den Fall, in dem das eingeschätzte Alter nicht mit dem Alter des Karteninhabers übereinstimmt, insbesondere ein Kind oder ein Jugendlicher erkannt wurde.

Dieses Szenario, bei dem durch die Sprecherklassifikation eine bestehende Hypothese überprüft werden soll, kann auf andere Anwendungen verallgemeinert werden: In der Stimm-Biometrie (vgl. Abschnitt 1.6.1) kann AGENDER beispielsweise als zusätzliches Verifikationsmodul eingesetzt werden.

AGENDER in eine Sprachplattform für telefonbasierte Dienste zu integrieren, ist Ziel eines jüngst gestarteten Projektes in Zusammenarbeit mit der Deutschen Telekom. Dabei werden in erster Linie neue Anforderungen an die Implementierung der Sprecherklassifikation gestellt, vor allem in Hinblick auf die Laufzeit. In Kapitel 13 wird darauf genauer eingegangen.

1.3.3 Verbesserung der Spracherkennung

Um den akustischen Eigenschaften des Sprachschalls eine Folge von Phonemen zuordnen zu können, greifen Spracherkenner auf *Akustikmodelle* zurück, die zuvor mithilfe von Sprachproben trainiert werden. Es wird dabei grundsätzlich zwischen *allgemeinen* und *individuellen* Akustikmodellen unterschieden. Erstere basieren auf den Sprachproben vieler unterschiedlicher Sprecher und werden zumeist zusammen mit dem Spracherkenner ausgeliefert, was den Vorteil hat, dass von Seiten des Benutzers kein Training durchgeführt werden muss. Allerdings ist unter Verwendung von allgemeinen Modellen die Erkennungsleistung, die anhand der so genannten *Word Error Rate* (WER) gemessen wird, also der prozentualen Anzahl der nicht korrekt erkannten Wörter, oftmals geringer als unter Verwendung von individuellen Modellen. Diese werden von Seiten des Benutzers erstellt, in einem Verfahren, das *Speaker Enrollment* genannt wird. Neben dem zusätzlichen Aufwand liegt der Nachteil der individuellen Modelle darin, dass sie unflexibel sind. In Anwendungen mit wechselnden Benutzern ist diese Variante nicht praktikabel, da Äußerungen modellfremder Sprecher schlechter erkannt werden als im Fall von allgemeinen Modellen. Man spricht daher auch von *sprecherabhängiger* Spracherkennung (individuelle Modelle) und *sprecherunabhängiger* Spracherkennung (allgemeine Modelle).

Um das trade-off zwischen Flexibilität und Erkennungsleistung zu optimieren, muss der Frage nachgegangen werden, warum die Fehlerrate bei Verwendung von allgemeinen Modellen so hoch ist. Es ist eine allgemein anerkannte Tatsache, dass die Unterschiede zwischen Frauen- und Männerstimmen dabei eine Rolle spielen. Nach Vergin, Farhat und Shaughnessy (1996) ist bei der sprecherunabhängigen Spracherkennung die Erkennungsleistung für weibliche Sprecher fast immer schlechter als die für männliche Sprecher. Vor dem Hintergrund der deutlichen Unterschiede zwischen den Stimmen von Menschen unterschiedlichen Lebensalters (z. B. Kinder, jüngere Erwachsene, Senioren), welche in Kapitel 3.4 der vorliegenden Arbeit zusammenfassend herausgestellt werden, liegt der Schluss nahe, dass auch die Variable Alter einen nicht unbedeutenden Anteil zur geringeren Erkennungsleistung mit allgemeinen Akustikmodellen leistet.

Wenn jeweils ein *spezifisches* Modell für eine Sprechergruppe erzeugt wird und bei der Erkennung die Mitgliedschaft des aktuellen Sprechers zu einer dieser Gruppen bekannt ist, sollte die Erkennungsleistung gegenüber dem allgemeinen Modell verbessert werden können, ohne einen Verlust der Flexibilität hinnehmen zu müssen. Vergin et al. (1996) berichten beispielsweise eine Verbesserung der Erkennungsleistung um $14\ \%$, die durch die Verwendung geschlechtsspezifischer Akustikmodelle erreicht werden konnte. Nix, Fairweather und Adams (1998) geben an, dass die Fehlerrate bei einem Spracherkenner für Kinder durch die Verwendung eines spezifischen Akustikmodells um 26 % verringert werden konnte.

Die Gruppenzugehörigkeit zu bestimmen kann als Anwendung von AGENDER angesehen werden: Unter der Voraussetzung, dass der zugrunde liegende Spracherkenner es gestattet, zur Lauf-

zeit dynamisch verschiedene Akustikmodelle zu aktivieren, kann die Auswahl des Modells also auf Basis des von AGENDER gelieferten Sprechermodells erfolgen (vgl. Müller, 2002).

Der Frage, inwieweit durch dieses Verfahren die Erkennungsleistung tatsächlich verbessert werden kann, wird derzeit im Rahmen einer an das Projekt angegliederten Bachelor-Arbeit nachgegangen. Als Spracherkenner dient dabei das flexible *Sphinx*-System von der Carnegie-Mellon Universität in Pittsburgh. Zum Programm-Umfang von Sphinx gehört ein Werkzeug für die Erstellung von Akustikmodellen, welches auf die nach Alter und Geschlecht ausbalancierte Datenbasis des AGENDER-Projektes angewendet wird, die auch Grundlage der Sprecherklassifikation ist. Obwohl die Fertigstellung der Arbeit erst im Januar 2006 erwartet wird, lassen die bereits vorliegenden Ergebnisse den Schluss zu, dass die in der Literatur berichteten Verbesserungen der Spracherkennungsgenauigkeit durch Verwendung spezifischer Akustikmodelle bestätigt werden konnten. Die letztendlichen Resultate werden in Germesin (2006) präsentiert.

1.4 Einordnung in ein interdisziplinäres Forschungsgebiet und Forschungsfragen

Wie in Abbildung 1.3 dargestellt wird, kann die Sprecherklassifikation in einem Überschneidungsbereich zwischen Human- und Ingenieurwissenschaften eingeordnet werden. Der humanwissenschaftliche Charakter entstammt der Tatsache, dass der hauptsächliche Untersuchungsgegenstand die menschliche Stimme und das menschliche Sprechverhalten ist. Somit kann die *Medizinische Phoniatrie*, wenn auch indirekt, als eine derjenigen Wissenschaften angesehen werden, auf deren Erkenntnissen die Ansätze von AGENDER basieren. Die Medizinische Phoniatrie untersucht pathologische (krankhafte) Stimmveränderungen, die entweder durch Regulations- und Steuerungsstörungen im zentralen Nervensystem oder durch Beeinträchtigungen der Mechanik der Stimmlippen-Schwingung hervorgerufen werden. Die Fragen, in welchem Maße solche Veränderungen von Alterungsprozessen beeinflusst werden und inwieweit es unterschiedliche Ausprägungen bei Frauen und Männern gibt, sind Gegenstand des Teilgebiets der *Stimmalterungsforschung* (engl. *vocal aging*). Das Ziel der in der Phoniatrie durchgeführten Untersuchungen ist die Schaffung einer Referenzbasis zur Normalstimme für die Diagnose von Stimmstörungen (vgl. Pützer, 2001). Die Veränderung der menschlichen Sprachproduktionsfähigkeit im Alter, insofern sie auf Gedächtnisverlust und Verlangsamung von Neurotransmissionen zurückgeführt wird, berührt darüber hinaus das Gebiet der *Kognitiven Psychologie*, die sich mit der menschlichen Informationsverarbeitung beschäftigt.

Wie durch die Anordnung der Pfeile in Abbildung 1.3 angedeutet wird, haben diese Disziplinen jedoch hauptsächlich einen indirekten Einfluss auf die vorliegende Arbeit. Als Mittlerin dient in zweifacher Hinsicht die *Phonetik*: Erstens kann ihr die Fragestellung zugeordnet werden, wie sich das Sprecheralter und -geschlecht in der Stimme und dem Sprechverhalten manifestiert, und zweitens bezeichnet sie den methodologischen Übergang zu den Ingenieurwissenschaften. Die *automatische* Sprecherklassifikation nämlich ist, wie in Kapitel 7 dargestellt wird, ein typisches *Mustererkennungsproblem*, also ein Gegenstand der *Künstlichen Intelligenz*. Wie in Abbildung 1.3

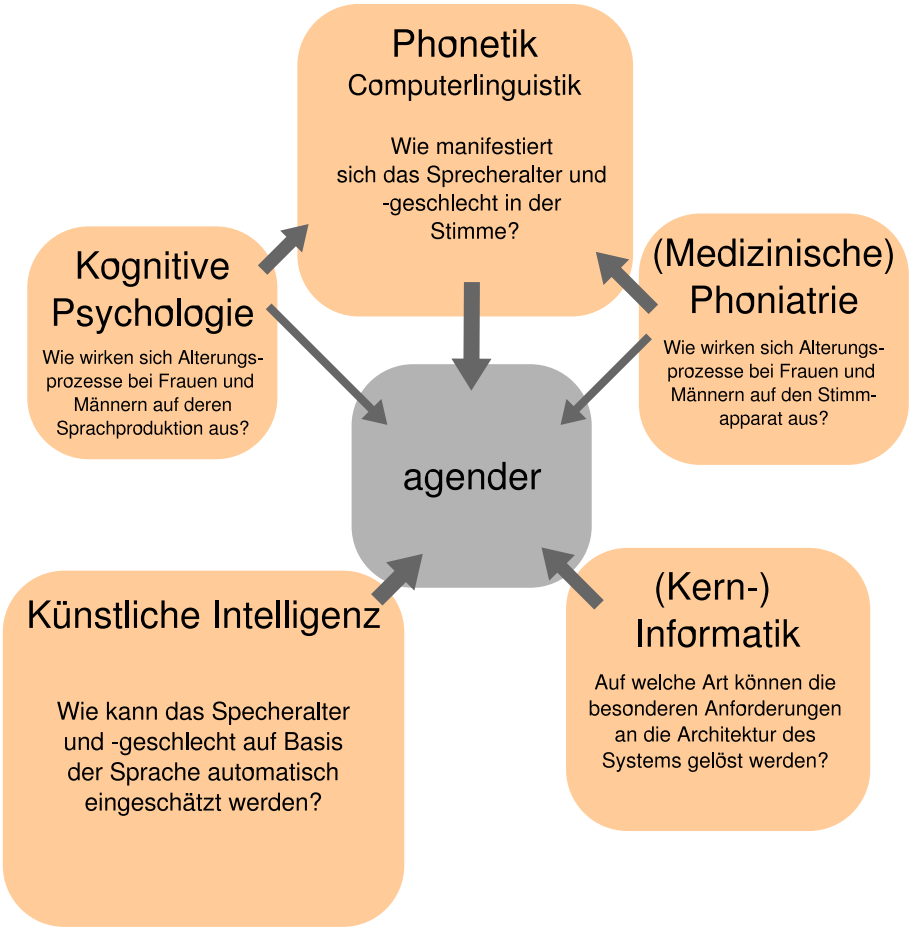


Abbildung 1.3: Einordnung der vorliegenden Arbeit in ein interdisziplinäres Forschungsgebiet.

durch die Größe der Rechtecke bereits angedeutet wird, stellt dies den Schwerpunkt der vorliegenden Arbeit dar.

Die wesentlichen Forschungsfragen, zu deren Beantwortung die vorliegende Arbeit einen Beitrag zu leisten versucht, spiegeln dementsprechend deren interdisziplinären Charakter wider:

Wie manifestiert sich das Specheralter und -geschlecht in der Stimme und dem Sprechverhalten?
Hierbei handelt es sich um eine humanwissenschaftliche Fragestellung, deren Beantwortung hauptsächlich mithilfe von umfangreichen Korpusanalysen angestrebt wird. Nach einer Einführung in die phonetischen Grundlagen in Kapitel 2 wird dazu in Kapitel 3 eine Übersicht über die Befundlage in der Literatur gegeben. Die daraus abgeleiteten Hypothesen und das Verfahren der Korpusanalysen werden in Kapitel 3 bzw. 4 beschrieben. Die in Kaptitel 5 präsentierten Ergebnisse können als ein Beitrag zum Aufbau einer Referenzbasis angesehen werden. Da es

sich bei diesen empirischen Untersuchungen um einen in sich abgeschlossenen Aspekt der Arbeit handelt, wurden die genannten Kapitel unter Teil I zusammengefügt.

Wie kann das Specheralter und -geschlecht von einem System automatisch erkannt werden? und allgemeiner: **Welches ist ein geeigneter Ansatz zur Erkennung von Sprechereigenschaften auf Basis der Sprache?**
Diese ingenieurwissenschaftlichen Fragestellungen, die der Künstlichen Intelligenz zugeordnet werden können, schließen sich unmittelbar an die zuvor genannte an und stellen gleichzeitig die Kernprobleme der vorliegenden Arbeit dar. Ihnen widmet sich der umfangreichere Teil II: In Kapitel 7 werden die Fragen zunächst als Mustererkennungsproblem formuliert. Der Lösungsvorschlag – ein zweistufiger Ansatz des maschinellen Lernens – wird in Kapitel 8 und 9 erläutert.

Welchen Einfluss auf die Erkennung von Sprechereigenschaften hat der auditive Kontext und auf welche Art und Weise kann dieser berücksichtigt werden?
Die Frage, die ebenfalls zu dem zentralen Teil II gehört, ergibt sich aus dem oben genannten Anwendungsszenario der mobilen sprachbasierten Dialogsysteme, bei welchen damit gerechnet werden muss, dass die Benutzung in unterschiedlichen Umgebungen stattfindet. In Kapitel 9 wird deutlich gemacht, inwiefern der vorgestellte Sprecherklassifikationsansatz kontextsensitiv ist.

Welche Anforderungen an ein Sprecherklassifikationssystem ergeben sich aus dem zugrunde liegenden Anwendungsszenario und wie kann das System diesen auf der Implementierungsebene gerecht werden?
Wie in Abbildung 1.3 bereits angedeutet wird, beinhaltet die vorliegende Arbeit auch Aspekte, die der (Kern-)Informatik zugeordnet werden können. Diese werden in Teil III behandelt, welcher wie folgt untergliedert ist: Kapitel 11 beschreibt ein Client/Server-System, welches eine prototypische Implementierung des vorgeschlagenen Sprecherklassifikationsansatzes darstellt. Kapitel 12 stellt darüber hinaus das Korpusanalyse-Werkzeug m3iCAT vor, welches die im Laufe des Projektes entstandenen Methoden der Korpusanalyse und Datensichtung hinter einem Web-Interface vereint und so flexibel gehalten wurde, dass es auch im Rahmen anderer datenbasierter Projekte eingesetzt werden kann.

1.5 Wesentliche methodologische Grundannahmen

Bevor die Stimmalterung und ihre geschlechtsspezifischen Ausprägungen dargestellt werden, soll zunächst das Problemfeld der Sprecherklassifikation deutlicher umrissen werden. Um eine Art vertikale Begrenzung zu schaffen, wird ein Ebenenmodell der Sprachmerkmale vorgestellt, welches – zumindest in ähnlicher Form – in der Literatur häufig vorzufinden ist. Die Grenze der hier betrachteten Merkmale wird dabei unterhalb der linguistischen Ebene gezogen, was in erster Linie durch technische Gründe gerechtfertigt wird. Es folgt eine, wenn man so will, horizontale Begrenzung in dem Sinne, dass diskutiert wird, inwiefern ein solches Verfahren sprachunabhängig (im Sinne von *language independant*) sein kann. Es wird proklamiert, dass eine *wesentliche* Sprachunabhängigkeit erreicht werden kann, wodurch ausgedrückt werden soll, dass die Modelle in der Sprache, in der sie trainiert wurden, zwar die beste Performanz erreichen werden, innerhalb vergleichbarer Sprachen jedoch, möglicherweise unter einer Anpassung von Schwellenwerten, ebenfalls eingesetzt werden können.

1.5.1 Ebenen sprachlicher Merkmale

Sprachliche Phänomene werden häufig anhand eines Ebenenmodells miteinander in Beziehung gesetzt. Auf niedriger Ebene stehen dabei Oberflächenphänomene, die einen direkten Bezug zu physikalischen Größen haben, wohingegen nach oben der Abstraktionsgrad immer größer wird. In der Sprachtechnologie findet ein solches Ebenenmodell seine Entsprechung in dem Begriff der Verarbeitungstiefe, die bereits in dem eingangs aufgestellten Vergleich zwischen einem reinen Spracherkenner und einem elaborierten System wie Verbmobil veranschaulicht worden ist.

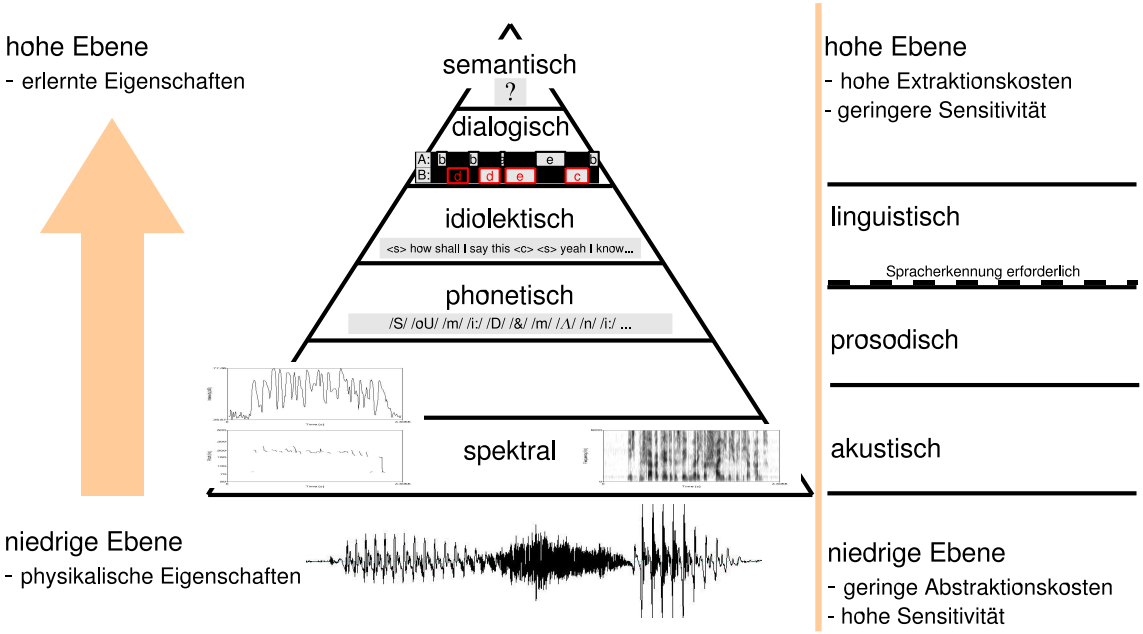


Abbildung 1.4: Hierarchische Modelle der Sprachmerkmale: links nach Reynolds et al. (2003), rechts nach Müller und Wittig (2003).

Für das Problem der *Sprecheridentifikation*, das dem der Sprecherklassifikation sehr ähnlich ist (vgl. Abschnitt 1.6.1), schlagen Reynolds et al. (2003) ein Ebenenmodell vor, wie es in Abbildung 1.4 links dargestellt wird. Die so genannten *spektralen Merkmale* haben offensichtlich den direktesten Bezug zu physikalischen Merkmalen, während die *semantischen Merkmale* am abstraktesten sind. Ein ähnliches, wenn auch etwas einfacheres Modell wurde von Müller und Wittig (2003) für das Problem der Sprecherklassifikation vorgeschlagen (vgl. Abbildung 1.4 rechts). Nach diesem Modell werden die Merkmale niedrigster Ebene als *akustische Merkmale* bezeichnet und diejenigen der höchsten Ebene als *linguistische* Merkmale. Durch die Nebeneinanderstellung wird in Abbildung 1.4 angedeutet, wie die Merkmale beider Modelle korrespondieren.

Merkmale verschiedener Ebenen zu betrachten ist für Mustererkennungsverfahren, zu denen sowohl die Sprecherklassifikation als auch die Sprecheridentifikation gehören, deshalb sinnvoll, weil diese oftmals komplementäre Informationen darstellen. Wie auf der Seite des Modells von Reynolds et al. in Abbildung 1.4 dargestellt wird, handelt es sich bei den Merkmalen niedriger Ebene beispielsweise um physikalische Eigenschaften der Stimme, während höhere Merkmale erlernte Eigenschaften repräsentieren.

Die Merkmale verschiedener Ebenen sind auch unterschiedlich anfällig gegenüber Veränderungen in der akustischen Umgebung: Merkmale niedrigerer Ebenen sind anfälliger gegenüber Hintergrundgeräuschen als Elemente höherer Ebenen. Darauf weisen auch Minematsu, Yamauchi und Hirose (2003) hin, die in ihrem Sprecherklassifikationsansatz ausschließlich akustische Merkmale verwenden. Sie schließen Sprachproben mit Hintergrundgeräuschen sowohl beim Training als auch beim Test der Modelle aus, um einen Einfluss auf die Klassifikationsgenauigkeit zu vermeiden.

Dieses Argument spricht dafür, dass auch Merkmale möglichst hoher Ebenen bei der Klassifikation betrachtet werden sollten. Dagegen sprechen die steigenden Kosten, die für die Extraktion der betreffenden Merkmale aufgewendet werden müssen: Je abstrakter ein Merkmal, desto tiefer muss die linguistische Verarbeitung sein, um das Merkmal zu erfassen, was höhere Kosten (im Sinne von Zeit und komputationeller Komplexität) bedeutet. Ein besonderer Schwellenwert innerhalb dieses trade-offs stellt nach dem Modell von Müller und Wittig (2003) die Grenze zwischen den *prosodischen* und den *linguistischen* Merkmalen dar, weil für die Extraktion der zuletzt genannten in jedem Fall eine wie auch immer geartete Interpretation der Äußerung notwendig wäre. Da im Hinblick auf die Anwendungen von AGENDER jedoch eine schnelle, der Spracherkennung vorgeschaltete Klassifikation notwendig ist, wurde die Betrachtung dieser Merkmale explizit ausgeklammert.

Ein Kritikpunkt an dem Modell von Müller und Wittig betrifft die Benennung der Zwischenebene als die *Prosodische Ebene*. Ursprünglich sollten dort Merkmale zusammengefasst werden, die den Sprachfluss betreffen, also etwa die Artikulationsgeschwindigkeit oder die Anzahl und Dauer von Sprechpausen. Wie jedoch in Kapitel 2 noch detaillierter ausgeführt wird, werden in der Phonetik jedoch Merkmale, welche die Grundfrequenz der Stimme (Pitch) betreffen, ebenfalls der Prosodie zugeschrieben. Diese fallen aber nach dem Modell von Müller und Wittig der *Akustischen Ebene* zu. Zur Vermeidung von Missverständnissen werden daher im Folgenden die Merkmale unterhalb der linguistischen Ebene in zwei vereinfachte Kategorien gefasst, nämlich

Stimmmerkmale und *Sprechverhaltensmerkmale*.

1.5.2 Sprachunabhängigkeit

Parris und Carey (1996) zeigen, dass der von ihnen verfolgte Ansatz zur Erkennung des Sprechergeschlechts auf der Basis akustischer Merkmale prinzipiell sprachunabhängig ist. Sie führten eine Serie von Experimenten durch, bei denen die mit britischem Englisch trainierten Modelle auf Testdaten unterschiedlicher Sprachen angewandt wurden. Im Einzelnen untersuchten sie: amerikanisches Englisch, Farsi, Französisch, Deutsch, Japanisch, Koreanisch, Hindi, Mandarin Chinesisch, Spanisch, Tamil und Vietnamesisch. Die Erkennungsleistung variierte je nach Sprache zwischen einer Fehlerrate von 5.2 % für Tamil und 0 % für Vietnamesisch und Mandarin Chinesisch. Bei den Sprachen mit der höchsten Fehlerrate war eine Neigung zu einem bestimmten Geschlecht zu verzeichnen. Bei Tamil und Hindi waren beispielsweise alle Fehler darauf zurückzuführen, dass fälschlicherweise männliche Sprecher als weiblich eingestuft worden sind, was die Autoren darauf zurückführen, dass diese Sprecher eine allgemein höhere Stimmtonlage besitzen. Analog dazu wurden amerikanische Sprecherinnen fälschlicherweise als männlich klassifiziert, offensichtlich weil in dieser Population eine Tendenz zu einer tieferen Stimmlage existiert. Kulturelle Unterschiede in der Stimmtonhöhe wurden auch von Yamazawa und Hollien (1992) beobachtet: Bei einer vergleichenden Studie zwischen den Stimmen japanischer und amerikanischer Frauen stellten sie fest, dass die Grundfrequenz der amerikanischer Frauen mit 205 Hz signifikant geringer war als die der japanischen Frauen (223 Hz).

Der Begriff der Sprachunabhängigkeit muss also differenzierter betrachtet werden. Zunächst einmal kann zwischen einer universellen, also für alle existierenden menschlichen Sprachen geltenden, und einer auf bestimmte Sprachfamilien, wie z. B. romanische oder indogermanische Sprachen beschränkten Unabhängigkeit unterschieden werden. Universelle Sprachunabhängigkeit ist nur sehr schwer nachzuweisen, da es praktisch unmöglich ist, geeignete Sprachproben in jeder denkbaren Sprache zu sammeln. Es ist jedoch fraglich, ob sie in einem Ansatz, der zwar als zur Grundlagenforschung zugehörig verstanden wird, jedoch konkrete Anwendungsimplikationen beinhaltet, überhaupt angestrebt werden sollte. Was AGENDER betrifft, werden die Anwendungen durch andere Faktoren, wie etwa den des Spracherkenners, auf eine, ggf. auf einige wenige Sprachen beschränkt. Um dennoch eine Integrierbarkeit in eine möglichst große Vielzahl von Applikationen zu gewährleisten, sollte eine *weitestgehende* Sprachunabhängigkeit angestrebt werden.

Der Begriff der weitestgehenden Sprachunabhängigkeit bezieht sich jedoch nicht nur auf die Menge der Sprachen, auf die der Ansatz angewendet werden kann, sondern auch auf die Äquivalenz der Modelle. Wie aus der Studie von Parris und Carey (1996) hervorgeht, darf angenommen werden, dass die Modelle mit der Sprache, mit der sie trainiert wurden, die beste Genauigkeit erreichen. Prinzipiell ist eine Sprachunabhängigkeit auch dann gewährleistet, wenn eine Klassifikation in einer anderen Sprache als der der Trainingssprache zwar eventuell suboptimale, aber aus Anwendungssicht befriedigende Ergebnisse erzielt, wie es bei Parris und Carey der Fall ist. Hinzu kommt, dass in einem solchen Fall die Möglichkeit besteht, durch Anpassung von Schwellenwerten oder *Baselines* die Performanz bezüglich einer bestimmten Zielsprache zu verbessern.

1.6 Verwandte Arbeiten

Aufgrund des interdisziplinären Charakters dieser Arbeit gibt es eine Vielzahl von verwandten Arbeiten aus unterschiedlichen Bereichen. Mit der Sprecherklassifikation eng verwandt ist die *Sprecheridentifikation* bzw. *-verifikation*, weshalb hier eine Abgrenzung der Problemfelder skizziert wird und einige Beispiele aus diesem Bereich aufgeführt werden. Als verwandte Arbeiten können auch diejenigen phonetischen Studien bezeichnet werden, die sich mit der Manifestation des Sprecheralters und -geschlechts auseinander setzen. Mit wenigen Ausnahmen, die in diesem Abschnitt Erwähnung finden, beschränken sich diese Studien jedoch auf eine Identifikation der Phänomene, ohne eine automatische Klassifikation darauf aufzubauen. Diejenigen Arbeiten, die dies tun, werden als unmittelbar verwandte Arbeiten gesondert aufgeführt. Eine besondere Art verwandter Arbeiten stellen darüber hinaus diejenigen Projekte dar, die, welche Fragestellung ihnen auch immer zugrunde liegen mag, Gebrauch von denselben Klassifikationsmethoden machen, wie sie in der vorliegenden Arbeit verwendet worden sind. Wie man sich leicht vorstellen kann, sind diese jedoch so zahlreich, dass eine Auswahl zwangsläufig willkürlich erscheinen muss. Auch würde ihre Auflistung in diesem vorweggenommenen Abschnitt die nötigen Erläuterungen der Funktionsweise der jeweiligen Methode vermissen lassen und sehr weit vom Thema wegführen. Es wurden daher an geeigneter Stelle, nämlich in den Abschnitten, in denen die Methoden diskutiert werden, dezentrale Abschnitte eingeführt, die den verwandten Arbeiten gewidmet sind. Die Auswahl, die dort getroffen wird, soll einerseits die konzeptuelle Nähe zu dieser Arbeit erkennen lassen, andererseits jedoch ein Bild von der Vielseitigkeit der Anwendungsbereiche der Mustererkennung zeichnen.

1.6.1 Sprechererkennung und Sprecherverifikation

Die Sprechererkennung und die Sprecherverifikation können nach Markowitz (2000) unter dem Begriff der *Stimmbiometrie* zusammengefasst werden (vgl. Abbildung 1.5). Systeme zur Sprecherverifikation überprüfen eine *Identitätsbehauptung* einer Person anhand eines *Stimmabdrucks* (*voice print*)[4]. Welche Form die Identitätsbehauptung hat, hängt von dem Ansatz ab. Wenn eine festgelegte Äußerung (z. B. ein Passwort) verlangt wird, spricht man von einer *textabhängigen Sprecherverifikation* (vgl. ebd.). In Beispiel 1.5 wird die Eingabe zusätzlich von einem Spracherkenner verarbeitet (Dekodierung der Zugangsnummer).

(1.5)

System:	Bitte geben Sie Ihre Kontonummer an.
Anrufer :	235167.
System:	[kann den Sprecher erfolgreich identifizieren]
	Danke.

Beispiel 1.6 illustriert eine Sprecherverifikations-Variante mit vorgegebenem Text (engl. *textpromted*). Bei diesem Verfahren wird der Sprecher aufgefordert, eine Folge von zufällig ausge-

[4] abgeleitet von Fingerabdruck (*fingerprint*).

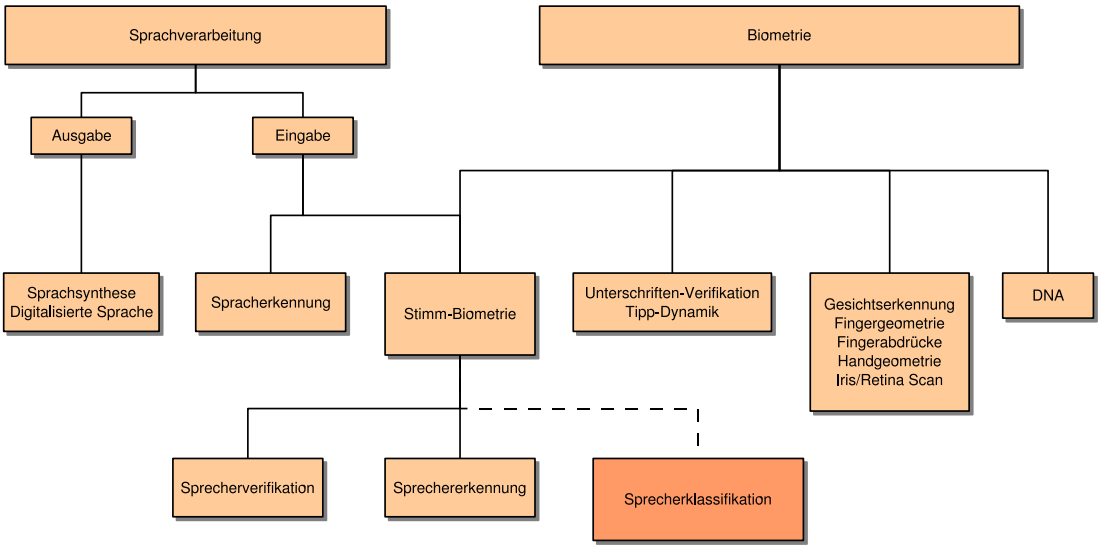


Abbildung 1.5: Um die Sprecherklassifikation erweiterter „Stammbaum“ der Stimmbiometrie nach Markowitz (2000).

wählten Ziffernfolgen, Wörtern oder Phrasen zu wiederholen. Hierfür ist in der Regel ein ausgiebigeres Training erforderlich als bei textabhängigen Systemen, da der Stimmabdruck alle Komponenten enthalten muss, die später für die Erzeugung der Abfragen verwendet werden können. Aus dem Beispiel wird darüber hinaus deutlich, dass die Verifikation länger dauert. Allerdings ist dieses Verfahren sicherer, da es nur sehr schwer mit zuvor aufgezeichneten Stimmproben des Zielsprechers getäuscht werden kann (vgl. Markowitz, 2000, S. 68).

(1.6)

System:	Bitte geben Sie mithilfe der Tastatur Ihres Telefons Ihre Kontonummer an.
Anrufer :	235167.
System:	Bitte sagen Sie 42-69.
Anrufer :	42-69.
System:	Bitte sagen Sie 83-24.
Anrufer :	83-24.
System:	Bitte sagen Sie 99-48.
Anrufer :	99-48.
System:	[kann den Sprecher erfolgreich identifizieren] Danke.

Die *textunabhängige* Variante schließlich akzeptiert jede Form von gesprochener Eingabe, wodurch eine nicht-intrusive Verifikation erreicht wird, d. h. es werden für den Prozess der Verifikation keine zusätzlichen Dialogschritte benötigt (vgl. Markowitz, 2000, S. 69). In Beispiel 1.7 wird eine solche textunabhängige Sprecherverifikation durch einen Bankangestellten gestartet, um den Kunden für die Transaktion zu authentifizieren. Die textunabhängige Technologie ist wesentlich schwieriger zu implementieren als die beiden anderen (vgl. ebd.). Sie erfordert längere Sprachproben und ist empfindlicher gegenüber der akustischen Qualität der Eingabe.

(1.7)

Bankangestellter:	Was kann ich für Sie tun, Frau Jones?
Anrufer :	Ich möchte einhunderttausend Dollar von meinem Sparbuch auf mein Konto in Bimini überweisen, das ich eben eröffnet habe.
Bankangestellter:	Lassen Sie mich das prüfen. Einen Moment bitte!
	[PAUSE während das System die Identität des Anrufers überprüft]
Bankangestellter:	Vielen Dank für Ihre Geduld. Ihr Auftrag wird nun ausgeführt.

Reynolds, Quatieri und Dunn (2000, S. 20) formulieren das Problem der Sprecherverifikation formal wie folgt: Gegeben sei ein Sprachsegment Y und ein hypothetischer Sprecher S. Das Problem der Sprecherverifikation besteht darin, zu bestimmen, ob Y von S gesprochen wurde (Hypothese H_0) oder nicht (Hypothese H_1). Implizit wird dabei angenommen, dass Y die Sprache von nur einem Sprecher enthält.

Bei der *Sprechererkennung* wird einem unbekannten Sprecher eine Identität zugewiesen, was in den meisten Fällen schwieriger ist als die Sprecherverifikation, da die initiale Identitätsbehauptung fehlt und das Verfahren fast immer textunabhängig ist (vgl. Markowitz, 2000, S. 69). Einige Systeme bewerten und sortieren die Menge der Sprachabdrücke nach ihrer Ähnlichkeit zu einer gegebenen Probe. Andere wählen eine oder mehrere mögliche Identitäten aus. Bei einigen Autoren, wie beispielsweise Cohen und Lapidus (1995), wird der Begriff der *Sprecherklassifikation* – anders als in der vorliegenden Arbeit – für diese Form der Stimmbiometrie verwendet.

Das Modell für H_0 ist sowohl für das Problem der Sprecherverifikation als auch das der Sprecheridentifikation wohl definiert und kann auf Basis von Trainingsdaten des Sprechers S eingeschätzt werden (vgl. Reynolds et al., 2000, S. 22). Das Modell für H_1 dagegen ist weniger gut definiert, da es potentiell den gesamten Raum von möglichen Alternativen repräsentieren muss. Einem gängigen Ansatz nach wird für H_1 ein einzelnes alternatives Modell aus einem *Pool* von verschiedenen Sprechern erzeugt, welches als *Universal Background Model* (UBM) bezeichnet wird. Das UBM repräsentiert die Population von Sprechern, die bei der Erkennung von S als Alternativen erwartet werden.

Bei der Entwicklung eines Sprecherverifikations oder -identifikationssystems kann durch die Wahl eines entsprechend einfachen UBMs das Entscheidungsproblem zwischen H_0 und H_1 be-

liebig vereinfacht werden. Im Minimalfall enthält das UBM nur die Sprache von einem einzelnen Alternativsprecher. Später kann das UBM dann schrittweise so erweitert werden, dass es gerade so komplex ist, wie es für die gegebene Anwendung erforderlich ist. Anders ist es bei der Sprecherklassifikation: Um S einer bestimmten Klasse von Sprechern zuweisen zu können, müssen die Modelle (hier der verschiedenen Klassen) von vornherein so generell wie möglich sein. Anders ausgedrückt: Während die Sprecherverifikation und -identifikation am besten mit einem UBM funktioniert, das auf Basis einer möglichst kleinen Population von Sprechern erzeugt wurde, funktioniert die Sprecherklassifikation am besten mit Modellen, die auf Basis einer möglichst großen Population von Mitgliedern der jeweiligen Klasse erzeugt wurden. Dieser methodologische Unterschied wirkt sich entscheidend auf die Art und Weise aus, wie die Modelle entwickelt werden.

1.6.2 Verwandte Arbeiten in der Phonetik

Bei den in der Phonetik durchgeführten Analysen zur Identifikation sprachlicher Indikatoren für das Alter und Geschlecht des Sprechers handelt es sich zumeist um *Perzeptionstests*. Dabei werden den Versuchspersonen entsprechende Stimuli präsentiert, die sie gemäß der untersuchten Sprechereigenschaft bewerten sollen. Ein Hauptunterschied dieser Tests betrifft die Art der Stimuli, bei denen es sich entweder um natürliche (nicht manipulierte) oder synthetisierte Sprache handeln kann.

Schötz (2004c) berichtet über ein Perzeptionsexperiment, bei welchem die Stimuli gezielt manipuliert wurden. Das Ziel der Untersuchung bestand darin, zu zeigen, ob prosodische Indikatoren bei der menschlichen Einschätzung des Sprecheralters wichtiger sind als spektrale. Prosodische Indikatoren wurden definiert als F0 (Grundfrequenz) und Dauer, spektrale Indikatoren als alle nicht-prosodischen, also alles außer F0 und Dauer. Insgesamt 28 Realisierungen des schwedischen Wortes *rasa* (zusammenbrechen) von 12 älteren Sprechern (60-82 Jahre) und 16 jüngeren Sprechern (18 bis 31 Jahre) wurden so manipuliert, dass vier Typen von Stimuli entstanden: A) Beide Attribute wurden so gewählt, dass die Werte auf einen älteren Sprecher hinweisen. B) Beide Attribute haben die Werte für einen jüngeren Sprecher. AB) Die spektralen Merkmale weisen die Werte für einen älteren Sprecher auf, aber F_0 und Dauer diejenigen für einen jüngeren Sprecher. BA) Die spektralen Merkmale weisen die Werte für einen jüngeren Sprecher auf, aber F0 und Dauer diejenigen für einen älteren Sprecher. Aus dieser Konstellation ergibt sich ein schematisches Diagramm, wie es in Abbildung 1.6 dargestellt wird.

In zwei Perzeptionstests (einem für Frauenstimmen und einem für Männerstimmen) präsentierte Schötz (2004c) Paare von AB- und BA-Stimuli in zufälliger Reihenfolge. Die Aufgabe der Versuchspersonen war, zu entscheiden, welcher der beiden Stimuli älter klang. Etwa 30 Studenten der Phonetik zwischen 18 und 36 Jahren nahmen an dem Experiment teil. Die Stimuli vom Typ AB (ältere Stimmen mit jüngerer F0 und Dauer) wurden signifikant häufiger als älter eingeschätzt als diejenigen vom Typ BA. Dabei wurden Unterschiede zwischen weiblichen und männlichen Stimmen festgestellt: Für weibliche Sprecher befanden mehr Versuchspersonen die Stimuli vom Typ BA als älter.

Abgesehen von dem globalen Ergebnis, dass F0 und Dauer relevante Indikatoren für das Spre-

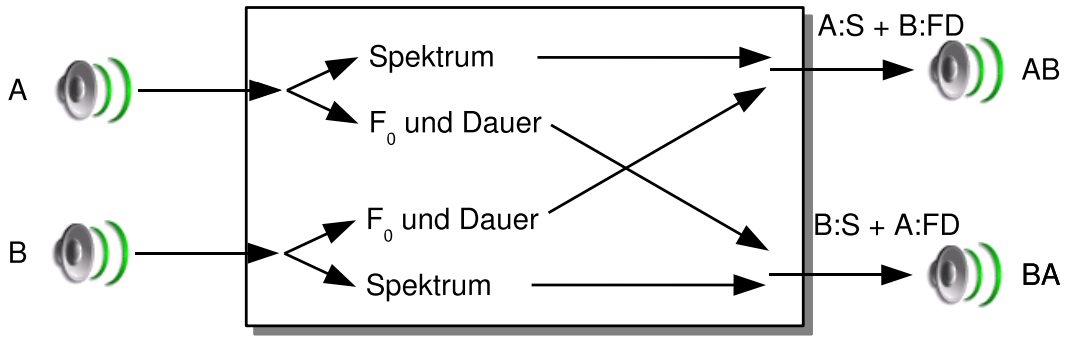


Abbildung 1.6: Schematisches Diagramm der Resynthese von Perzeptionsstimuli bei Schötz (2004c).

cheralter zu sein scheinen, und dass in diesem Zusammenhang auch Unterschiede zwischen Frauen und Männer zu beobachten sind, ist hier vor allem die Methode bemerkenswert, die als „Analyse durch Synthese“ bezeichnet werden kann, und sich algorithmisch wie folgt beschreiben ließe: (1) Finde heraus, welche Merkmale Menschen benutzen, um das Alter zu schätzen. (2) Erstelle daraus eine Menge von Merkmalskandidaten. (3) Erzeuge diese Merkmale künstlich durch eine Simulation der Veränderung an dem Sprachsignal. (3) Überprüfe, ob die Versuchspersonen die künstlich erzeugte Tendenz korrekt einschätzen. (4) Messe die Qualität der Methode durch die Differenz zwischen simuliertem Alter und eingeschätztem Alter. (5) Verbessere inkrementell die Merkmalskandidaten.

Diese Methode als Grundlage zur Identifikation einer Merkmalsmenge für die automatische Klassifikation zu verwenden, würde bedeuten, die menschliche Fähigkeit zur Einschätzung des Alters so gut wie möglich simulieren zu wollen, und dabei bewusst diejenigen Eigenschaften zu verwenden, die Menschen möglicherweise unbewusst benutzen. Der Vorteil wäre, dass – im Gegensatz zur Korpusanalyse – keine großen Datenmengen erforderlich wären. Allerdings ist zu bedenken, dass die Veränderung möglicherweise einen anderen Effekt hat, z. B. dass die Stimme kränklich klingt, die Versuchspersonen jedoch, weil sie danach nicht gefragt wurden, dazu tendieren, die Stimme älter einzuschätzen. Darüber hinaus ist fraglich, ob eine Maschine in jedem Fall dieselben Merkmale zur Klassifikation benutzen sollte wie der Mensch. Möglicherweise gibt es Eigenschaften in der Sprache, anhand derer das Alter gut bestimmt werden kann, die dem Menschen jedoch verborgen bleiben – andere Merkmale, die der Mensch benutzt, könnten wiederum nicht maschinell erfassbar sein. Der inkrementelle Verbesserungsprozess ist darüber hinaus arbeits- und zeitintensiv. Unter Umständen wäre eine Kombination aus beiden Methoden zielführend: Eine Korpusanalyse könnte für die Zusammenstellung der initialen Merkmale eingesetzt werden. Zur Optimierung des Merkmalsets könnten dann Perzeptionsexperimente durchgeführt werden.

1.6.3 Unmittelbar verwandte Arbeiten

Parris und Carey (1996) beschreiben einen Ansatz zur sprachunabhängigen Erkennung des Sprechergeschlechts auf der Basis der Stimmtonhöhe (Pitch) sowie einem Merkmal, dass die Autoren „akustische Analyse“ nennen. Die akustischen Merkmale wurden zunächst mithilfe zweier *Hidden Markov Modelle* klassifiziert (vgl. Abschnitt 8.1), von denen eines für Männer und eines für Frauen trainiert wurde. Die Pitch-Information wurde unabhängig davon von einem linearen Klassifizierer verarbeitet (vgl. Abschnitt 8.6). Die Ergebnisse der beiden Klassifikatoren wurden mithilfe der *Maximummethode* miteinander kombiniert (vgl. Abschnitt 9.3.3). Die für Britisches Englisch (die Trainingssprache) erreichte Genauigkeit wird in Tabelle 1.1 zusammengefasst. Was die Untersuchung mit anderen Sprachen betrifft, so kann das System als *weitestgehend sprachunabhängig* bezeichnet werden, da die Modelle mit anderen Sprachen als der Trainingssprache zwar suboptimale, aber zufrieden stellende Ergebnisse erbrachten.

Merkmal	Genauigkeit
Akustische Analyse	91.65 %
Pitch	97.20 %
Kombination	99.30 %

Tabelle 1.1: Genauigkeit der Geschlechtsklassifikation bei Parris und Carey (1996) für Britisches Englisch (Trainingssprache).

Vergin et al. (1996) verwenden neben der Stimmtonhöhe die Lage der ersten beiden Formanten (vgl. Abschnitt 3.3.2) als Merkmale zur Klassifikation des Sprechergeschlechts. Diese Information wird in einem Spracherkennungssystem ausgenutzt, um die Erkennungsleistung durch spezifische Akustikmodelle zu verbessern. Für die Geschlechtsklassifikation werden die ersten beiden Äußerungen eines Sprechers verwendet (bei einer durchschnittlichen Länge von sieben Sekunden pro Äußerung). Die zugrunde liegende Methode ist idiosynkratisch: Für jede der beiden Äußerungen erfolgt eine Bewertung der Formanten bezüglich der Ähnlichkeit zu den männlichen bzw. weiblichen Referenzwerten aus der Literatur. Die Bewertungen aus beiden Äußerungen werden anschließend summiert. Das Ergebnis entspricht derjenigen Klasse, die die höchste Gesamtbewertung erhalten hat. Vergin et al. geben die erreichte Genauigkeit mit 85 % an. Durch eine Integration der Sprecherklassifikation in ein Spracherkennungssystem konnte dessen Fehlerrate erheblich reduziert werden. Trotz der eher geringen Genauigkeit der Geschlechtsklassifikation beziffern die Autoren die Verbesserung der Spracherkennung auf beachtliche 14 %.

Minematsu et al. (2003) führten ein Perzeptionsexperiment durch, bei dem die Versuchspersonen das Alter von etwa 500 männlichen Sprechern schätzen sollten. Die Verteilung des biologischen Alters belief sich auf 20 bis 90 Jahre. Außerdem waren in der Datenbasis Sprachproben von 140 Jungen zwischen sechs und zwölf Jahren enthalten. Zusätzlich zur Einschätzung des Alters sollten die Versuchspersonen auch die Qualität der Aufnahme einschätzen. Diejenigen Sprachproben, die Hintergrundlärm enthielten, wurden anschließend aus der Datenbasis entfernt. Von den übrig gebliebenen 407 Sprechern wurden *Gaussian Mixture Models* (GMMs) trainiert. Das Alter eines Sprechers wurde auf Basis einer gewichteten Summe der GMMs aller restlichen Sprecher

geschätzt. Dabei wurde eine Korrelation zwischen dem geschätzten Alter aus dem Perzeptionsexperiment und dem automatisch geschätzten Alter von durchschnittlich 91 % erreicht.

Trotz dieses vielversprechenden Ergebnisses ist die unmittelbare Anwendbarkeit dieses Verfahrens fraglich. Der erste und offensichtlichste Kritikpunkt ist der, dass ausschließlich männliche Sprecher betrachtet worden sind, was eine starke Vereinfachung der Alterseinschätzung darstellt (vgl. Kapitel 3). Zweitens stellt auch die Einschätzung des *perzeptiven* statt des *biologischen* Alters eine Vereinfachung dar. Minematsu et al. rechtfertigen dies damit, dass in der zwischenmenschlichen Kommunikation die Anpassung an einen Gesprächspartner ebenfalls auf Basis des perzeptiven Alters erfolgt. Dass dadurch eine starke Korrelation mit den akustischen Indikatoren des Sprecheralters hergestellt wird, ist offensichtlich. Im Hinblick auf die Anwendung sind jedoch die daraus ableitbaren Inferenzen schwächer, da die besonderen Bedürfnisse einer Altersgruppe nicht damit zusammenhängen, wie alt ein Mensch *klingt*, sondern wie alt ein Mensch *ist*. In Tabelle 1.2 wird der Vergleich der vorliegenden Arbeit mit den unmittelbar verwandten Arbeiten zusammenfassend dargestellt. Die genannten Methoden werden in Kapitel 8 erläutert.

Referenz	Geschlecht (Genauigkeit in Prozent)	Alter			Verfahren
		Anzahl der Klassen	perzeptiv / biologisch	Genauigkeit in Prozent	
Abdulla und Kasabov (2001)	100	keine Berücksichtigung			Schwellenwertmethode auf Basis von Korpusanalysen
Minematsu et al. (2003)	keine Unterscheidung; bei der Altersklassifikation ausschließlich Betrachtung männlicher Sprecher	numerisches Alter	perzeptiv	91	Gaussian Mixture Models
Parris und Carey (1996)	99.30	keine Berücksichtigung			Hidden-Markov-Models
Schötz (2004a)	keine Unterscheidung, aber Berücksichtigung bei der Altersklassifikation	numerisches Alter	perzeptiv	81.17	CART
Vergin et al. (1996)	85	keine Berücksichtigung			Schwellenwertmethode auf Basis von Referenzwerten
vorliegende Arbeit	93	2-8*	biologisch	64 – 91*	zweistufiges Verfahren; verschiedene Klassifikationsmethoden plus Dynamische Bayes'sche Netze
		*inklusive Geschlecht			

Tabelle 1.2: Vergleich der vorliegenden Arbeit mit den unmittelbar verwandten Arbeiten. Die genannten Verfahren werden in Kapitel 8 erläutert.

Teil I

Empirische Studien zur Manifestation von Sprecheralter und -geschlecht in Stimme und Sprecherverhalten

2 Phonetische Grundlagen

Es gibt zwei Disziplinen, die sich mit den lautlichen Aspekten der Sprache befassen: Phonetik und Phonologie. Gegenstand der Phonologie ist die Beschreibung von Lautsystemen und von systematischen Prozessen innerhalb von Lautsystemen. Die Phonetik interessiert sich dagegen mehr für die physikalischen Aspekte der Lautsprache: Wie werden Laute produziert, wie unterscheiden sich Laute akustisch und wie werden akustische Ereignisse wahrgenommen? Der Unterschied soll an folgendem Beispiel verdeutlicht werden: Eine elementare Methode der Phonologie ist der so genannte Minimalpaartest. Die Lautketten /lift/ und /luft/ haben im Deutschen eine unterschiedliche Bedeutung, die ausschließlich dadurch hergestellt wird, dass ein Laut ausgetauscht wird. Die Laute /i/ und /u/ sind daher Minimalpaare und können als solche als bedeutungstragende Laute (*Phoneme*) identifiziert werden. Für die Phonetik typisch ist dagegen die Analyse der /i/-Realisierungen zweier Sprecher des Deutschen, z. B. die einer Frau und die eines Mannes, wobei sich in der Regel erhebliche Unterschiede in der Höhe des Stimmtons (Sprachgrundfrequenz) feststellen lassen. Auch die /i/-Realisierungen zweier männlicher Sprecher klingen nicht identisch. Die mitunter subtilen Eigenheiten der individuellen Lautproduktion sind eine Grundlage der *Sprechererkennung*, die in der *forensischen Phonetik* eine wichtige Rolle spielt. Für die allgemeine Phonetik lässt sich daraus die folgende zentrale Fragestellung ableiten: Welche Faktoren beeinflussen auf welche Weise die Artikulation von Sprachlauten und welche Konsequenzen hat dies für die akustische Qualität der Sprachlaute? Zu phonetischen Fragestellungen gibt es verschiedene Zugänge: Die älteste Art, Phonetik zu betreiben, ist die so genannte *Ohrenphonetik*, bei der sich der Phonetiker Äußerungen anhört und sie zu beschreiben versucht. Eine andere Art ist die Instrumental- oder Signalphonetik, bei der die Signale analysiert werden, die von der Lautsprache erzeugt werden.

Artikulatorische Prozesse lassen sich mit speziellen Geräten messen, wie beispielsweise einem Elektroglottographen. Dieses nicht-invasive Verfahren misst die Impedanzveränderung des Kehlkopfes während der Phonation. Dazu werden zwei Elektroden auf die Haut über den Schildknorpelplatten angebracht, zwischen denen ein hochfrequenter Wechselstrom geleitet wird. Die durch die Respiration und Phonation hervorgerufenen Impedanzveränderungen werden amplitudenmodelliert, wobei die Amplitude des Signals in etwa linear abhängig ist von der Kontaktfläche der Stimmlippen (vgl. Pützer und Marasek, 2000). Zu den invasiven Methoden gehört der Elek-

tropalatograph, bei welchem dem Sprecher ein künstlicher Gaumen eingesetzt wird, der mit einer Reihe von Elektroden versehen ist, um den Zungen-Gaumen-Kontakt zu messen. Beim ebenfalls invasiven Elektromagnetischen Artikulographen werden mithilfe von elektromagnetischen Spulen Bewegungen an bis zu fünf Positionen gleichzeitig (Zunge, Lippen, Unterkiefer, Gaumen) aufgezeichnet. Der Elektromyograph wiederum kann als nicht-invasive Methode angesehen werden, da mithilfe von Oberflächenelektroden die Potenzialdifferenzen in Muskelfasergruppen gemessen werden und auf diese Art und Weise die unmittelbare Muskelaktivität bei der Artikulation aufgezeichnet wird.

Aus Gründen, die durch die technischen Rahmenbedingungen gegeben sind, beschränkt sich die vorliegende Arbeit auf die Signale, die über die Luft (oder andere Materialien) übertragen werden, auf die *Akustik*. Neben den Grundlagen der *Artikulatorischen Phonetik*, die deshalb relevant sind, weil sie den grundsätzlichen Zusammenhang zwischen den Eigenschaften des Sprechers und den messbaren akustischen Signalen beschreiben, werden daher einige Aspekte der *Akustischen Phonetik* thematisiert. Diese bilden die Grundlage für das *Merkmalsinventar*, das aus Ergebnissen relevanter phonetischer Studien zusammengestellt wurde.

Dieser Abschnitt wurde bewusst nicht mit „Grundlagen der Phonetik“ betitelt, da es weniger darum geht, eine umfassende Einführung in diese Disziplin zu geben, als darum, diejenigen Konzepte aus Artikulatorischer und Akustischer Phonetik zu umreißen, die die Basis für die Zusammenstellung des *Merkmalsinventars* bilden. Als Merkmalsinventar werden diejenigen messbaren Eigenschaften der Lautsprache bezeichnet, die für die Einschätzung von Alter und Geschlecht sowie, bei entsprechender Verallgemeinerung, weiterer Sprechereigenschaften relevant sind.

2.1 Artikulatorische Phonetik

Der menschliche Sprachapparat ist ein äußerst komplexes System, an dem eine Reihe von Organen und Muskeln beteiligt ist: das Zwerchfell, die Lunge, die Brustkorbmuskulatur, der Kehlkopf, die Zunge sowie die Schlund- und Mundmuskulatur. Das Zwerchfell und die Brustkorbmuskulatur üben Druck auf die in der Lunge gespeicherte Luft aus, so dass diese durch Luftröhre, den Kehlkopf und den Nasen- und Rachentrakt entströmt.

2.1.1 Phonation

Für die Erzeugung von Stimmlauten (*Phonation*) ist der Kehlkopf (Larynx) verantwortlich. Er besteht aus gelenkig miteinander verbundenen Knorpelstrukturen, Muskeln, Bändern und Schleimhäuten (vgl. Pompino-Marschall, 2003, S. 31).

Zwischen *Schildknorpel* und den beiden so genannten *Stellknorpeln* spannen sich die *Stimmlippen*, bestehend aus *Stimmbändern*, (Vocalis-)Muskeln und umgebenden Schleimhäuten (vgl. Abbildung 2.1). Der Spalt zwischen den Stimmlippen wird *Glottis* genannt. Durch entsprechende Konstellation der Stellknorpel kann die Glottis (ganz oder teilweise) geschlossen oder geöffnet werden (vgl. Abbildung 2.2).

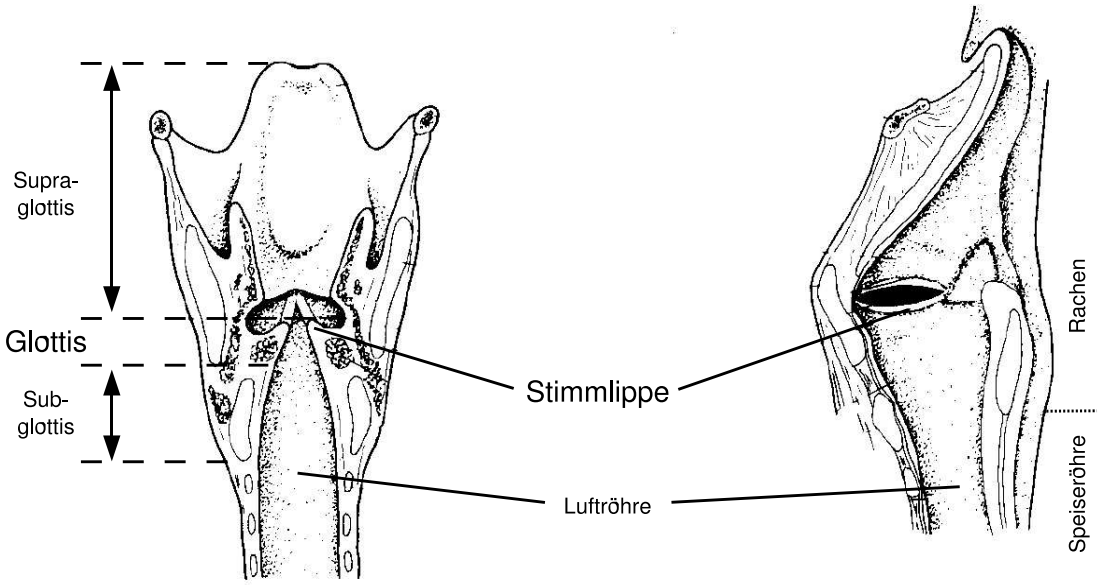


Abbildung 2.1: Kehlkopfinneres in Frontal- und Seitenansicht nach Petursson und Neppert (1991, S. 64).

Bei der Phonation werden die Stimmlippen in regelmäßige Schwingungen versetzt. Dabei wird zunächst (1) unterhalb der Glottis Druck aufgebaut, der zu (2) deren Sprengung führt. Aufgrund des in der Lunge herrschenden Überdrucks strömt (3) Luft durch die Glottis. Da der Spalt eine Verengung darstellt, die zu einer höheren Fließgeschwindigkeit führt, entsteht (4) ein Sog senkrecht zur Fließrichtung (*Bernoulli-Effekt*) und die Stimmlippen schließen sich wieder. Der Zyklus beginnt wieder von vorn (vgl. Petursson und Neppert, 1991, S. 70 f) . Das akustische Resultat der Stimmlippenschwingungen wird *Anregungssignal* genannt. Das Anregungssignal wäre eine geeignete Quelle, um Rückschlüsse auf die physiologischen Eigenschaften des Sprechers zu ziehen, da es unmittelbar von den biologischen Faktoren von Kehlkopf und Stimmlippen abhängig ist. Leider ist es nicht mit dem Mikrofon messbar, da es auf dem Weg durch den Vokaltrakt stark verändert wird. Die *Frequenz* der Stimmlippenbewegungen korreliert mit der wahrgenommenen Tonhöhe und wird stark geprägt durch die Länge der Stimmlippen. Die Stimmlippen von Männern sind 17 bis 24 mm lang, was einer Frequenz von ca. 120 Hz entspricht. Frauen haben dagegen in der Regel 13 bis 17 mm lange Stimmlippen (ca. 220 Hz), die von Säuglingen sind 5 mm lang (400 Hz). Es gibt jedoch noch andere, steuerbare Faktoren, die die Höhe des Stimmtons beeinflussen. Dazu zählen die Steifheit der Stimmlippen, durch die der schwingungsfähige Teil verändert wird, sowie die Stärke des subglottalen Luftdrucks. Von der Höhe dieses Drucks hängt auch die *Lautstärke* des Stimmtons ab. Die *Stimmqualität* hängt unter anderem davon ab, ob die Glottis bei der Phonation komplett geschlossen ist und ob die Stimmlippen steif genug sind, um dem subglottalen Druck

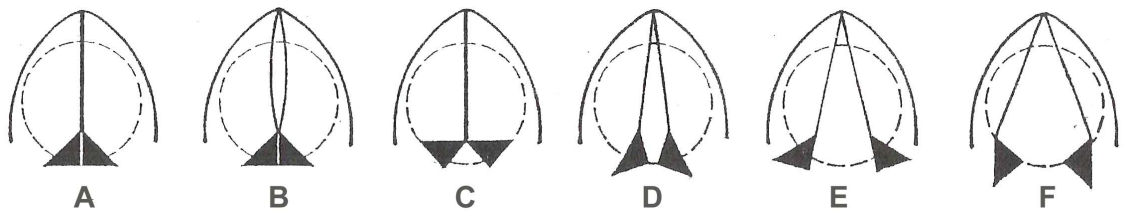


Abbildung 2.2: Schema der verschiedenen Stellungen von Stellknorpel und Stimmlippen; A: Glottisverschluss, B: Phonationsstellung, C: Flüsterstellung, D: Hauchstellung; E: Atmungsstellung oder Ruhestellung; F: Tiefatmungstellung (vgl. Petursson und Neppert, 1991, S. 73 ff).

ausreichenden Widerstand entgegenzusetzen. Sowohl ein unvollständiger Verschluss als auch eine mangelnde Steifheit führen zu einer *behauchten Stimme*. Unregelmäßigkeiten (Mikrovariationen) in den Stimmlippenbewegungen, sowohl in der Amplitude (*Shimmer*) als auch in der Frequenz (*Jitter*), können als *raue Stimme* wahrgenommen werden (vgl. Abbildung 2.3).

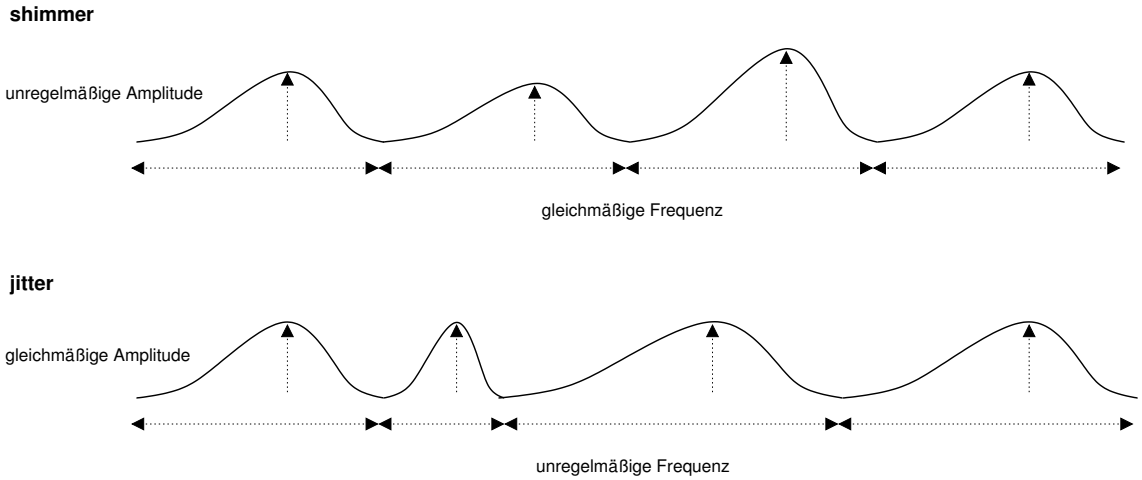


Abbildung 2.3: Unregelmäßigkeiten (Mikrovariationen) bei den Stimmlippenbewegungen lassen sich im Signal durch Shimmer (Amplitudenvariation) und Jitter (Frequenzvariation) messen.

2.1.2 Das Ansatzrohr

Der Vokaltrakt kann als eine Art Filter betrachtet werden, der je nach Konfiguration bestimmte Frequenzen des Anregungssignals verstärkt bzw. dämpft. Um die Konfiguration zu beschreiben, bedient man sich des Begriffs des *Ansatzrohres* aus der Instrumentenkunde, welches der Primärschall (Anregungssignal) mit einer bestimmten Klangqualität verlässt. Wir verfügen über zwei *Resonatoren*, den oralen (Mundraum) und den nasalen (Nasenraum). Letzterer kann zur Artikulation bestimmter Laute wahlweise zugeschaltet werden oder nicht. Die Resonanz wird gesteuert

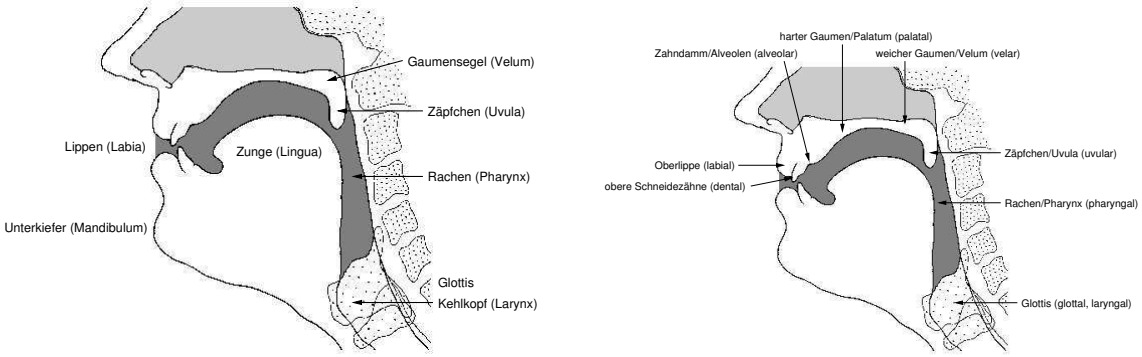


Abbildung 2.4: Links: Artikulatoren, bewegliche Teile des Vokaltrakts. Rechts: Artikulationsorte, anatomische Fixpunkte (vgl. Pompino-Marschall, 2003, S. 44).

mithilfe des *weichen Gaumens*, des *Velums* (vgl. Pompino-Marschall, 2003, S. 43).

2.1.3 Artikulation

Die Artikulation, hier im engeren Sinne verstanden als die Variation des Vokaltrakts während des Sprechens, wird bestimmt durch den *Artikulationsort* und den *Artikulationsmodus*. Der räumliche Aspekt kann als eine Positionsveränderung der beweglichen Teile des Vokaltrakts, der *Artikulatoren* (vgl. Abbildung 2.4 links), in Bezug auf die anatomischen Fixpunkte, die *Artikulationsorte* (vgl. Abbildung 2.4 rechts), betrachtet werden.

Was den Modus betrifft, unterscheidet man zwischen zwei Grundkonstellationen: dem *vokalischen Modus* und dem *konsonantischen Modus*. Der vokalische Modus ist insbesondere dadurch gekennzeichnet, dass die Luft den Vokaltrakt mehr oder weniger ungehindert passieren kann, so dass das Anregungssignal ausschließlich durch globale Veränderungen des Ansatzrohres verändert (*moduliert*) wird. Durch Vorstülpen der Lippen wird z. B. das Ansatzrohr verlängert, durch Absenken des Kiefers oder der Zunge oder durch Vor- und Zurückbewegen derselben wird dessen Querschnitt verändert. Im konsonantischen Modus dagegen wird das Ansatzrohr lokal verengt oder verschlossen. Je nach Grad, Dauer oder Form der Verschlussbildung und der damit verbundenen Blockierung des Luftstroms, unterscheidet man verschiedene Lautklassen.

Auf die verschiedenen Konstellationen und die Benennung der damit verbundenen Laute, aus denen sich das *Internationale Phonetische Alphabet* (IPA) zusammensetzt, wird hier nicht eingegangen. Dennoch ist zum Verständnis des Ursprungs der in dieser Arbeit verwendeten phonetischen Merkmale die Unterscheidung einiger Lautarten notwendig: Bei *Plosiven* (Verschlusslauten) findet ein kompletter oraler und velarer Verschluss statt (z. B. /t/, /p/). *Nasale* kommen dadurch zu Stande, dass der Mundraum zwar blockiert ist, das Velum jedoch abgesenkt ist, wodurch Luft durch die Nase entweichen kann (z. B. /n/, /m/). *Vibranten* zeichnen sich durch intermittierende orale Verschlüsse aus (zwei bis drei in fließender Rede, z. B. /r/). *Frikative* entstehen durch eine starke Verengung, durch welche Turbulenzen entstehen, die für das Geräusch verantwort-

lich sind (z. B. /s/, /f/). *Approximanten* sind gekennzeichnet durch eine schwache Verengung ohne Geräuschbildung, da der Luftstrom nahezu ungehindert passieren kann. Daher werden sie auch *Halbvokale* oder *Vokoide* genannt (z. B. /j/).

Die Vokale werden zunächst nach horizontaler und vertikaler Zungenposition differenziert. Im *Vokalviereck* repräsentiert die vertikale Dimension die *Höhe* der Vokale und die horizontale Dimension den Artikulationsort. Hohe Vokale sind z. B. /i/ und /u/, während /a/ ein tiefer Vokal ist; /i/ wird vorne realisiert, /u/ hinten.

2.2 Akustische Phonetik

Das Schallsignal, das von der Artikulation erzeugt wird, ist Untersuchungsgegenstand der *Akustischen Phonetik*, wobei diese sich sowohl für die produktiven Aspekte des Sprachschalls interessiert, d. h. den Zusammenhang zwischen sprechmotorischen Vorgängen und bestimmten Schallformen, als auch für rezeptive Aspekte, d. h. wie der Hörer bestimmte Schallformen interpretiert (vgl. Neppert und Petursson, 1986, S. 5 ff). Obwohl die Rezeption nicht ganz ausgeklammert werden kann, da die *Wahrnehmung* des Sprecheralters und -geschlechts eine Rolle spielen wird, liegt das Hauptinteresse im Rahmen dieser Arbeit natürlich auf der Produktion.

Der Sprachschall (wie der Schall im Allgemeinen) kann als ein Schwingungsvorgang beschrieben werden, genauer gesagt als auditiv wahrnehmbare Luftdruckschwankungen. Um das Trommelfell in Schwingungen zu versetzen und somit für den Hörer wahrnehmbar zu sein, müssen die Luftdruckschwankungen bestimmten Anforderungen genügen, was die Amplitude (Stärke der Schwankung) und die Frequenz (Geschwindigkeit der Schwankungen) betrifft. Schwingungen, die weniger als 20-mal in der Sekunde und häufiger als 20 000-mal in der Sekunde ablaufen, sind nicht auditiv wahrnehmbar. Für die Amplitude gilt, dass sie mindestens 0.000 000 000 1 bar betragen muss, um wahrnehmbar zu sein. Die Schmerzgrenze liegt bei 0.001 bar. Die Frequenz wird in Hertz (Hz) angegeben (Schwingungen pro Sekunde), die Amplitude in dem logarithmischen Maß Dezibel (dB).

Reine Sinusschwingungen heißen *Töne* und kommen in der Natur praktisch nicht vor. Komplexe Schwingungen, die sich aus Tönen verschiedener Frequenz zusammensetzen, heißen *Klänge*. Voraussetzung für die Entstehung eines Klangs ist jedoch, dass die verschiedenen Frequenzen in einem harmonischen Verhältnis zueinander stehen, d. h. sie müssen jeweils ganzzahlige Vielfache der niedrigsten Frequenz (Grundfrequenz) sein. Ist das Frequenzverhältnis der einzelnen Sinuskomponenten nicht ganzzahlig oder folgen die Amplitudenwerte gar aperiodisch aufeinander, so spricht man von *Geräuschen*. Einen Spezialfall aperiodischer Schwingungen stellen die *Tansienten* oder *Impulse* dar, welche durch plötzlich auftretende, sich nicht wiederholende Luftdruckschwankungen erzeugt werden. Das Prinzip der Zusammensetzung einzelner Töne zu einem Klang wird nach seinem Entdecker, dem französischen Mathematiker Jean Baptiste Joseph Fourier (1768–1830), *Fourier-Synthese* genannt (vgl. Abbildung 2.5).

Ein wichtiges Darstellungsmittel in der Phonetik ist das *Amplitudenspektrum*, d. h. die Analyse der Frequenzkomponenten eines Signal(ausschnitt)s. Da es sich dabei um die Umkehrung des beschriebenen Additionsprinzips handelt, wird dieser Vorgang *Fourier-Analyse* genannt. Das

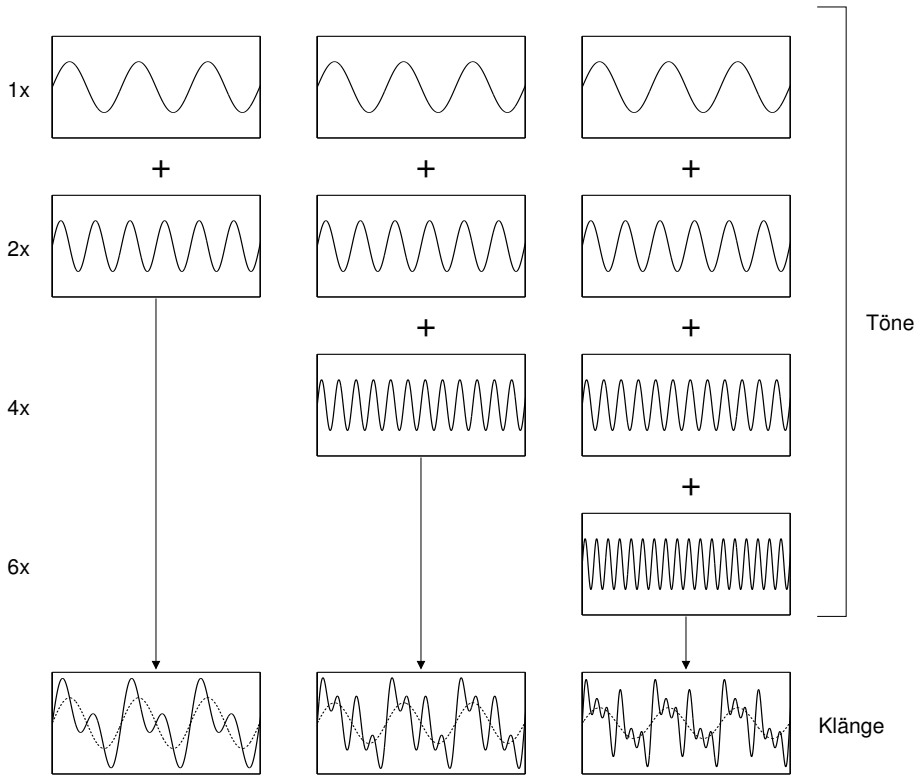


Abbildung 2.5: Fourier-Synthese (vgl. Neppert und Petursson, 1986, S. 29).

Resultat stellt die Amplitude (y-Achse) über der Frequenz (x-Achse) dar (vgl. Abbildung 2.6). Informationen über den zeitlichen Verlauf sind in dieser Darstellungsform nicht mehr enthalten.

2.2.1 Sprachschall

Sprachschall kann vereinfachend in vier Grundschallformen unterteilt werden (vgl. Abbildung 2.7): *Explosionsschall* entsteht bei der Sprengung eines oralen oder glottalen Verschlusses infolge von Überdruck. Diese kurze Schallform ist charakteristisch für alle Arten von Verschlusslauten, z. B. Plosive (vgl. Neppert und Petursson, 1986, S. 70). Das *Frikationsrauschen* wird verursacht durch Turbulenzen, die entstehen, wenn Luft durch eine Engbildung strömt. Es ist charakteristisch für Frikative (vgl. ebd., S.72). Der *Klang*, der auf die Phonation zurückzuführen ist, stellt die Schallform der Vokale, Approximanten und Nasale dar (vgl. ebd., S.73f). Beim *stummen Schall* ist die Signalamplitude gleich Null, d. h. es ist kein Nutzschall vorhanden. Von diesen verschiedenen Schallformen enthalten Klang und Frikationsrauschen die meiste Information über den Sprecher.

Als Modell für sprachlich-akustische Ereignisse hat sich in der Phonetik das *Quelle-Filter-Modell* durchgesetzt, nach dem in einem zweistufigen Prozess zunächst *Rohschall* erzeugt wird, der anschließend modifiziert wird (vgl. Pompino-Marschall, 2003, S. 99 ff). Neben der Phonation

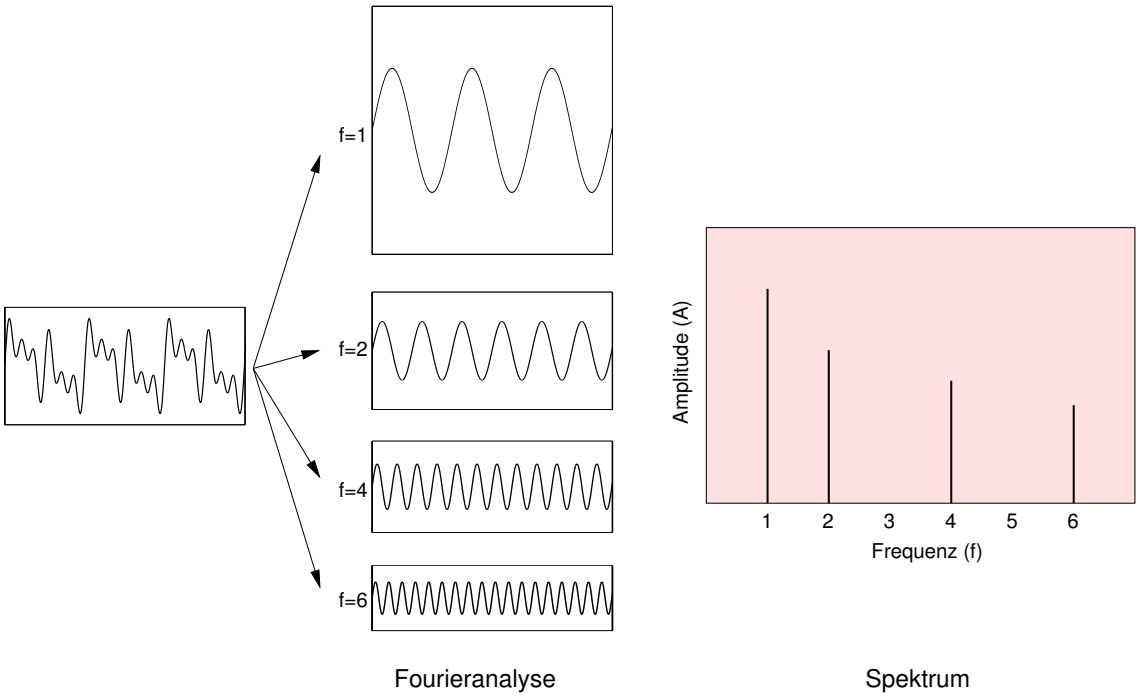


Abbildung 2.6: Fourier-Analyse (vgl. Neppert und Petursson, 1986, S. 31).

kann der Rohschall durch Geräuschbildung infolge einer glottalen Engbildung erzeugt werden, der dann *geräuschförmiger Rohschall* genannt wird. Die Bezeichnung „Rohschall“ kommt daher, dass dieser Schall nie in seiner reinen Form wahrgenommen oder aufgezeichnet werden kann, da das Signal auf seinem Weg durch den Rachen und den Mundraum erheblich verändert wird. In der zweiten Phase der Schallerzeugung, der Modifikation, werden durch einen *Filter* bestimmte Frequenzbereiche (Frequenzbänder) verstärkt und andere abgeschwächt. Welche Frequenzen verstärkt werden, hängt von der Geometrie des Ansatzrohres (Bewegung der Zunge, des Kiefer, der Lippen etc.) ab. Im Klangspektrum nennt man diese Frequenzen *Formanten* (vgl. Abbildung 2.8).

Die Lage dieser lokalen Energiemaxima im Spektrum ist ein wichtiges Merkmal zur Unterscheidung von Vokalen, während zur Unterscheidung von Frikativen eher die globale spektrale Form eine Rolle spielt, d. h. die Energieverteilung in relativ breiten Frequenzbändern. Im Fall der Frikative hat schon das Quellsignal einen wichtigen Anteil an der Form des ausgehenden Spektrums, worin auch der Grund zu sehen ist, warum Frikative neben Vokalen besonders interessant für die Einschätzung von Sprechereigenschaften sind.

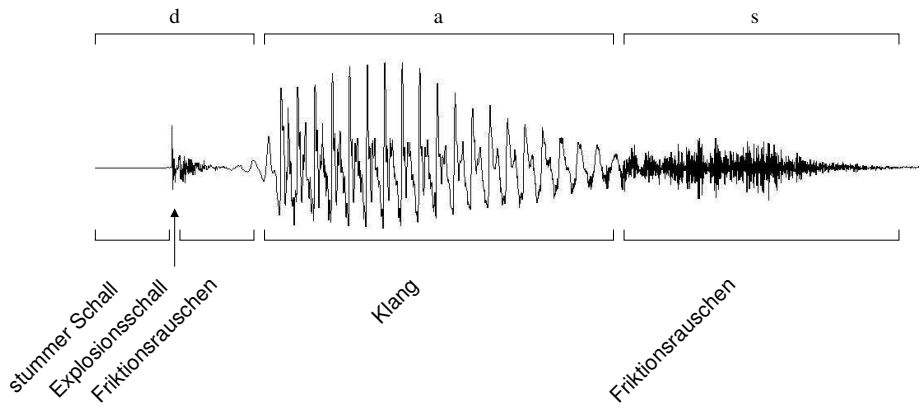


Abbildung 2.7: Die Grundschallformen bei der Äußerung [das].

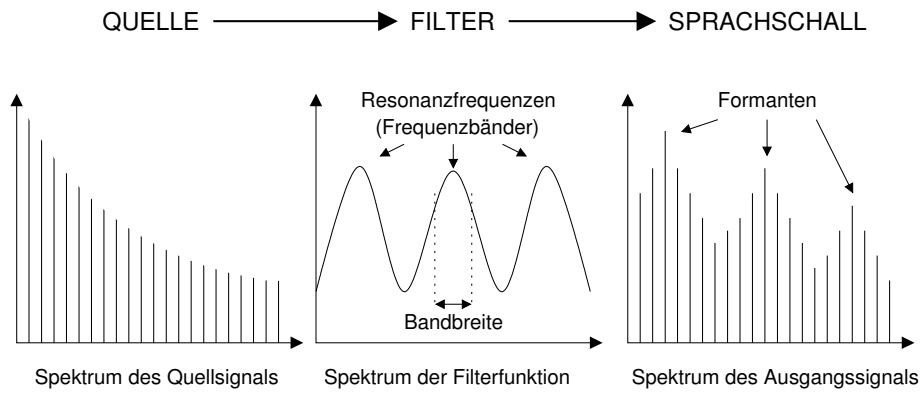


Abbildung 2.8: Formantenbildung im Quelle-Filter-Modell (vgl. Pompino-Marschall, 2003, S. 103).

2.2.2 Digitale Signalverarbeitung

Das Abtasttheorem

Die Digitalisierung des ursprünglich analogen Sprachsignals bringt eine tief greifende Manipulation des Untersuchungsgegenstandes mit sich. Sie erfolgt im Wesentlichen aus zwei Schritten: 1. Das Signal wird in regelmäßigen Zeitabständen abgetastet (*Sampling*), wobei Zeitwerte mit unendlich vielen Nachkommastellen in Zeitwerte mit endlich vielen Nachkommastellen konvertiert werden. Wie oft pro Sekunde abgetastet wird, bestimmt die *Samplingrate*. 2. Das Signal wird *quantisiert*, d. h. Amplitudenwerte mit unendlich vielen Nachkommastellen werden in Amplitudenwerte mit endlich vielen Nachkommastellen konvertiert. Wie genau die Amplitudenwerte konvertiert werden, bestimmt die *Abtasttiefe*. Bei der Wahl der Samplingrate ist zu beachten, dass, um die Periodizität eines analogen Signals mit einer bestimmten Frequenz digital zu erfassen, die

Abtastfrequenz mindestens doppelt so hoch sein muss (vgl. Abbildung 2.9). Die höchste messbare Frequenz wird *Nyquist-Frequenz* genannt und entspricht folglich der Hälfte der Samplingrate.

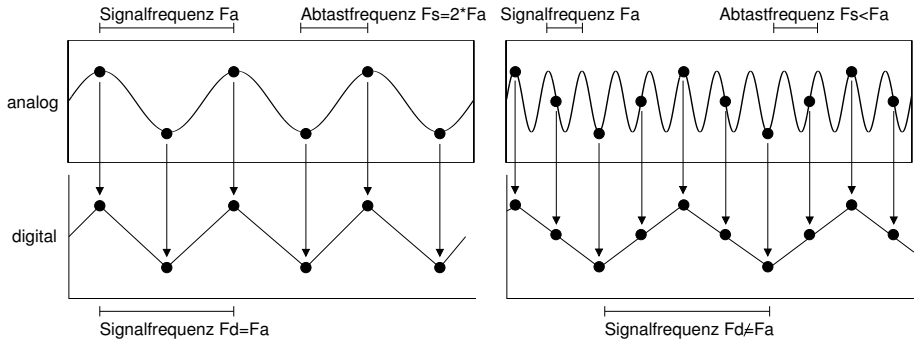


Abbildung 2.9: Links: Digitalisierung eines analogen Signals. Rechts: Die maximale darstellbare Frequenz (Nyquist-Frequenz) entspricht der Hälfte der Abtastfrequenz (Samplingrate). Alle darüber hinausgehenden Frequenzen führen zu Artefakten (Aliasing).

Ob eine gegebene Samplingrate – und damit verbunden die Nyquist-Frequenz – ausreicht, hängt davon ab, welche Komponenten eines komplexen analogen Signals erfasst werden sollen. Frequenzen über 20 KHz müssen für die Untersuchung der Sprache nicht berücksichtigt werden, da unser Gehör diese nicht wahrnehmen kann und sie daher bei der lautsprachlichen Kommunikation sicherlich keine Rolle spielen. Außerdem sehen die technischen Rahmenbedingungen einer automatischen Sprecherklassifikation vor, dass für die Aufzeichnung der Sprache dieselben Geräte benutzt werden wie für die Applikation, in die sie eingebettet wird. Das Telefon z. B. überträgt sogar nur Frequenzen bis 4 KHz, was zwar zu einem qualitativ schlechten, aber durchaus verständlichen Sprachsignal führt.

Frequenzen, die über die Nyquist-Frequenz hinausgehen, müssen vor der Digitalisierung aus dem Signal herausgefiltert werden, da sie ansonsten zu Artefakten führen: Angenommen die Frequenz eines analogen Signals beträgt 15 KHz und die Abtastfrequenz 14 KHz. In so einem Fall würde im digitalen Signal eine Frequenz von 1 KHz auftauchen. Dieses Phänomen wird als *Aliasing* bezeichnet (vgl. Abbildung 2.9). Das Entfernen von hohen Frequenzen erfolgt mithilfe von *Tiefpassfiltern*, die in diesem Fall alle Frequenzen unterhalb der Nyquist-Frequenz passieren lassen, während sie höhere Frequenzen blockieren. Erst nach der Filterung wird das analoge Signal abgetastet und quantisiert. Tiefpassfilter, deren Effekt in Abbildung 2.10 exemplarisch dargestellt wird, spielen nicht nur beim *Anti-Aliasing*, sondern auch bei der sprachlich-akustischen Analyse eine wichtige Rolle.

Die Quantisierung eines analogen Audiosignals bedeutet die Übersetzung einer kontinuierlichen Amplitudenskala in eine diskrete Amplitudenskala mit einer endlichen Menge von Werten. Die Größe der Amplitudenskala, die Abtasttiefe, wird in Bit angegeben. Eine Quantisierung von 16 Bit (65 536 Stufen) ergibt eine sehr fein aufgelöste Skala.

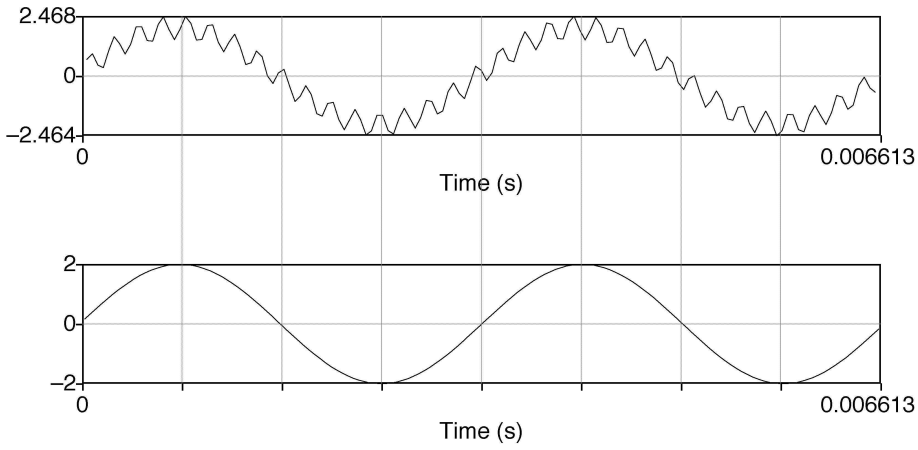


Abbildung 2.10: Tiefpassfilter. Der komplexe Klang oben besteht aus zwei Tönen mit einer niedrigen Frequenz (300 Hz) und einer hohen Frequenz (5 KHz). Unten wird dasselbe Signal nach Anwendung eines Filters mit einer Grenzfrequenz von 3 KHz dargestellt.

Fast Fourier Transformation

Die *Fourier-Analyse* stellt eine der wichtigsten Methoden zur akustischen Analyse von Sprachschall dar. Sie zerlegt das komplexe Sprachsignal in seine Frequenzbestandteile (siehe oben). Bei der *Fast Fourier Transformation* FFT handelt es sich um einen Algorithmus, der die Fourier-Analyse digitaler, d. h. diskreter Signale besonders effizient implementiert. Für die praktische Anwendung ist eine Eigenschaft der FFT von besonderer Bedeutung, nämlich die wechselseitige Abhängigkeit von zeitlicher Auflösung und Frequenzauflösung. Soll nämlich eine von beiden verbessert werden, müssen dabei unweigerlich Einbußen bei der Qualität der zweiten in Kauf genommen werden. Um eine FFT durchzuführen, muss ein *Analysefenster* gewählt werden, da Frequenzen nicht auf Basis einer Momentaufnahme gemessen werden können. Wählt man das Fenster entsprechend groß, können alle Frequenzen von 0 bis zur Nyquist-Frequenz dargestellt werden. Allerdings bedeutet dies, dass alle zeitlichen Aspekte in diesem Fenster verloren gehen. Um dynamische Aspekte berücksichtigen zu können, muss ein kleineres Analysefenster gewählt werden, wodurch die Frequenzauflösung grober wird. Eine gute Zeitauflösung erreicht man z. B. durch die Wahl eines Fensters mit einer Größe von 128 Punkten (bei 22 KHz entspricht dies 5.8 ms), während für eine feine Frequenzauflösung ein Fenster mit 1 024 Punkten geeignet ist (46.6 ms bei 22 KHz).

Berechnung der Grundfrequenz F0

Dieses trade-off bei der Wahl eines geeigneten Analysefensters spielt auch bei der Berechnung der Grundfrequenz F0 eine Rolle, wofür häufig die *Autokorrelationsmethode* angewandt wird (vgl. Boersma, 1993). Dabei wird für eine große Anzahl von Punkten innerhalb des Sprachsignals (stan-

dardmäßig im Abstand von je 0,01 Sekunden) ein kleiner Bereich einer festen Größe betrachtet, der um diesen Punkt zentriert ist. Die Größe des Bereichs hängt von der minimalen zu erkennenden Frequenz ab. Auf das Signal in diesem Bereich wird dann ein *Hanning-Window* angewendet, eine Funktion, die den gleichen Definitionsbereich wie das Teilsignal besitzt und zu den Rändern hin gegen Null konvergiert. Bei Multiplikation des Hanning-Windows mit dem Audiosignal wird dieses an den Rändern abgeschwächt und im Zentrum hervorgehoben. Anschließend wird von beiden Signalen, dem modifizierten Audiosignal und dem Hanning-Window die normalisierte Autokorrelation berechnet und ein Quotient aus beiden gebildet. Das größte der lokalen Maxima des Quotienten kann als ein Bezugspunkt bei der Suche der Grundfrequenz genutzt werden. Da das Signal aber nicht als Funktion vorliegt, sondern aus diskreten Samples besteht, kann diese nur durch eine Interpolation der Bezugspunkte ermittelt werden. Da die Form der Kurve einer sin(x)/x-Funktion entspricht, wird diese zur Berechnung benutzt. Die Tiefe der Interpolation bestimmt dabei die Genauigkeit des Ergebnisses.

2.2.3 Grundlegende Analysemethoden

Spektrographie

Das Spektrum an sich ist eine statische Darstellungsform und enthält keine Informationen über den zeitlichen Verlauf (vgl. Pompino-Marschall, 2003, S. 108 f). In einem Spektrogramm werden jedoch drei Dimensionen (Zeit, Frequenz und Amplitude) in einer zweidimensionalen Grafik dargestellt, indem die Amplitudenwerte durch unterschiedliche Farbgebung gekennzeichnet werden: hohe Amplitudenwerte entsprechen dunklen Grautönen bzw. intensiven Farben (vgl. Abbildung 2.11).

In einem Spektrogramm lassen sich Lautgrenzen an der abrupten Veränderung der Amplituden und der spektralen Struktur erkennen. Vokale zeigen eine harmonische Struktur mit schmalen, horizontalen Schwärzungen, welche den Formanten entsprechen. Vokalische Spektren sind daher als Linienspektren darstellbar. Die erste Linie entspricht der Grundfrequenz (F0), mit der der Laut produziert wurde. Diese korreliert mit der wahrgenommenen Tonhöhe. Die weiteren Linien (die Formanten F_1, F_2, F_3 etc.) entsprechen ganzzahligen Vielfachen der Grundfrequenz. Ihre Lage im Spektrum, die *Formantenstruktur* ist entscheidend für die Vokalqualität. F_1 und F_2 sind die wichtigsten Formanten zur perzeptiven Unterscheidung von Vokalen. Sie liegen bei einem erwachsenen Sprecher unterhalb von etwa 3-3.5 KHz. Höhere Formanten spielen zwar für die Lauterkennung eine untergeordnete Rolle, enthalten jedoch Informationen über individuelle Charakteristika der Stimme. Die genaue Lage einzelner Formanten kann jedoch nur als ungefährer Anhaltspunkt verstanden werden. Wichtiger als die absoluten Werte ist z. B. für die Vokalidentifikation das Verhältnis von F_1 und F_2 zueinander. In Abbildung 2.12 werden die typischen Formantenpositionen für die Vokale /a/ bzw. /i/ im Vergleich zu der „neutralen“ Konfiguration bei der Artikulation eines Schwa-Lautes dargestellt.

Frikative erkennt man an einer breitbandigen Graufärbung, die keine horizontale Struktur aufweist. Verschiedene Frikative unterscheiden sich in ihrer Gesamtintensität: Das alveolare /s/ weist z. B. eine relativ große Gesamtintensität auf, während das labiodentale /f/ und das glottale /h/ eher

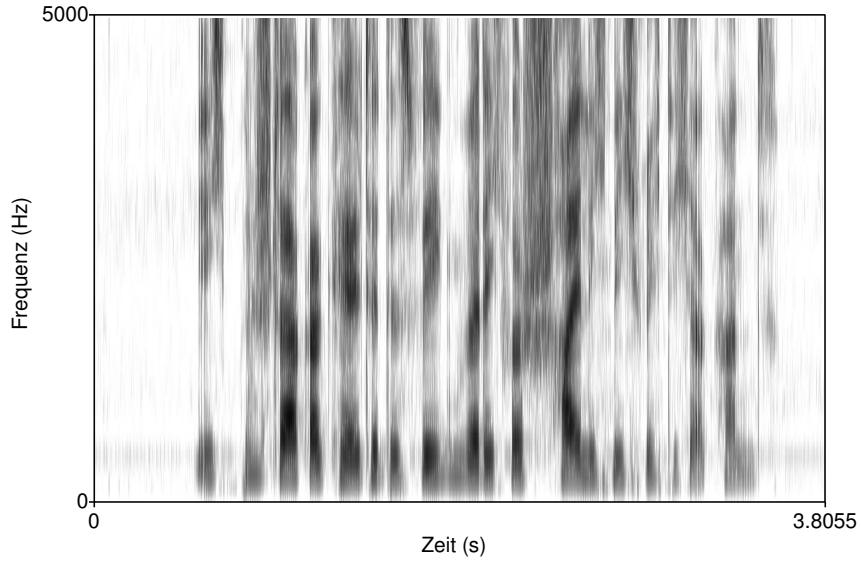


Abbildung 2.11: Spektrogramm: die x-Achse stellt die Zeit dar, die y-Achse die Frequenz; die Färbung entspricht der Amplitude.

schwach ausgeprägt sind. Dieses Merkmal hängt jedoch von vielen anderen Faktoren ab, unter anderem von individuellen Eigenarten des Sprechers. Das akustische Hauptunterscheidungsmerkmal der Frikative ist die Lage der breitbandigen Maxima im Rauschspektrum. Das /s/ weist z. B. unterhalb von ca. 3 KHz eine geringere Intensität auf, die bei etwa 3.5 KHz ansteigt (vgl. Abbildung 2.13 links). Die niedrigere Gesamtenergie des /f/ dagegen ist relativ gleichmäßig verteilt und fällt lediglich zu den höheren Frequenzen leicht ab (vgl. Abbildung 2.13 rechts).

Stimmhafte Frikative weisen prinzipiell das gleiche Rauschspektrum auf wie ihre jeweilige

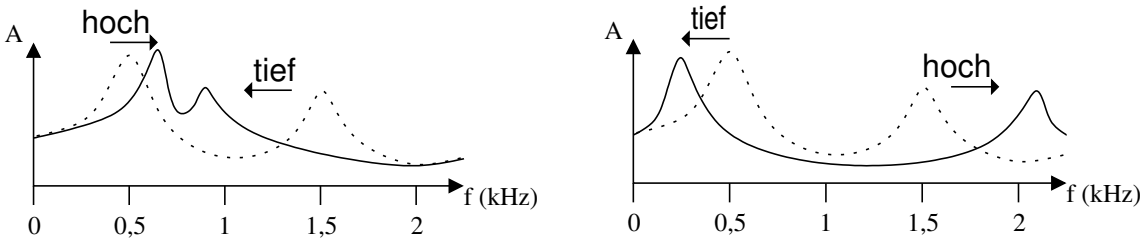


Abbildung 2.12: Links: typische Formantenposition eines /a/. Rechts: typische Formantenposition eines /i/. Die gestrichelten Linien stellen die „neutrale“ Konfiguration dar, die bei der Äußerung eines Schwa-Lautes gegeben ist (vgl. Neppert und Petursson, 1986, S. 122).

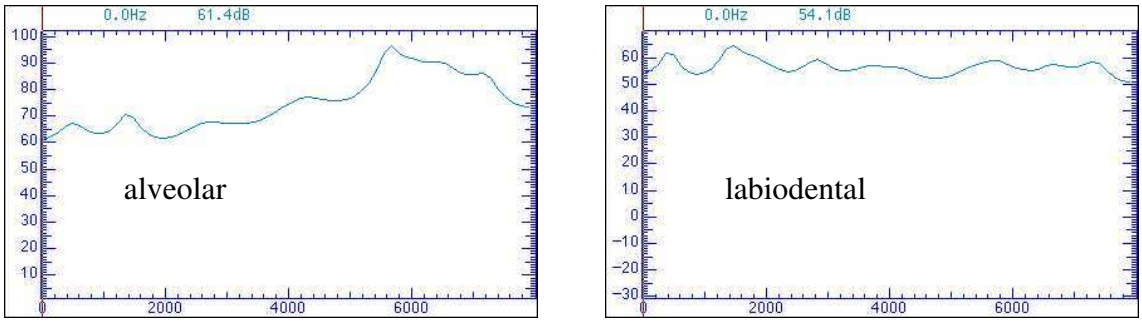


Abbildung 2.13: Links: typisches Spektrum eines alveolaren Frikativs wie z. B. /s/. Rechts: Typisches Spektrum eines labiodentalen Frikativs wie z. B. /f/ (vgl. Neppert und Petursson, 1986, S. 195).

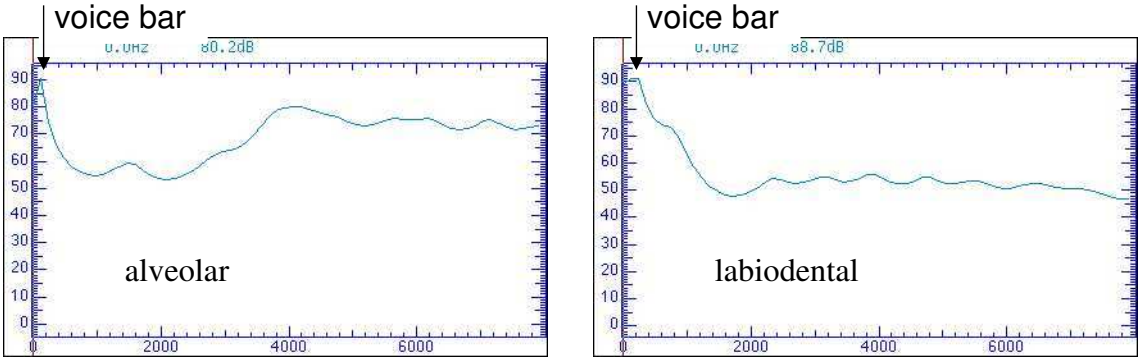


Abbildung 2.14: Links: Typisches Spektrum eines stimmhaften alveolaren Frikativs wie z. B. /s/. Rechts Typisches Spektrum eines stimmhaften labiodentalen Frikativs wie z. B. /f/ (vgl. Neppert und Petursson, 1986, S. 195).

stimmlose Variante. Im unteren Frequenzbereich zeigt sich jedoch ein schmalbandiges Maximum, die *voice bar* (vgl. Abbildung 2.14). Man spricht in diesem Fall von einem Mischspektrum aus harmonischen und nichtharmonischen Teilschwingungen.

Mel-Cepstrum

Der Begriff *cepstrum* ist ein Anagramm des Wortes *spectrum* und bezeichnet ein Maß, welches in der Regel als das „Spektrum eines Spektrums “ paraphrasiert wird. Das cepstrum kann als die Variabilität in den verschiedenen Spektralbändern angesehen werden. Es wurde ursprünglich zur Charakterisierung seismischer Echos entwickelt, kommt heute jedoch hauptsächlich in der Sprachtechnologie als ein Maß zur Repräsentation der menschlichen Stimme zum Einsatz. In dem Fall wird das Spektrum in der Regel zunächst auf die *Mel-Skala* übertragen. Das Resultat wird *mel frequency cepstral coefficients* (MFCCs) genannt (vgl. Holmes und Holmes, 2001, S. 164).

Mel ist eine Maßeinheit für die Tonheit, also die vom Menschen wahrgenommene Tonhöhe (vgl. Zwicker und Fastl, 1999, S. 112). Sie unterscheidet sich von der akustischen Frequenz dahingehend, dass mit ansteigender Frequenz größere Intervalle nötig sind, um eine Verdopplung der wahrgenommenen Tonhöhe zu erreichen. Abbildung 2.15 stellt die wahrgenommene Tonhöhe in Mel als eine Funktion der akustischen Frequenz in Hertz dar. Die Beziehung wurde experimentell ermittelt und entspricht in etwa der Gleichung

$$Mel = 1127.01048 \quad log(1 + Hz/700). \tag{2.1}$$

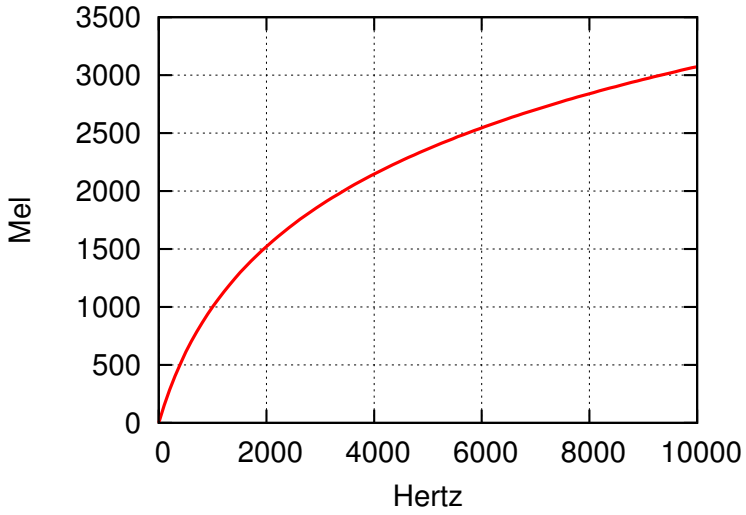


Abbildung 2.15: Die wahrgenommene Tonhöhe in Mel als eine Funktion der akustischen Frequenz in Hertz.

Grundfrequenzkonturen

Die *Sprachmelodie* kann sichtbar gemacht werden, indem aus dem gegebenen Signal in regelmäßigen Zeitabständen Grundfrequenzwerte berechnet und über der Zeitachse dargestellt werden. Dieser Vorgang wird *pitch tracking* genannt. Anhand der entstehenden Kontur kann unter anderem der Stimmumfang als Umfang der Grundfrequenzvariation eines Sprechers gemessen werden. Neben der Unterscheidung von weiblichen und männlichen Sprechern ist der Stimmumfang ein wichtiges Merkmal bei der Einschätzung des Sprecheralters (vgl. Abschnitt 3.3).

2.3 Relevanz für die vorliegende Studie

Der menschliche Stimmapparat weist sowohl geschlechtsspezifische als auch altersabhängige Unterschiede auf, die den gesamten Prozess des Sprechens betreffen, von der Phonation bis zur Artikulation. Diese Unterschiede manifestieren sich im Sprachschall und können somit zur Unterscheidung von Sprecherklassen auf Basis akustischer Merkmale genutzt werden. Die Unterschiede bezüglich der stimmlichen Charakteristika und des Sprechverhaltens zwischen Menschen unterschiedlichen Lebensalters und Geschlechts werden Thema des nachfolgenden Kapitels sein. Anhand einer Literaturübersicht wird ein Inventar von Merkmalen zusammengestellt, anhand derer eine automatische Sprecherklassifikation möglich sein könnte. Zu diesen Merkmalen gehören unter anderem die Grundfrequenz F0, Mikrovariationen von Frequenz (Jitter) und Amplitude (Shimmer) sowie die Formantenstruktur.

Bei der im Anschluss beschriebenen empirischen Studie werden digital aufgezeichnete Sprachproben einer großen Anzahl von Sprechern auf diese Merkmale hin untersucht. Dabei sind die Parameter Abtastfrequenz und Abtasttiefe von Bedeutung, da Merkmale, die Frequenzen oberhalb der Nyquist-Frequenz betreffen, nicht gemessen werden können. Im Hinblick auf die eingangs erwähnten telefonbasierten Anwendungsszenarien wurde die Abtastfrequenz auf 8 KHz und die Abtasttiefe auf 16 Bit beschränkt.

In der Studie wird neben den Sprechercharakteristika auch der (Sprech-) Kontext untersucht. Dies betrifft in erster Linie die Anwendungsszenarien der mobilen sprachbasierten Dialogsysteme, wie z. B. das des *m3i Personal Navigator*, bei dem der Dialog mit dem Benutzer in sehr unterschiedlichen Anwendungsumgebungen stattfinden kann. Ziel dieses Teils der Studie ist es, zu überprüfen, ob anhand der akustischen Merkmale im Signal, die nicht unmittelbar zur Äußerung gehören, Informationen über den Kontext ableitbar sind.

Die übrigen Kapitel von Teil I sind wie folgt gegliedert: In Kapitel 3 wird zunächst eine Übersicht über die zu betrachtenden Phänomene der Stimmentwicklung (bei Kindern), der Stimmalterung (bei Senioren) sowie geschlechtsspezifische Unterschiede gegeben. Auf dieser Grundlage werden in Kapitel 3.4 Hypothesen bezüglich der eigenen Korpusanalysen aufgestellt. In Kapitel 4 werden das verwendete Analyseverfahren sowie die betrachteten Merkmale im Detail beschrieben. Die Ergebnisse werden in Kapitel 5 präsentiert. Kapitel 6 schließlich fasst die gewonnenen Erkenntnisse zusammen und diskutiert ihre Anwendbarkeit für die automatische Sprecherklassifikation, die Thema von Teil II sein wird.

3 Hypothesenbildung

3.1 Kinder und Jugendliche

Nach Fitch und Giedd (1999)[1] unterscheidet sich der Aufbau des Stimmapparats von Kindern unter acht Jahren nicht wesentlich zwischen den Geschlechtern. Somit können auch keine signifikanten Effekte in der Grundfrequenz erwartet werden. Die anatomischen Unterschiede zwischen Frauen- und Männerstimmen entstehen erst in der Pubertät. Diese These wird von Klein (2004) unterstützt, die in einer Untersuchung der Stimmen von drei- bis fünfjährigen Kindern keine signifikanten Unterschiede in der Grundfrequenz zwischen Mädchen und Jungen fand. Obgleich es nach Nairn (1997) bezüglich dieses Ergebnisses einen Konsens zu geben scheint, geht Tanner (1962) davon aus, dass Unterschiede in der anatomischen Struktur zwischen den Geschlechtern von Geburt an vorhanden sind und bis zur Pubertät und darüber hinaus stärker werden. Tanner räumt jedoch ein, dass (in den 1960er Jahren) kein eindeutiger Beleg dieser Geschlechtsunterschiede gefunden werden konnte. Sachs, Liebermann und Erickson (1973) fanden F0 Unterschiede bei Kindern im vorpubertären Alter. Bei einer Untersuchung angehaltener Vokale mit Jungen und Mädchen zwischen vier und vierzehn Jahren fanden sie eine signifikante Differenz: Die Mädchen hatten durchschnittlich eine um 25 Hz geringere Grundfrequenz als die Jungen. Dieses Ergebnis wird von Robb und Simmons (1990) bestätigt, die bei den Mädchen eine durchschnittlich um 10 Hz tiefere Grundfrequenz feststellten. Darüber hinaus untersuchten Sachs et al. die ersten beiden Formantenfrequenzen, wobei sie feststellten, dass die Gruppe der Mädchen durchschnittlich höhere Werte aufwies als die Gruppe der Jungen. Beide Studien wurden allerdings nicht anhand spontansprachlicher Äußerungen evaluiert. Nach Hasek und Singh (1980), die die Aussprache des Vokals /a/ bei 180 Kindern untersuchten, treten geschlechtsspezifische Unterschiede in der Grundfrequenz ab einem Alter von sieben Jahren auf. Bennett und Weinberg (1979) untersuchten neben isolierten Vokalen auch vollständige Sätze bei 73 Kindern. Ihren Ergebnissen zufolge ist die Grundfrequenz der Mädchen um durchschnittlich 16 Hz höher als die der Jungen.

Die Einwirkung verschiedener Hormone bewirken zu Beginn der Pubertät ein Wachstum des Körpers an sich und damit auch ein Anwachsen des Stimmapparates, besonders des Kehlkopfes und der Stimmbänder. Da die Jugendlichen die sich plötzlich verändernde Anatomie nicht un-

[1] Die in diesem Kapitel aufgeführte Literaturübersicht basiert zum Großteil auf der Zusammenstellung, die in Klein (2004) vorgenommen wurde.

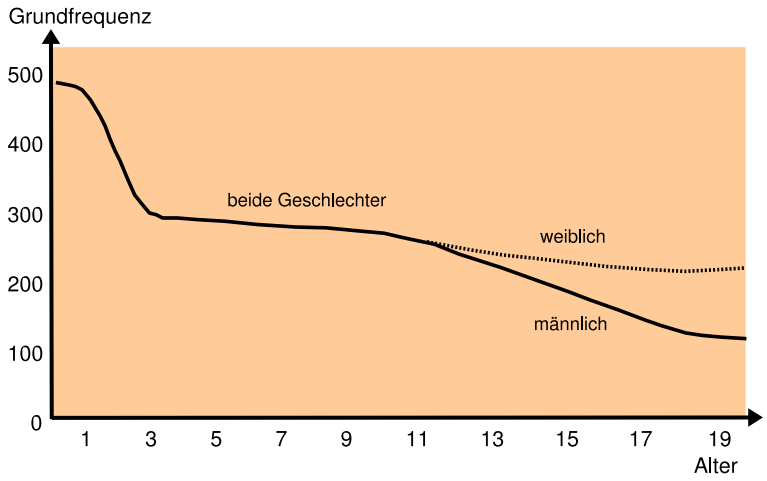


Abbildung 3.1: Entwicklung der Grundfrequenz nach Kent (1976).

mittelbar kontrollieren können, kommt es zum Stimmbruch (vgl. Kent, 1976). Obwohl diese Veränderungen bei beiden Geschlechtern vorkommen, sind sie bei Jungen schwerwiegender als bei Mädchen und können daher auch besser wahrgenommen werden: Während die Grundfrequenz bei den Mädchen nur um etwa ein Drittel sinkt, unterliegt sie bei Jungen einem Abfall von einer Oktave (vgl. Abbildung 3.1). Über die Entwicklung der Grundfrequenz mit zunehmendem Alter gibt es jedoch keinen Konsens, vielmehr konkurrieren nach Nairn (1997) zwei Theorien miteinander: die *lineare Theorie* und die *quantale Theorie* des Wachstums. Die lineare Theorie sagt eine stetige Wachstumsrate der Stimmbänder voraus. Neugeborene haben danach etwa 3 mm lange Stimmbänder, die bei Mädchen 0.4 mm im Jahr und bei Jungen 0.7 mm im Jahr wachsen, bis sie in beiden Fällen mit 20 Jahren die ausgewachsene Länge erreichen. Die quantale Theorie dagegen geht von der Annahme aus, dass die Stimmbänder in den ersten drei Jahren einem sprunghaften Anstieg unterliegen, der für beide Geschlechter gleich ist. Nach einem eher geringen Wachstum in den darauf folgenden Jahren ist ein zweiter Sprung in der Pubertät zu erwarten, der im Gegensatz zu dem ersten bei Jungen ausgeprägter ist als bei Mädchen. Verglichen mit der linearen Theorie sagt die quantale Theorie bis zu einem Alter von etwa acht Jahren etwas längere Stimmbänder voraus. Bei Mädchen ist diese Kurve bis zu einem Alter von etwa fünfzehn Jahren steiler als es nach der linearen Theorie anzunehmen wäre. Für die Jungen sagt die quantale Theorie kein Wachstum zwischen acht und fünfzehn Jahren voraus. Bis zu einem Alter von zwanzig Jahren ist dagegen ein sprunghaftes Anwachsen der Stimmlippen auf die doppelte Länge zu erwarten.

Bezüglich der Artikulationsgeschwindigkeit fanden Karlsson (1987) und Haselager, Slis und Rietveld (1991) keine geschlechtsspezifischen Unterschiede. Karlsson untersuchte die Sprache von Drei- bis Achtjährigen, Haselager et al. die von fünf, sieben, neun und elf Jahre alten Kindern. Allerdings gaben letztere an, dass die Artikulationsgeschwindigkeit bei Kindern mit dem Alter zunimmt.

3.2 Jüngere Erwachsene

Der Kehlkopf von Männern ist in der Regel größer und schwerer als der von Frauen und weist darüber hinaus eine andere Form auf (vgl. Pompino-Marschall, 2003). Die Stimmbänder von Männern und Frauen sind ebenfalls nicht identisch: Die Stimmlippenlänge variiert bei Frauen zwischen 13 und 17 mm und bei Männern zwischen 17 und 24 mm. Infolgedessen weisen weibliche Stimmen eine höhere Grundfrequenz auf als männliche: Im Allgemeinen sprechen Frauen mit einer Grundfrequenz von 120 bis 200 Hz, Männer mit einer Grundfrequenz von 60 bis 120 Hz (vgl. ebd.). Diese Angaben gelten für jüngere, erwachsene Sprecher. Nach Klein (2004, S. 6) bleibt die Grundfrequenz bei Frauen zwischen 16 Jahren und dem Beginn der Wechseljahre (zwischen 40 und 58 Jahren) und bei Männern zwischen 20 und 55 Jahren weitestgehend stabil. Neppert und Petursson (1986) weisen jedoch darauf hin, dass Männer abhängig vom Kontext und emotionalen Zustand durchaus in der Lage sind, mit einer Stimmtonhöhe im Bereich von Frauenstimmen zu artikulieren. Ebenso können Frauen in Abhängigkeit der genannten Faktoren tiefere Töne erzeugen, die in den Frequenzbereich der Männer fallen.

Vergin et al. (1996) fanden Unterschiede zwischen Männer- und Frauenstimmen auch auf Basis der Konstellation der ersten drei Vokalformanten. In Tabelle 3.1 werden die Formantenfrequenzen weiblicher und männlicher Erwachsener für die Vokale /a/, /i/ und /u/ dargestellt. Aus den Daten ist ersichtlich, dass die Frequenzen zwar prinzipiell zur Unterscheidung der Geschlechter geeignet sind, dies jedoch in hohem Maße von der korrekten Identifikation des jeweiligen Vokals abhängig ist.

	F1		F2		F3	
	weiblich	männlich	weiblich	männlich	weiblich	männlich
/a/	850	730	1 220	1 090	2 810	2 440
/i/	310	270	2 790	2 290	3 310	3 010
/u/	370	300	950	870	2 670	2 240

Tabelle 3.1: Formantenfrequenzen weiblicher und männlicher Erwachsener für die Vokale /a/, /i/ und /u/ nach Vergin et al. (1996).

In einer vergleichenden Studie männlicher und weiblicher Normalstimmen fand Pützer (2001) signifikante Unterschiede sowohl in einer Reihe akustischer als auch elektroglottographischer Maße. Die untersuchten akustischen Maße sind im Einzelnen: die Standardabweichung der Grundfrequenz (Frequenztremor), Jitter (Mikrovariation der Frequenz der Stimmlippen-Schwingung), Shimmer (Mikrovariation der Amplitude) und die Harmonicity-to-Noise-Ratio (Verhältnis von harmonischen und nicht harmonischen Anteilen im Signal). Diese Maße betreffen sämtlich die Stimmqualität und werden zu den Indikatoren alternder Stimmen gezählt (vgl. Abschnitt 3.3). In Tabelle 3.2 werden die Mittelwerte einiger der von Pützer untersuchten statistischen Derivate aufgeführt. Bei dem Maß für Jitter handelt es sich um das so genannte *RAP*-Maß (*Relative Average Perturbation*), der Shimmer wird als *APQ* (*Average Perturbation Quotient*) angegeben (vgl. Abschnitt 4.2.1).

	weiblich	männlich
Standardabweichung der Grundfrequenz	3.73	1.87
Jitter (RAP)	0.8	0.52
Shimmer (APQ)	1.91	2.07
Harmonicity-to-Noise-Ratio	0.11	0.12

Tabelle 3.2: Akustische Parameter männlicher und weiblicher Normalstimmen nach Pützer (2001).

Zusammenfassend können diese Werte dahingehend interpretiert werden, dass männliche Stimmen eine höhere Qualität aufweisen als weibliche Stimmen: Die Standardabweichung der Grundfrequenz und die Jitter-Werte sind geringer und der harmonische Anteil ist höher. Lediglich die Amplitudenperturbation ist bei männlichen Stimmen höher als bei weiblichen Stimmen. Sämtliche Tendenzen wurden von Pützer als signifikant angegeben.

In Tabelle 3.3 werden die Werte der untersuchten elektroglottographischen Maße aufgeführt. Diese betreffen zwar nicht unmittelbar akustische Eigenschaften der Stimme, lassen jedoch Rückschlüsse auf die stimmlich-anatomischen Unterschiede zwischen Frauen und Männern zu, z. B. den subglottalen Druck und die Beschaffenheit der Stimmlippen betreffend. Im Einzelnen wurden gemessen: 1. der *Kontaktquotient*, d. h. die Dauer der Verschlussphase im Verhältnis zur Dauer der Periode; 2. der *Skewingquotient*, d. h. die Dauer der Verschlussphase im Verhältnis zur Dauer der Öffnungsphase; 3. der Periodizitätsfaktor, welcher die Gesamtamplitude der periodischen Signalanteile darstellt.

	weiblich	männlich
Mittlerer Kontaktquotient	45.35	44.79
Mittlerer Skewingquotient	45.35	37.79
Periodizitätsfaktor	4.17	3.09

Tabelle 3.3: Elektroglottographische Parameter männlicher und weiblicher Normalstimmen nach Pützer (2001).

Auch in diesem Fall konnten signifikante Unterschiede zwischen männlichen und weiblichen Stimmen nachgewiesen werden, allerdings sind die Tendenzen weniger stark als bei den akustischen Maßen. Pützer bewertet diese Ergebnisse dahingehend, dass akustische Parameter besser als die elektroglottographischen Parameter zur Differenzierung und Interpretation normalstimmlicher Signale geeignet sind.

Nach Byrd (1992) bestehen geschlechtsspezifische Unterschiede bei erwachsenen Sprechern auch bezüglich der Artikulationsgeschwindigkeit. Die Studie wurde auf Basis des *Timit*-Korpus (vgl. Abschnitt 4.1) durchgeführt, welcher zwei so genannte *Kalibrierungssätze* enthält, die für alle Sprecher gleich sind. Die durchschnittliche Dauer dieser Sätze beträgt bei den Männern 2849 ms und bei den Frauen 2676 ms. Die Differenz von 173 ms ist nach Byrd statistisch signifikant.

3.3 Senioren

„[...] and his big manly voice, turning again toward childish treble, pipes and whistles in his sound.“ (Seven stages in a man's life, aus Shakespeare's 'As you like it')

3.3.1 Anatomische und physiologische Veränderungen des Vokaltraktes

Sämtliche Subsysteme des Vokaltraktes sind anatomischen und physiologischen Veränderungen unterworfen, die von dem normalen Prozess des Alterns verursacht werden. Obwohl diese typischen Veränderungen Individuen in unterschiedlichem Maße betreffen, ist es zu erwarten, dass die Stimmen aller Individuen mit zunehmendem Alter anatomische Veränderungen aufweisen (Benjamin, 1988, S. 163).

Das zentrale Nervensystem

Das zentrale Nervensystem ist der Ursprung aller motorischen Bewegungen, inklusive der Sprache. Als Folge dessen spiegeln sich neurologische Veränderungen, die mit dem Fortschreiten des Alters verbunden sind, in der Sprache wider (vgl. Benjamin, 1988, S. 163). Die gesamte Masse des Gehirns schrumpft im Alter um zehn bis fünfzehn Prozent, wobei der hauptsächliche Gewebeverlust Teile der Kortex betrifft, die primär für die Sprache verantwortlich sind (vgl. Benjamin, 1988, S. 164). Es ist sehr wahrscheinlich, dass diese Veränderungen sich auf die Sprachproduktion auswirken.

Des Weiteren wird angenommen, dass mit dem Alter einhergehende Veränderungen im peripheren Nervensystem die Kontrolle der Sprechmotorik älterer Sprecher beeinflussen. Dies wird als eine Ursache der wohl dokumentierten Verringerung der Sprechgeschwindigkeit angesehen (vgl. Linville, 2001, S. 95). Außerdem könnte die verminderte Leitungsfähigkeit der Nerven die Koordination zwischen den Artikulatoren stören (vgl. ebd.).

Sprechatmung

Die Atmungsfunktionen nehmen mit zunehmendem Alter sowohl bei Frauen als auch bei Männern stufenweise ab (vgl. Green, 1982, S. 64). Die Lungen älterer Menschen sind leichter und weniger elastisch und weisen eine nach unten veränderte Lage im Brustkorb auf. Der Elastizitätsverlust des Lungengewebes gilt als eine der signifikantesten altersbedingten Veränderungen der Lunge (vgl. ebd.). Der Brustkorb (Thorax) wird zunehmend steifer, was zu einer geringeren Beweglichkeit führt, insbesondere in Kombination mit einer schwächer werdenden Atemmuskulatur (vgl. ebd.). Als eine Folge dieser Veränderungen ist zu beobachten, dass der expiratorische Fluss und das expiratorische Volumen sowie der intraorale Luftdruck (vgl. Benjamin, 1988, S. 164) mit zunehmendem Alter abnahmen. Des Weiteren ist das Volumen der Residualluft, d. h. der Luft, die nach dem erschöpfenden Ausatmen in der Lunge verbleibt, bei älteren Menschen größer. Die Verminderung der Effizienz der Atmung führt zu Veränderungen der Sprechatmung, die

bereits im mittleren Alter beginnt (vgl. Linville, 2001, S. 117). Dies ist besonders unter extremen Sprechbedingungen erkennbar, in denen eine exzessive, anhaltende Expiration notwendig ist (vgl. Benjamin, 1988, S. 164). Als eine Folge werden oftmals häufiger auftretende Sprechpausen und eine verminderte Lautstärke beobachtet.

Larynx

Der Kehlkopf ist alterungsbedingten anatomischen Veränderungen unterworfen, die sich auf die Stimme älterer Menschen auswirkt (vgl. z. B. Green, 1982, S. 63). Sowohl Verkalkungs- als auch Verknöcherungsvorgänge des Knorpelgewebes wurden festgestellt, ein Vorgang, der nach Benjamin (1988, S. 164) ab einem Alter von 65 Jahren extensiv einsetzt. Ein unvollständiger glottaler Verschluss ist unter bestimmten phonatorischen Bedingungen auch bei jüngeren Erwachsenen zu beobachten, z. B. bei hohen Tonlagen und bei weicher Phonation (Linville, 2001, S. 134). Bei älteren Männern verstärkt der Schwund der Stimmlippen die allgemeine Häufigkeit von glottalen Lücken. Bei älteren Frauen gibt es bezüglich der Häufigkeit der Lücken keinen Unterschied im Vergleich zu jüngeren Frauen, wohl aber in deren Konfiguration: Ältere Frauen tendieren zu vorderen Lücken (vgl. Abbildung 3.2 links), während bei jüngeren Frauen eher hintere Lücken zu beobachten sind (vgl. Abbildung 3.2 rechts).

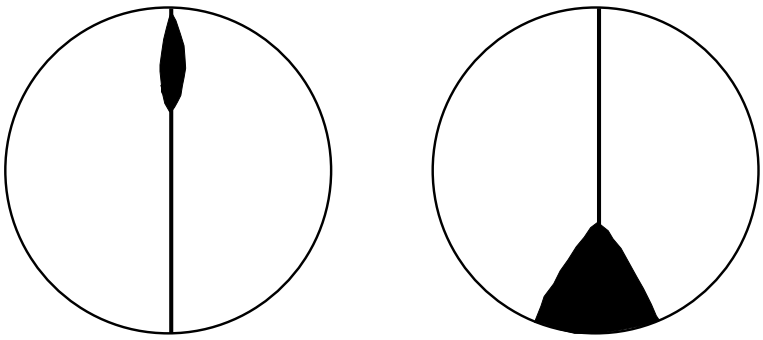


Abbildung 3.2: Glottale Lücken bei älteren Frauen (links) und jüngeren Frauen (rechts) nach Linville (2001, S. 122).

Supralaryngale Mechanismen

Auch im supralaryngalen Bereich, also oberhalb des Kehlkopfes, sind altersbedingte Veränderungen zu verzeichnen: Sowohl strukturelle Veränderungen der Gesichtsknochen als auch der Gesichtsmuskeln wurden festgestellt (vgl. Linville, 2001, S. 72). Obwohl muskelinhärente Faktoren wie Muskelschwächung und -rückbildung, Verlust von Elastizität, verminderte Durchblutung und Bindegewebsrisse sowie umweltbedingte Faktoren zur „Alterung“ der Muskeln beitragen, werden Faktoren, die die Muskelkontrolle durch das neuronale System betreffen, oftmals als die prinzipielle Ursache für die Degeneration angesehen (vgl. Linville, 2001, S. 94). Artikulatorische Ungenauigkeiten wurden im Zusammenhang mit dem Altern in einer Reihe von Studien beobachtet (vgl.

Linville, 2001, S. 153). Männer zeigen dabei einen stärkeren Abfall als Frauen, vor allen Dingen, wenn so genannte stimmlose Intervalle als eine unmittelbare Fehlfunktion des Kehlkopfes gemessen werden (vgl. ebd.). Des Weiteren ist sowohl bei Frauen als auch bei Männern ein Rückgang der oralen Motorkontrolle festzustellen. Anhand von EMG-Messungen konnte festgestellt werden, dass die Lippenmuskelaktivität im Alter zurückgeht, verbunden mit einer reduzierten Kontrolle der Feinmotorik und einer spatiotemporalen Instabilität. Bei beiden Geschlechtern wurden eine verlängerte Phonemdauer sowie langsamere artikulatorische Bewegungen gemessen, was auf die oben genannte verminderte Leistungsfähigkeit der Neurotransmitter zurückgeführt wird. Ältere Sprecher, zumindest ältere Frauen, benutzen einen größeren F0 Bereich, wenn sie betonte Wörter aussprechen, als jüngere. Diese stärkere Kontrastierung wird unter anderem darauf zurückgeführt, dass ältere Sprecher die eigene verminderte Perzeptionsfähigkeit mit ihrer eigenen Artikulation zu kompensieren versuchen (vgl. ebd.). Eine ganze Reihe von Studien haben gezeigt, dass ältere Menschen langsamer sprechen als jüngere (vgl. Linville, 2001, S. 157). Dabei verwenden sowohl jüngere Erwachsene als auch ältere eine langsamere Sprechgeschwindigkeit, wenn die kognitive Belastung ansteigt (z. B. bei der Bildbeschreibung) im Vergleich zu mehr automatischen Sprechaufgaben, wie lautes Lesen. Wenn sie mit ihrer normalen Artikulationsgeschwindigkeit sprechen, sind ältere Sprecher 20-25% langsamer als jüngere Sprecher. Darüber hinaus weisen ältere Sprecher eine um 55% höhere Variabilität auf. Ältere Sprecher produzieren eine konsistent längere Segmentdauer als jüngere Erwachsene. Das gilt für die Satzlänge, die Silbenlänge und die Phonemlänge. Außerdem machen ältere Sprecher mehr Atempausen als jüngere. Bei Frauen wird das Sprechen in geringerem Maße langsamer als bei Männern (vgl. Linville, 2001, S. 158). Die Verlangsamung der Performanz wird als eine der am weitest verbreiteten Veränderungen im Alter angesehen, welche sich sowohl auf die motorische Produktion als auch auf die sensorische Rezeption auswirkt. Der Rückgang der Artikulationsgeschwindigkeit wird zumeist mit Veränderungen im zentralen Nervensystem in Verbindung gebracht. Es wird davon ausgegangen, dass die verringerte Leitungsfähigkeit der Nerven die Phonemrate verringert und die Koordination zwischen den Artikulatoren unterbricht. Eine vermehrte Steifheit der Ausatmungsmuskulatur vergrößert die Pausenzeiten während der Sprachproduktion (vgl. Linville, 2001, S. 159). Eine weitere Hypothese ist die, dass ältere Menschen langsamer sprechen, um die Defizite bei der Verarbeitung der Sprache (Perzeption) zu kompensieren.

3.3.2 Akustische Aspekte alternder Stimmen

Aufgrund der durchgehend stabilen Effekte gilt das Forschungsinteresse hingegen zunehmend den akustischen Veränderungen, die mit dem Altern der Stimme einhergehen (vgl. Linville, 2001, S. 169). In der Stimmforschung ist es in diesem Zusammenhang besonders die Unterscheidung zwischen stimmlichen Veränderungen, die für ältere Stimmen normal sind, und denjenigen, die durch Krankheitsprozesse hervorgerufen wurden (vgl. Ferrand, 2002).

Akustische Stimmmerkmale werden sowohl von phonatorischen als auch von resonatorischen Ereignissen beeinflusst. Die phonatorische Qualität geht einher mit Veränderungen im Kehlkopf oder des Atmungssystems. Besonders akustische Parameter wie die Grundfrequenz F0, die Stabilität der Grundfrequenz und Amplitude, die Intensität der Phonation und spektrales Rauschen

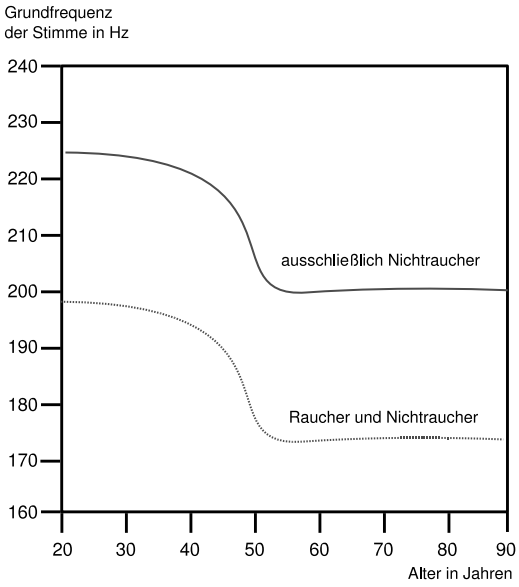


Abbildung 3.3: Grundfrequenz der Stimme bei Frauen als eine Funktion des Alters. Idealisierte Darstellung nach Linville (2001, S. 171).

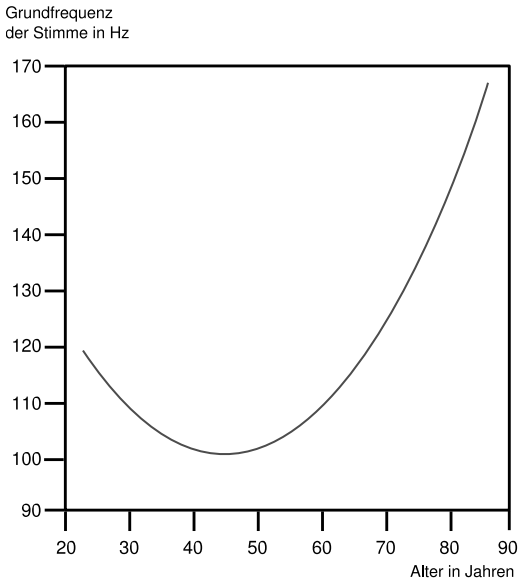


Abbildung 3.4: Grundfrequenz der Stimme bei Männern als eine Funktion des Alters. Idealisierte Darstellung nach Linville (2001, S. 173).

reflektieren die Atmungsfunktionen und/oder phonatorische Mechanismen. Veränderungen des Kehlkopfes und des Atmungssystems mit dem Alter beeinflussen auch die maximale und minimale Performanzfähigkeit der vokalen Mechanismen. Die oberen und unteren Grenzen der Grundfrequenz und Intensität weisen z.T. subtile Veränderungen auf. Durch Messungen dieser Werte lassen sich Alterungsprozesse aufdecken (vgl. Linville, 2001, S. 170).

Die mittlere Sprachgrundfrequenz

Die Grundfrequenz sowohl bei Frauen als auch bei Männern verändert sich von jungen Jahren bis zum Alter. Die Muster sind jedoch bei den beiden Geschlechtern sehr unterschiedlich (vgl. Green, 1982, S. 64).

Die Grundfrequenz bleibt bei Frauen bis zu den Wechseljahren relativ konstant. Dann erfolgt ein Abfall von etwa 10 Hz - 15 Hz. Es wird angenommen, dass dieser Abfall auf hormonelle Veränderungen in den Wechseljahren zurückgeht, wodurch Veränderungen im Kehlkopf ausgelöst werden. Es ist jedoch zu beachten, dass Zigarettenkonsum denselben Effekt hat: Bei Rauchern ist das Niveau der Grundfrequenz über alle Altersgruppen tiefer als bei Nichtrauchern (vgl. Abbildung 3.3).

Bei Männern fällt die Grundfrequenz der Stimme vom jungen Erwachsenenalter bis zum mittleren Alter um etwa 10 Hz ab. Es wird angenommen, dass dies auf normale Abnutzungserschei-

nungen der Stimme zurückzuführen ist (vgl. Linville, 2001, S. 172). Danach steigt die Grundfrequenz signifikant an, und zwar um etwa 35 Hz. Der höchste Punkt wird demnach erst in einem sehr hohen Alter erreicht (vgl. Abbildung 3.4). Dieses Ansteigen wird auf Muskelschwund und eine zunehmende Steifheit der Stimmlippen zurückgeführt (vgl. ebd.). Es gibt jedoch auch gegenteilige Befunde: Hartmann und Danhauer (1976) stellten eine geringere Grundfrequenz bei älteren Männern im Vergleich zu jüngeren fest.

Leistungsmerkmale

Der maximale Stimmumfang (*Maximal Phonational Frequency Range*, MPFR) ist definiert als die komplette Frequenzbandbreite, die ein Mensch produzieren kann, vom tiefsten bis zum höchsten Ton. Alterungsprozesse beeinflussen den Stimmumfang bei beiden Geschlechtern (vgl. Linville, 2001, S. 173). Bei Frauen wird die untere Grenze während der Wechseljahre nach unten verändert, was bedeutet, dass Frauen mittleren Alters tiefere Töne erzeugen können als jüngere und auch als ältere Frauen. Dieser Effekt wird durch hormonelle Veränderungen hervorgerufen, die die Masse der Stimmlippen vergrößern. Größer sind nach Linville jedoch die Veränderungen an der oberen Grenze des Stimmumfangs: Mit zunehmendem Alter können Frauen weniger hohe Töne erzeugen. Männer verlieren im Wesentlichen die Fähigkeit zur Phonation von hohen Tönen.

Jitter und Shimmer

Die Stabilität von Grundfrequenz (F0) und Amplitude (Amp) reflektiert die Regulation und Kontrolle der Stimme und kann somit als ein Aspekt der Stimmalterung angesehen werden. Es wird angenommen, dass die Abnahme der Funktionen mit zunehmendem Alter in einer zunehmenden Instabilität der Stimmlippen-Schwingung resultiert. Dazu tragen Änderungen im Kehlkopf und im Resonanzsystem bei: Degeneration der Kehlkopfmuskulatur und des -bindegewebes sowie Verkalkung und Verknöcherung der Kehlkopfgefäße. Ältere Sprecher zeigen daher eine größere Instabilität in der Grundfrequenz als jüngere Sprecher. Die Maße der Stabilität der Stimmlippen-Schwingung lassen sich in zwei Gruppen aufteilen: Erstens Mikrovariationen zwischen zwei oder mehr Zyklen der Schwingungen (Jitter, Shimmer) und zweitens Makrovariationen, die über eine größere Einheit (z. B. Äußerung) gemessen werden (vgl. Linville, 2001, S. 175).

Wilcox und Horii (1980) fanden in einer Studie mit 20 jüngeren, männlichen Sprechern (Durchschnittsalter 25.8 Jahre) und 20 älteren, männlichen Sprechern (Durchschnittsalter 69.5 Jahre) einen signifikanten Anstieg der Jitterwerte bei der gehaltener Phonation der Vokale /i/, /a/ und /u/. Gleichzeitig wurden jedoch auch starke interindividuelle Unterschiede und vor allem Unterschiede zwischen den einzelnen Vokalen gemessen. Darüber hinaus berichten Huang, Minifie, Kasuya und Lin (1995), dass das Jitter-Niveau bei einer größeren Anstrengung der Stimme sinkt.

Tremor

Die Standardabweichungen von Grundfrequenz und Amplitude reflektieren Makrovariationen und werden häufig als *Frequenz-* bzw. *Amplitudentremor* bezeichnet. Linville (2001, S. 176) sieht diese Maße als geeignetere Diskriminatoren für das Stimmalter an als Jitter und Shimmer. Sowohl

bei Frauen als auch bei Männern wurden zwischen jungen und alten Sprechern substantielle Unterschiede in der Standardabweichung der Frequenz gemessen. Besonders bemerkenswert ist, dass einige der jüngeren Sprecher in den entsprechenden Studien einen Grad an Stabilität erreicht haben, der von keinem der älteren Sprecher erzeugt werden konnte. Bei den Messungen von der Standardabweichung der Amplitude wurden ebenfalls Unterschiede zwischen den Altersgruppen festgestellt, zumindest für die Männer (vgl. ebd.).

Spektrales Rauschen / Harmonizität

Als spektrales Rauschen wird das Vorhandensein von Rauschkomponenten im Spektrum von Vokalen bezeichnet (vgl. Linville, 2001, S. 178). Es wird angenommen, dass spektrales Rauschen von Luftturbulenzen stammt, die durch einen unvollständigen glottalen Verschluss oder irreguläre glottale Zyklen hervorgerufen werden. Das spektrale Rauschen steigt, wenn sich die Stimmlippen während eines Vibrationszyklus nicht mehr vollständig auseinander ziehen.

Das spektrale Rauschen wird in der Regel anhand des Grades der akustischen Periodizität gemessen, wofür häufig der Begriff *Harmonicity-to-Noise-Ratio* (HNR) verwendet wird (vgl. Ferrand, 2002). Die HNR wird in dB ausgedrückt: Wenn 99 % der Energie des Signals periodisch und 1 % Rauschen ist, dann gilt $HNR = 10 * log10(99/1) = 20dB$. Eine HNR von 0 dB bedeutet, dass gleich viele harmonische und nicht harmonische Energie vorhanden ist. Bezüglich des Zusammenhangs von HNR und Alter macht Ferrand (2002) in einer Studie die folgenden Beobachtungen: Junge Sprecher (21-34 Jahre) und Sprecher mittleren Alters (40-63 Jahre) erreichen bei gehalten (verlängert) realisierten Vokalen einen durchschnittlichen HNR-Wert von 7.82 dB bzw. 7.86 dB. Ältere Sprecher (70-90 Jahre) erreichen dagegen einen Wert von nur 5.54 dB.

Bei einem reinen Sprachsignal bzw. einem bekannten Hintergrund-Anteil, kann die HNR als Maß für die Qualität der Stimme herangezogen werden. Boersma (2001) gibt sogar an, dass ein gesunder, junger Sprecher ein gehaltenes /a/ oder /i/ mit einer HNR von etwa 20 dB produzieren kann und ein /u/ mit etwa 40 dB. Nach Boersma entsteht der Unterschied durch die vergleichsweise hohen Frequenzen in /a/ und /i/, welche die HNR anfälliger für Schwankungen macht.

Ferrand postuliert, dass die HNR als Diskriminator für das Sprecheralter besser geeignet ist als beispielsweise Jitter. Als ein Nachteil dieses Maßes kann jedoch die isolierte Betrachtung einzelner, gehaltener Vokale angesehen werden. Es steht zu bezweifeln, dass die Effekte auch dann gemessen werden können, wenn eine vollständige Äußerung betrachtet wird. Selbst bei einer Separation der Vokale liegen diese in der natürlich vorkommenden Länge und nicht in der gehaltenen Variante vor. Weit schwerwiegender ist jedoch der folgende Nachteil: Die Untersuchung des Rauschanteils ist nur unter laborgleichen Bedingungen möglich, da Hintergrundlärm die sprecherabhängigen Effekte eliminiert.

Formantenfrequenzverschiebung

Als eine Folge der oben erwähnten Veränderungen im supraglottalen Vokaltrakt werden auch diejenigen akustischen Eigenschaften der Stimme beeinflusst, die von der Resonanz abhängig sind. Die Formantenfrequenz der Vokale ist ein solches Maß, das von Faktoren wie den Dimensionen

und der Konfiguration des Vokaltraktes beeinflusst wird (vgl. Linville, 2001, S. 179). Es wird angenommen, dass ältere Männer dazu tendieren, die Zungenposition bei der Produktion von Vokalen zu zentralisieren. Dieser Effekt wird als *Schwaisierung* bezeichnet, da die Vokale dem von der Stellung der Artikulatoren her neutralen „Schwa“ ähnlicher werden (vgl. ebd.). Die Formantenfrequenzen älterer Sprecher (inklusive Frauen) reflektieren darüber hinaus eine Verlängerung des Vokaltraktes, d. h. die Formanten F_1 und F_2 sind tiefer als die bei jüngeren Sprechern (vgl. Linville und Fisher, 1985). Insgesamt lässt die Befundlage nach Linville (2001, S. 181) den Schluss zu, dass die Formantenfrequenzen eines derjenigen Charakteristika sind, anhand derer sich ein geschlechtsspezifisches Modell des alternden Vokaltraktes erstellen lässt (vgl. Abbildung 3.5).

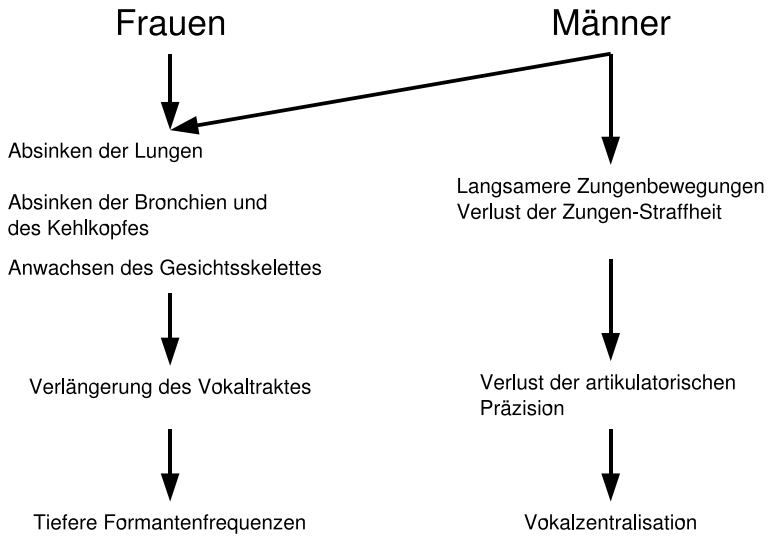


Abbildung 3.5: Modell des alternden Vokaltraktes nach Linville (2001, S. 182).

3.3.3 Linguistische Merkmale des Alterns

Benjamin (1988, S. 171) zählt eine Reihe von Studien auf, die eine Reduktion der syntaktischen Komplexität sowohl bei der Perzeption als auch bei der Produktion von Sprache bei älteren Menschen nachweisen. Auf der semantischen Ebene ist demnach eine Zunahme von syntagmatischen und idiosynkratischen Assoziationen zu verzeichnen. Des Weiteren sind bei älteren Menschen zunehmend Wortfindungsprobleme zu beobachten (vgl. ebd.). Als Ursache für diese Phänomene wird übereinstimmend der bereits erwähnte Rückgang der kognitiven Leistungsfähigkeit mit zunehmendem Alter angenommen.

Linville (2001) führt eine Reihe von Studien über altersbedingte Veränderungen im Sprachfluss auf. Aus diesen geht hervor, dass „formulative“ Störungen des Sprachflusses, d. h. Selbstkorrekturen, Füllwörter und Wiederholungen von Wörtern und Phrasen, insgesamt häufiger und mit dem Alter zunehmend auftreten. „Motorische“ Brüche dagegen, zu denen Wiederholungen von Phonemen und Silben, Phonemverlängerungen und disrhythmische Phonationen gehören, treten

in den untersuchten Gruppen seltener auf, und das Vorkommen ändert sich nicht mit dem Alter (Linville, 2001, S. 162). Diese These wird unterstützt von den Ergebnissen einer Studie sprachlicher Disfluenzen bei Müller et al. (2001), die feststellten, dass die Anzahl der formulativen Störungen des Sprachflusses unter kognitiver Belastung zunimmt, welche, bedenkt man den allgemeinen Rückgang der kognitiven Leistungsfähigkeit im Alter, zu ähnlichen Effekten führen sollte wie die Altersdegeneration.

3.3.4 Methodologische Schwierigkeiten in der Stimmaltersforschung

Trotz der umfangreichen Sammlung von Merkmalen, die Rückschlüsse auf das Sprecheralter zulassen und die in jeder phonetischen Dimension aufzufinden sind, darf nicht außer Acht gelassen werden, dass die deskriptiven Untersuchungen der Sprache älterer Menschen aufgrund methodologischer Restriktionen nur eingeschränkt generalisierbar sind. Benjamin (1988, S. 162) weist dabei auf den so genannten *age cohort effect* (Altersgruppen- oder Jahrgangseffekt) hin. Unter einer Altersgruppe versteht die Autorin Menschen, die in einem bestimmten Zeitraum geboren wurden und daher im Laufe ihres Lebens ähnliche Ereignisse miterlebt haben. Ein Altersgruppeneffekt ist demnach ein Effekt, der nicht durch das biologische Alter eines Individuums, sondern durch die Zugehörigkeit zu der Gruppe erzeugt wurde. Zu den Ereignissen, die so einschneidend sind, dass sie sich auf eine ganze „Generation“ von Menschen auswirken können, zählen z. B. Kriege, Zeiten wirtschaftlicher Depression, Phasen starker Umweltverschmutzung usw. Die Möglichkeit eines Altersgruppeneffektes sollte dann berücksichtigt werden, wenn die Effekte, die gemessen worden sind, direkt auf das biologische Alter abgebildet werden sollen. Eine Verallgemeinerung der Beobachtungen zum Zeitpunkt der Untersuchung auf zukünftige Generationen ist unter Umständen problematisch. Die Inferenzen, die im Rahmen der vorliegenden Arbeit gezogen werden, sind jedoch andere: Im Vordergrund steht weniger der Gewinn an Erkenntnis über stimmliche Faktoren des biologischen Alterns, sondern vielmehr die Zuordnung eines gegebenen Sprechers zu einer bestimmten Altersklasse.

Als weitaus problematischer als der Altersgruppeneffekt ist ein weiteres Problem anzusehen, welches ebenfalls von Benjamin (1988, S. 162) aufgezeigt wird: Die Identifikation von Gemeinsamkeiten innerhalb der Gruppe älterer Sprecher ist deshalb schwierig, weil innerhalb dieser große individuelle Unterschiede auftreten. Unterschiede im Gesundheitszustand und der physischen Kondition tragen zu einer erhöhten Variabilität in der Gruppe der älteren Sprecher bei. Des Weiteren ist es weniger wahrscheinlich, dass sich ältere Sprecher bezüglich ihres Verhaltens ähneln (vgl. ebd.). Es ist also zu erwarten, dass die altersabhängigen Maße, auch wenn sie Unterschiede zwischen den Gruppen erkennen lassen, von starken Abweichungen innerhalb der Gruppe der älteren Sprecher geprägt sind, während die Gruppe der jüngeren Sprecher homogener sein sollte. Die Frage, wie schwerwiegend dieses Problem letztlich für die Aufgabe der Sprecherklassifikation sein wird, müssen die Daten beantworten. Benjamin gibt als Drittes zu bedenken, dass das Kriterium des biologischen bzw. chronologischen Alters bei der Bestimmung der Sprechergruppen möglicherweise nicht angemessen ist, da eine Vielzahl der Effekte durch eine Verminderung des Gesundheitszustandes hervorgerufen wird, die zwar mit zunehmendem Alter voranschreitet, jedoch durch andere Einflüsse verstärkt oder vermindert werden kann. Um diese These

zu untermauern, gibt die Autorin eine Reihe von Studien an, bei denen die physische Kondition einer Versuchsperson einen stärkeren Effekt auf die sprachlichen Eigenschaften hatte als das biologische Alter. Sie schlägt daher vor, die Gruppen eher auf Basis des *wahrgenommenen* Sprecheralters zu definieren, welches durch Perzeptionstests oder Selbsteinschätzung des Sprechers bestimmt wird. Bedenkt man jedoch die Grundfragestellungen der vorliegenden Arbeit, so wird deutlich, dass dies ein unzulässige Vereinfachung darstellt, da z. B. die oben genannten individuellen Unterschiede vermindert werden. Es ist durchaus zu erwarten, dass die resultierenden Modelle eine bessere Genauigkeit in Bezug auf die Trainingsdaten zeigen. Es stellt sich jedoch die Frage, ob eine angemessene Generalisierbarkeit erreicht werden kann, um eine korrekte Klassifikation weiterer Sprecher zu ermöglichen, da das wahrgenommene Alter für die gegebenen Anwendungsszenarien kein geeignetes Charakteristikum darstellt. Die Konfusion von biologischem Alter und physischer Kondition bleibt ein inhärentes Problem bei der Betrachtung des Sprecheralters auf der Basis der Stimme. Wie bei der Diskussion des Altersgruppeneffektes kommt jedoch auch hier das Anwendungsszenario zu Hilfe: Solange keine allgemeinen Schlussfolgerungen auf Alterungsprozesse angestrebt werden, ist die Konfusion von der methodologischen Seite her tolerierbar, da gleichfalls der Gesundheitszustand mit dem biologischen Alter korreliert. Ob die anwendungsspezifischen Implikationen jeweils zulässig sind oder nicht, muss von Fall zu Fall geprüft werden.

3.4 Hypothesen

Die Fragestellungen, die der nachfolgend beschriebenen Studie zugrunde liegen, sind erstens: Welche Unterschiede manifestieren sich in den akustischen Eigenschaften von Sprechern unterschiedlicher Gruppen, die im Einzelnen gebildet werden durch vorpubertäre Kinder unterschiedlichen Geschlechts, vorpubertäre Kinder und Jugendliche; jugendliche Mädchen und Jungen; Kinder/Jugendliche und (jüngere) Erwachsene; Frauen und Männer; jüngere Erwachsene und Senioren; Frauen und Männer im Seniorenalter? Zweitens: Können anhand akustischer Merkmale die Eigenschaften des Sprechkontextes ermittelt werden, und lassen diese Eigenschaften Rückschlüsse darauf zu, in welcher Umgebung sich der Sprecher zum Zeitpunkt der Äußerung befand?

3.4.1 Hypothesen bezüglich des Sprecheralters und -geschlechts

Die Ergebnisse aus den bisherigen Studien zu sprachlichen Indikatoren des Sprecheralters und -geschlechts führen zu einer teilweise widersprüchlichen Befundlage. Dies ist zum Teil darin begründet, dass in den einzelnen Studien sowohl bei der Analyse von Stimmentwicklungsprozessen bei Kindern als auch Stimmalterungsprozessen bei Senioren unterschiedliche Altersspannen betrachtet wurden. Des Weiteren unterscheiden sich die Studien auch im Hinblick auf das verwendete Sprachmaterial: Zum Teil wurden gehalten artikulierte Vokale betrachtet, zum Teil vollständige Äußerungen. Schließlich muss aufgrund der Komplexität der Phänomene ungeachtet dieser methodologischen Unterschiede mit einer hohen inhärenten Varianz gerechnet werden. Dennoch lassen sich in Bezug auf die Stimmentwicklung und -alterung und deren unterschiedliche Ausprägung in Abhängigkeit vom Geschlecht eine Reihe von merkmalsbezogenen Hypothesen ableiten.

Kinder und Jugendliche

Geschlechtsspezifische Unterschiede zwischen Kindern im vorpubertären Alter sind nicht oder nur in geringem Maße zu erwarten. Mit zunehmendem Alter sollte die *Grundfrequenz* sowohl bei Mädchen als auch bei Jungen absinken. Über den *Verlauf der F0-Entwicklung* gibt es jedoch aufgrund konkurrierender Ansichten keine konkreten Hypothesen. Ein Konsens besteht darin, dass die Stimmen von jugendlichen Mädchen und Jungen sich in ihrer Grundfrequenz unterscheiden. Es ist darüber hinaus zu erwarten, dass die *Artikulationsgeschwindigkeit* zunimmt. Zwischen Jugendlichen und Erwachsenen sollte wiederum ein Absinken der Stimmtonhöhe feststellbar sein, wobei davon ausgegangen werden kann, dass sowohl Mädchen als auch Jungen in einem ungefähren Alter von 17 bis 18 Jahren das Erwachsenen-Niveau des jeweiligen Geschlechts erreichen.

Inwiefern die akustischen Maße bezüglich der *Stimmqualität* als Unterscheidungskriterium herangezogen werden können, ist unklar. Die Tatsache, dass die Stimmen von Kindern und Jugendlichen einem Reifungsprozess unterlegen sind, spricht dafür, dass ähnliche Effekte auftreten, wie sie bei älteren Menschen im Zuge der Degeneration des Stimmapparates zu beobachten sind. Über den Vergleich von *Jitter*, *Shimmer* und *Harmonicity-to-Noise-Ratio* von Kindern und Jugendlichen im Vergleich zu Erwachsenen sind jedoch keine Studien bekannt.

Jüngere Erwachsene

Als Hauptunterscheidungskriterium weiblicher und männlicher Erwachsener ist die *Grundfrequenz* zu erwarten, die bei Frauen signifikant höher sein sollte als bei Männern. Darüber hinaus kann angenommen werden, dass sich Frauen und Männer anhand der Merkmale *Jitter*, *Shimmer* und *Harmonicity-to-Noise-Ratio* unterscheiden. Im Einzelnen sollten männliche Stimmen weniger Jitter und eine höhere HNR, allerdings mehr Shimmer aufweisen als weibliche Stimmen.

Senioren

Die Stimmalterungsprozesse sind in einer Vielzahl von Studien analysiert worden und daher können – obgleich die Befundlage z.T. widersprüchlich ist – bezüglich der Tendenzen zwischen jüngeren Erwachsenen und Senioren beider Geschlechter konkrete Hypothesen abgeleitet werden. Bei weiblichen Sprechern sollte die *mittlere Stimmtonhöhe* absinken, wohingegen bei männlichen Sprechern ein Anstieg zu erwarten ist. Was die Verhältnisse dieser Tendenzen betrifft, ist anzunehmen, dass die Effekte bei Frauen stärker sind als die bei Männern. Der Stimmumfang sollte bei den Sprecherinnen geringer werden, da das Absinken der oberen Grenze stärker eingeschätzt wird als das der unteren Grenze. Ein geringerer Stimmumfang ist aufgrund des Verlustes der Fähigkeit zur Phonation hoher Töne auch bei den Männern zu erwarten.

Die *Stimmqualität* bei älteren Sprechern sollte generell geringer sein als bei jüngeren Erwachsenen, wobei bei Männern stärkere Effekte zu erwarten sind als bei Frauen. Das bedeutet im Einzelnen: höhere *Jitter*- und *Shimmerwerte*, eine höhere *Standardabweichung der Grundfrequenz* (Frequenztremor) und eine geringere *Harmonicity-to-Noise-Ratio*.

Die *Artikulationsgeschwindigkeit* sollte bei älteren Sprechern beider Geschlechter geringer sein als bei jüngeren Erwachsenen. Es ist anzunehmen, dass diese Tendenz bei Männern stärker

ausgeprägt ist als bei Frauen.

3.4.2 Hypothesen bezüglich des Sprechkontextes

Die Auswahl der Merkmale, anhand derer die verschiedenen Kontexte untersucht wurden, basiert auf recht einfachen Annahmen: Reine Sprache besteht überwiegend aus periodischen Signalanteilen, was darauf zurückzuführen ist, dass sämtliche Vokale, aber auch einige Konsonanten (vor allem Frikative) harmonische Anteile haben. Die meisten Hintergrundgeräusche dagegen enthalten vorwiegend aperiodische Anteile, was dazu führen sollte, dass bei Anwesenheit eines entsprechenden Hintergrundes die mittlere *Harmonicity-to-Noise-Ratio* geringer ist. Reine Sprache enthält darüber hinaus stets Pausen, seien es Atempausen, Pausen zwischen sprachlichen Einheiten oder Pausen zur Kontrastierung oder Emphasen. Diese stellen Passagen dar, in denen die Intensität sehr gering (im Optimalfall gleich Null) ist. Überlagert die Sprache dagegen einen Kontext, so sollten keine oder zumindest weniger Stellen mit einer solch geringen Intensität vorhanden sein. Als Maß kann hierfür die *minimale Intensität* herangezogen werden bzw. – um Effekte der Grundlautstärke zu vermeiden – das Verhältnis von maximaler und minimaler Intensität.

4 Verfahren

4.1 Datenbasis

Die Analysen wurden auf Basis dreier Korpora durchgeführt: der deutschen Korpora *BAS* (Schiel, 1998) und *Timit* (Garofolo, 1998) sowie des englischen Korpus *Scansoft*. Bei letzterem handelt es sich um eine Datenbasis, die nicht frei verfügbar ist. Sie wurde für dieses Projekt freundlicherweise von der Firma Scansoft[1] zur Verfügung gestellt.

Die Ausgangsdatenbasis, die sich aus den drei genannten Korpora zusammensetzt, umfasst Daten von 1 164 Sprechern (507 weiblichen und 657 männlichen) mit insgesamt 38 202 Äußerungen, von denen 19 839 Äußerungen von weiblichen Sprechern und 18 363 von männlichen Sprechern stammen. In Abbildung 4.1 wird ein Histogramm mit der Anzahl der Äußerungen der Sprecher dargestellt, aus dem hervorgeht, dass von den meisten Sprechern zehn Äußerungen zur Verfügung stehen. In Abbildung 4.2 wird die Anzahl der Äußerungen pro Lebensalter dargestellt.

Aus der Zusammensetzung der Korpora ergeben sich die folgenden methodologischen Probleme: Erstens handelt es sich um Daten unterschiedlicher Sprachen, so dass trotz der zu erwartenden weitestgehenden Sprachunabhängigkeit der betrachteten Merkmale davon ausgegangen werden muss, dass die gemessenen Effekte von Sprecheralter und -geschlecht zumindest teilweise überschrieben werden. Dasselbe gilt für die Tatsache, dass es sich um eine Mischung aus spontansprachlichen Äußerungen und vorgelesenen Sätzen handelt. Hinzu kommt außerdem, dass die Daten aus den einzelnen Korpora mit unterschiedlichen Aufnahmeverfahren gesammelt worden sind, was sich auf die Qualität der Sprachproben auswirkt – wobei diese Unterschiede durch die einheitliche Vorverarbeitung verringert werden (vgl. Abschnitt 4). Bei der Interpretation der Ergebnisse müssen die genannten Gegebenheiten sicherlich berücksichtigt werden, jedoch in der Art, dass bei der Verwendung einer einheitlichen Datenbasis deutlichere Resultate zu erwarten sind. Was die Anzahl der Sprachproben betrifft, wurde für die Korpusanalysen eine ausbalancierte Auswahl getroffen, d. h. eine Auswahl, bei der die Anzahl der Äußerungen pro Altersklasse und Geschlecht identisch ist. In den Fällen, in denen mehr Daten vorhanden waren, wurde eine zufällige Selektion durchgeführt, wodurch eine ungefähre Gleichverteilung der Lebensalter sowie anderer Faktoren wie Äußerungslänge und Dialogsituation (spontan / gelesen) gewährleistet ist.

Die Definition der Altersklassen ist wie folgt: Als KINDER werden Sprecher bis einschließ-

[1] http://www.scansoft.com (01.10.2005).

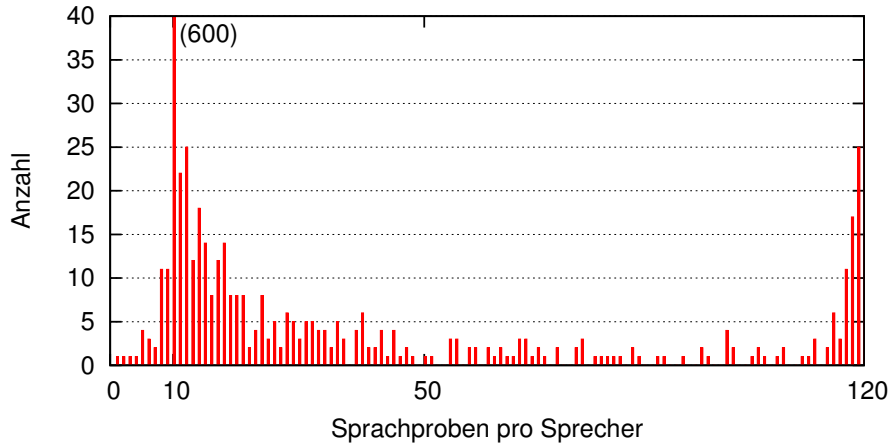


Abbildung 4.1: Histogramm der Anzahl der Sprachproben pro Sprecher.

lich 12 Jahre bezeichnet. Die Klasse JUGENDLICHE umfasst Sprecher von 13 bis einschließlich 19 Jahren. Sprecher zwischen 20 und einschließlich 64 Jahren werden als (jüngere) ERWACHSENE bezeichnet. Ab 65 Jahren gehören die Sprecher der Klasse SENIOREN an. Bei der Interpretation der Ergebnisse bezüglich der Kinder ist jedoch zu bedenken, dass in der zugrunde liegenden Datenbasis keine Sprecher enthalten sind, die jünger als 10 Jahre sind.

4.1.1 Kontext

Anders als bei der Sprecherklassifikation standen für die Untersuchungen des Kontextes keine geeigneten Korpora zur Verfügung. Um dennoch Analysen durchführen zu können, deren Ergebnisse mit denen der Sprecherklassifikation vergleichbar sind, wurden die zugrunde liegenden Sprachdaten künstlich erzeugt. Das bedeutet, dass Sprachproben ohne Kontext (Hintergrund) mit einem reinen Kontext (ohne Sprache) überlagert wurden. Hierfür wurden zunächst in sechs verschiedenen anwendungsnahen Situationen Aufnahmen von mehreren Minuten Länge durchgeführt: an einer Straßenkreuzung (KREUZUNG), auf einem Autobahnparkplatz (AUTOBAHN), in einer Buchhandlung (BUCHHANDLUNG), in einer Bibliothek (BIBLIOTHEK), an einer Baustelle, an der mit einem Presslufthammer gearbeitet wurde (KOMPRESSOR), und in einem Raum, in dem sich mehrere Personen in einiger Entfernung vom Mikrofon unterhielten (STIMMEN). Für die Aufnahmen wurde ein *PDA* der Marke *Hewlett Packard* (Modell *Jornada 560*) verwendet.

In einem ersten Schritt wurde die Intensität der aufgezeichneten Kontexte auf dasselbe Niveau gebracht. Die Überlagerung mit den Sprachproben erfolgte mithilfe von m3i CAT unter Verwendung eines entsprechenden PRAAT-Skriptes (vgl. Kapitel 12). Dabei wurden fünf verschiedene so genannte Überlagerungsfaktoren ausgewählt, mit denen die Intensität des Kontextes multipliziert wurde, bevor die Sprachproben hinzugefügt wurden. Verwendet wurden die Faktoren: 0.05, 0.1, 0.15, 0.2 und 0.25.

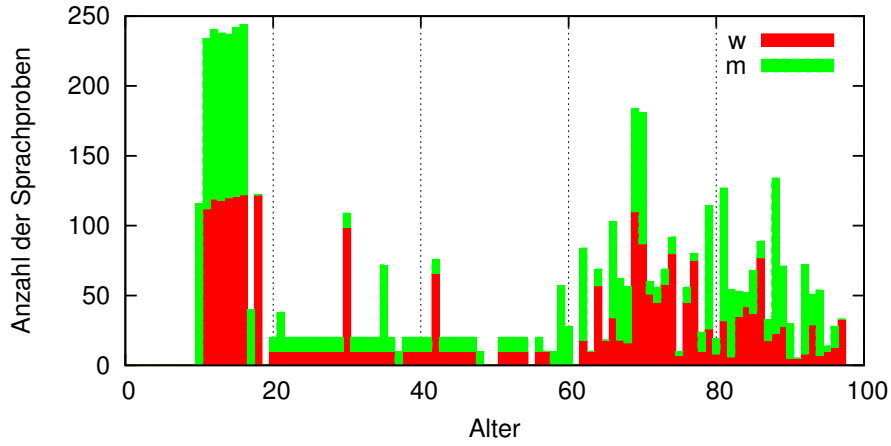


Abbildung 4.2: Anzahl der Sprachproben von Sprecherinnen (rot) und Sprechern (grün) nach Lebensalter.

Mit sechs Kontexten und fünf Überlagerungsfaktoren wurden bei einer Ausgangsmenge von 6300 Sprachproben demnach insgesamt 18 9000 Dateien erzeugt. Dabei wurde intensiv von der Möglichkeit Gebrauch gemacht, mit m3i CAT die Generierung der Dateien sowie die spätere Analyse parallel auf mehreren Rechnern durchzuführen.

4.2 Apparatur

Die Korpora wurden mithilfe des frei verfügbaren Programms SOX auf die Qualität von 8 KHz Abtastfrequenz und 16 Bit Abtasttiefe konvertiert und in dem Format *Sun u-law* abgespeichert. Die Korpusanalyse wurde auf einem *Cluster-Rechner* mit dem Betriebssystem LINUX durchgeführt. Ein Cluster stellt einen Zusammenschluss von gleichartigen Einzelrechnern dar, auf denen eine spezielle Software zur Verteilung der Last installiert ist. Das verwendete LINUX-Cluster bestand aus zehn Knoten mit jeweils zwei Prozessoren (AMD Athlon mit einer Taktfrequenz von 1.5 GHz), so dass die Analysen 20-fach parallel ausgeführt werden konnten. Für die Verteilung der Last wurde das Korpusanalyse-Werkzeug m3i CAT eingesetzt, das im Rahmen dieses Projektes entwickelt wurde und in Kapitel 12 im Detail beschrieben wird. Ein wesentlicher Vorteil von m3i CAT besteht darin, dass eine heterogene Sammlung von Analyseskripten integriert werden kann, deren Ergebnisse dann in einer homogen strukturierten Datenbank gesammelt werden.

Die meisten der betrachteten Sprachmerkmale wurde mithilfe von PRAAT analysiert. PRAAT ist eine frei verfügbare Anwendung zur Sprachanalyse, die 1992 von Paul Boersma und David Weenink vom *Department of Phonetics* an der Universität von Amsterdam geschrieben und seitdem fortlaufend weiterentwickelt wurde. Sie beinhaltet umfangreiche Funktionen und Algorithmen zur Arbeit mit Sprachdateien und erlaubt darüber hinaus das Schreiben eigener Skripte. Mithilfe dieser können neue Funktionen definiert und Sprachproben automatisiert bzw. stapelweise abgearbeitet

werden, wodurch sie sehr gut zur Integration in m3i CAT geeignet sind. Verwendet wurde die LINUX-Version von PRAAT mit der Versionkennzeichnung 4.2.12.
Neben PRAAT wurden die Werkzeuge SRSAD (vgl. Smith et al., 1999) und ENRATE (vgl. Morgan, Fosler und Mirghafori, 1997) verwendet. Mithilfe von SRSAD kann festgestellt werden, ob in einer Aufnahme Sprache enthalten ist oder nicht. Dazu gibt SRSAD eine Liste von Sample-Nummern aus, an denen Sprache beginnt bzw. endet. Tests haben ergeben, dass das System sehr stabil gegenüber Hintergrundlärm ist. ENRATE ist ein Programm, mithilfe dessen die Äußerungsgeschwindigkeit ermittelt werden kann.

4.2.1 Praat-basierte Maße

Pitch

Das Merkmal *Pitch* beschreibt die Grundfrequenz, die in PRAAT mithilfe der Autokorrelationsmethode ermittelt wird (vgl. 2.2.2). Die Angabe des Pitch-Wertes erfolgt in Hertz (Hz). Die folgenden statistischen Derivate der Pitch-Kontur wurden untersucht:

$pitch\_mean$ Das Maß *Mean Pitch* bezeichnet die durchschnittliche Grundfrequenz, die über die gesamte Äußerung ermittelt wird.

$pitch\_min$ *Minimal Pitch* bezeichnet den geringsten Wert der Grundfrequenz.

$pitch\_max$ *Maximal Pitch* bezeichnet den höchsten Wert der Grundfrequenz.

$pitch\_stddev$ *Standard Deviation of Pitch* bezeichnet die Standardabweichung der Grundfrequenz.

$pitch\_mas$ *Mean Absolute Slope* bezeichnet die mittlere Geschwindigkeit der Stimmtonhöhen-Veränderungen. Die Angabe erfolgt in Hz pro Sekunde.

$pitch\_swoj$ *Slope Without Octave Jumps*. Bezeichnet die mittlere Geschwindigkeit der Stimmtonhöhen-Veränderungen, wobei Oktavensprünge nicht berücksichtigt werden, da diese in der Regel auf Artefakte zurückzuführen sind. Die Angabe erfolgt in Halbtönen pro Sekunde.

Jitter und Shimmer

Unter *Jitter* versteht man Mikrovariationen der Frequenz, *Shimmer* bezeichnet Mikrovariationen der Amplitude. Dabei werden direkt benachbarte Schwingungen (Perioden) gemessen und mit einem Durchschnittswert verglichen, welcher je nach Maß aus einem mehr oder weniger großen Bereich vor und nach der aktuellen Periode berechnet wird. Die Maße stammen aus der medizinischen Phoniatrie und werden zur Bestimmung der laryngalen Stimmqualität eingesetzt, wobei in der Regel lang angehaltene Vokale untersucht werden. Bei der Messung von Jitter und

Shimmer auf Basis vollständiger Äußerungen besteht das Problem, dass diese Frequenz- und Amplitudenschwankungen enthalten, die linguistischen Ursprungs sind: Die Intonation ist variant und führt daher zu unterschiedlichen Tonhöhen, was sich in der Dauer der Perioden auswirkt. Dasselbe gilt für die Amplitude: Der Öffnungsgrad des Kiefers bewirkt eine unterschiedlich starke Schallabstrahlung. Es ist daher zu erwarten, dass die relativen Algorithmen zur Einschätzung von Sprechereigenschaften die validesten Maße erzeugen (vgl. Baken und Orlikoff, 2000).

Im Einzelnen bietet PRAAT die folgenden Jitter- und Shimmer-Algorithmen an:

$jitt\_la$	Der *Local Absolute Jitter* bezeichnet die durchschnittliche (absolute) Differenz zwischen zwei aufeinander folgenden Perioden und wird in Mikrosekunden angegeben.
$jitt\_l$	Der *Local Jitter* bezeichnet den absoluten lokalen Jitter geteilt durch die durchschnittliche Periodendauer. Die Angabe erfolgt in Prozent.
$jitt\_rap$	Die *Relative Average Perturbation* bezeichnet die durchschnittliche Differenz einer Periode p und dem Mittelwert aus $p-1$ und $p+1$ geteilt durch die mittlere Periodendauer. Die Angabe erfolgt in Prozent.
$jitt\_ppq$	Der *Period Perturbation Quotient* bezeichnet die durchschnittliche Differenz einer Periode p und dem Mittelwert aus $p-2, p-1, p+1, p+2$ geteilt durch die mittlere Periodendauer. Die Angabe erfolgt in Prozent.
$jitt\_ddp$	Die *Differences of Differences of Periods* bezeichnen die durchschnittlichen Differenzen zwischen aufeinander folgenden Differenzen aufeinander folgender Perioden. Die Angabe erfolgt in Prozent.
$shim\_l$	Der *Local Shimmer* bezeichnet die durchschnittliche (absolute) Amplitudendifferenz aufeinander folgender Perioden geteilt durch die mittlere Amplitude. Die Angabe erfolgt in Prozent.
$shim\_ldb$	Der *Local Shimmer (dB)* bezeichnet den durchschnittlichen Logarithmus zur Basis 10 der Amplitudendifferenz aufeinander folgender Perioden multipliziert mit 20. Die Angabe erfolgt in dB.
$shim\_apq3$	Der *Three-Point Amplitude Perturbation Quotient* bezeichnet die durchschnittliche Amplitudendifferenz einer Periode p und dem Mittelwert aus $p-1$ und $p+1$ geteilt durch die mittlere Amplitude. Die Angabe erfolgt in Prozent.
$shim\_apq11$	Der *Eleven-Point Amplitude Perturbation Quotient* bezeichnet die durchschnittliche Amplitudendifferenz einer Periode p und dem Mittelwert aus $p-5, p-4, ..., p-1, p+1, ..., p+4, p+5$ geteilt durch die mittlere Amplitude. Die Angabe erfolgt in Prozent.
$shim\_ddp$	Die *Differences of Differences of Periods* bezeichnen die durchschnittlichen Differenzen zwischen aufeinander folgenden Differenzen aufeinander folgender Perioden. Die Angabe erfolgt in Prozent.

Harmonicity-to-Noise-Ratio

Bezügliche der Harmonicity-to-Noise-Ratio (HNR), die den Grad der akustischen Periodizität ausdrückt (vgl. Abschnitt 3.3.2), wurden die folgenden statistischen Derivate betrachtet:

$harm\_mean$	Die *Mean Harmonicity-to-Noise-Ratio* bezeichnet den durchschnittlichen Wert über die gesamte Äußerung. Die Angabe erfolgt in dB.
$harm\_min$	Die *Minimal Harmonicity-to-Noise-Ratio* bezeichnet den Minimalwert über die gesamte Äußerung, d. h. den höchsten Rauschanteil. Die Angabe erfolgt in dB.
$harm\_max$	Die *Maximal Harmonicity-to-Noise-Ratio* bezeichnet den Maximalwert über die gesamte Äußerung, d. h. den geringsten Rauschanteil. Die Angabe erfolgt in dB.

4.2.2 Sprechaktivität und abgeleitete Maße

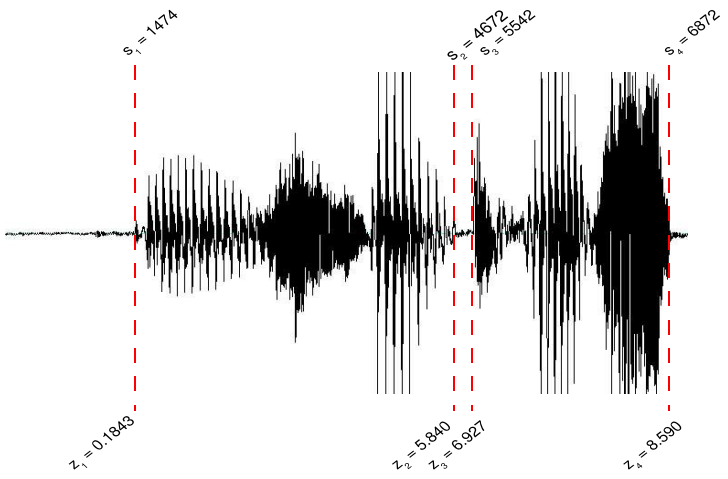


Abbildung 4.3: Oben (s_1 bis s_4): Samplenummern, die von SRSAD ausgegeben werden. Unten (z_1 bis z_4): Angabe in Sekunden.

Die Maße Anzahl und Dauer von Sprechpausen sowie die Einsatzlatenz basieren auf der Erkennung von *Sprechaktivität*, wofür das in Smith et al. (1999) beschriebene System SRSAD (*Syllable Rate Speech Activity Detector*) verwendet wurde, das die Autoren freundlicherweise zu diesem Zweck zur Verfügung stellten. SRSAD klassifiziert Segmente als Sprache bzw. Nicht-Sprache auf der Basis der Silbenrate. Dazu wird das Spektrogramm einer Äußerung nach Energiekonzentrationen im Bereich von 5 KHz untersucht, da die Abfolge von Silben mit ihrem vokalischen Kern in der Regel in dieser Frequenz erfolgt. Bei der verwendeten SRSAD-Version handelt es sich um ein LINUX-Binary, welches für eine gegebene Sprachprobe eine Liste von Sample-Nummern ausgibt,

deren Einträge abwechselnd den Beginn und das Ende eines Sprachabschnittes markieren (vgl. Abbildung 4.3). Auf dieser Basis wurden die folgenden Maße abgeleitet:

$pauses\_onset$ Die *Onset Latency* (Einsatzlatenz) bezeichnet den Wert des ersten Eintrages geteilt durch die Abtastfrequenz. Die Angabe erfolgt in Sekunden.

$pauses\_num$ Die *Number of Pauses* (Anzahl der Sprechpausen) bezeichnet die Anzahl der Bereiche zwischen geraden und ungeraden Einträgen abzüglich des ersten und des letzten geteilt durch die Abtastfrequenz. Es wurden allerdings nur Pausen gezählt, die länger als 200 ms sind, da kürzere Pausen nach Levelt (1989, S. 127) nicht als kognitive Sprechpausen gewertet werden können.

$pauses\_dur$ Die *Duration of Pauses* (Dauer der Sprechpausen) bezeichnet die Dauer der nach dem oben genannten Prinzip gezählten Pausen. Die Angabe erfolgt in Sekunden.

ar_time Die *Articulation Time* (Artikulationszeit) bezeichnet die Dauer der Bereiche zwischen ungeraden und geraden Einträgen. Dieses Maß wird nicht unmittelbar zur Untersuchung von Sprechereigenschaften verwendet, sondern dient als Test der Korrelation anderer Maße mit der Länge der Äußerung. Die Angabe erfolgt in Sekunden.

4.2.3 Äußerungsgeschwindigkeit

Die Äußerungsgeschwindigkeit wurde mithilfe des Systems ENRATE (*energy rate*) ermittelt (vgl. Morgan et al., 1997), das ursprünglich entwickelt wurde, um eine Verringerung der Fehlerrate bei der Spracherkennung zu erreichen. Der ENRATE-Algorithmus basiert auf Merkmalen, die unmittelbar aus dem akustischen Signal abgeleitet werden und stellt somit eine Alternative zu den so genannten *lexikalischen* Maßen der Äußerunggeschwindigkeit dar, die auf der Phonem- oder Silbenrate basieren und erst nach einem initialen Erkennungsdurchlauf angewendet werden können (vgl. z. B. Mirghafori, Fosler und Morgan, 1996). Ein weiterer Nachteil dieser Systeme besteht darin, dass ihre Genauigkeit von der Qualität der Spracherkennung abhängig ist. ENRATE dagegen stellt einen erkennungsunabhängigen Prädiktor der Äußerungsgeschwindigkeit dar, dessen Ergebnisse dennoch in zufriedenstellendem Maße mit denen der lexikalischen Systeme korrelieren.

Der Algorithmus basiert auf der Feststellung, dass das Sprachsignal sich in Abhängigkeit von der Artikulationsgeschwindigkeit signifikant verändert: Die Amplitudenumhüllende variiert beispielsweise bei einer höheren Artikulationsgeschwindigkeit deutlich mehr als bei einer niedrigen, was mithilfe eines Kurzzeitspektrums gemessen werden kann. Dazu führt ENRATE die folgenden Verarbeitungsschritte durch:

1. Das Signal wird mithilfe eines Einweggleichrichters bearbeitet, der die negative Halbwelle unterdrückt.

2. Auf die gleichgerichtete *Waveform* wird ein 16KHz-Tiefpassfilter angewendet.
3. Die Abtastfrequenz wird auf 100 Hz reduziert.
4. Ein Hanning-Fenster von ein bis zwei Sekunden Länge wird auf das Signal angewendet. Die Vorwärtsbewegung erfolgt dabei mit einer deutlichen Überlappung.
5. Durch Anwendung einer Fourier-Transformation wird von jedem Fenster ein Kurzzeitspektrum ermittelt.
6. Durch eine Index-Gewichtung wird die spektrale Bewegung errechnet.

Zur Evaluation des Systems ermittelten Morgan et al. die Korrelation der Ergebnisse von ENRATE mit denjenigen lexikalischer Systeme zur Ermittlung der Äußerungsgeschwindigkeit. Die Analyse wurde auf Basis von 451 segmentierten Äußerungen durchgeführt, die zuvor manuell annotiert wurden, so dass ein phonetischer Abgleich möglich war. Auf dieser Basis wurden sowohl die Phonem- als auch Silbenrate gemessen. Der ermittelte Korrelationskoeffizient beträgt im ersten Fall 0.50 und im zweiten 0.42. Obwohl dies von den Autoren als eine mittlere Korrelation eingeschätzt wird, konnte in einem zweiten Experiment nachgewiesen werden, dass auf Basis der Ergebnisse von ENRATE eine deutliche Verbesserung der Spracherkennung erreicht werden konnte.

4.2.4 Verfahren bei der Kontextanalyse und konkretisierte Hypothesen

In einer Vorstudie wurden alle erzeugten Korpora bezüglich des Merkmals $harm\_mean$ untersucht. Jeder Punkt der in Abbildung 4.4 dargestellten Graphen basiert auf der Analyse von 6 300 Sprachproben.

Wie zu erwarten war, nimmt die *Harmonicity-to-Noise-Ratio* mit zunehmendem Überlagerungsfaktor bei allen Kontexten ab. Der auf der y-Achse markierte Wert von 8.83 dB bezeichnet die *Baseline*, also die durchschnittliche Harmonicity-to-Noise-Ratio der Sprachproben ohne Kontext. Darüber hinaus kann festgestellt werden, dass die verschiedenen Kontexte deutliche Unterschiede aufweisen: Je „geräuschhafter“ der Kontext ist, desto steiler fällt die Kurve ab. Dabei bilden sich zwei Gruppen heraus: STIMMEN, BIBLIOTHEK und BUCHHANDLUNG scheinen eher „harmonische“ Kontexte zu sein, während KOMPRESSOR, AUTOBAHN und KREUZUNG eher unharmonisch sind. Im Folgenden werden die drei zuerst genannten Kontexte unter dem Begriff „voicy“ (von „stimmhaft“ oder „stimmenähnlich“) und die drei zuletzt genannten unter dem Begriff „noisy“ (von „geräuschhaft“) zusammengefasst.

In Abbildung 4.5 werden die Hypothesen bezüglich dieser Kontextklassen und der beiden oben genannten Merkmale zusammengefasst: Ein ruhiger Kontext (QUIET), also reine Sprache, sollte eine vergleichsweise hohe mittlere Harmonicity-to-Noise-Ratio aufweisen sowie eine hohe Intensity-Ratio. Was die Harmonizität betrifft, so sollte sich ein stimmhafter Kontext (VOICY) nicht wesentlich von dem ruhigen Kontext unterscheiden. Dagegen sollte das Intensitätsverhältnis deutlich geringer sein. Ein geräuschhafter Kontext schließlich sollte sowohl eine geringe Harmonicity-to-Noise-Ratio als auch eine geringe Intensity-Ratio aufweisen.

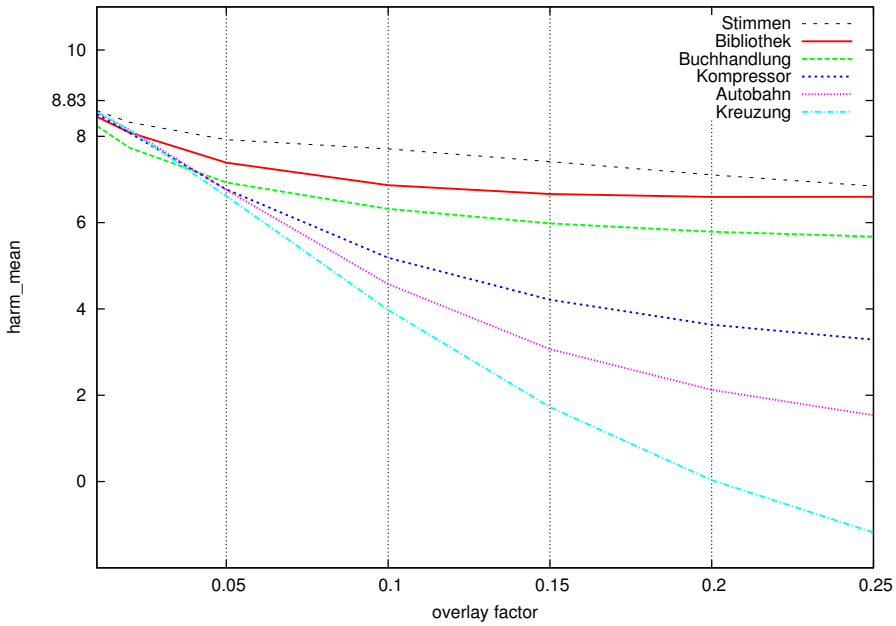


Abbildung 4.4: Analyse der mittleren Harmonicity-to-Noise-Ratio verschiedener Kontexte als Funktion des Überlagerungsfaktors. Der Mittelwert für die Sprachproben ohne Kontext liegt bei 8.83 dB.

Abgesehen von der künstlichen Erzeugung des Korpus war das Verfahren zur Analyse des Kontextes dasselbe, wie es für die Sprechercharakteristika angewendet wurde. Gemessen wurde neben $harm\_mean$ (siehe oben) eine Reihe von statistischen Derivaten der Intensität, die mithilfe von PRAAT ermittelt wurden:

$intens\_mean$ Die *Mean Intensity* bezeichnet den Mittelwert der Intensität. Die Angabe erfolgt in dB.

$intens\_min$ Die *Minimal Intensity* bezeichnet den Minimalwert der Intensität. Die Angabe erfolgt in dB.

$intens\_max$ Die *Maximal Intensity* bezeichnet den Maximalwert der Intensität. Die Angabe erfolgt in dB.

$intens\_ratio$ Die *Intensity Ratio* bezeichnet das Verhältnis von Maximalwert und Minimalwert. Die Angabe erfolgt in dB.

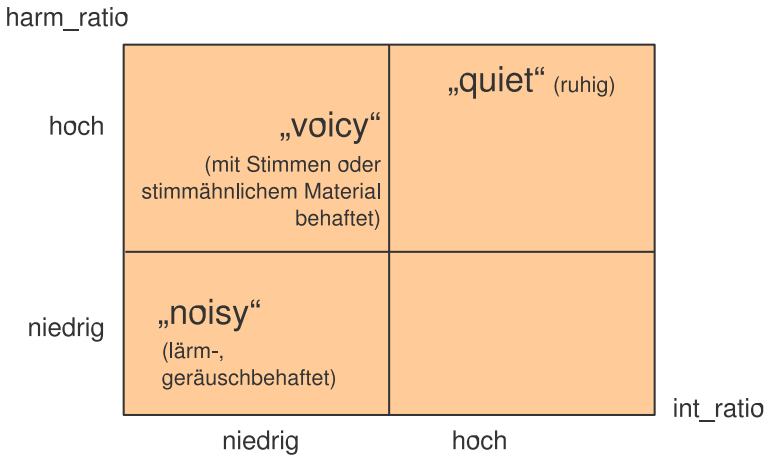


Abbildung 4.5: Kontextklassen.

4.3 Nachbereitung der Daten

Normalisierung

Die Ergebnisse der Korpusanalysen werden unter anderem in Form von normierten Mittelwerttendenzen angegeben, was den Vorteil hat, dass die Effekte verschiedener Maße direkt miteinander verglichen werden können, auch wenn diese eine unterschiedliche Skalierung aufweisen. Die Normierung der Daten wurde nach der folgenden Gleichung vorgenommen:

$$a_i = \frac{v_i - minv_i}{maxv_i - minv_i}, \tag{4.1}$$

wobei v_i den eigentlichen Wert repräsentiert, und $maxv_i$ und $minv_i$ den Maximal- bzw. Minimalwert über die gesamte Trainingsdatenbank. Durch die Normierung werden sämtliche Werte in einen Bereich zwischen -1 und 1 überführt.

Interkorrelationen

Um den statistischen Zusammenhang zwischen den gewonnenen Daten zu ermitteln, wurde der empirische Korrelationskoeffizient (EK) berechnet. Der EK der Stichproben $(x_1, \ldots, x_n)$ und $(y_1, ..., y_n)$ wird beschrieben durch:

$$\rho = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}_n\overline{y}_n}{\sqrt{(\sum_{i=1}^{n} x_i^2 - n\overline{x}_n^2)(\sum_{i=1}^{n} y_i^2 - n\overline{y}_n^2)}}, \tag{4.2}$$

wobei für alle x, y gilt $-1 < EK_{xy} < 1$. Ein maximaler Zusammenhang zwischen x und y besteht, wenn $|EK_{xy}| = 1$; x und y sind statistisch voneinander unabhängig, wenn $|EK_{xy}| = 0$.

Eine (grobe) Klassifikation des Zusammenhanges der Merkmale x und y kann wie folgt beschrieben werden: Ein „schwacher Zusammenhang“ besteht, falls $|EK_{xy}| < 0.5$, ein „mittlerer Zusammenhang“ besteht, falls $0.5 \leq |EK_{xy}| < 0.8$, und ein „starker Zusammenhang“, falls $|EK_{xy}| \geq 0.8$ (vgl. Bosch, 1987).

Der Korrelationskoeffizient von $pitch\_mean$ und $pitch\_quant$ liegt bei 0.97. Bei der Gruppe der älteren Sprecher ist der Zusammenhang zwar mit 0.95 für die Frauen und 0.96 für die Männer etwas geringer als bei der Gruppe der jüngeren Sprecher (0.99 Frauen, 0.97 Männer), insgesamt besteht jedoch erwartungsgemäß ein starker positiver Zusammenhang. Aufgrund der besseren Vergleichbarkeit mit den in der Literatur genannten Maßen wird $pitch\_mean$ in die Auswahl aufgenommen.

	$jitt\_l$	$jitt\_la$	$jitt\_ppq$	$jitt\_rap$	$jitt\_ddp$
$jitt\_l$	1	0.91	0.97	0.99	0.99
$jitt\_la$		1	0.85	0.87	0.87
$jitt\_ppq$			1	0.98	0.98
$jitt\_rap$				1	1
$jitt\_ddp$					1

Tabelle 4.1: Korrelationskoeffizienten der Jitter-Maße auf Basis aller Sprecherklassen.

In Tabelle 4.1 werden die Korrelationskoeffizienten der verschiedenen Jitter-Maße dargestellt. Es ist festzustellen, dass zwischen den Maßen durchweg eine starke Korrelation besteht. Lediglich zwischen dem absoluten $jitter\_la$ und den übrigen Maßen besteht ein vergleichsweise geringer Zusammenhang, was darauf zurückzuführen ist, dass letztere relative Werte repräsentieren. Die relativen Werte sind jedoch aufgrund der geringeren Abhängigkeit von globalen Frequenzschwankungen dem absoluten $jitter\_la$ vorzuziehen. Die Diskussion der Ergebnisse konzentriert sich auf $jitt\_rap$, weil dieses als das bekannteste Jitter-Maß gilt (vgl. Baken und Orlikoff, 2000).

	$shim\_l$	$shim\_ldb$	$shim\_apq3$	$shim\_apq11$	$shim\_ddp$
$shim\_l$	1	0.91	0.91	0.85	0.91
$shim\_ldb$		1	0.77	0.81	0.77
$shim\_apq3$			1	0.67	1
$shim\_apq11$				1	0.67
$shim\_ddp$					1

Tabelle 4.2: Korrelationskoeffizienten der Shimmer-Maße auf Basis aller Sprecherklassen.

Tabelle 4.2 stellt die Korrelationskoeffizienten der Shimmer-Werte dar, die mit durchschnittlich 0.83 geringer sind als diejenigen der Jitter-Werte (0.94). Die geringste Korrelation besteht zwischen $shim\_apq11$ und $shim\_apq3$ bzw. $shim\_ddp$. Im Hauptteil werden die Ergebnisse bezüglich $shim\_aqp3$ präsentiert, die übrigen Shimmer-Maße werden im Anhang aufgeführt.

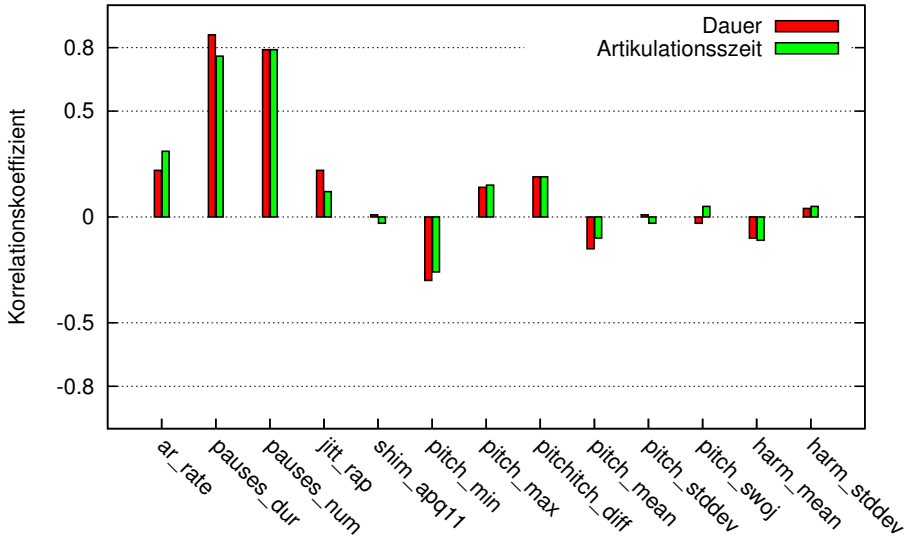


Abbildung 4.6: Korrelation der Merkmale mit der Äußerungslänge (rot) und der Artikulationszeit (grün).

Korrelation mit der Äußerungslänge

Da die Sprachproben, die zur Analyse ausgewählt wurden, unterschiedlich lang sind, sollte der Einfluss der Äußerungslänge auf die Merkmale im Optimalfall sehr gering sein. Dieser Einfluss kann ebenfalls anhand des Korrelationskoeffizienten bestimmt werden: Eine positive Korrelation bedeutet, dass die Werte des entsprechenden Maßes mit zunehmender Äußerungslänge größer werden. Bei einer negativen Korrelation werden sie kleiner. Besteht eine starke (positive oder negative) Korrelation, sollte das Maß nur dann für die Unterscheidung der Klassen herangezogen werden, wenn die Länge der Sprachproben zuvor normiert worden ist. Alternativ dazu kann auch ein Quotient aus Wert und Äußerungslänge gebildet werden.

In Abbildung 4.6 wird die Korrelation aller Merkmale mit der Äußerungslänge (rote Balken) und der Artikulationszeit (grüne Balken) dargestellt. Die Artikulationszeit berechnet sich aus Äußerungslänge abzüglich der Einsatzlatenz und Pausen. Alle Stimmmerkmale sowie die Artikulationsgeschwindigkeit sind nur in geringem Maße von der Äußerungslänge abhängig. Die Anzahl und Dauer der Pausen zeigen jedoch eine mittlere positive Korrelation und müssen daher in Relation zur Äußerungslänge betrachtet werden. Das Merkmal $pauses\_num$ wird durch die Dauer der Äußerung geteilt und ergibt $pauses\_numps$ („number per second“). Das Merkmal $pauses\_dur$ wird durch die Anzahl der Pausen geteilt und ergibt $pauses\_durpp$ („duration per pause“). Aus Abbildung 4.7 ist ersichtlich, dass die Korrelation der neuen Merkmale mit der Äußerungslänge respektive der Artikulationszeit deutlich geringer ist.

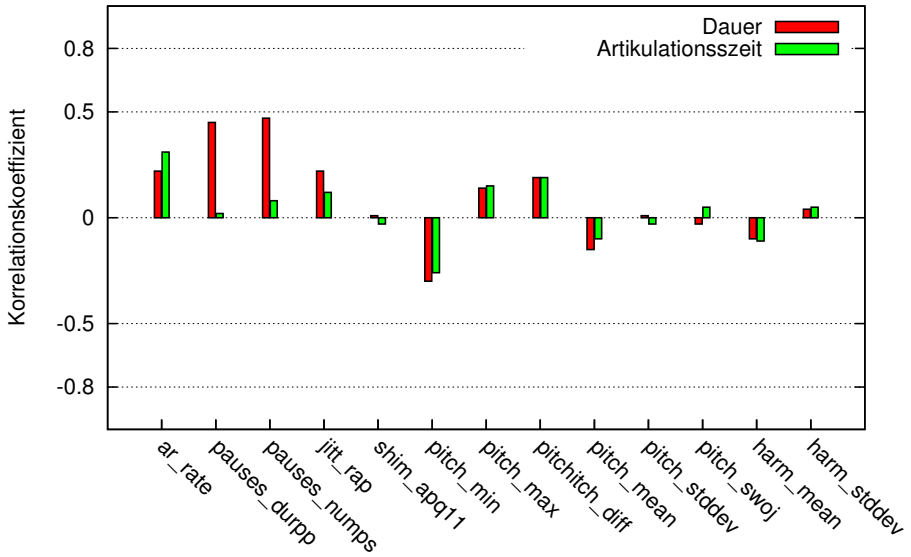


Abbildung 4.7: Korrelation der Merkmale mit der Äußerungslänge nachdem die Anzahl und Dauer der Pausen in Relation zur Äußerungslänge gesetzt wurden.

5 Ergebnisse

Die Ergebnisse der Korpusanalysen werden in verschiedenen Darstellungsformen präsentiert. Die Unterschiede zwischen den einzelnen Klassen – z. B. zwischen Erwachsenen und Senioren oder zwischen Frauen und Männern – werden in Form von normierten Mittelwerttendenzen dargestellt. Auf diese Art und Weise kann die Stärke der Effekte unmittelbar an der Steigung bzw. dem Gefälle der Linien abgelesen werden. Da die normierten Werte in einem Bereich zwischen -1 und 1 liegen, beträgt die theoretisch stärkste Tendenz $\pm 1.\bar{9}$. Da dies jedoch bei echten Daten nicht vorkommt, können Tendenzen ab einem Betrag von etwa 0.8 als stark eingestuft werden. Falls nicht anders angegeben, sind sämtliche Mittelwertunterschiede statistisch signifikant (t-Test, $p < 0.01$). Falls das nicht zutrifft, werden sie mit „(ns)“ markiert.

Darüber hinaus werden für jedes Merkmal die klassenspezifischen univariaten Gauß'schen Wahrscheinlichkeitsdichten dargestellt, da diese eine Beurteilung der Unterscheidbarkeit der Klassen ermöglichen. Die probabilistischen Eigenschaften der Wahrscheinlichkeitsdichten werden in Abschnitt 7.2 beschrieben. Die Spitze der Glockenkurve stellt den klassenspezifischen Mittelwert des betrachteten Merkmals dar. Die oberen und unteren Ränder entsprechen etwa dem Mittelwert $\pm$ der Standardabweichung multipliziert mit zwei.

5.1 Sprecheralter

5.1.1 Jitter

Unterschiede in den Jitter-Werten – hier am Beispiel von $jitt\_rap$ – gibt es hauptsächlich zwischen der Klasse ERWACHSENE und der Klasse SENIOREN, wo für beide Geschlechter eine starke positive Tendenz zu beobachten ist, obgleich diese bei den Frauen mit 1.022 etwas stärker ist als bei den Männern (0.805). Bei den Jugendlichen ist ebenfalls ein leichtes Ansteigen des Jitters gegenüber den Kindern zu verzeichnen. In diesem Fall sind es jedoch die männlichen Sprecher, die mit 0.281 eine stärkere Tendenz aufweisen als die weiblichen mit 0.219. Möglicherweise wirkt sich der Umstand, dass die Sprecher die starken anatomischen Veränderungen während der Pubertät nicht kontrollieren können, dahingehend auf die Stimme aus, dass ein erhöhter Jitter-Wert messbar ist. Für diese Annahme spricht erstens, dass der Effekt bei den Jungen stärker ausgeprägt ist als bei den Mädchen, und zweitens, dass bei beiden Geschlechtern die Werte bei der Gruppe

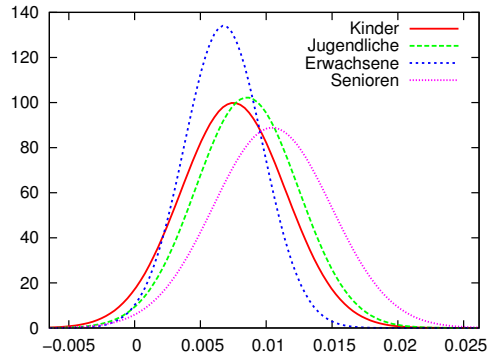
der Erwachsenen geringer sind.

Der Blick auf die Gauß-Verteilungen verrät allerdings, dass es eine starke Überlappung der klassenspezifischen Wahrscheinlichkeitsdichten gibt: Sowohl in der allgemeinen Sicht als auch in den beiden nach Geschlecht differenzierten Bildern ist zu erkennen, dass die Klasse der jüngeren Erwachsenen nicht nur die der Kinder fast vollständig überdeckt, was anhand der Mittelwertunterschiede zu erwarten war, sondern auch diejenige der Jugendlichen. Lediglich die Glockenkurve der männlichen Jugendlichen rückt etwas nach rechts heraus, gerät dort aber in den Überschneidungsbereich der Klasse SENIOREN.

Diese Ergebnisse legen den Schluss nahe, dass das Merkmal Jitter sinnvoll nur zur Unterscheidung der Gruppe KINDER/JUGENDLICHE/ERWACHSENE von der Klasse SENIOREN zu verwenden ist. Das entspricht den Erwartungen, da die Frequenzperturbation als ein Anzeichen der mit dem zunehmenden Alter einhergehenden Degeneration der Stimme angesehen wird. Die Tendenzen für die genannte Gruppierung werden in Anhang A (Seite 267) dargestellt: Zwischen der Altersgruppe KINDER/JUGENDLICHE/ERWACHSENE und der Klasse SENIOREN besteht eine mittlere positive Tendenz. Entgegen den Erwartungen ist die Tendenz allerdings bei den Frauen mit 0.776 höher als bei den Männern (0.544).

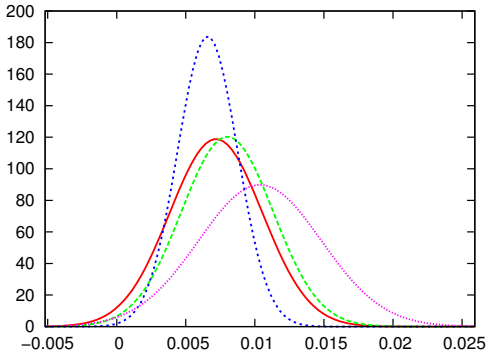
Die Tendenzen der übrigen Jitter-Werte, deren Analysen in Anhang A (Seite 267) aufgeführt werden, sind ähnlich: Bei $jitt\_l$ sind die geschlechtsspezifischen Unterschiede etwas deutlicher ausgeprägt, genauso bei $jitt\_ppq$. Bei $jitt\_ddp$ gibt es keinen nennenswerten Unterschied, was aufgrund des Korrelationskoeffizienten von 1 mit $jitt\_rap$ nicht überraschend ist. Eine Ausnahme stellt der absolute lokale Jitter $jitt\_la$ dar, der bei den weiblichen Sprechern eine schwache positive Tendenz von KINDER über JUGENDLICHE nach ERWACHSENE aufweist und dann eine starke positive Tendenz nach SENIOREN. Bei den männlichen Sprechern sind die Tendenzen ebenfalls durchweg positiv, allerdings stärker zwischen den Kindern und den Jugendlichen und schwächer zwischen den jüngeren Erwachsenen und den Senioren. Bei der Interpretation von $jitt\_la$ muss jedoch berücksichtigt werden, dass es sich um ein absolutes Maß handelt, welches in stärkerem Maße mit globalen Veränderungen der Frequenz korreliert.

5.1 Jitter — $jitt\_rap$

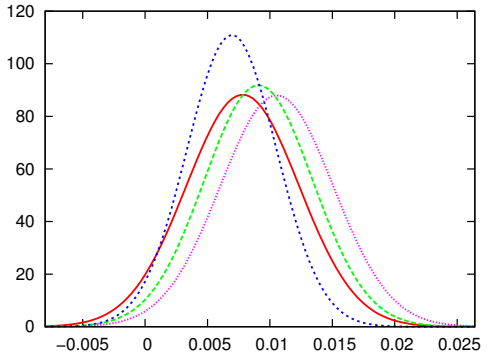


	μ	Σ	$\bar{\mu}$
gesamt	0.0083	0.0041	0
Kinder	0.0075	0.004	-0.197
Jugendliche	0.0085	0.0039	0.0547
Erwachsene	0.0068	0.003	-0.3774
Senioren	0.0105	0.0045	0.5198

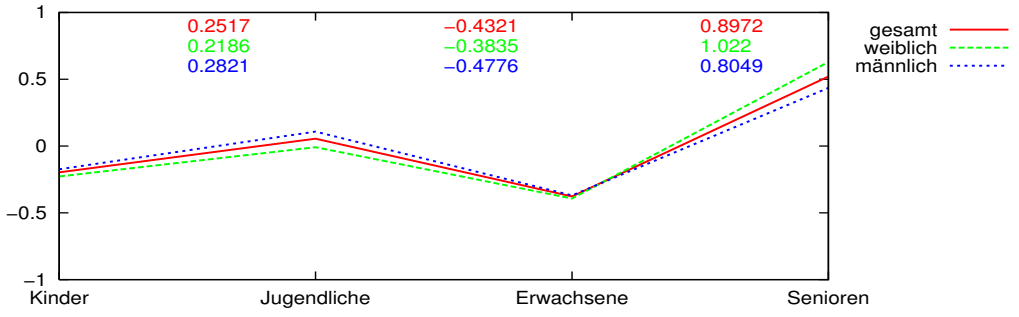
5.1 Jitter (Fortsetzung) *jitt_rap*



	weibliche Sprecher		
gesamt	0.008	0.0037	0
Kinder	0.0072	0.0034	-0.2277
Jugendliche	0.008	0.0033	-0.0091
Erwachsene	0.0066	0.0022	-0.3926
Senioren	0.0104	0.0044	0.6294



	männliche Sprecher		
gesamt	0.0086	0.0045	0
Kinder	0.0078	0.0045	-0.174
Jugendliche	0.0091	0.0043	0.1081
Erwachsene	0.0069	0.0036	-0.3695
Senioren	0.0105	0.0045	0.4354



5.1.2 Shimmer

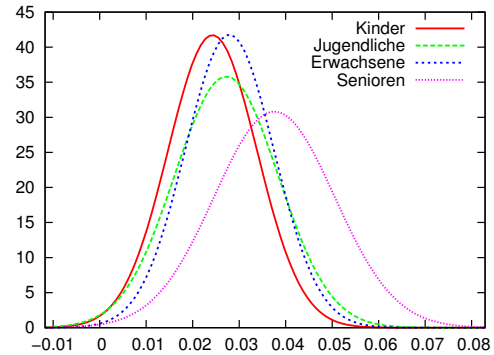
Die Ergebnisse bezüglich der Amplitudenperturbation $shim\_apq3$ bestätigen die Erwartungen: Bei den männlichen Sprechern gibt es einen linearen Anstieg von KINDER bis SENIOREN, der mit zunehmendem Alter stärker wird. Bei den weiblichen Sprechern ist eine mit -0.27 verhältnismäßig schwache negative Tendenz zwischen JUGENDLICHE und ERWACHSENE zu verzeichnen, der eine mit 1.14 recht starke positive Tendenz zur Klasse SENIOREN folgt.

Die Wahrscheinlichkeitsdichte für weibliche Sprecher verheißt eine gute Unterscheidbarkeit zwischen der Gruppe KINDER/JUGENDLICHE/ERWACHSENE und der Klasse SENIOREN. Bei den

männlichen Sprechern ist darüber hinaus die Kurve der Klasse KINDER von derjenigen der Klasse ERWACHSENE deutlich zu unterscheiden. Daher kann das Merkmal auch zur Differenzierung von (männlichen) Kindern, Erwachsenen und Senioren einen Beitrag leisten.

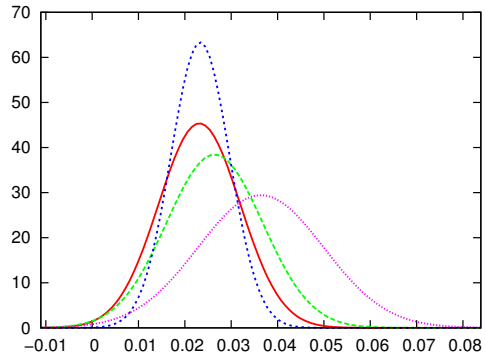
Während das (stark mit diesem korrelierende) Maß $shim\_ddp$ ähnliche Tendenzen aufweist, zeigt sich im Fall von $shim\_aq11$, $shim\_l$ und $shim\_ldb$ ein anderes Bild (vgl. Anhang A, Seite 274). Während es bei den Frauen auch dort einen gleichmäßigen Anstieg gibt, wird dieser bei den Männern durch eine Spitze bei den jüngeren Erwachsenen unterbrochen. Ein erhöhter Shimmer-Wert für jüngere, männliche Erwachsene wird auch von Pützer (2001) beobachtet, wenngleich wenn bei dieser Studie der Unterschied zu den Sprecherinnen nicht so stark ist wie im vorliegenden Fall. Die Gauß'schen Wahrscheinlichkeitsdichten für die weiblichen Sprecher zeigen für die besagten Shimmer-Maße eine wesentliche Überdeckung der Klassen. In dem geschlechtsunspezifischen Fall kann das Maß lediglich zur Unterscheidung von KINDER und SENIOREN herangezogen werden, die Kurven von JUGENDLICHE und ERWACHSENE werden fast vollständig überdeckt. Betrachtet man nur die weiblichen Sprecher, entzerrt sich das Bild etwas: Sowohl nach rechts für die Senioren als auch nach links für die Kinder ergibt sich ein kleiner freier Bereich. Die Glockenkurve der Jugendlichen wird jedoch auch in diesem Fall vollständig überdeckt. Bei den Männern ist die Reihenfolge der Kurven gemäß der erwähnten unerwarteten Tendenzen umgekehrt wie bei den Frauen. Es ist eine relativ scharfe Trennung zwischen SENIOREN und KINDER zu erkennen, während die Kurve der jüngeren Erwachsenen zwar deutlich nach rechts versetzt ist, gleichzeitig jedoch eine breite Basis hat.

5.2 Shimmer — $shim\_apq3$

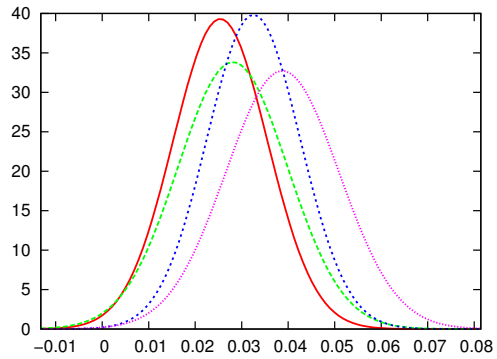


	μ	Σ	$\bar{\mu}$
gesamt	0.0293	0.012	0
Kinder	0.0243	0.0096	-0.4162
Jugendliche	0.0273	0.0111	-0.1672
Erwachsene	0.0279	0.0096	-0.1107
Senioren	0.0376	0.0129	0.6941

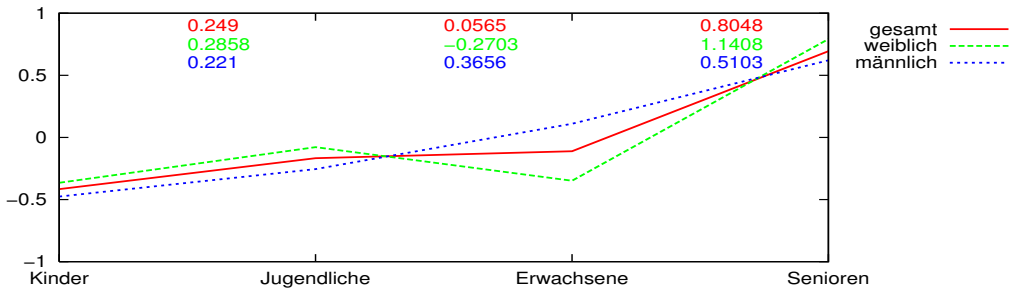
5.2 Shimmer (Fortsetzung) *shim_apq3*



	weibliche Sprecher		
gesamt	0.0273	0.0115	0
Kinder	0.0231	0.0088	-0.3644
Jugendliche	0.0264	0.0104	-0.0786
Erwachsene	0.0233	0.0063	-0.3489
Senioren	0.0364	0.0136	0.7919



	männliche Sprecher		
gesamt	0.0312	0.0122	0
Kinder	0.0254	0.0101	-0.4761
Jugendliche	0.0281	0.0118	-0.2551
Erwachsene	0.0325	0.01	0.1105
Senioren	0.0388	0.0122	0.6208



5.1.3 Stimmumfang

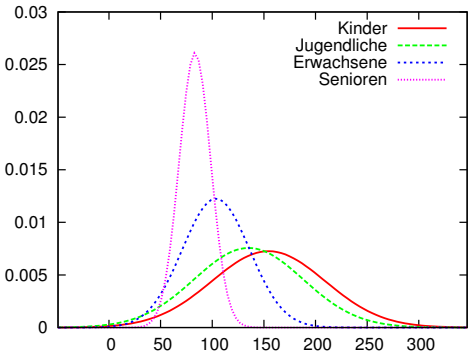
Die folgenden drei Merkmale beziehen sich auf Veränderungen im Stimmumfang zwischen den Altersklassen. Im Einzelnen werden betrachtet: der durchschnittlich tiefste Ton (*pitch_min*), der durchschnittlich höchste Ton (*pitch_max*) und das Spektrum (*pitch_diff*), welches eine Differenz von *pitch_max* und *pitch_min* darstellt.

Sowohl bei weiblichen als auch bei männlichen Sprechern nimmt die minimale Stimmtonhöhe (*pitch_min*) zwischen den Altersklassen von KINDER bis SENIOREN kontinuierlich ab. Bei den Frauen ist zunächst eine mit -0.19 sehr schwache (aber statistisch signifikante) negative Ten-

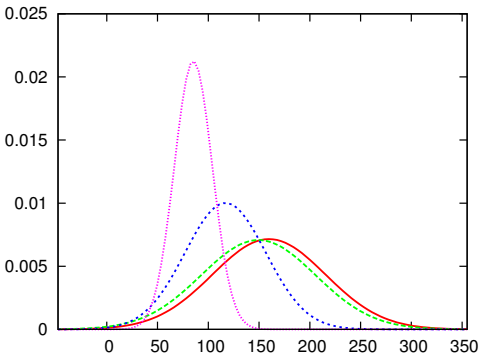
denz zwischen KINDER und JUGENDLICHE zu erkennen, die dann an Stärke zunimmt und bis zu SENIOREN auf einem Niveau von ungefähr -0.6 bleibt. Diese Tendenzen entsprechen den Erwartungen. Für die Männer trifft das nicht zu, da erwartet wurde, dass das Niveau zwischen jüngeren Erwachsenen und Senioren gleicht bleibt. Die negative Tendenz wird allerdings immerhin abgeschwächt.

Insgesamt sind große Unterschiede zwischen den Geschlechtern zu erkennen, was ebenfalls erwartet wurde. Kinder und Jugendliche zeigen bei beiden Geschlechtern eine wesentlich höhere Standardabweichung als die Erwachsenen und Senioren, wie anhand der Gauß'schen Wahrscheinlichkeitsdichten gut zu erkennen ist. Die Bilder sind insgesamt sehr ähnlich und unterscheiden sich nur dadurch, dass die Effekte bei den Männern stärker sind als bei den Frauen. Das Maß kann daher für beide Geschlechter zur Unterscheidung der Gruppe KINDER/JUGENDLICHE und der Klassen ERWACHSENE und SENIOREN herangezogen werden (vgl. Anhang A, Seite 276).

5.3 Minimale Stimmtonhöhe — *pitch_min*



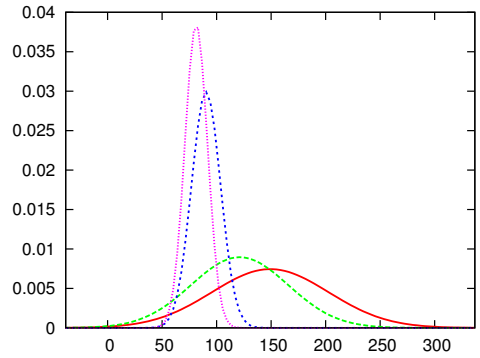
	μ	Σ	$\bar{\mu}$
gesamt	119.15	50.33	0
Kinder	154.49	54.95	0.7
Jugendliche	135.23	52.73	0.32
Erwachsene	103.6	32.56	-0.31
Senioren	83.27	15.28	-0.71



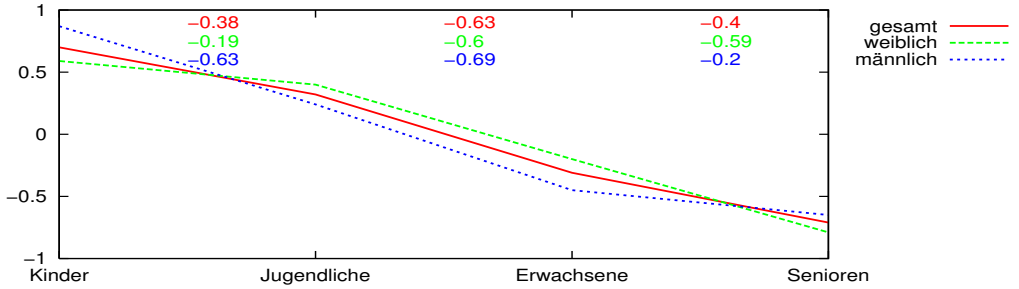
weibliche Sprecher

gesamt	127.83	53.95	0
Kinder	159.78	55.77	0.59
Jugendliche	149.36	56.38	0.4
Erwachsene	116.89	39.85	-0.2
Senioren	85.28	18.76	-0.79

5.3 Minimale Stimmtonhöhe (Fortsetzung) *pitch_min*



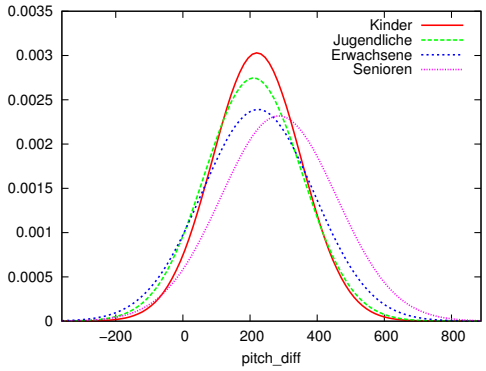
	männliche Sprecher		
gesamt	110.47	44.77	0
Kinder	149.21	53.61	0.87
Jugendliche	121.11	44.54	0.24
Erwachsene	90.31	13.36	-0.45
Senioren	81.26	10.35	-0.65



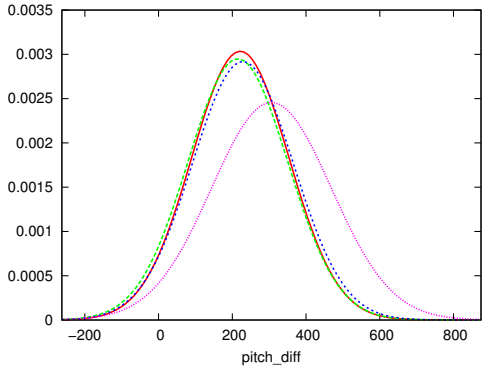
Was die maximale Stimmtonhöhe (*pitch_max*) betrifft, sind die Tendenzen insgesamt schwächer (vgl. Anhang A, Seite 276). Bei beiden Geschlechtern gibt es eine mittlere bis geringe negative Tendenz zwischen Kindern und Jugendlichen von -0.13 für die Sprecherinnen und -0.3 für die Sprecher. Zwischen der Klasse JUGENDLICHE und der Klasse ERWACHSENE ist bei den Frauen eine geringe negative Tendenz von 0.08 feststellbar. Die Mittelwertunterschiede bei den Männern sind dagegen nicht signifikant. Von ERWACHSENE nach SENIOREN gibt es dann wieder eine mittlere positive Tendenz, sowohl bei den Frauen (0.36) als auch bei den Männern (0.25), die allerdings nicht den Hypothesen entspricht. Die Wahrscheinlichkeitsdichte macht allerdings deutlich, dass dieses Merkmal nicht zur Unterscheidung der Altersklassen verwendet werden kann: Zu groß sind die Überschneidungsbereiche der Gauß-Kurven.

Der Stimmumfang *pitch_diff* lässt die stärksten Tendenzen ebenfalls zwischen Erwachsenen und Senioren erkennen: Die Frauen verzeichnen hier einen Anstieg von 0.53 und die Männer einen Anstieg von 0.3. Die übrigen Tendenzen sind eher gering: Ein leichter Abfall zwischen Kindern und Jugendlichen von -0.05 für die Sprecherinnen und -0.12 für die Sprecher, dann ein etwas deutlicherer Anstieg hin zu den jüngeren Erwachsenen, der mit 0.17 bei den Männern höher ist als bei den Frauen (0.15). Die Verwendbarkeit des Maßes zur Unterscheidung der Altersklassen beschränkt sich daher auf die Differenzierung zwischen der Gruppe KINDER/JUGENDLICHE/ERWACHSENE und der Klasse SENIOREN (vgl. Anhang A, Seite 276).

5.4 Stimmumfang *pitch_diff*

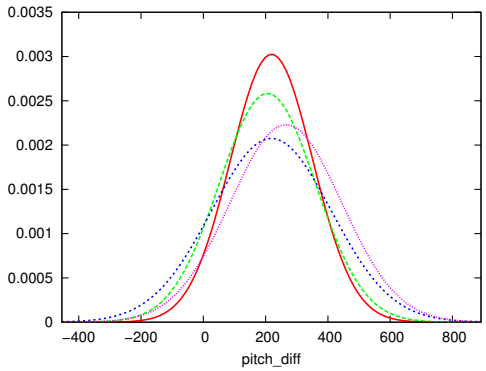


	μ	Σ	$\bar{\mu}$
gesamt	233.43	159.53	0
Kinder	214.82	133.41	-0.12
Jugendliche	201.01	146.2	-0.2
Erwachsene	226.53	167.77	-0.04
Senioren	291.37	172.25	0.36



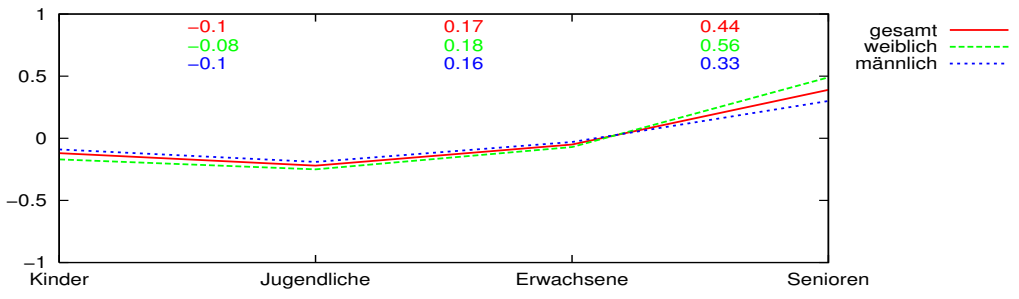
weibliche Sprecher

gesamt	239.18	148.44	0
Kinder	213.56	132.67	-0.17
Jugendliche	206.69	137.41	-0.22
Erwachsene	228.53	136.69	-0.07
Senioren	307.96	162.53	0.46



männliche Sprecher

gesamt	227.68	169.69	0
Kinder	216.09	134.14	-0.07
Jugendliche	195.33	154.28	-0.19
Erwachsene	224.53	193.91	-0.02
Senioren	274.78	179.93	0.28



5.1.4 Mittlere Stimmtonhöhe

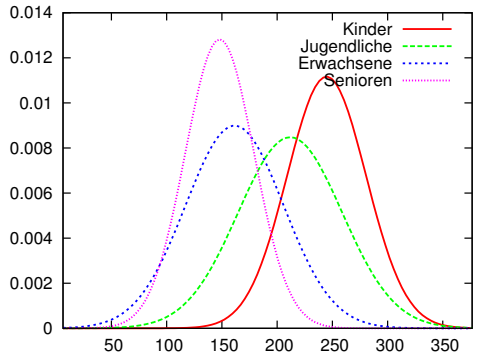
Die Notwendigkeit einer geschlechtsdifferenzierten Sicht wird besonders bei der Betrachtung der mittleren Stimmtonhöhe $pitch\_mean$ deutlich: Bei den männlichen Sprechern ist das Absinken des Wertes zwischen KINDER und JUGENDLICHE mit -0.72 deutlich stärker als bei den Sprecherinnen mit -0.3. Die Tendenzen zwischen Jugendlichen und Erwachsenen sind dagegen bei beiden Geschlechtern weitestgehend gleich: Die Sprecherinnen verzeichnen ein Absinken von -0.88, die Sprecher von -1.05.

Zwischen den jüngeren Erwachsenen und den Senioren ist das zu erwartende komplementäre Bild deutlich zu erkennen: Während die mittlere Stimmtonhöhe der Frauen mit einer Tendenz von -0.66 weiter sinkt, steigt die der Männer zwar nur leicht, aber statistisch signifikant an (0.07).

Die Wahrscheinlichkeitsdichten der Sprecherinnen verschiedener Altersklassen reihen sich entlang der x-Achse in fast gleichmäßiger Art und Weise auf. Dank der verhältnismäßig geringen Standardabweichungen und der damit verbundenen schmalen Glockenformen, gibt es für jede der Altersklassen einen Bereich, in dem eine eindeutige Zuordnung erfolgen kann. Bei den Männern ist sehr deutlich eine Gruppenbildung von Erwachsenen und Senioren zu erkennen, von der sowohl die Klassen KINDER als auch JUGENDLICHE gut zu unterscheiden sind. Das Merkmal kann daher bei den Frauen zur Unterscheidung aller Altersklassen verwendet werden und bei den Männern zur Unterscheidung der Klassen KINDER und JUGENDLICHE von der Gruppe ERWACHSENE/SENIOREN. In der undifferenzierten Sicht ist dagegen lediglich eine Unterscheidung von KINDER/JUGENDLICHE und ERWACHSENE/SENIOREN möglich.

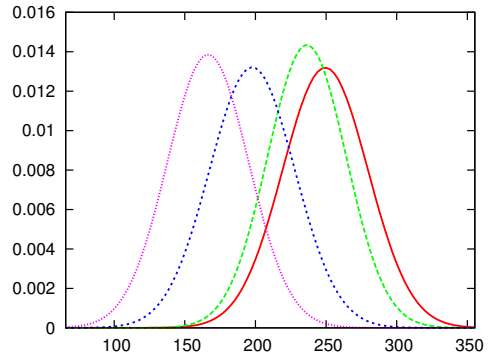
Das Bild des Merkmals $pitch\_quant$ (vgl. Anhang A, Seite 276) ist dem von $pitch\_mean$ erwartungsgemäß sehr ähnlich, abgesehen davon, dass die Mittelwerttendenzen durchweg deutlicher sind: Die positive Tendenz zwischen den jüngeren, erwachsenen Männern und den Senioren beläuft sich in diesem Fall beispielsweise auf 0.14. Die Anordnungen der Wahrscheinlichkeitsdichten sind ebenfalls dieselben.

5.5 Mittlere Stimmtonhöhe — $pitch\_mean$

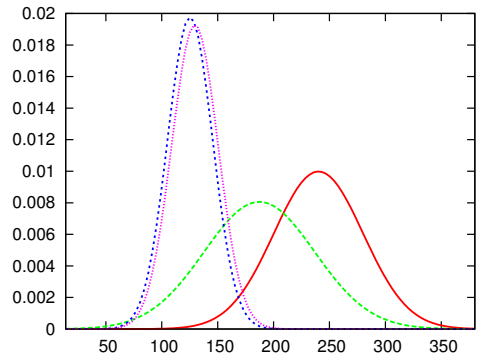


	μ	Σ	$\bar{\mu}$
gesamt	191.55	55.83	0
Kinder	244.8	35.74	0.95
Jugendliche	211.86	47.06	0.36
Erwachsene	161.51	44.36	-0.54
Senioren	148.04	31.11	-0.78

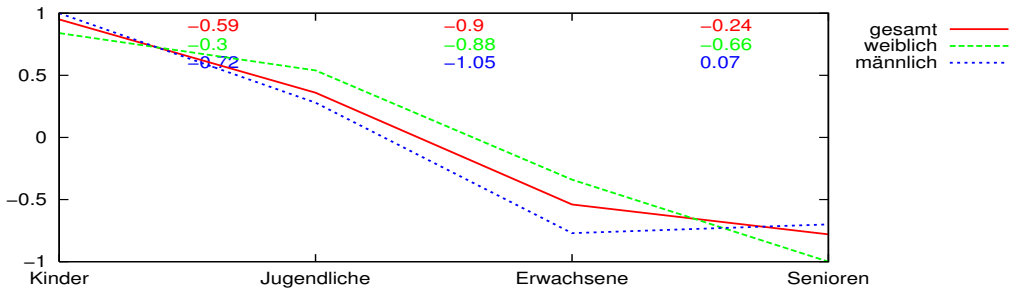
5.5 Mittlere Stimmtonhöhe (Fortsetzung) *pitch_mean*



	weibliche Sprecher		
gesamt	212.49	43.88	0
Kinder	249.43	30.24	0.84
Jugendliche	236.4	27.82	0.54
Erwachsene	197.68	30.2	-0.34
Senioren	166.46	28.79	-1



	männliche Sprecher		
gesamt	170.61	58.59	0
Kinder	240.17	39.96	1
Jugendliche	187.31	49.5	0.28
Erwachsene	125.35	20.19	-0.77
Senioren	129.63	20.71	-0.7

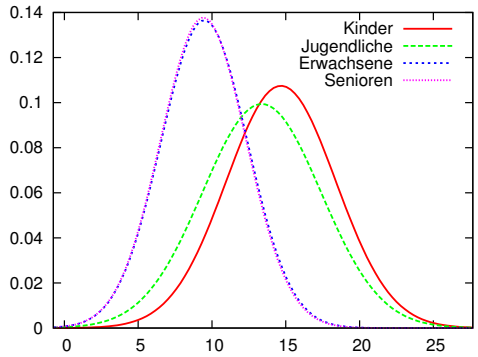


5.1.5 Harmonizität

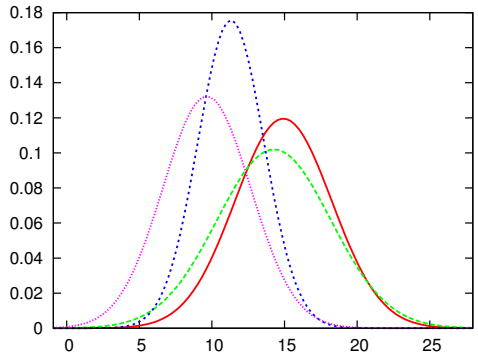
Das Verhältnis von stimmhaften zu stimmlosen Anteilen in der Sprache – $harm\_mean$ – lässt bei den Frauen erwartungsgemäß einen linearen Abfall von der Klasse KINDER bis zur Klasse SENIOREN erkennen, der zunächst zwischen KINDER und JUGENDLICHE mit -0.168 verhältnismäßig gering ist, dann von JUGENDLICHE auf ERWACHSENE auf -0.77 anwächst und zwischen ERWACHSENE und SENIOREN wieder auf -0.432 zurückgeht. Bei den jüngeren, erwachsenen Männern wird der lineare Abfall durch einen tiefen Wert unterbrochen, der mit einem hohen Shimmer-Wert (z. B. $shim\_apq11$) korrespondiert. Im Gegensatz zu diesem entspricht der gerin-

ge $harm\_mean$-Wert jedoch nicht den Ergebnissen von Pützer (2001). Trotz dieser unerwarteten Tendenz lassen die Bilder der Wahrscheinlichkeitsdichten vermuten, dass das Merkmal zur Unterscheidung der Gruppen KINDER/JUGENDLICHE und ERWACHSENE/SENIOREN geeignet ist. Die Analyse der entsprechend gruppierten Ergebnisse bestätigt diese Vermutung: Unabhängig vom Geschlecht ist eine verhältnismäßig starke negative Tendenz zwischen den beiden Gruppen zu erkennen, die bei den Sprecherinnen -1.07 und bei den Sprechern -1.18 beträgt (vgl. Anhang A, Seite 280).

5.6 Harmonizität — $harm\_mean$

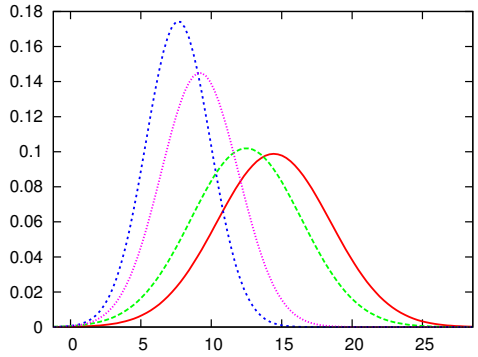


	μ	Σ	$\bar{\mu}$
gesamt	11.737	4.142	0
Kinder	14.679	3.712	0.71
Jugendliche	13.375	4.014	0.395
Erwachsene	9.481	2.922	-0.545
Senioren	9.415	2.897	-0.561



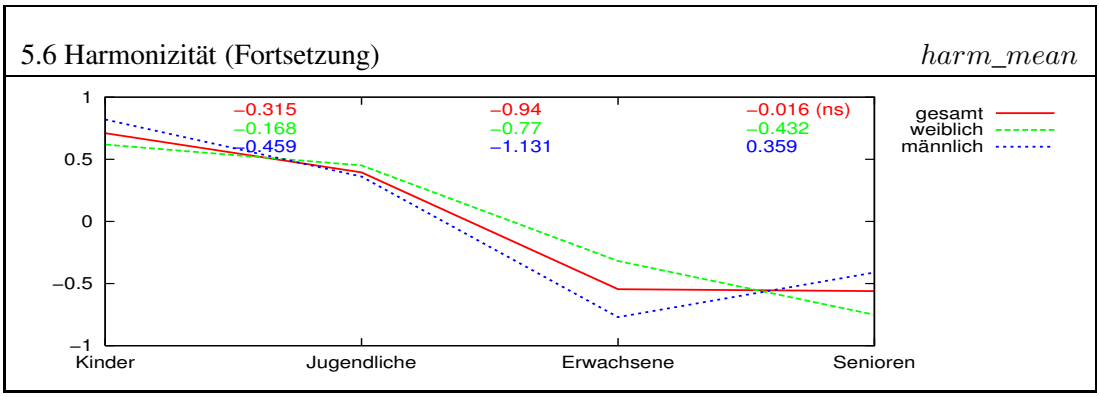
weibliche Sprecher

	μ	Σ	$\bar{\mu}$
gesamt	12.536	3.849	0
Kinder	14.919	3.338	0.619
Jugendliche	14.27	3.912	0.451
Erwachsene	11.309	2.271	-0.319
Senioren	9.645	3.018	-0.751



männliche Sprecher

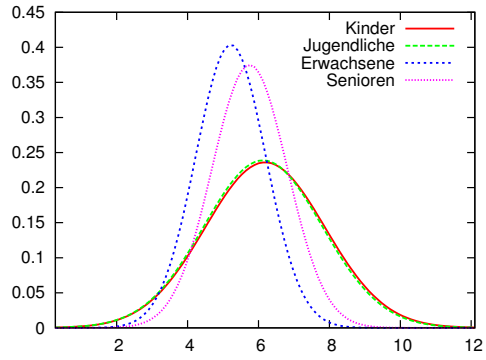
	μ	Σ	$\bar{\mu}$
gesamt	10.939	4.269	0
Kinder	14.438	4.038	0.82
Jugendliche	12.48	3.913	0.361
Erwachsene	7.652	2.287	-0.77
Senioren	9.185	2.751	-0.411



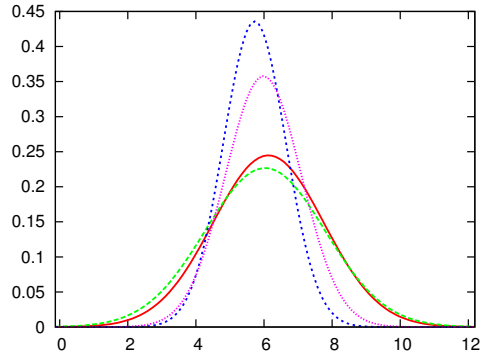
Weitere Statistiken der Harmonizität sind weniger gut zur Unterscheidung der Altersklassen geeignet: Der Minimalwert $harm\_min$ weist keine signifikanten Mittelwertdifferenzen auf. Der Maximalwert $harm\_max$ lässt zwischen den Klassen KINDER und JUGENDLICHE eine leichte negative Tendenz erkennen, die jedoch für die männlichen Sprecher nicht signifikant ist. Für die Sprecherinnen beläuft sich diese Tendenz auf -0.09. Für beide Geschlechter ist zwischen den Klassen JUGENDLICHE und ERWACHSENE eine positive Tendenz zu verzeichnen, die mit 0.33 bei den Frauen stärker ist als bei den Männern (0.08). Zwischen den Klassen ERWACHSENE und SENIOREN schwächt die positive Tendenz bei den Frauen allerdings auf 0.13 ab, während sie bei den Männern auf 0.34 ansteigt (vgl. Anhang A, Seite 280). Der Umstand, dass die Wahrscheinlichkeitsdichten eine weitgehende Überdeckung aufweisen, lässt dieses Maß insgesamt für die Differenzierung ungeeignet erscheinen.

Lediglich die Standardabweichung $harm\_stdev$ ist ein Maß, welches bei geeigneter Gruppierung zu einer Differenzierung der Klassen beitragen könnte. Qualitativ ist das Bild ähnlich dem des Jitter: Die Klasse SENIOREN und die beiden Klassen KINDER und JUGENDLICHE weisen einen verhältnismäßig hohen Wert auf, wohingegen die Klasse ERWACHSENE sowohl bei den Männern als auch bei den Frauen darunter liegt. Der Unterschied zwischen den Kindern und den Jugendlichen ist dabei zunächst lediglich für die Frauen signifikant und beträgt -0.069. Die negative Tendenz setzt sich zu den Erwachsenen hin fort, nimmt aber an Stärke zu (-0.218). Darauf folgt zu den Senioren hin eine leichte positive Tendenz von 0.182. Bei den Männern sind die Tendenzen zwischen den drei oberen Altersklassen insgesamt stärker: Zwischen den Jugendlichen und den Erwachsenen beträgt sie -1.033 und zwischen den Erwachsenen und den Senioren 0.553. Die geeigneten Gruppierungen wären somit KINDER/JUGENDLICHE, ERWACHSENE und SENIOREN. Die entsprechende Analyse zeigt, dass für die Männer eine Differenzierung dieser Gruppen möglich sein sollte (vgl. Anhang A, Seite 280).

5.7 Harmonizität $harm\_stddev$

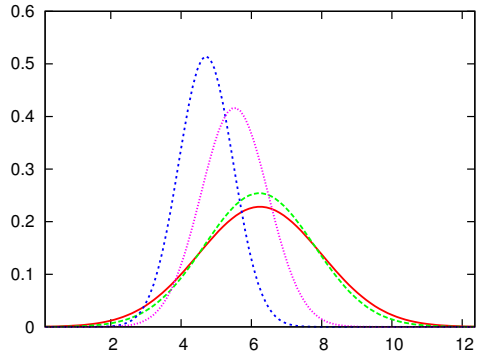


	μ	Σ	$\bar{\mu}$
gesamt	5.821	1.445	0
Kinder	6.187	1.69	0.253
Jugendliche	6.129	1.67	0.213
Erwachsene	5.217	0.988	-0.418
Senioren	5.751	1.065	-0.048



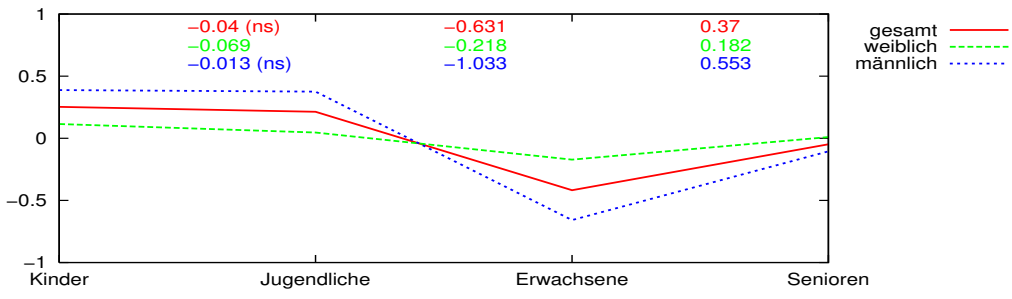
weibliche Sprecher

gesamt	5.967	1.407	0
Kinder	6.129	1.63	0.115
Jugendliche	6.032	1.76	0.046
Erwachsene	5.725	0.915	-0.172
Senioren	5.981	1.115	0.01



männliche Sprecher

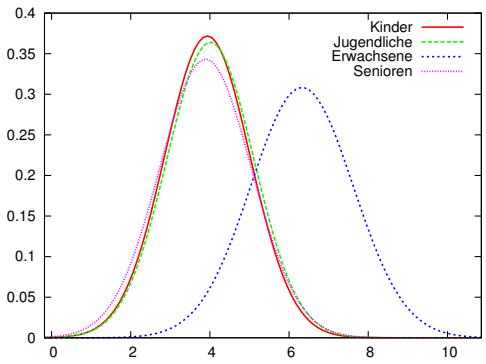
gesamt	5.676	1.468	0
Kinder	6.245	1.747	0.388
Jugendliche	6.226	1.57	0.375
Erwachsene	4.709	0.775	-0.658
Senioren	5.522	0.958	-0.105



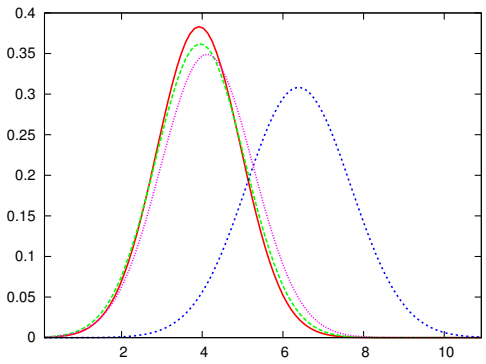
5.1.6 Artikulationsgeschwindigkeit

Die Artikulationsgeschwindigkeit der jüngeren Erwachsenen beider Geschlechter ist deutlich höher als die der Kinder, Jugendlichen oder Senioren: Zwischen den Jugendlichen und den Erwachsenen ist bei den Sprecherinnen eine starke positive Tendenz von 1.42 zu beobachten, die bei den Männern mit 1.26 nur unwesentlich geringer ist. Der Abfall der Artikulationsgeschwindigkeit zu der Klasse der Senioren ist mit -1.30 bzw. -1.53 ebenso deutlich. Zwischen den Kindern und den Jugendlichen besteht im Fall der Mädchen kein signifikanter Unterschied. Die Jungen verzeichnen zwar eine deutliche positive Tendenz, die jedoch mit 0.08 verhältnismäßig gering ist. Die Wahrscheinlichkeitsdichten erlauben ebenfalls die Bildung der Gruppe KINDER/JUGENDLICHE/SENIOREN, die deutlich von der Klasse ERWACHSENE zu unterscheiden ist.

5.8 Artikulationsgeschwindigkeit — ar_rate



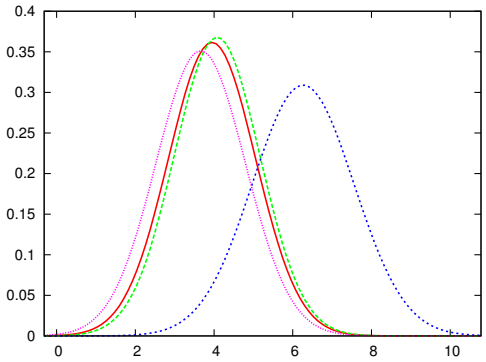
	μ	Σ	$\bar{\mu}$
gesamt	4.541	1.552	0
Kinder	3.935	1.073	-0.39
Jugendliche	4.012	1.096	-0.34
Erwachsene	6.325	1.294	1
Senioren	3.89	1.162	-0.419



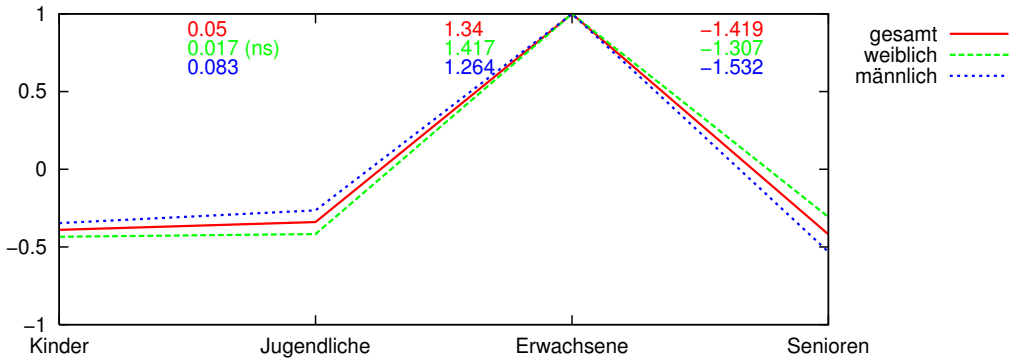
weibliche Sprecher

gesamt	4.595	1.549	0
Kinder	3.923	1.041	-0.434
Jugendliche	3.948	1.102	-0.417
Erwachsene	6.388	1.294	1
Senioren	4.12	1.143	-0.307

5.8 Artikulationsgeschwindigkeit (Fortsetzung) ar_rate



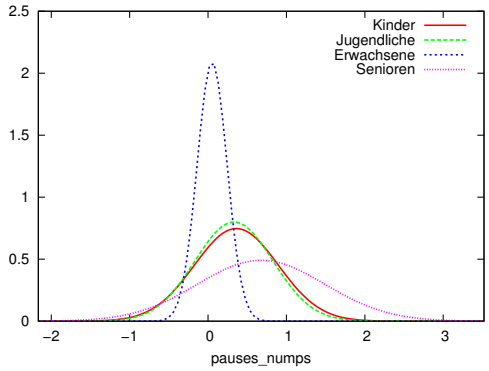
	männliche Sprecher		
gesamt	4.487	1.553	0
Kinder	3.948	1.103	-0.347
Jugendliche	4.077	1.085	-0.264
Erwachsene	6.262	1.291	1
Senioren	3.66	1.136	-0.532



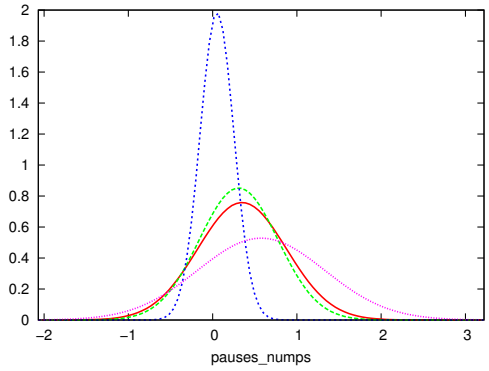
5.1.7 Sprechpausen

Die Ergebnisse bezüglich der Sprechpausen korrespondieren mit denjenigen der Artikulationsgeschwindigkeit: Die jüngeren Erwachsenen machen die wenigsten Pausen bei der Artikulation. Zwischen den weiblichen Kindern und Jugendlichen besteht eine sehr geringe negative Tendenz von -0.073. Bei den Jungen ist die Tendenz nicht signifikant. Zwischen den Jugendlichen und den jüngeren Erwachsenen besteht bei beiden Geschlechtern eine mittlere negative Tendenz, die bei den männlichen Sprechern mit -0.476 etwas deutlicher ist als bei den Sprecherinnen (-0.459). Die stärkste Tendenz ist zwischen der Klasse ERWACHSENE und der Klasse SENIOREN zu erkennen: Bei den Frauen ist hier eine Zunahme der Anzahl der Pausen von 0.938 zu verzeichnen. Die positive Tendenz bei den Männern ist mit 1.151 sogar noch deutlicher. Die Wahrscheinlichkeitsdichten zeigen jedoch eine hohe Varianz bei den Senioren beider Geschlechter. Insgesamt lassen diese Ergebnisse den Schluss zu, dass, ähnlich wie bei der Artikulationsgeschwindigkeit, auf Basis der Anzahl der Sprechpausen die Gruppe KINDER/JUGENDLICHE/SENIOREN von der Klasse ERWACHSENE unterschieden werden kann.

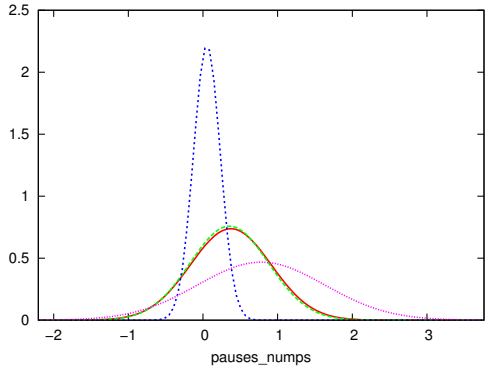
5.9 Sprechpausen *pauses_numps*



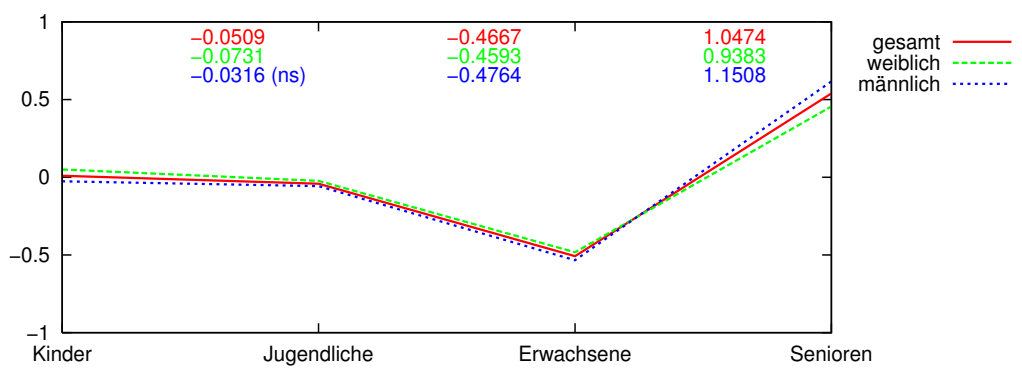
	μ	Σ	$\bar{\mu}$
gesamt	0.3554	0.5969	0
Kinder	0.3612	0.5336	0.0097
Jugendliche	0.3308	0.4985	-0.0412
Erwachsene	0.0522	0.1907	-0.5079
Senioren	0.6774	0.8121	0.5395



weibliche Sprecher			
gesamt	0.3214	0.5581	0
Kinder	0.3493	0.5262	0.0499
Jugendliche	0.3085	0.4685	-0.0232
Erwachsene	0.0521	0.2009	-0.4825
Senioren	0.5758	0.7556	0.4558

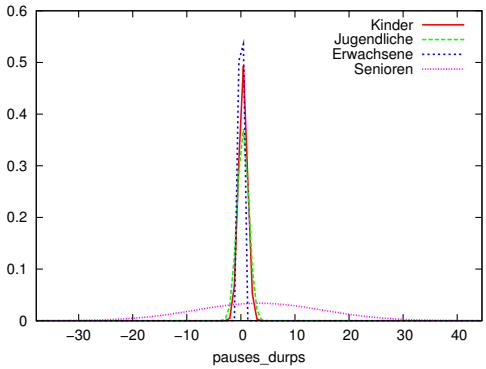


männliche Sprecher			
gesamt	0.3894	0.6316	0
Kinder	0.3731	0.5406	-0.0258
Jugendliche	0.3531	0.5259	-0.0574
Erwachsene	0.0522	0.1799	-0.5338
Senioren	0.7791	0.8528	0.617

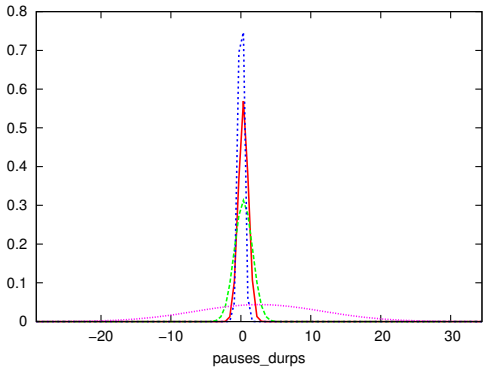


Das Merkmal Dauer der Sprechpausen zeigt ein etwas anderes Bild: Zwischen den Kindern, Jugendlichen und jüngeren Erwachsenen gibt es, falls überhaupt, nur sehr schwache (negative) Tendenzen, die darüber hinaus so gut wie keine Unterschiede zwischen den Geschlechtern aufweisen. Zwischen der Klasse ERWACHSENE und der Klasse SENIOREN besteht allerdings eine mittlere positive Tendenz von 0.541 bei den Frauen und 0.57 bei den Männern. Die Wahrscheinlichkeitsdichten zeigen einen deutlichen Unterschied: Die Klasse SENIOREN weist eine um ein Vielfaches breitere Glockenform auf als die übrigen Klassen. Eine Unterscheidung der Klassen auf Basis der Dauer der Sprechpausen kann daher sinnvoll nur mit einer Gruppierung KINDER/JUGENDLICHE/ERWACHSENE erfolgen.

5.10 Sprechpausen — *pauses_durps*

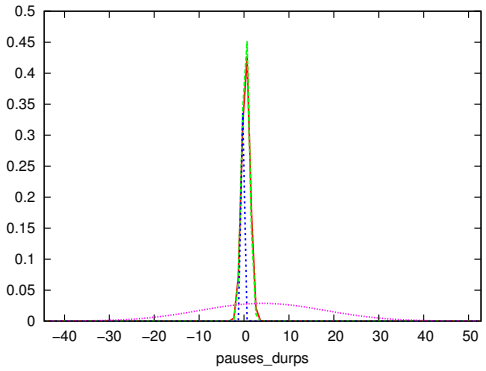


	μ	Σ	$\bar{\mu}$
gesamt	1.048	6.078	0
Kinder	0.4	0.807	-0.107
Jugendliche	0.379	1.075	-0.11
Erwachsene	0.045	0.308	-0.165
Senioren	3.367	11.773	0.382

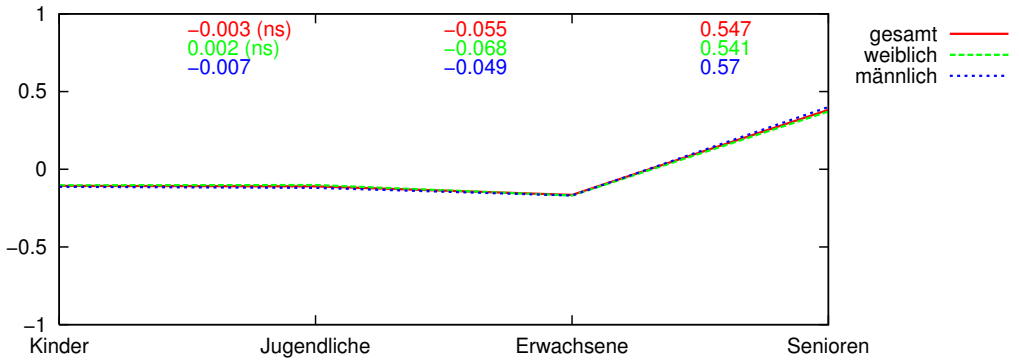


	weibliche Sprecher		
gesamt	0.852	4.727	0
Kinder	0.366	0.701	-0.103
Jugendliche	0.374	1.275	-0.101
Erwachsene	0.054	0.401	-0.169
Senioren	2.611	9.105	0.372

5.10 Sprechpausen (Fortsetzung) *pauses_durps*



	männliche Sprecher		
gesamt	1.244	7.173	0
Kinder	0.434	0.899	-0.113
Jugendliche	0.384	0.828	-0.12
Erwachsene	0.035	0.169	-0.169
Senioren	4.123	13.898	0.401



5.2 Sprechergeschlecht

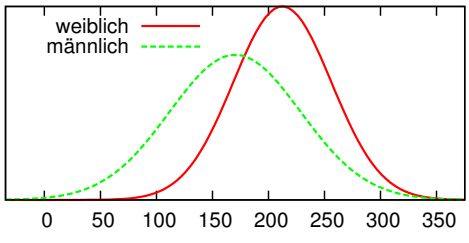
5.2.1 Mittlere Stimmtonhöhe

Dass die mittlere Stimmtonhöhe $pitch\_mean$ als das Hauptunterscheidungsmerkmal von weiblichen und männlichen Stimmen angesehen werden kann, bestätigt sich in den Ergebnissen: Zwischen Sprecherinnen und Sprechern ist eine negative Tendenz von -0.76 zu verzeichnen. Von Altersklasse zu Altersklasse ist jedoch eine Entwicklung auszumachen, die am deutlichsten anhand der klassenspezifischen Wahrscheinlichkeitsdichten zu erkennen ist: Während die Unterschiede zwischen den Geschlechtern bei den Kindern eher gering sind, tritt die Kurve der Sprecher bildlich gesprochen mit zunehmendem Alter unter der Kurve der Sprecherinnen heraus und formt sich, indem sie schlanker wird. Die Basis der roten Kurve, welche die Sprecherinnen repräsentiert, wird dagegen breiter, was mit einer höheren Varianz weiblicher Stimmen gleichbedeutend ist.

Dies bestätigt sich auch anhand der normalisierten Mittelwerttendenzen: In der Altersklasse KINDER ist lediglich eine schwache negative Tendenz von -0.28 zwischen Frauen und Männern zu verzeichnen. Bei der Klasse JUGENDLICHE ist die Tendenz mit -1.04 bereits deutlich stärker und wächst zur Klasse ERWACHSENE noch einmal auf -1.62 an. Bei der Klasse SENIOREN ist die Tendenz dann wieder mit -1.18 etwas geringer.

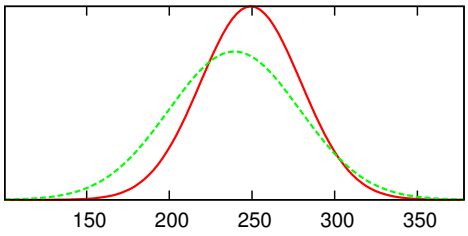
Zur sinnvollen Unterscheidung des Geschlechts anhand der mittleren Stimmtonhöhe muss eine altersklassenspezifische Analyse erfolgen. Lediglich die Klassen ERWACHSENE und SENIOREN können in einer Gruppe zusammengefasst werden. Diese Annahme wird in der entsprechenden Analyse, die in Anhang A (Seite 283) dargestellt wird, bekräftigt: Die Gruppe ERWACHSENE/SENIOREN weist mit -1.4 eine starke negative Tendenz zwischen Frauen und Männern auf; bei den Jugendlichen ist die Tendenz zwar schwächer, aber mit -1.02 immer noch deutlich; die Tendenz bei den Kindern dagegen ist mit -0.28 eher schwach, auch wenn sie statistisch signifikant ist.

5.11 Mittlere Stimmtonhöhe — $pitch\_mean$

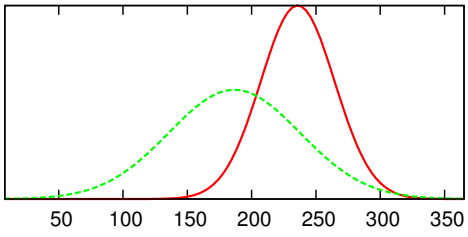


	μ	Σ	$\bar{\mu}$
gesamt	191.31	55.87	0
weiblich	212.38	43.9	0.38
männlich	170.24	58.55	-0.38

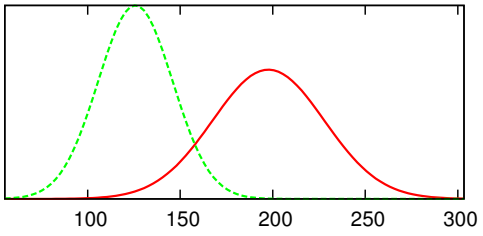
5.11 Mittlere Stimmtonhöhe (Fortsetzung) *pitch_mean*



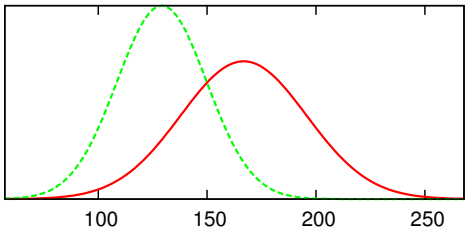
	Kinder		
gesamt	244.31	35.68	0
weiblich	249.26	30.39	0.14
männlich	239.37	39.67	-0.14



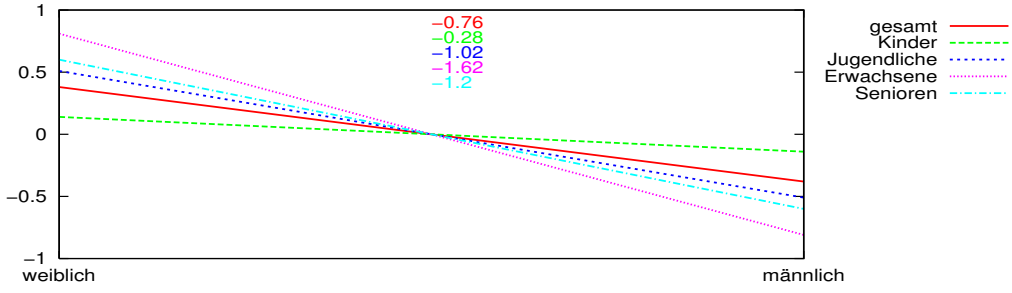
	Jugendliche		
gesamt	211.21	48.17	0
weiblich	235.81	28.79	0.51
männlich	186.61	51.01	-0.51



	Erwachsene		
gesamt	161.73	44.18	0
weiblich	197.68	30.2	0.81
männlich	125.78	20.17	-0.81



	Senioren		
gesamt	147.98	31.35	0
weiblich	166.75	28.91	0.6
männlich	129.21	20.6	-0.6



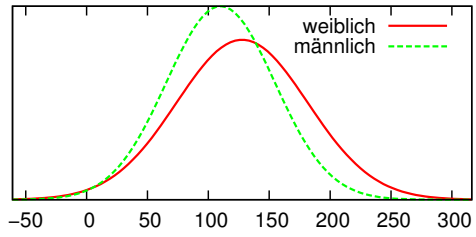
5.2.2 Stimmumfang

Die Effekte bei der minimalen Stimmtonhöhe $pitch\_min$ ähneln denjenigen von $pitch\_mean$: Die stärkste Tendenz ist bei den jüngeren Erwachsenen zu verzeichnen, wo sie bei 0.82 liegt. Allerdings ist der Abfall der Werte zwischen der Klasse WEIBLICH und der Klasse MÄNNLICH bei den Senioren in diesem Fall mit -0.24 fast so gering wie bei den Kindern (-0.2). Die Jugendlichen zeigen eine mittlere negative Tendenz von -0.56.

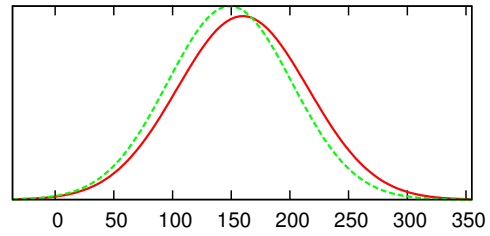
Die Wahrscheinlichkeitsdichten zeigen bei den Männern erneut das Bild der „Formierung“ und „Deformierung“, d. h. die Grundflächen der Kurven werden von der Klasse KINDER bis zu der Klasse ERWACHSENE zunehmend kleiner und zu der Klasse SENIOREN wieder größer. Bei den Frauen dagegen ist das umgekehrte Bild zu beobachten. Der Vergleich mit den zahlenmäßigen Angaben der korrespondierenden Standardabweichungen zeigt allerdings, dass diese Effekte sich lediglich in einer Relation zwischen den Geschlechtern äußern.

Was die Unterscheidbarkeit von Frauen und Männern betrifft, so beschränkt sich diese bezüglich $pitch\_min$ auf die Erwachsenen und Jugendlichen, allerdings ohne eine Gruppierung zuzulassen.

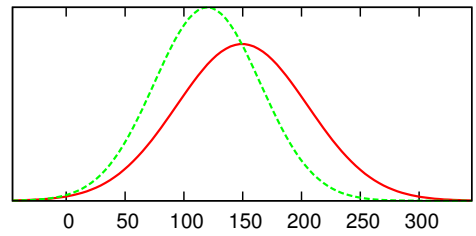
5.12 Minimale Stimmtonhöhe $pitch\_min$



	μ	Σ	$\bar{\mu}$
gesamt	118.89	50.22	0
weiblich	127.74	53.88	0.18
männlich	110.04	44.54	-0.18

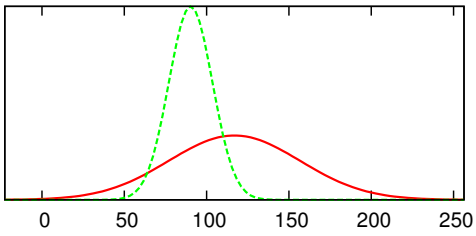


Kinder			
gesamt	154.28	54.61	0
weiblich	159.81	55.74	0.1
männlich	148.75	52.88	-0.1

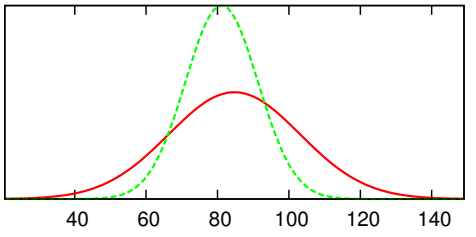


Jugendliche			
gesamt	134.71	52.87	0
weiblich	149.52	55.77	0.28
männlich	119.89	45.19	-0.28

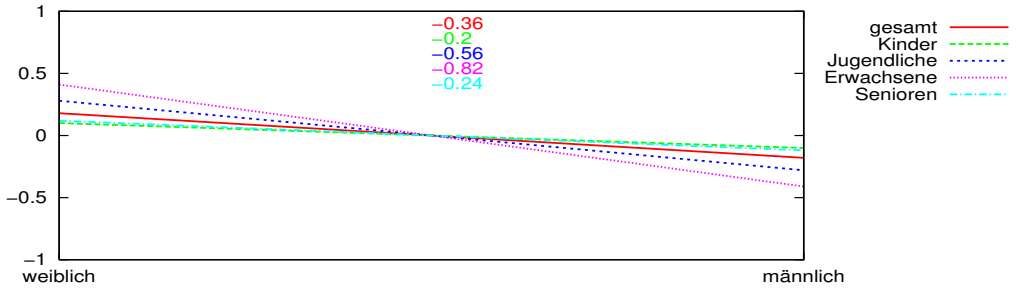
5.12 Minimale Stimmtonhöhe (Fortsetzung) *pitch_min*



	Erwachsene		
gesamt	103.62	32.53	0
weiblich	116.89	39.85	0.41
männlich	90.34	13.25	-0.41



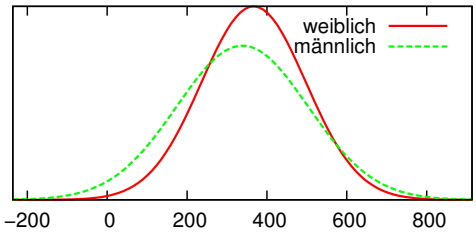
	Senioren		
gesamt	82.95	14.95	0
weiblich	84.73	18.38	0.12
männlich	81.17	10.14	-0.12



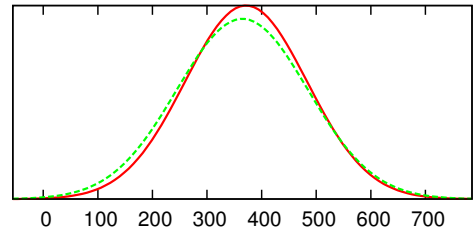
Die maximale Stimmtonhöhe *pitch_max* zeigt einen über die Altersklassen schwachen, aber stabilen Effekt: Abgesehen von der Klasse KINDER mit -0.06 liegt die negative Tendenz zwischen Frauen und Männern im Bereich von -0.2. Die Varianz ist allerdings so stark, dass die Kurven sich weitgehend überschneiden. Zusammenfassend kann das Merkmal dahingehend eingeschätzt werden, dass es in der Altersklassengruppe JUGENDLICHE/ERWACHSENE/SENIOREN einen Beitrag zur Unterscheidung der Geschlechter leisten kann.

Das Stimmumfang *pitch_diff* zeigt noch geringere Tendenzen: Bei Kindern, Jugendlichen und Erwachsenen sind keine signifikanten Mittelwertunterschiede feststellbar. Lediglich bei den Senioren gibt es eine schwache negative Tendenz von -0.02 (vgl. Anhang A, Seite 278).

5.13 Maximale Stimmtonhöhe *pitch_max*

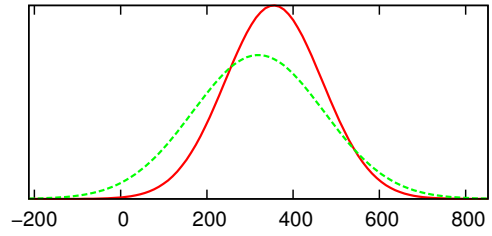


	μ	Σ	$\bar{\mu}$
gesamt	352.69	149.19	0
weiblich	366.83	130.88	0.09
männlich	338.55	164.28	-0.09



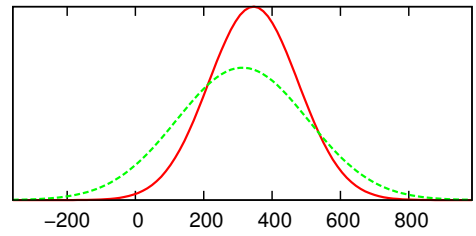
Kinder

gesamt	368.12	116.14	0
weiblich	371.47	111.87	0.03
männlich	364.77	120.16	-0.03



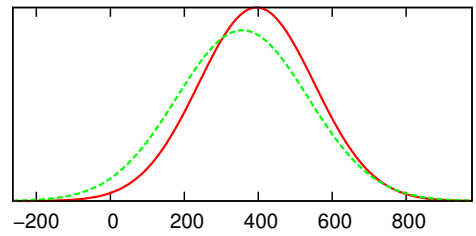
Jugendliche

gesamt	337.18	135.21	0
weiblich	354.96	113.12	0.13
männlich	319.4	152.11	-0.13



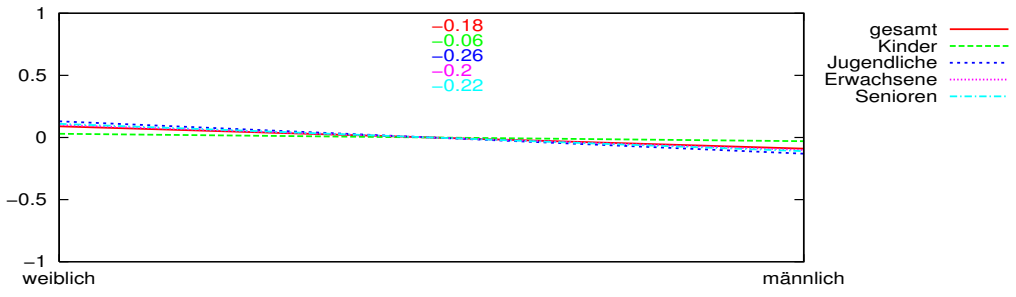
Erwachsene

gesamt	329.17	165.29	0
weiblich	345.42	131.41	0.1
männlich	312.92	191.96	-0.1



Senioren

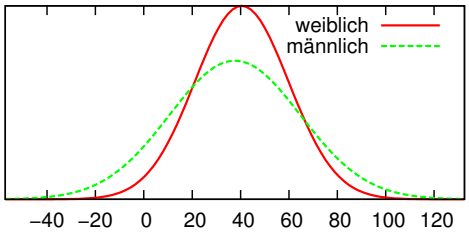
gesamt	376.29	168.39	0
weiblich	395.47	156.53	0.11
männlich	357.11	177.4	-0.11



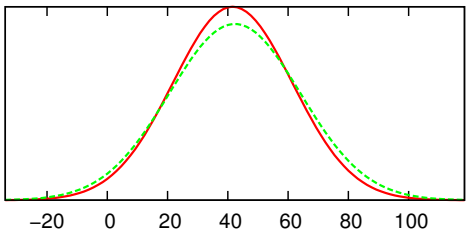
5.2.3 Frequenztremor

Die Standardabweichung der Grundfrequenz $pitch\_stddev$, die als Maß für die globalen Frequenzschwankungen (Frequenztremor) herangezogen werden kann, zeigt in der Altersklasse KINDER und in der Altersklasse ERWACHSENE zwischen Frauen und Männern keine signifikanten Unterschiede. In der Altersklasse JUGENDLICHE ist die Tendenz zwar signifikant, aber mit -0.12 eher schwach. Bei den Senioren allerdings ist nicht nur eine signifikante, sondern eine mit -0.52 darüber hinaus verhältnismäßig starke negative Tendenz feststellbar. Der Frequenztremor ist demnach ein Maß, anhand dessen sich Seniorinnen von Senioren unterscheiden lassen.

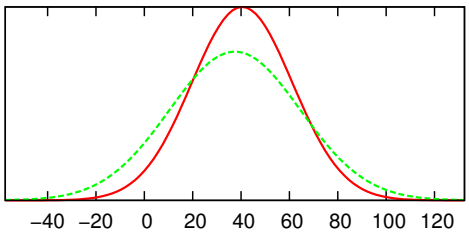
5.14 Frequenztremor — $pitch\_stddev$



	μ	Σ	$\bar{\mu}$
gesamt	39.06	23.64	0
weiblich	40.47	19.44	0.06
männlich	37.66	27.13	-0.06

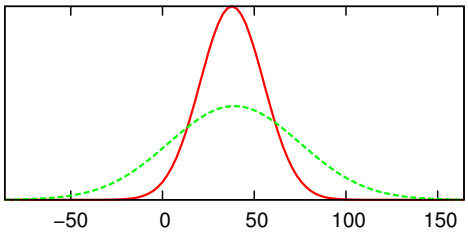


Kinder			
gesamt	41.97	20.84	0
weiblich	41.6	19.84	-0.02
männlich	42.34	21.78	0.02

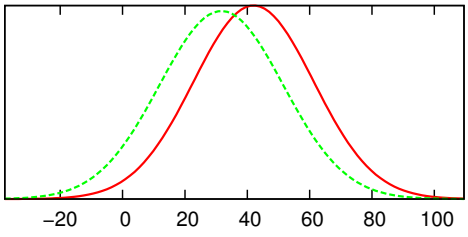


Jugendliche			
gesamt	38.93	24.25	0
weiblich	40.42	20.86	0.06
männlich	37.44	27.14	-0.06

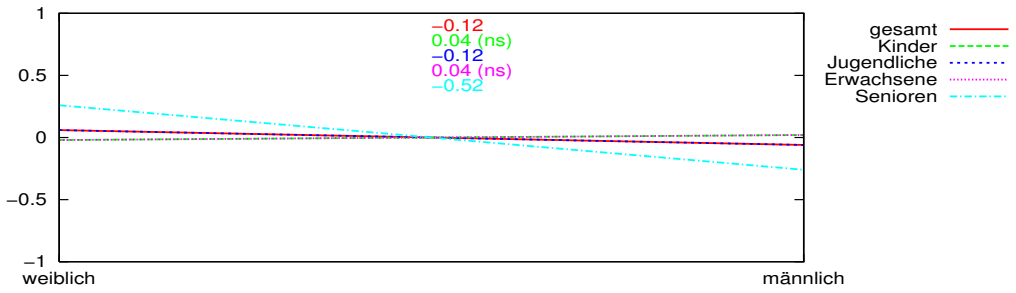
5.14 Frequenztremor (Fortsetzung) *pitch_stddev*



	Erwachsene		
gesamt	38.5	28.11	0
weiblich	37.84	17.33	-0.02
männlich	39.17	35.76	0.02



	Senioren		
gesamt	36.85	20.24	0
weiblich	42.03	19.28	0.26
männlich	31.68	19.85	-0.26

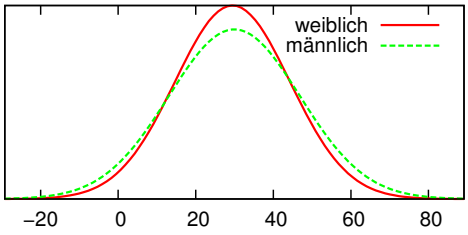


5.2.4 Geschwindigkeit der Stimmtonhöhen-Veränderungen

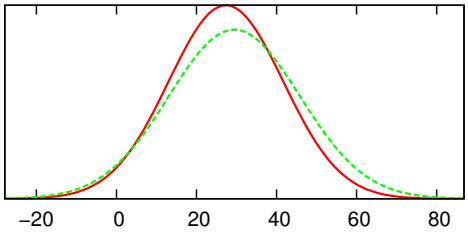
Die mittlere Geschwindigkeit der Stimmtonhöhen-Veränderungen $pitch\_mas$ zeigt ein ähnliches Bild: Während die drei ersten Altersklassen eine schwache positive Tendenz zwischen Sprecherinnen und Sprechern erkennen lassen, gibt es bei den Senioren eine mittlere negative Tendenz von -0.498. Die Wahrscheinlichkeitsdichten zeigen entsprechend in den jüngeren Altersklassen eine weitgehende Überdeckung der Glockenkurven. Die geschlechtsmäßige Differenzierung anhand von $pitch\_mas$ beschränkt sich daher ebenfalls auf die Seniorinnen und Senioren.

Lässt man die Oktavensprünge unbeachtet und betrachtet die Stimmtonhöhen-Veränderungen anhand des Maßes $pitch\_swoj$, ändern sich die Tendenzen qualitativ zwar nicht, jedoch ist die negative Tendenz bei der Altersklasse der Senioren mit -0.054 wesentlich geringer. Demgegenüber sind die positiven Tendenzen sämtlicher jüngeren Altersklassen stärker, besonders diejenigen der Jugendlichen, die auf einen Wert von 0.401 ansteigt (vgl. Anhang A, Seite 286).

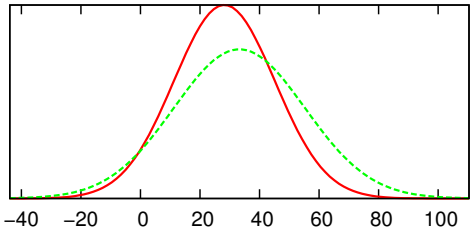
5.15 Stimmtonhöhen-Veränderungen *pitch_mas*



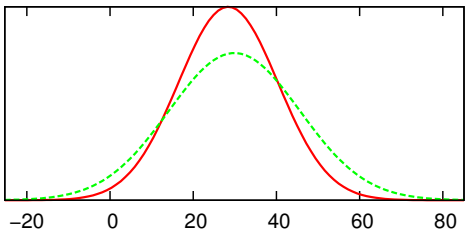
	μ	Σ	$\bar{\mu}$
gesamt	29.721	15.905	0
weiblich	29.477	14.824	-0.015
männlich	29.965	16.913	0.015



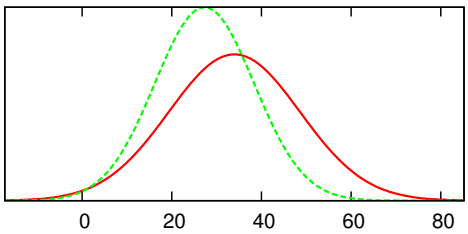
Kinder			
gesamt	28.417	15.423	0
weiblich	27.363	14.313	-0.068
männlich	29.47	16.391	0.068



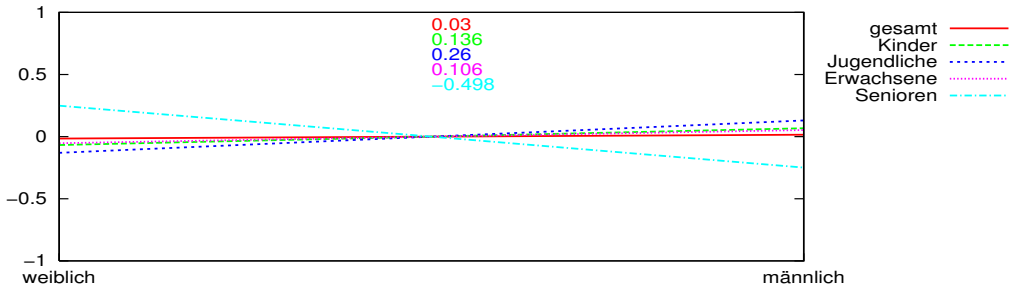
Jugendliche			
gesamt	30.71	19.84	0
weiblich	28.19	16.99	-0.13
männlich	33.23	22.04	0.13



Erwachsene			
gesamt	29.14	14.032	0
weiblich	28.399	11.998	-0.053
männlich	29.881	15.771	0.053



Senioren			
gesamt	30.618	13.401	0
weiblich	33.959	14.635	0.249
männlich	27.277	11.075	-0.249

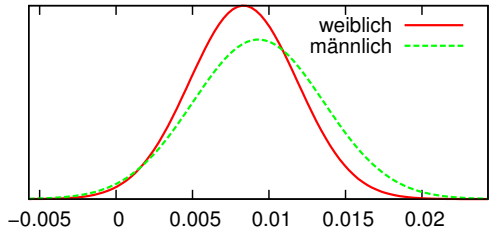


5.2.5 Jitter

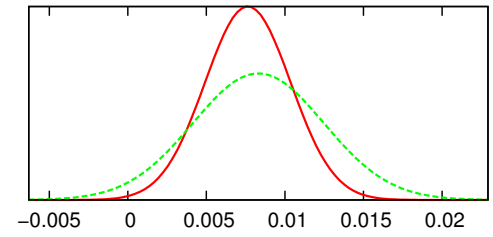
Die Ergebnisse bezüglich Jitter entsprechen nicht den Ergebnissen von Pützer (2001): Die Sprecherinnen weisen in allen Altersklassen geringere Werte auf als die Sprecher. Die stärksten positiven Tendenzen bestehen mit 0.38 und 0.41 bei der Klasse der Jugendlichen bzw. der Klasse der Erwachsenen. Für den erhöhten Jitter-Wert bei den heranwachsenden Jungen im Vergleich zu den Mädchen wurde bereits im Zuge der Diskussion der Ergebnisse bezüglich des Sprecheralters eine Interpretation angeboten: Möglicherweise ist der Jitter ein Indikator des Verlustes der Stimmkontrolle aufgrund der plötzlichen anatomischen Veränderungen in der Pubertät, die bei den Jungen stärker ausgeprägt sind als bei den Mädchen.

Zwischen Mädchen und Jungen und Seniorinnen und Senioren besteht eine eher schwache Tendenz von 0.19 bzw. 0.01. Aufgrund der hohen Standardabweichung insbesondere bei den Männern ist eine Unterscheidung des Geschlechts in allen Altersklassen nur bedingt möglich. Bei dem Maß $jitt\_la$ sind die Tendenzen insgesamt deutlich, und auch die Unterscheidbarkeit, gemessen an den Gauß'schen Wahrscheinlichkeitsdichten, ist insbesondere für die Klassen ERWACHSENE und SENIOREN besser. Aufgrund der zu erwartenden Korrelation des Maßes mit globalen Frequenzschwankungen ist dieses Ergebnis allerdings mit Vorsicht zu bewerten.

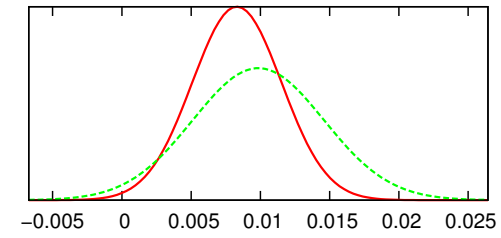
5.16 Jitter — $jitt\_ppq$



	μ	Σ	$\bar{\mu}$
gesamt	0.0088	0.004	0
weiblich	0.0083	0.0035	-0.1229
männlich	0.0093	0.0043	0.1229

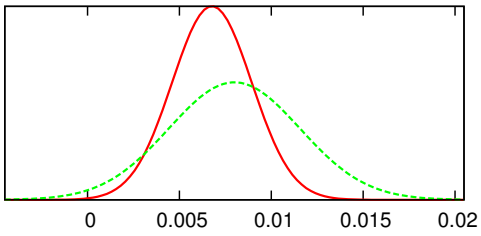


Kinder			
gesamt	0.008	0.0035	0
weiblich	0.0076	0.0027	-0.0947
männlich	0.0083	0.0042	0.0947

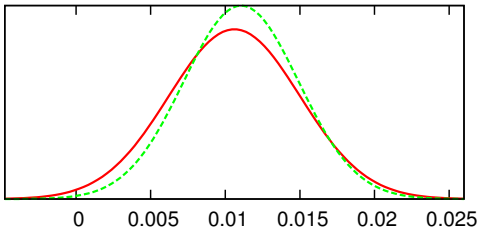


Jugendliche			
gesamt	0.0091	0.0041	0
weiblich	0.0083	0.0032	-0.1876
männlich	0.0099	0.0047	0.1876

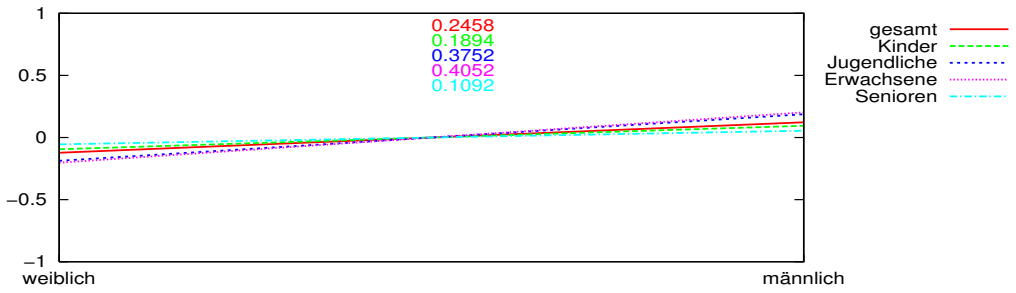
5.16 Jitter (Fortsetzung) *jitt_ppq*



	Erwachsene		
gesamt	0.0074	0.003	0
weiblich	0.0068	0.0022	-0.2026
männlich	0.008	0.0036	0.2026



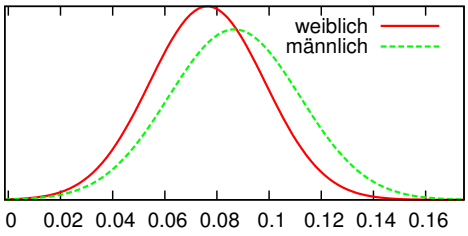
	Senioren		
gesamt	0.0108	0.0041	0
weiblich	0.0106	0.0044	-0.0546
männlich	0.0111	0.0039	0.0546



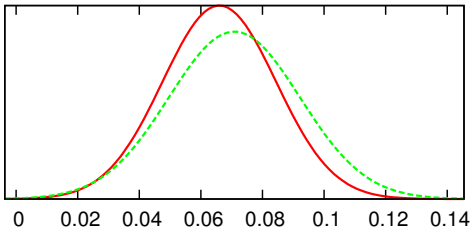
5.2.6 Shimmer

Der herausragende geschlechtsspezifische Unterschied bezüglich $shim\_l$ betrifft die Klasse der Erwachsenen, bei der zwischen Frauen und Männern eine positive Tendenz von 1.2 zu beobachten ist. Dieses Ergebnis ist mit demjenigen von Pützer (2001) konform, wenngleich der Effekt dort nicht so stark war. Im Übrigen bestehen zwar ebenfalls durchweg positive Tendenzen, die jedoch weniger deutlich sind: Bei der Klasse KINDER beträgt die Tendenz 0.24, bei der Klasse JUGENDLICHE 0.32 und bei der Klasse SENIOREN lediglich 0.12. Die Wahrscheinlichkeitsdichten lassen den Schluss zu, dass das Geschlecht auf der Basis von $shim\_la$ lediglich bei den jüngeren Erwachsenen zu unterscheiden ist. Das gilt auch für die übrigen Shimmer-Maße, auch wenn die Unterscheidbarkeit insbesondere in der Klasse ERWACHSENE etwas geringer ausfällt.

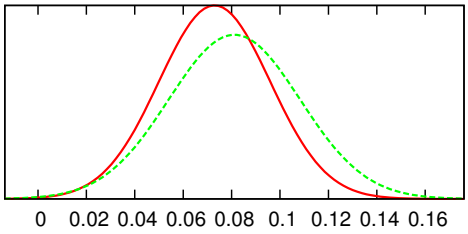
5.17 Shimmer *shim_l*



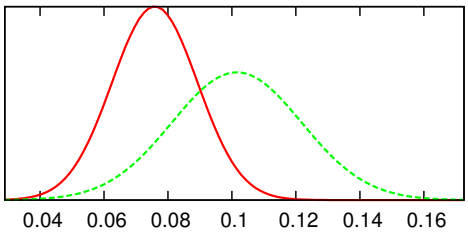
	μ	Σ	$\bar{\mu}$
gesamt	0.0815	0.0243	0
weiblich	0.0763	0.0222	-0.2133
männlich	0.0867	0.0252	0.2133



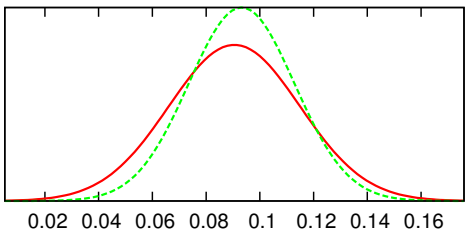
Kinder			
gesamt	0.0684	0.0201	0
weiblich	0.066	0.0184	-0.1221
männlich	0.0709	0.0213	0.1221



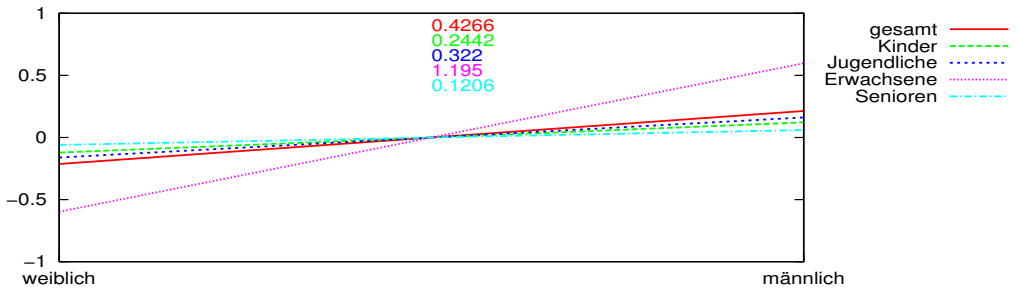
Jugendliche			
gesamt	0.0771	0.0255	0
weiblich	0.073	0.023	-0.161
männlich	0.0812	0.0271	0.161



Erwachsene			
gesamt	0.0887	0.0215	0
weiblich	0.0759	0.0134	-0.5975
männlich	0.1015	0.0203	0.5975



Senioren			
gesamt	0.0919	0.0222	0
weiblich	0.0905	0.0244	-0.0603
männlich	0.0932	0.0197	0.0603

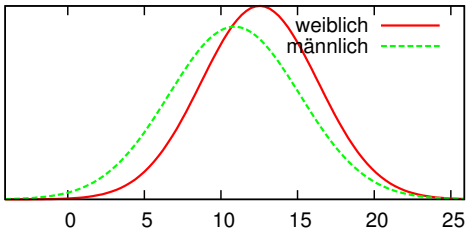


5.2.7 Harmonizität

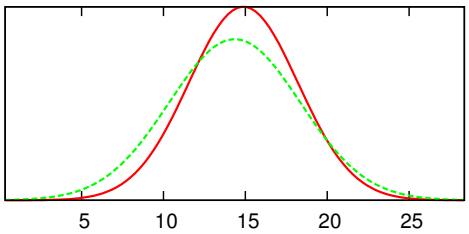
Wie schon beim Shimmer, ist auch die mittlere Harmonicity-to-Noise-Ratio ein Maß, welches die Geschlechter am deutlichsten im Fall der jüngeren Erwachsenen diskriminiert. Die Tendenz zwischen Frauen und Männern ist allerdings entgegen Pützer (2001) mit -1.26 stark negativ. Mit -0.48 zeigen die Jugendlichen einen schwächeren, aber immer noch deutlichen Abfall. Bei den Kindern und den Senioren ist die Tendenz mit -0.14 bzw. -0.16 eher schwach. Die Wahrscheinlichkeitsdichten bestätigen dieses Bild: Bei den jüngeren Erwachsenen lassen sie eine gute Unterscheidbarkeit vermuten, die bei den Jugendlichen geringer ist, durchaus aber besteht. Dagegen sind auf der Basis von $harm\_mean$ weder Mädchen von Jungen noch Seniorinnen von Senioren zu unterscheiden.

Anders sieht es bei der Standardabweichung der Harmonicity-to-Noise-Ratio aus (vgl. Anhang A, Seite 289): Neben den jüngeren Erwachsenen sind es hier die Senioren, die nach dem Geschlecht unterschieden werden können. Bei den Kindern und Jugendlichen überlagern sich nicht nur die Gauß-Kurven, sondern auch die Mittelwerttendenzen sind mit 0.07 bzw. 0.06 zwar statistisch signifikant, aber sehr schwach. Bei den Erwachsenen ist eine starke negative Tendenz von -1.02 und bei den Senioren eine mittlere negative Tendenz von -0.38 zu verzeichnen.

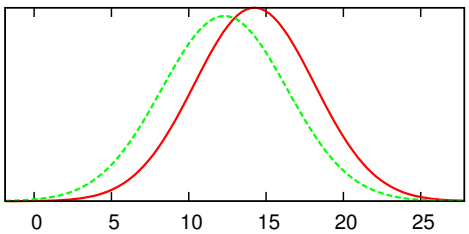
5.18 Harmonizität $harm\_mean$



	μ	Σ	$\bar{\mu}$
gesamt	11.702	4.135	0
weiblich	12.529	3.816	0.2
männlich	10.875	4.274	-0.2

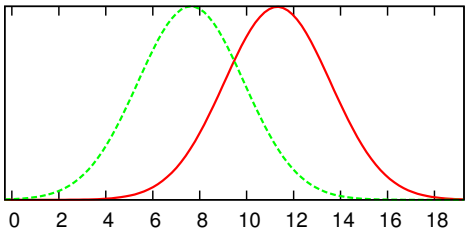


Kinder			
gesamt	14.617	3.697	0
weiblich	14.88	3.336	0.071
männlich	14.353	4.008	-0.071

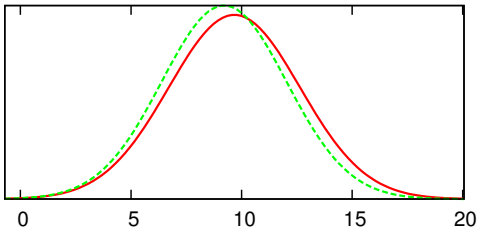


Jugendliche			
gesamt	13.28	4.083	0
weiblich	14.25	3.878	0.238
männlich	12.31	4.051	-0.238

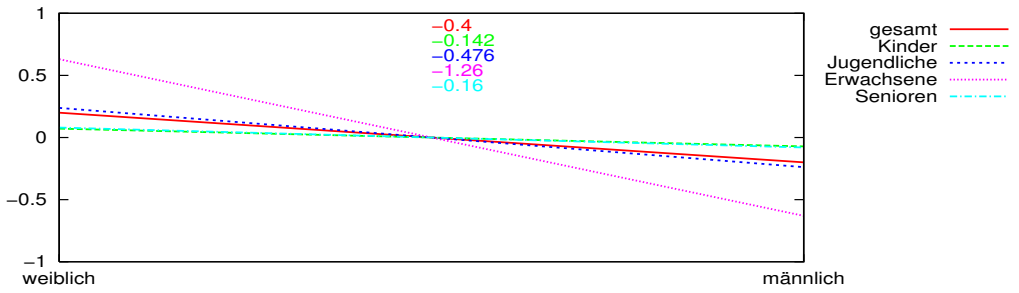
5.18 Harmonizität (Fortsetzung) $harm\_mean$



	Erwachsene		
gesamt	9.468	2.921	0
weiblich	11.309	2.271	0.63
männlich	7.627	2.265	-0.63



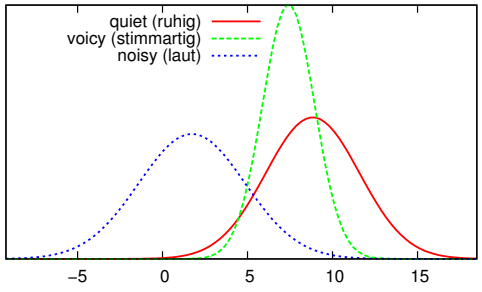
	Senioren		
gesamt	9.443	2.905	0
weiblich	9.675	2.966	0.08
männlich	9.211	2.823	-0.08



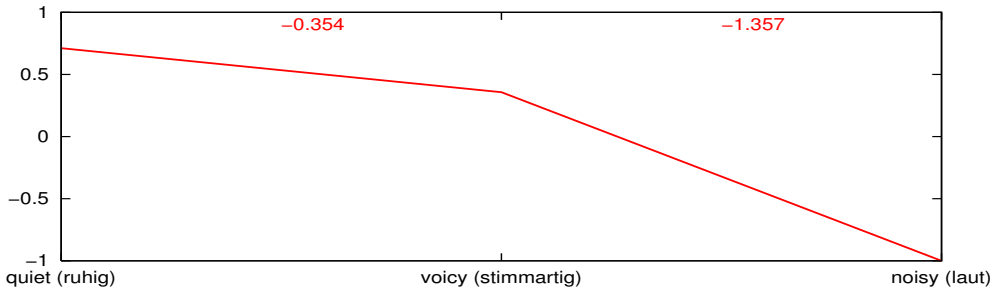
5.3 Kontext

Die mittlere Harmonicity-to-Noise-Ratio ($harm\_mean$) ist erwartungsgemäß bei den Äußerungen ohne Kontext am höchsten. Zwischen QUIET und VOICY gibt es eine mittlere negative Tendenz von -0.35 und zwischen VOICY und NOISY eine starke negative Tendenz von -1.36. Diese Effekte bestätigen die Hypothesen aus Abbildung 4.5, und die Wahrscheinlichkeitsdichten lassen die Annahme einer guten Unterscheidbarkeit zu. Sämtliche Mittelwerttendenzen sind – wie es bei der Beschreibung der Ergebnisse bezüglich der Sprechercharakteristika der Fall war – statistisch signifikant mit $p \leq 0.01$ (t-Test), falls nicht anders angegeben.

5.19 Harmonizität $harm\_mean$

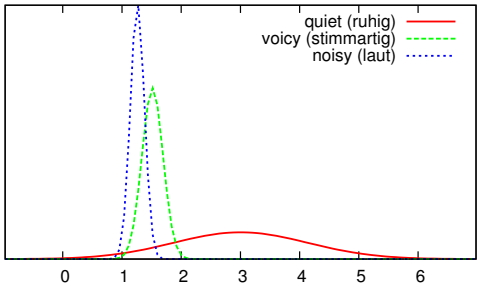


	μ	Σ	$\bar{\mu}$
gesamt	5.984	4.001	0
quiet (ruhig)	8.83	2.761	0.711
voicy (stimmartig)	7.414	1.535	0.357
noisy (laut)	1.726	3.126	-1

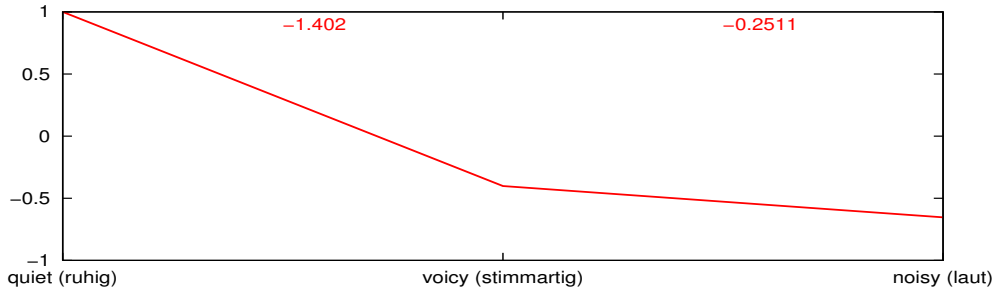


Bezüglich der Intensität ist erwartungsgemäß ebenfalls eine durchgehende negative Tendenz zu beobachten, die allerdings – im Gegensatz zu $harm\_mean$ – zwischen QUIET und VOICY stärker ist als zwischen VOICY und NOISY. Dies gilt sowohl für die $inten\_ratio$ mit den Tendenzen von -1.40 bzw. -0.25 als auch für $intens\_stddev$ mit -1.28 bzw. -0.54. In beiden Fällen sind die Klassen anhand der Wahrscheinlichkeitsdichten deutlich voneinander zu unterscheiden.

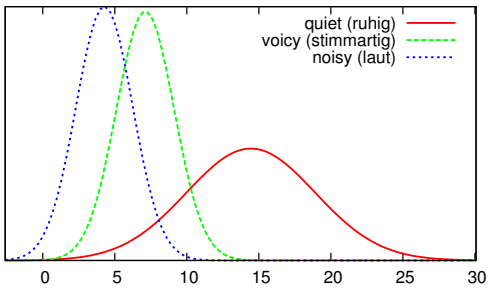
5.20 Intensität *intens_ratio*



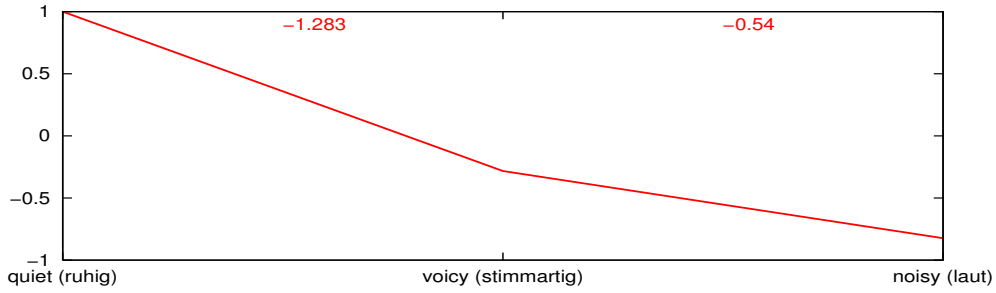
	μ	Σ	$\bar{\mu}$
gesamt	1.9237	1.0165	0
quiet (ruhig)	3.0029	1.136	1
voicy (stimmartig)	1.5152	0.1778	-0.402
noisy (laut)	1.2599	0.1186	-0.6531



intens_stddev



	μ	Σ	$\bar{\mu}$
gesamt	8.602	5.251	0
quiet (ruhig)	14.445	4.467	1
voicy (stimmartig)	7.117	2.008	-0.283
noisy (laut)	4.279	1.968	-0.823



6 Zusammenfassung der empirischen Studien

Bezüglich der Stimmen von Kindern im vorpubertären Alter konnte in der vorliegenden Studie ein geschlechtsspezifischer Unterschied in der mittleren Höhe der Grundfrequenz festgestellt werden: Der gemessene Mittelwert bei den Mädchen beträgt 249.26 Hz, während er bei den Jungen um etwa 10 Hz darunter liegt (239.37 Hz). Die normierte Tendenz ist mit -0.28 zwar recht schwach, aber dennoch signifikant. Bei der Interpretation dieses Ergebnisses ist aber zu bedenken, dass keine Sprecher unter zehn Jahren betrachtet wurden.

Die Stimmen der Jugendlichen zeigen erhöhte Jitter- und Shimmerwerte, was möglicherweise auf den Kontrollverlust aufgrund der rapiden anatomischen Veränderungen während der Pubertät zurückzuführen ist. Für diese Interpretation spricht, dass die Werte sowohl bei den Kindern als auch bei den jüngeren Erwachsenen geringer sind. Kinder und Jugendliche unterscheiden sich von Erwachsenen und Senioren durch harmonischere Stimmen: Zusammengenommen beträgt die Harmonicity-to-Noise-Ratio für Kinder und Jugendliche durchschnittlich 14.0 dB, während sie bei Erwachsenen und Senioren im Mittel nur bei 9.4 dB liegt.

Die Grundfrequenz als das Hauptunterscheidungskriterium zwischen erwachsenen Frauen und Männern bestätigt sich auch in dieser Studie: Bei den Frauen ist der Durchschnittswert bei 197.68 Hz angesiedelt und bei den Männern bei 125.78 Hz. Was die Maße der Stimmqualität betrifft, zeigt sich hier jedoch ein komplementäres Bild im Vergleich zu der Studie von Pützer (2001): Bei den Sprecherinnen ist weniger Jitter, weniger Shimmer und eine höhere Harmonicity-to-Noise-Ratio zu beobachten. Die Unterschiede in der Standardabweichung der Grundfrequenz sind nicht signifikant.

Die Stimmen der Senioren enthalten erwartungsgemäß mehr Jitter und mehr Shimmer als alle anderen Altersgruppen, was sowohl für die Frauen als auch für die Männer gilt. Beide Maße können hauptsächlich als ein Symptom der Stimmalterung angesehen werden. Die leistungsbezogenen Maße entsprechen nur zum Teil den Vorhersagen: Die untere Grenze des Stimmumfangs bei Frauen geht erwartungsgemäß nach unten. Allerdings ist das – wenn auch in geringerem Maße – bei den Männern ebenso der Fall, was nicht mit den Hypothesen konform geht. An der oberen Grenze des Stimmumfangs sind bei beiden Geschlechtern nur sehr schwache Veränderungen messbar. Was die altersbedingten Veränderungen der mittleren Stimmtonhöhe betrifft, werden die Hypothesen bestätigt: Die Grundfrequenz sinkt bei den Frauen ab, während sie bei den Männern ansteigt. Letztere Tendenz ist zwar schwach, aber statistisch signifikant.

Die Artikulationsgeschwindigkeit der Senioren beider Geschlechter ist erwartungsgemäß deutlich geringer als die der jüngeren Erwachsenen. Dennoch kann sie nicht ausschließlich als ein Symptom der Veränderungen des Sprechverhaltens im Alter angesehen werden, da sie bei Kindern und Jugendlichen auf einem ähnlichen Niveau liegt. Was die Sprechpausen betrifft, werden die Hypothesen ebenfalls bestätigt: Die meisten Pausen sind bei den Senioren zu verzeichnen, die wenigsten bei den jüngeren Erwachsenen. Das Niveau der Kinder und Jugendlichen liegt dazwischen. Die Dauer der Sprechpausen ist dagegen bei Kindern, Jugendlichen und Erwachsenen in etwa gleich, während sie bei den Senioren durchschnittlich länger ist.

Die hier präsentierten Resultate sind geeignet, um als Beitrag zum Aufbau einer Referenzbasis für die Analyse der Stimme und des Sprechverhalten zu dienen. Die Daten bezüglich der Artikulationsgeschwindigkeit stellen beispielsweise eine zusätzliche Informationsquelle für Studien zu Tempo(variationen) in der Sprachproduktion dar, wie sie von Trouvain (2004) beschrieben werden. Im Rahmen der vorliegenden Arbeit bilden sie jedoch in erster Linie die Basis für ein System zur automatischen Sprecherklassifikation. Hierzu konnten insgesamt in genügendem Maße sowohl alters- als auch geschlechtsspezifische Unterschiede bei den betrachteten Merkmalen gemessen werden. Allerdings wurde ebenso deutlich, dass eine wechselseitige Abhängigkeit zwischen den beiden Sprechercharakteristika besteht, so dass davon ausgegangen werden kann, dass bei bekanntem Geschlecht eine zuverlässige Einschätzung der Altersklasse möglich ist und umgekehrt. Bevor in Teil II daher ein Vergleich verschiedener Klassifikationsmethoden bezüglich ihrer Eignung für die Sprecherklassifikation angestellt wird, werden die Ergebnisse der Korpusanalysen zunächst in einer für das maschinelle Lernen geeigneten Art und Weise aufbereitet. Dabei besteht die Rolle der Kontextklassifikation darin, erstens der Applikation Informationen über die Umgebung des Sprechers zur Verfügung zu stellen und zweitens die Sprecherklassifikation an sich zu verbessern, indem eine geeignete Auswahl von Merkmalen getroffen wird. Die diesbezüglich präsentierten Resultate belegen eine prinzipielle Unterscheidbarkeit von ruhigen, lauten und stimmartigen Kontexten.

Das primäre Ziel zukünftiger Korpusanalysen wird eine genauere Differenzierung der Altersklassen sein. Um dies zu erreichen, bieten sich eine Reihe von Möglichkeiten zur Verbesserung des Verfahrens an. Einer der Ansatzpunkte betrifft beispielsweise die Verwendung multilingualer Spachproben (Englisch und Deutsch). Trotz der Annahme einer weitestgehenden Sprachunabhängigkeit der untersuchten Merkmale (vgl. Abschnitt 1.5), muss davon ausgegangen werden, dass die sprecherabhängigen Effekte zumindest teilweise durch sprachabhängige Effekte überlagert werden. Im Zuge der Weiterentwicklung des AGENDER-Ansatzes, auf die in Kapitel 10.1 genauer eingegangen wird, sind Analysen geplant, denen ausschließlich unilinguale Sprachproben zugrunde liegen werden.

Des Weiteren ist das Potenzial zur Erweiterung der Menge der Merkmale bisher nicht erschöpfend genutzt worden. Es bestehen Pläne, in Zusammenarbeit mit der Universität von Lund (Schweden) die Möglichkeiten einer automatischen Extraktion weiterer akustischer Manifestationen des Sprecheralters zu eruieren. Dazu gehören *cepstrale* Merkmale (vgl. Abschnitt 2.2.3) und vor allen Dingen die Formantenfrequenzen F1 und F2, deren Relevanz von schwedischer Seite be-

reits nachgewiesen wurde (vgl. z. B. Schötz, 2003 und Schötz, 2004a). Die Betrachtung subtilerer Eigenschaften der Sprache stellt jedoch neue Anforderungen an das zugrunde liegende Verfahren: Statt eine Erweiterung der Datenbasis vorzunehmen, wird diesbezüglich zunächst eine kleine, besser zu kontrollierende Menge von Daten betrachtet. Auf diese Art und Weise können zusätzliche Probleme bei der automatischen Extraktion besser gelöst werden, wie z. B. im Fall der Formantenfrequenzen die Segmentierung des Eingabesignals, die nötig ist, um die informativen Bereiche (Vokale und Frikative) zu identifizieren. Die initiale Studie wird sich daher auf die Aussprache einzelner, für alle Sprecher gleicher Wörter beschränken, wobei als Quelle das entsprechend aufbereitete *Swedia*-Korpus vorgesehen ist.

Teil II

Ein zweistufiger Ansatz zur automatischen Sprecherklassifikation

7 Automatische Sprecherklassifikation als Mustererkennungsproblem

7.1 Einführung

Im vorangegangenen Kapitel wurden die Ergebnisse von Korpusanalysen präsentiert, die den Schluss zulassen, dass auf Basis der Sprache sowohl das Sprechergeschlecht als auch das Sprecheralter unterschieden werden können – mit der Einschränkung auf die Altersklassen KINDER (Alter $<$ 12), JUGENDLICHE (13 $<=$ Alter $<$ 20), ERWACHSENE (20 $<=$ Alter $<$ 65) und SENIOREN (65 $<=$ Alter). Dieser Schluss ist deshalb gerechtfertigt, weil für eine Reihe von Merkmalen signifikante Mittelwertdifferenzen zwischen den Sprecherklassen gefunden wurden, die zu einem großen Teil die Hypothesen bestätigen, die auf Basis von Studien über den Einfluss des Sprecheralters bzw. -geschlechts auf die Stimme und die Sprache aufgestellt worden sind.

Obgleich diese Ergebnisse an sich wertvolle Erkenntnisse zur Schaffung einer Referenzbasis erbringen, besteht das hauptsächliche Thema dieser Arbeit darin, ein maschinelles System zu beschreiben, welches in der Lage ist, auf Basis dieser Merkmale automatisch das Geschlecht und die Altersklasse des Sprechers einzuschätzen. Hierbei handelt es sich um ein so genanntes *Mustererkennungsproblem*, d. h. es wird versucht, aus einer Kombination einzelner Merkmale (Muster) eine diskrete Klasse abzuleiten.

Die Mustererkennung wird im Allgemeinen als ein Teilgebiet der *Künstlichen Intelligenz* angesehen. Zu ihren Anwendungsgebieten gehört die Spracherkennung, bei der wie hier aus elementaren Merkmalen eine Klasse abgeleitet wird, in diesem Fall ein Laut, der sich mit anderen zu einer Lautfolge und schließlich zu einer Äußerung zusammensetzt. Auf die Besonderheiten der Spracherkennung als Mustererkennungsproblem wird noch etwas genauer in Abschnitt 8.1 eingegangen. Weitere Anwendungen sind beispielsweise die optische Zeichenerkennung, bei der aus einem Muster von Hell-Dunkel-Unterschieden ein Buchstabe erkannt werden soll.

Nach Duda, Hart und Stork (2000, S. 14) umfasst der Entwurf eines Mustererkennungssystems die Wiederholung einer Anzahl von verschiedenen Aktivitäten: (1) Datensammlung, (2) Merkmalsauswahl, (3) Modell-Auswahl, (4) Training und (5) Evaluation. Im Zuge der Auswertung der in Teil I präsentierten Korpusanalysen wurden bereits mehrere Iterationen der Phasen eins und zwei durchgeführt, so dass die Datenbasis und die Menge der Merkmale zunächst als gegeben betrachtet werden können. Im Fokus stehen dementsprechend nun die Phasen drei bis fünf.

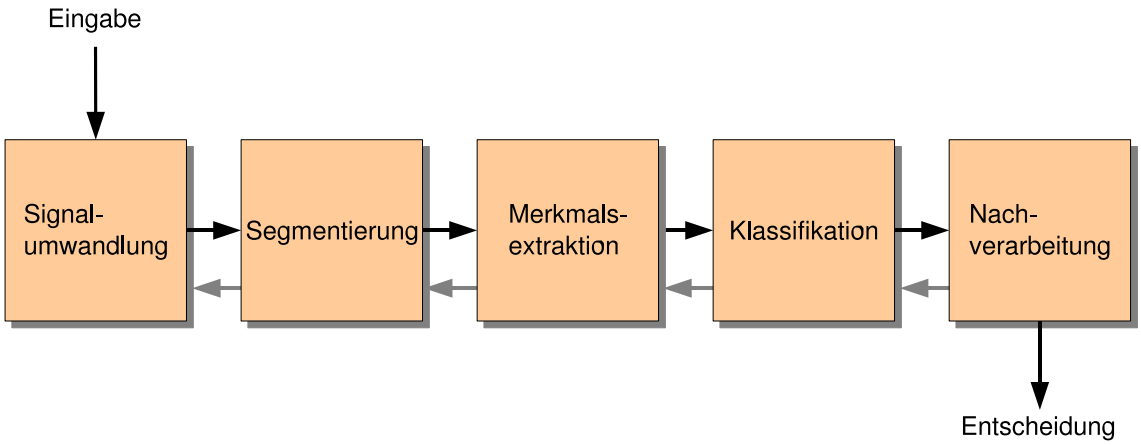


Abbildung 7.1: Typischer Ablauf eines Mustererkennungssystems nach Duda et al. (2000, S. 10).

In Abbildung 7.1 wird der typische Ablauf eines Mustererkennungssystems dargestellt (vgl. Duda et al., 2000, S. 10). Die Eingabe wird zunächst von einem Signalumwandler verarbeitet, wie z. B. von einer Kamera oder einem Mikrofon. Häufig hängen die Schwierigkeiten eines Mustererkennungsproblems mit den Einschränkungen des Umwandlers zusammen: Bandbreite, Auflösung, Empfindlichkeit und Störungen spielen eine Rolle bei der Entscheidung, ob die Qualität der vom Umwandler erzeugten Daten für die Klassifikation ausreichend ist. Im Fall von AGENDER erfolgt die Eingabe ausschließlich über Mikrofon, und daher wird die Qualität des Signals gemessen an der Abtastfrequenz, der Abtasttiefe und dem Anteil von Störgeräuschen. Die Eigenschaften des Eingabesignals wurden auf die Anforderungen der Anwendung abgestimmt, d. h. sie entsprechen mit 8 KHz Abtastfrequenz und 16 Bit Abtasttiefe einer „Telefonqualität“.

7.1.1 Segmentierung und Gruppierung

Zur Vorverarbeitung gehört nach Duda et al. (2000, S. 9) das Problem der *Segmentierung*, zu dessen Veranschaulichung sie ein einfaches Mustererkennungssystem beschreiben: In einer Fischfabrik rollen über ein Förderband sowohl Lachse als auch Seebarsche. Das System soll anhand von Daten einer Kamera eine Sortierung vornehmen, wozu Merkmale wie Größe, Farbe, Maserung der Haut usw. abgeleitet werden. Da die Fische nacheinander über das Band rollen, besteht das Problem der Segmentierung zunächst darin zu entscheiden, welche Teile des von der Kamera gelieferten Bildes zu einem Fisch und welche zu dem nächsten gehören. Bei dem oben genannten Beispiel der Spracherkennung bedeutet Segmentierung die Zuordnung von Teilen des kontinuierlichen Sprachsignals zu Phonemen, und bei der optischen Zeichenerkennung müssen Grenzen der Buchstaben erkannt werden. Bei der Sprecherklassifikation, wie sie in AGENDER durchgeführt wird, ist die Segmentierung natürlich gegeben, da eine Einheit einer Äußerung gleichkommt. Es wird also davon ausgegangen, dass eine Äußerung jeweils nur von einem Sprecher gemacht wird, was ohne weiteres als eine zulässige Annahme gelten kann, wenn man das Problem der Hinter-

grundstimmen zunächst einmal außer Acht lässt.

Von der Segmentierung kommen wir jedoch gleich zu dem Problem der *Gruppierung* eines zusammengesetzten Objektes. In dem Fischfabrik-Beispiel können Seebarsche beliebig mit Lachsen gemischt auf dem Förderband liegen – es gibt daher keinen Zusammenhang zwischen der Klasse, für die sich das System gerade entschieden hat, und derjenigen, die als nächstes kommt. Die Abfolge der Fische setzt sich nicht zu einem „großen Bild“ zusammen. Bei der Spracherkennung dagegen wird davon ausgegangen, dass sich die erkannten Phoneme zu (bekannten) Wörtern und diese zu (interpretierbaren) Äußerungen zusammensetzen. Über statistische Modelle der zugrunde liegenden Sprache lassen sich Wahrscheinlichkeiten zur Vorhersage von Kombinationen (Gruppen) von Klassen berechnen. Eine ähnliche Situation liegt bei der optischen Zeichenerkennung vor: Die erkannten Buchstaben setzen sich zu Wörtern, Sätzen, Absätzen und schließlich zu einem Text zusammen. Anhand dieser Beispiele ist erkennbar, dass je nach Domäne das Problem der Gruppierung sehr komplex werden kann, so dass mehr als nur Vorverarbeitungsprozesse davon beeinflusst werden. Wie in Abbildung 7.1 durch die nach links gerichteten Pfeile angedeutet wird, kann es in einem Mustererkennungssystem durchaus vorkommen, dass es zu einer Art *Backtracking* zwischen höheren und niedrigeren Prozessen kommt. In AGENDER wird davon ausgegangen, dass aufeinander folgende Äußerungen solange ein und demselben Sprecher zugeordnet werden können, bis ein Sprecherwechsel signalisiert wird. Die Kombination der Einzelergebnisse erfolgt in einer späteren Phase, die in Abbildung 7.1 als *Nachverarbeitung* bezeichnet wird.

7.1.2 Merkmalsextraktion

Als nächster und sehr wichtiger Schritt erfolgt die *Merkmalsextraktion*, d. h. die Merkmale, anhand derer das System Seebarsche von Lachsen unterscheiden soll, müssen aus dem (segmentierten) Eingabesignal abgeleitet werden. Nach Duda et al. (2000, S. 11) sind die konzeptuellen Grenzen zwischen Merkmalsextraktion und Klassifikation nicht eindeutig: Ein idealer Merkmalsextraktor würde zu einer Repräsentation führen, die die Klassifikation zu einem trivialen Problem macht – auf der anderen Seite würde ein omnipotenter Klassifikator nicht die Hilfe eines hoch entwickelten Merkmalsextraktors benötigen. Die Unterscheidung wird eher aus praktischen als aus theoretischen Gründen benötigt: Die traditionelle Aufgabe eines Merkmalsextraktors besteht darin, ein Objekt so zu charakterisieren, dass die entsprechenden Werte für Objekte derselben Kategorie sehr ähnlich und für Objekte unterschiedlicher Kategorien sehr verschieden sind. Verfahren, die diese Voraussetzungen mit sich bringen, um die Aufgabe der Merkmalsextraktion im AGENDER-System zu übernehmen, wurden in Abschnitt 4.2 beschrieben. Im Folgenden werden daher die Verfahren, die bei der Korpusanalyse zur Ableitung von $pitch\_mean$, $jitt\_rap$, $shim\_apq11$ usw. verwendet wurden, als *Merkmalsextraktoren* bezeichnet. Wenn – wie in diesem Fall – mehrere Merkmale berücksichtigt werden, werden diese als *Merkmalsraum* bezeichnet. Bei der Klassifikation wird versucht, Regionen in diesem Merkmalsraum zu identifizieren, die den jeweiligen Klassen zugeordnet werden können.

In praktischen multikategorialen Anwendungen kann es nach Duda et al. (2000, S. 107) vorkommen, dass die Menge der Merkmale eine Größe von fünfzig oder gar hundert erreicht. Dabei trägt theoretisch jedes weitere Merkmal etwas zu der Unterscheidbarkeit der Klassen bei, wobei

nicht notwendigerweise jedes der Merkmale eine unabhängige Information darstellt. Die nützlichsten Merkmale sind diejenigen, bei denen die Mittelwerttendenzen der verschiedenen Klassen im Vergleich zu der Standardabweichung groß sind. Solange jedoch die Mittelwerte verschieden sind, ist kein Merkmal nutzlos. Auf der anderen Seite verursacht die Hinzunahme eines weiteren Merkmals Kosten (Zeit und Komplexität), sowohl bei der Merkmalsextraktion als auch bei der Klassifikation (vgl. ebd.). Obwohl immer noch angenommen werden darf, dass eine gewisse Verbesserung der Klassifikationsperformanz erreicht werden kann, sollte demnach die Menge der *Merkmalskandidaten* einer Prüfung unterzogen werden, um offensichtlich überflüssige Elemente eliminieren zu können.

7.1.3 Klassifikation

Die Aufgabe des Klassifikators ist es, den Merkmalsvektor zu nehmen und das dazugehörige Objekt – d. h. in dem Fall den Sprecher – einer Klasse zuzuweisen, z. B. ERWACHSENE WEIBLICH. Da eine perfekte Klassifikation oft unmöglich ist, muss das Problem allgemeiner formuliert werden: Wie hoch sind die Wahrscheinlichkeiten, dass das Objekt zu einer der alternativen Klassen gehört? Die Abstraktion, die durch die Merkmalsextraktion erfolgt ist, erlaubt eine weitestgehend domänenunabhängige Theorie der Klassifikation (vgl. Duda et al., 2000, S. 12).

Der Schwierigkeitsgrad bei der Klassifikation hängt von der Variabilität der Merkmalswerte für Objekte derselben Klasse ab, einem Wert, der im Fall der Altersklassifikation sehr hoch ist (vgl. Kapitel 3.4). Eines der Probleme, die in der Praxis auftauchen, besteht darin, dass es nicht immer möglich ist, für eine bestimmte Eingabe alle Merkmale zu bestimmen. Die Frage ist, wie der Klassifizierer dies kompensieren soll. Es könnte angenommen werden, dass der fehlende Wert Null ist oder der Durchschnitt der Werte, die bereits beobachtet wurden. Diese Methode ist jedoch nicht optimal (ebd.). Die verschiedenen Klassifikationsmethoden, die in Abschnitt 8 miteinander verglichen werden, unterscheiden sich unter anderem durch ihre Stabilität gegenüber fehlenden Daten oder Rauschen. Dies ist jedoch eines derjenigen Probleme, die durch den in der vorliegenden Arbeit vorgeschlagenen zweistufigen Klassifikationsansatz gelöst werden können. In Abschnitt 9.3.1 wird darauf genauer eingegangen.

Ein Klassifizierer wird häufig anhand einer Menge von *Diskriminantenfunktionen* $g_i(\mathbf{x}), i = 1, \ldots, c$ beschrieben, wobei c die Menge der Klassen repräsentiert (vgl. Duda et al., 2000). Somit kann der Klassifizierer als eine Art Netzwerk angesehen werden, das c Diskriminantenfunktionen berechnet und die Kategorie anhand des größten Diskriminanten wählt. Abbildung 7.2 stellt die funktionale Struktur eines allgemeinen statistischen Klassifizierers dar. Er beinhaltet d Eingaben und c Diskriminantenfunktionen oder Entscheidungsregeln $g_i(x)$. Ein nachfolgender Schritt bestimmt, welcher der Diskriminantenwerte der maximale ist und klassifiziert die Eingabemuster entsprechend. Die Entscheidungsregeln unterteilen den Merkmalsraum in c *Entscheidungsregionen* $\Re_1, \ldots, \Re_c$. Wenn $g_i(\mathbf{x}) > \mathbf{g_j}(\mathbf{x})$ für alle $j \neq i$, dann ist $\mathbf{x}$ in $\Re_i$ und die Entscheidungsregel weist $\mathbf{x}$ die Klasse ω_i zu. Die Grenzen zwischen den Entscheidungsregionen werden *Entscheidungsgrenzen* genannt.

Bei der Entwicklung der Entscheidungsregeln, dem Training eines Klassifizierers, wird nach einer *Generalisierung* gesucht. In Abbildung 7.3 und 7.4 werden zwei alternative Entscheidungs-

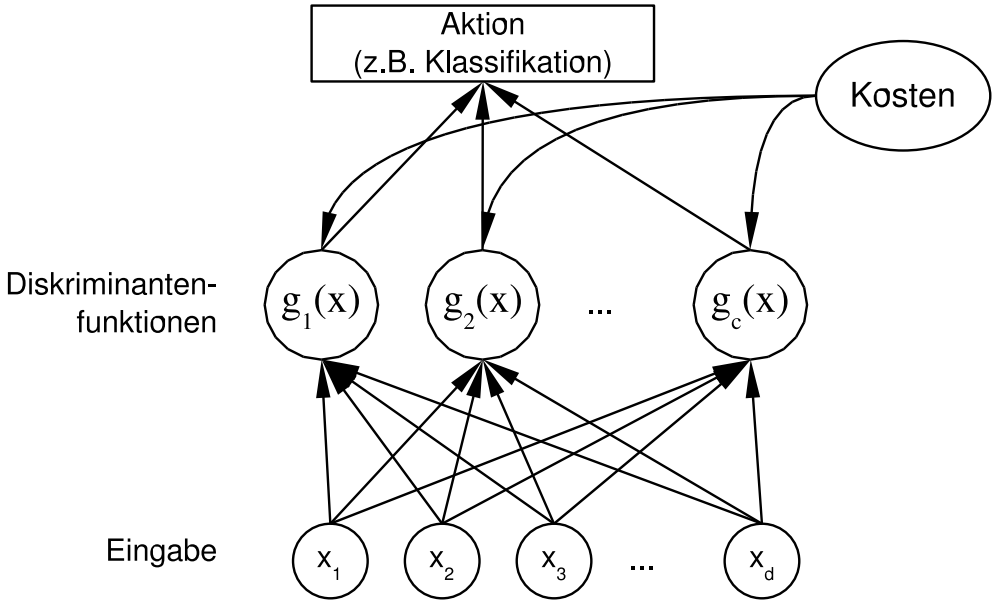


Abbildung 7.2: Funktionale Struktur eines allgemeinen statistischen Klassifizierers nach Duda et al. (2000, S. 30).

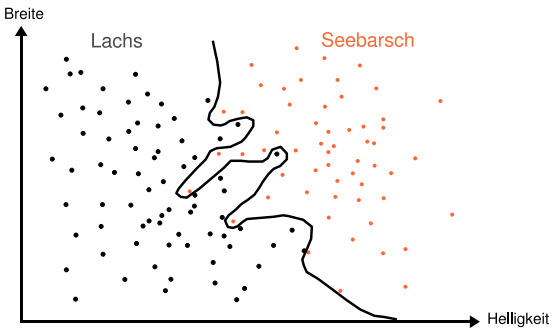


Abbildung 7.3: Übermäßig komplexe Entscheidungsgrenze (Overfitting).

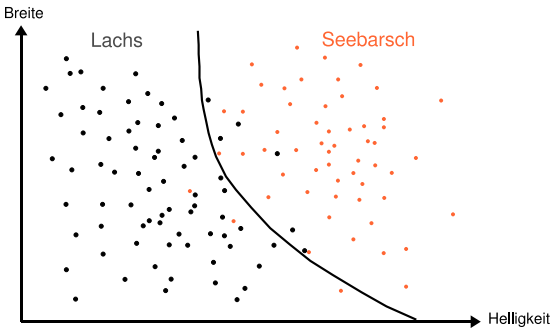


Abbildung 7.4: Entscheidungsgrenze mit optimaler Balance zwischen Performanz auf dem Trainingsset und Performanz auf neuen Beispielen.

grenzen für das Fischfabrik-Beispiel dargestellt. Bei der Variante in Abbildung 7.3 wird eine perfekte Trennung der Trainingsbeispiele erreicht. Die Wahrscheinlichkeit jedoch, dass neu hinzukommende, bisher ungesehene Instanzen korrekt klassifiziert werden, kann als eher gering eingeschätzt werden. Dieses Problem wird *Overfitting*-Problem genannt. Bei der Entscheidungsgrenze in Abbildung 7.4 dagegen wurde durch eine Generalisierung ein gewisser Fehler auf den Trainingsbeispielen zugunsten einer besseren Performanz für neue Instanzen in Kauf genommen.

7.2 Probabilistische Grundlagen der Klassifikation

Nach Duda et al. (2000, S. 20) kann die *Bayes'sche Entscheidungstheorie* als der fundamentale statistische Ansatz für das Problem der Mustererkennung angesehen werden. Da die Theorie auch in AGENDER eine wichtige Rolle spielt, beispielsweise bei der Zusammenstellung der Merkmalsets für die Klassifizierer, soll zunächst eine kurze Einführung in ihre Grundlagen gegeben werden. Die Tatsache, dass es sich bei dem gegebenen Sprecher beispielsweise um einen Mann handelt, wird *Weltzustand* genannt und durch ω_1 repräsentiert, der alternative Weltzustand (‚dass es sich um eine Frau handelt,) als ω_2. Die denkbar einfachste Entscheidungsregel basiert auf den A-priori-Wahrscheinlichkeiten $P(\omega_1)$ und $P(\omega_2)$ und besagt: Wenn keine andere Information zur Verfügung steht, entscheide ω_1, falls $P(\omega_1) > P(\omega_2)$, und ansonsten umgekehrt. Die Genauigkeit dieser Entscheidungsregel entspricht dem kleineren Wert von $P(\omega_1)$ und $P(\omega_2)$.

Eine genauere Zuweisung eines Musters zu einer Klasse kann auf der Basis der klassenspezifischen Wahrscheinlichkeitsdichte $p(x|\omega)$ erreicht werden. Der Unterschied zwischen $p(x|\omega_1)$ und $p(x|\omega_2)$ beschreibt die Unterschiede zwischen Frauen und Männern über die gesamte Population bezüglich der Eigenschaft x. Das Problem ist jedoch, dass $p(x|\omega)$ in der Regel unbekannt ist und daher aufgrund einer Sammlung von Beispielinstanzen geschätzt werden muss. Die Gauß'sche Wahrscheinlichkeitsdichte stellt in vielen praktischen Fällen eine gute Annäherung dar, und zwar immer dann, wenn davon ausgegangen werden kann, dass es für jede der Klassen eine Art prototypisches Muster μ_i gibt und dass die Muster, die in der Trainingsdatenbank zu finden sind, durch eine Reihe von zufälligen Prozessen korrumpierte Varianten dieses Prototyps sind. Zufallsexperimente, wenn sie sehr häufig wiederholt werden, erzeugen nämlich Histogramme, die die Glockenform einer *Gauß'schen Normalverteilung* approximieren. Man werfe beispielsweise hundert Mal eine Münze, zähle die Häufigkeit, mit der „Kopf“ oben liegt, und wiederhole dieses Experiment sehr häufig. Unter normalen Bedingungen wird die Zahl Fünfzig am häufigsten das Ergebnis sein, Hundert oder Null werden dagegen so gut wie nie vorkommen. Insgesamt erhält man eine Verteilung der Ergebnisse, wie sie in Abbildung 7.5 dargestellt wird: eine Gauß-Verteilung. Die treppenartige Struktur der Kurve entsteht dadurch, dass die Ergebnisse in diesem Fall nur ganzzahlig sein können.

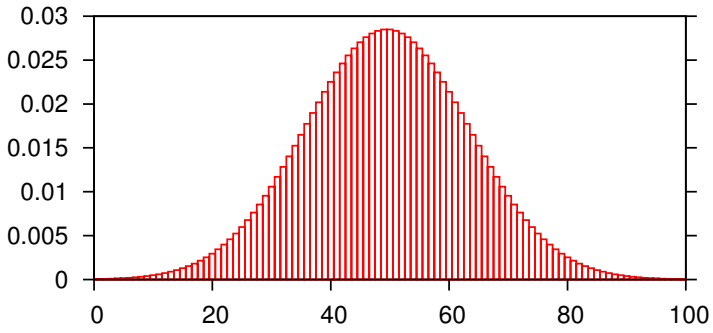


Abbildung 7.5: Anzahl der „Köpfe“ von hundert Münzwürfen bei ausreichend häufiger Wiederholung.

Die reellwertige univariate Gauß'sche Normalverteilung wird beschrieben durch

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \quad exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \tag{7.1}$$

und wird vollständig spezifiziert durch zwei Parameter: den Mittelwert μ und die Varianz σ^2. Häufig wird die Abkürzung $p(x) \sim N(\mu, \sigma^2)$ verwendet, um auszudrücken, dass x normal verteilt ist mit einem Mittelwert von μ und einer Varianz von σ^2.

Bereits bei der Diskussion der Ergebnisse der Korpusanalysen in Kapitel 5 wurde von der univariaten klassenspezifischen Gauß'schen Wahrscheinlichkeitsdichte Gebrauch gemacht, da sie ein zweckmäßiges Verfahren zur Beurteilung der Unterscheidbarkeit von Alters- bzw. Geschlechtsklassen auf Basis eines gegebenen Merkmals darstellt.

7.2.1 Diskriminantenfunktionen eines Bayes'schen Klassifizierers

Die Diskriminantenfunktion eines Bayes'schen Klassifizierers basiert auf der klassenspezifischen Wahrscheinlichkeitsdichte: $g_i(\mathbf{x}) = p(\omega_i|\mathbf{x})$. Das $\mathbf{x}$ ist fett gedruckt, um anzuzeigen, dass es sich um einen Merkmalsvektor handelt. Falls dieser jedoch aus mehr als nur einem Element besteht, können auf Basis einer univariaten Wahrscheinlichkeitsdichte, wie sie in Gleichung 7.1 beschrieben wird, keine Entscheidungsregionen $\Re_i$ berechnet werden. Die dazu notwendige multivariate Gauß'sche Wahrscheinlichkeitsdichte wird beschrieben durch

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}\,|\Sigma|^{1/2}} \quad exp\left[-\frac{1}{2}\,(\mathbf{x}-\mu)^t\,\Sigma^{-1}\,(\mathbf{x}-\mu)\right], \tag{7.2}$$

wobei $\mathbf{x}$ ein Spaltenvektor mit d Elementen ist, μ der Mittelwertvektor, Σ die d-mal-d-Kovarianzmatrix, $|\Sigma|$ ihre Determinante und Σ^{-1} die inverse Matrix. Des Weiteren ist $(\mathbf{x}-\mu)^t$ die Transponente von $(\mathbf{x}-\mu)$.

Duda et al. (2000, S. 41) leiten daraus die folgenden Diskriminantenfunktionen für einen Bayes'schen Klassifizierer ab:

$$g_i(\mathbf{x}) = \mathbf{x}^t \mathbf{W}_i \mathbf{x} + \mathbf{w}_i^t \mathbf{x} + \omega_{i0}, \tag{7.3}$$

wobei

$$\mathbf{W}_i = -\frac{1}{2}\Sigma_i^{-1}, \tag{7.4}$$

$$\mathbf{w}_i = \Sigma_i^{-1}\mu_i \tag{7.5}$$

und

$$\omega_{i0} = -\frac{1}{2}\mu_i^t \Sigma_i^{-1}\mu_i - \frac{1}{2}\ln|\Sigma_i| + lnP(\omega_i). \tag{7.6}$$

Der Bayes'sche Klassifizierer weist dem Muster $\mathbf{x}$ die Kategorie i zu, wenn $g_i(\mathbf{x}) > g_j(\mathbf{x})$ für alle $i \neq j$. Durch die Gleichsetzung von $g_i(\mathbf{x})$ und $g_j(\mathbf{x})$ ergeben sich demnach die *Entscheidungsgrenzen*.

Die Bayes'sche Entscheidungsregel kann auch als so genannte *Likelihood-Ratio* formuliert werden:

$$\frac{p(\mathbf{x}|\omega_0)}{p(\mathbf{x}|\omega_1)} \begin{cases} \geq \Theta & \text{weise } \mathbf{x} \text{ die Kategorie } \omega_0 \text{ zu} \\ < \Theta & \text{weise } \mathbf{x} \text{ die Kategorie } \omega_1 \text{ zu,} \end{cases} \tag{7.7}$$

wobei $p(\mathbf{x}|\omega_i)$ die Wahrscheinlichkeitsdichte für ω_i repräsentiert, die auch als *Likelihood von* ω_i *gegeben das Muster* $\mathbf{x}$ bezeichnet werden kann. Θ stellt den *Entscheidungsschwellenwert* dar (vgl. Reynolds et al., 2000, S. 21). Zur Klassifikation wird häufig die *Log-Likelihood-Ratio* angewendet (vgl. Reynolds et al., 2000, S. 22):

$$logp(\mathbf{x}|\omega_0) - logp(\mathbf{x}|\omega_1). \tag{7.8}$$

7.2.2 Bestimmung der Parameter

Auf Basis der A-priori-Wahrscheinlichkeit $P(w_i)$ und der klassenspezifischen Wahrscheinlichkeitsdichten $p(x|w_i)$ kann ein optimaler Klassifizierer entworfen werden (vgl. Duda et al., 2000, S. 84). Da dieser Parameter jedoch unbekannt ist, muss er auf der Basis einer Sammlung von Beispielen, den *Trainingsdaten*, ermittelt werden. Wenn dabei von einer multivariaten Gauß'schen

Wahrscheinlichkeitsdichte der Daten ausgegangen werden kann, vereinfacht sich dieses Problem: Gesucht wird nicht die unbekannte Funktion $p(x|w_i)$, sondern die Parameter-Vektoren $\theta_1, \ldots, \theta_c$ für c Kategorien. In unserem Fall besteht θ_i aus dem Mittelwert μ und der Kovarianzmatrix Σ_i (vgl. Gleichung 7.2).

Nehmen wir an, D sei eine Menge von Trainingsdaten, bestehend aus n Problemen $x_1, \ldots, x_n$. Dann gilt:

$$p(D|\theta) = \prod_{k=1}^{n} p(\mathbf{x_k}|\theta). \tag{7.9}$$

Wir nennen $p(D|\theta)$ *Likelihood* von θ in Bezug auf D. Gesucht wird dasjenige $\hat{\theta}$, das am ehesten mit den tatsächlichen Trainingsdaten übereinstimmt, das also $p(D|\theta)$ maximiert. Diese Methode zur Parameter-Bestimmung wird daher auch *Maximum-Likelihood-Methode* genannt. Abbildung 7.6 stellt einige Trainingsdaten des Merkmals $pitch\_mean$ für die Klasse WEIBLICH dar. Da es sich um den univariaten Fall handelt, entspricht das gesuchte θ den Werten μ (Mittelwert) und σ (Standardabweichung) für die gegebene Klasse. Wir nehmen an, σ ist bekannt (39.9), μ allerdings nicht. Vier der unendlich vielen potentiellen Verteilungen werden mit gepunkteten Linien darüber abgebildet. Um die richtige zu finden, wird Gleichung 7.9 auf alle „Kandidaten“ für μ angewendet. Abbildung 7.7 stellt das Ergebnis für eine zufällige Auswahl von tausend Werten aus der Trainingsdatenbank dar. Der Wert, der die *Likelihood* maximiert, ist 177.13. Man beachte, dass die Kurve in Abbildung 7.7 zwar eine (sehr enge) Gauß-Glocke beschreibt, jedoch eine Funktion von θ ist und somit keine Wahrscheinlichkeitsdichte darstellt. Die Fläche unter der Kurve ist irrelevant (vgl. ebd. S. 86). Je mehr Trainingsdaten für die Berechnung zugrunde gelegt werden, desto enger wird die *Likelihood*-Verteilung.

Bereits auf der Basis von tausend Datensätzen nähert sich der Wert $\hat{\theta}$, d. h. in dem Fall der fehlender Parameter μ, stark dem arithmetischen Mittel der $pitch\_mean$-Werte von 177.42. Tatsächlich kann man zeigen, dass μ für eine unbekannte Population dem arithmetischen Mittelwert der Trainingsdaten entspricht, dem so genannten *Probenmittel*. Es bietet sich an, hierfür die Notation $\hat{\mu}_n$ zu verwenden, um die Abhängigkeit von der Anzahl der Trainingsdaten anzudeuten (vgl. ebd. S. 88). Darüber hinaus lässt sich zeigen, dass auch für ein unbekanntes Σ die Einschätzung mithilfe der *Maximum-Likelihood*-Methode dem arithmetischen Mittel aus n Matrizen $(\mathbf{x} - \hat{\mu})(\mathbf{x}_k - \hat{\mu})^t$ entspricht (vgl. ebd. S. 89).

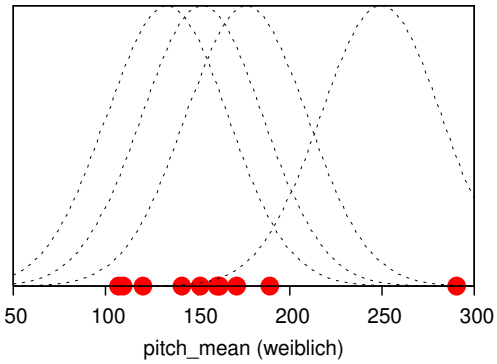


Abbildung 7.6: Einige Trainingsdaten des Merkmals $pitch\_mean$ für die Klasse WEIBLICH. Darüber: vier der unendlich vielen potentiellen Normalverteilungen. Gesucht werden die Parameter μ und σ, die eine Verteilung ergeben, die den Trainingsdaten am ehesten entspricht.

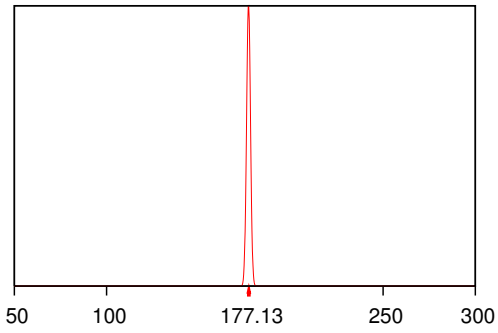


Abbildung 7.7: Anwendung der *Maximum-Likelihood*-Methode auf den in Abbildung 7.6 dargestellten Fall bei bekanntem σ und unbekanntem μ. Das Ergebnis entspricht demjenigen Wert, für den die Kurve maximal ist.

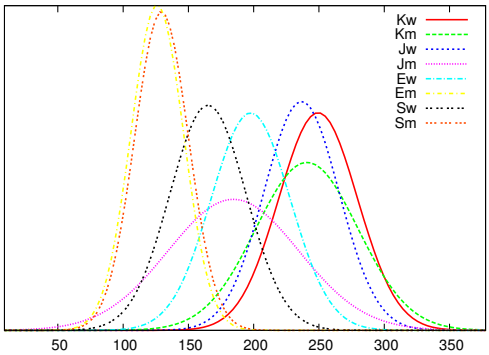
7.3 Das Acht-Klassen-Problem

Bei der Diskussion der Ergebnisse der Korpusanalysen wurde bereits eine Einschätzung der Unterscheidbarkeit der Klassen auf Basis der einzelnen Merkmale vorgenommen. Besonders auffällig war dabei, dass erstens die Merkmalsmengen zur Unterscheidung von Sprecheralter und -geschlecht sich größtenteils überschneiden und zweitens eine wechselseitige Abhängigkeit zwischen beiden Charakteristika besteht: Die korrekte Einschätzung der Altersklasse ist abhängig von der Kenntnis des Geschlechts und umgekehrt. Zur Lösung dieses Problems bietet es sich an, statt eines Klassifikationsproblems mit zwei mal vier Klassen (zwei Geschlechts- und vier Altersklassen) ein Acht-Klassen-Problem zu betrachten: weibliche Kinder KW, männliche Kinder KM, weibliche Jugendliche JW, männliche Jugendliche JM, weibliche Erwachsene EW, männliche Erwachsene EM, weibliche Senioren SW und männliche Senioren SM.

Mit den Mitteln der Bayes'schen Entscheidungstheorie, wie beispielsweise den Gauß'schen Wahrscheinlichkeitsdichten, die implizit bereits in Teil I zur Darstellung der Ergebnisse verwendet worden sind, können auf Basis dieser acht Grundklassen neue Gruppierungen erstellt werden, die – jeweils mit einem eigenen Merkmalset – von einem Klassifizierer vorhergesagt werden können.

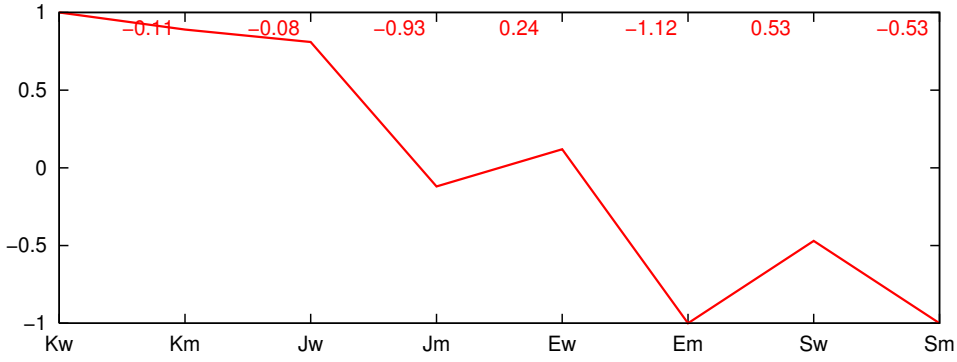
Das folgende Beispiel zeigt die Tendenzen des Merkmals $pitch\_mean$ für die acht Klassen. Die Gesamttendenz der Mittelwerte ist negativ und zeigt dabei eine charakteristische Treppenform, die wie folgt interpretiert werden kann: Zwischen Kindern und Jugendlichen überwiegt der Alterseffekt, was daran zu erkennen ist, dass der Abfall der Linie nicht unterbrochen ist. Der Effekt des Geschlechts ist bei der Klasse der Erwachsenen und Senioren größer, was zu einer positiven Tendenz zwischen JM und EW und zwischen EM und SW führt. Die Gauß'schen Wahrscheinlichkeitsdichten (oben) erscheinen aufgrund der großen Anzahl von Klassen zunächst chaotisch, lassen jedoch Gruppierungen erkennen: Eine Gruppe wird durch EM und SM gebildet, eine zweite durch KW, KM und JW. Die weiblichen Erwachsenen und Senioren bleiben einzelne Klassen. Die Klasse der männlichen Jugendlichen ist aufgrund ihrer hohen Standardabweichung auf Basis dieses Merkmals nicht klassifizierbar.

7.1 Mittlere Stimmtonhöhe bei acht Klassen — $pitch\_mean$



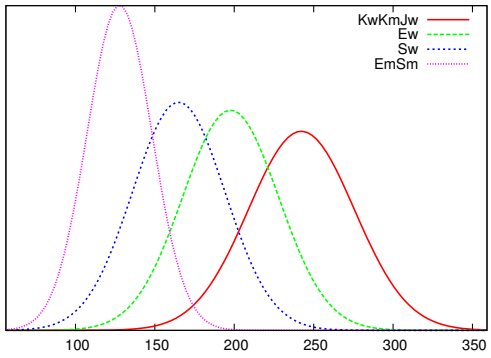
	μ	Σ	$\bar{\mu}$
gesamt	191.03	56.01	0
Kw	249.11	29.51	1
Km	240.47	38.9	0.88
Jw	235.91	27.59	0.8
Jm	185.28	51.95	-0.1
Ew	197.68	30.2	0.12
Em	125.38	19.91	-1
Sw	165.39	28.16	-0.46
Sm	129.01	20.59	-1

7.1 Mittlere Stimmtonhöhe bei acht Klassen (Fortsetzung) *pitch_mean*

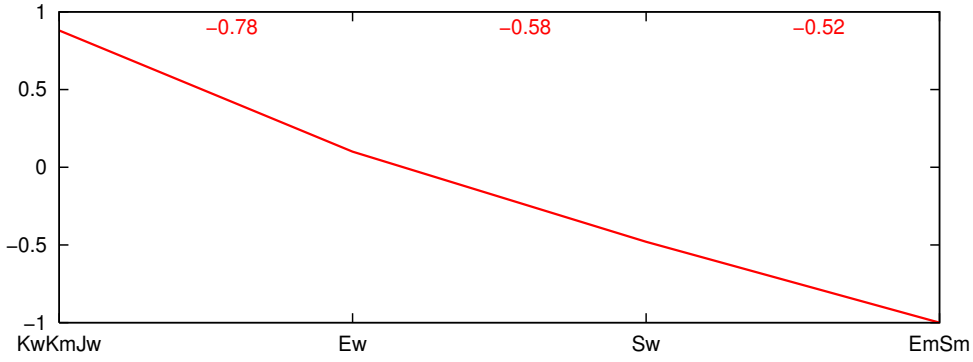


Bei entsprechender Gruppierung der Klassen und Entfernung von JM ist eine durchgängig negative Tendenz zu erkennen, die darüber hinaus fast gleichmäßig ist. Die Wahrscheinlichkeitsdichten erlauben die Annahme, dass das Merkmal *pitch_mean* zur Unterscheidung von KWKMJW, EW, SW und EMSM geeignet ist.

7.2 Mittlere Stimmtonhöhe bei geeigneter Gruppierung *pitch_mean*



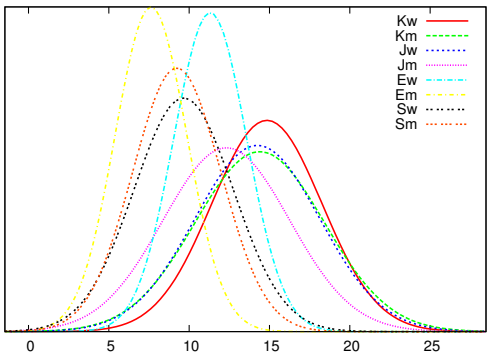
	μ	Σ	$\bar{\mu}$
KwKmJw	242.12	33.55	0.88
Ew	197.68	30.2	0.1
Sw	166.76	28.32	-0.45
EmSm	126.99	20.21	-1



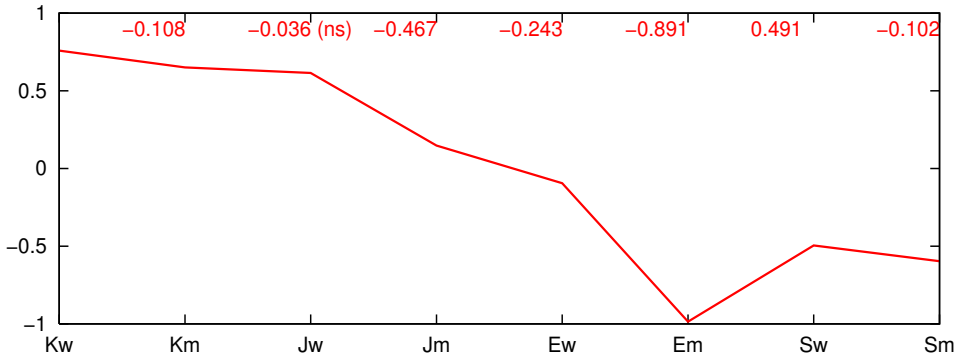
Um das Prinzip an einem weiteren Beispiel zu verdeutlichen, wird nachfolgend die Acht-Klassen-Analyse des Merkmals *harm_mean* dargestellt. Die Tendenzen sind durchweg negativ

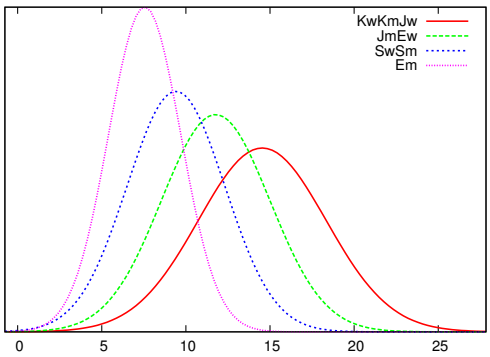
mit Ausnahme derjenigen zwischen EM und SW, die durch die sehr niedrigen Werte bei der Klasse der männlichen Erwachsenen verursacht wird. Die Wahrscheinlichkeitsdichten erlauben die folgenden Gruppierungen: KWKMJW, JMEW, SWSM und EM.

7.3 Harmonizität bei acht Klassen und geeigneter Gruppierung $harm\_mean$



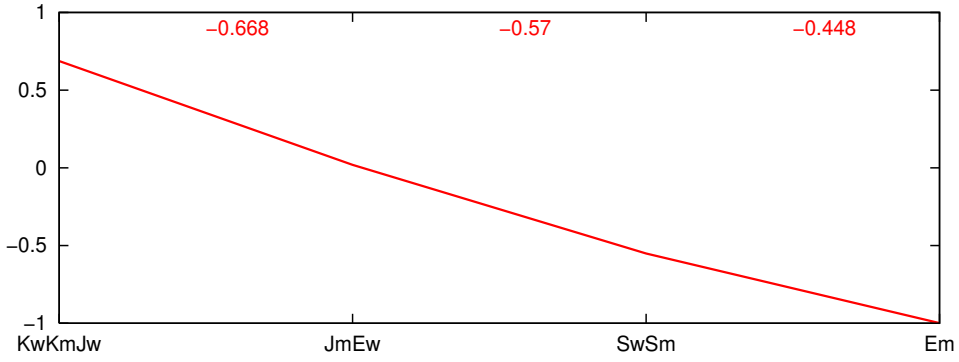
	μ	Σ	$\bar{\mu}$
gesamt	11.702	4.137	0
Kw	14.841	3.428	0.759
Km	14.396	4.025	0.651
Jw	14.247	3.888	0.615
Jm	12.316	3.939	0.148
Ew	11.309	2.271	-0.095
Em	7.625	2.23	-0.986
Sw	9.655	3.1	-0.495
Sm	9.231	2.746	-0.597





	μ	Σ	$\bar{\mu}$
gesamt	11.691	4.144	0
KwKmJw	14.538	3.801	0.687
JmEw	11.771	3.216	0.019
SwSm	9.41	2.905	-0.551
Em	7.552	2.151	-0.999

7.3 Harmonizität bei geeigneter Gruppierung (Fortsetzung) $harm\_mean$



Eine entsprechende Analyse aller Merkmale ergab insgesamt fünf solcher Gruppierungen, die in Tabelle 7.4 schematisch dargestellt werden. Die Gruppierungen (1) und (2) entsprechen den oben aufgeführten Beispielen, die von $pitch\_mean$ bzw. $harm\_mean$ gebildet wurden. In Gruppierung (3) werden die Kinder und Jugendlichen den erwachsenen Altersklassen (inklusive der Senioren) gegenübergestellt, wobei die Klasse KM ausgeklammert wurde. Die Gruppierung (5) stellt die jüngeren Erwachsenen (Frauen und Männer) allen anderen Altersklassen gegenüber; die Gruppierung (4) ist ähnlich, unterscheidet aber zusätzlich die Klasse EM. Tabelle 7.5 stellt eine Übersicht dar, welche Merkmale welche Gruppierungen unterstützen. Eine Unterstützung ist dann gegeben, wenn eine eindeutige Unterscheidbarkeit möglich ist. Dies wird in der Tabelle durch einen Kreis gekennzeichnet. Sie ist „bedingt“ (Halbkreis), wenn eine oder mehrere Klassen zum Großteil überlagert werden (vgl. Abbildung 7.8).

Aus Tabelle 7.5 lassen sich Merkmalsets ableiten, auf deren Basis in Kapitel 8 alternative Modellierungsmethoden miteinander verglichen werden. Die Gruppierung (0), die aus den acht Einzelklassen besteht, bildet dabei die Referenz. Es ist zu erwarten, dass vor allem die parametrischen Methoden von einer geeigneten Gruppierung der Klassen und einer entsprechenden Zusammenstellung der Merkmalsets profitieren, während Methoden mit aufwendigeren Lernverfahren diese implizit selbst vornehmen. Die Merkmale, die sich auf das Sprechverhalten beziehen, d. h. ar_rate, $pauses\_numps$ und $pauses\_durpp$, bilden separate Merkmalsets. Es ist anzunehmen, dass Klassifizierer, die auf diesen Merkmalsets basieren, weniger anfällig gegenüber Hintergrundlärm sind (vgl. Kapitel 1.5.1).

Eine multiple Klassifikation auf Basis der Gruppierungen führt zu ambigen Ergebnissen. Angenommen, die Sprachprobe stammt von einer jüngeren, erwachsenen Frau. Die einzelnen Klassifizierer würden dann – falls sie alle richtig klassifizieren – die Ergebnisse liefern, die in Tabelle 7.4 farbig hinterlegt sind. Da die korrekte Klasse EW die einzige ist, die in allen Ergebnissen enthalten ist, ist die Disambiguierung trivial. Aber auch wenn eines oder mehrere der Klassifikationsergebnisse falsch sind, ist es dennoch möglich, zu einem richtigen Gesamtergebnis zu kommen: Dies gehört zu den Aufgaben der *Nachverarbeitung*, die in der vorliegenden Arbeit *Zweite Ebene* genannt wird (vgl. Kapitel 9).

Gr. 0	Gr. 1	Gr. 2	Gr. 3	Gr. 4	Gr. 5
KW	KW	KW	KW	KW	KW
KM	KM	KM		KM	KM
JW	JW	JW	JW	JW	JW
JM		JM		JM	JM
EW	EW	EW	EW	EW	SW
EM	SW	EM	EM	EM	SM
SW	EM	SW	SW	SW	EW
SM	SM	SM	SM	SM	EM

Tabelle 7.4: Gruppierungen von Klassen.

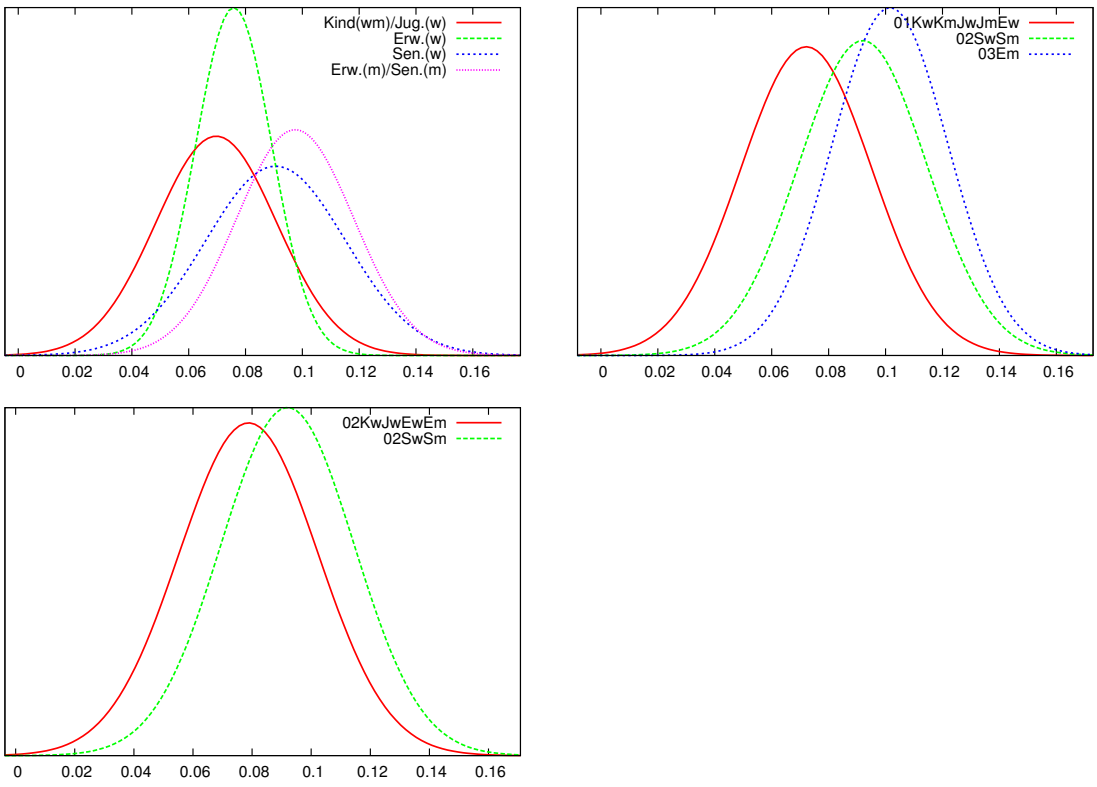


Abbildung 7.8: Oben links: Gruppierung 1 wird von dem Merkmal $shim\_l$ nicht unterstützt. Oben rechts: Gruppierung 4 wird bedingt unterstützt. Unten rechts: Gruppierung 3 wird voll unterstützt.

	Gruppierung					
	(1)	(2)	(3)	(4)	(5)	Summe
ar_rate			●		●	2 / 5
pauses_numps			●		●	2 / 5
pauses_durpp			●			1 / 5
intens_mean			●		●	2 / 5
intens_ratio			●		●	2 / 5
intens_stddev			●		●	2 / 5
jitt_l			●		◗	1.5 / 5
jitt_la			●	●		2 / 5
jitt_rap			●	●	◗	2.5 / 5
jitt_ddp			●	●	◗	2.5 / 5
jitt_ppq	●		●		◗	2.5 / 5
shim_l		●	●	◗	●	3.5 / 5
shim_ldb		●	●	●	●	4 / 5
shim_apq3		◗	●	●	◗	3 / 5
shim_apq11	◗		◗	●	●	3 / 5
shim_ddp		◗	●	●	◗	3 / 5
harm_mean		●	●	●	●	4 / 5
harm_stddev		◗	●		◗	2 / 5
pitch_min	◗	◗	◗		◗	2 / 5
pitch_max					●	1 / 5
pitch_diff			●			1 / 5
pitch_quant	●	●	●	●	●	5 / 5
pitch_mean	●	●	●	●	●	5 / 5
pitch_stddev						0 / 5
pitch_mas			●			1 / 5
pitch_swoj			●			1 / 5
Summe	4 / 26	7 / 28	23 / 26	10.5 / 26	16 / 26	

Tabelle 7.5: Unterstützung der Gruppierungen durch die einzelnen Merkmale.

8 Alternative Methoden des maschinellen Lernens (Erste Ebene)

Der parametrische Ansatz des Bayes'schen Klassifizierers stellt ein zweckmäßiges Mittel zur Sichtung der Trainingsdaten dar, die über die Betrachtung von Mittelwertdifferenzen und statistischer Signifikanz hinausgeht. Gleichzeitig stellt er eine Ausgangsbasis dar, um verschiedene alternative Klassifikationsmethoden miteinander zu vergleichen, beispielsweise was die Komplexität der Entscheidungsgrenzen betrifft.

Ein solcher Vergleich wird im folgenden Abschnitt angestrebt, mit dem Ziel, eine Klassifikationsmethode zu identifizieren, die jeweils für die verschiedenen Probleme der Sprecherklassifikation sowie der Kontextklassifikation am besten geeignet erscheint. Gemäß dem allgemeinen Schema, das in Abbildung 7.1 dargestellt ist, betrifft diese Frage nach wie vor den Punkt „Klassifikation". Jedes der Kandidaten-Verfahren dient dazu, eine oder mehrere Diskriminantenfunktionen zu erzeugen, wie sie in Abbildung 7.2 dargestellt werden. Auf deren Basis wird ein gegebener Merkmalsvektor $\mathbf{x}$ einer Klasse ω_i zugewiesen (beispielsweise WEIBLICH).

Eines der hauptsächlichen Bewertungskriterien eines jeden Verfahrens wird die Klassifikationsgenauigkeit bzw. das möglichst geringe Vorkommen von Klassifikationsfehlern sein. Nach Duda et al. (2000, S. 101) gibt es in einem fertigen System drei Quellen von Klassifikationsfehlern: (1) Der Bayes'sche Fehler oder Nichtunterscheidbarkeitsfehler, der dadurch entsteht, dass sich die Wahrscheinlichkeitsdichten $p(\mathbf{x}|\omega_i)$ für verschiedene i überlappen. Dieser Fehler ist eine inhärente Eigenschaft eines Klassifikationsproblems und kann nicht eliminiert werden. (2) Der Modell-Fehler, d. h. ein inadäquates Modell wurde ausgewählt. (3) Der Einschätzungsfehler. Dieser Fehler entsteht durch die (unumgängliche) Tatsache, dass die Parameter auf Basis einer endlichen Menge von Trainingsdaten geschätzt wurden. Dieser Fehler kann durch eine Vergrößerung der Menge der Trainingsdaten reduziert werden.

Der Entwickler eines Mustererkennungssystems ist nicht selten gezwungen, in einem „Trial and Error"-Verfahren Modelle auszuwählen und zu testen, um das für das gegebene Problem am besten geeignete Verfahren zu identifizieren. Doch selbst wenn der Fehler auf den Trainingsdaten minimiert wurde, ist dies letztlich keine Garantie dafür, dass eine ausreichende Generalisierung erreicht worden ist, um neue, bisher ungesehene Instanzen korrekt zu klassifizieren (*Overfitting*). Nach Duda et al. (2000, S. 6) ist die Anpassung der Komplexität des Modells eines der wichtigsten Forschungsgebiete in der statistischen Mustererkennung: Es darf nicht so einfach sein, dass es die

Unterschiede zwischen den Klassen nicht erklären kann, und doch nicht so komplex, dass neue Muster schlecht erkannt werden (vgl. ebd.). Eine Methode, um die Genauigkeit eines Klassifizierers zu ermitteln, die das Overfitting-Problem bestmöglich berücksichtigt, ist die so genannte *Kreuzvalidierung* (*cross validation*). Dabei wird ein (größerer) Teil der Daten zum Trainieren und ein (kleinerer) Teil zum Testen des Modells verwendet. Dieser Vorgang wird so oft wiederholt, bis alle Daten einmal in der Testmenge enthalten waren. Eine gängige Größe für die Testmenge ist 10 Prozent (mit entsprechender zehnmaliger Wiederholung). Diese Variante, die *zehnfache Kreuzvalidierung* (*tenfold cross validation*) genannt wird, ist bei der Evaluierung der hier verglichenen Modelle zum Einsatz gekommen.

Die untersuchten Klassifikationsverfahren weisen jedoch auch Gemeinsamkeiten auf. Die herausragende Gemeinsamkeit ist die allgemeine Form des Lernens. Nach Duda et al. (2000, S. 16 f) werden drei verschiedene Formen unterschieden: 1. Beim *überwachten Lernen* wird eine Trainingsdatenbank zur Verfügung gestellt, die nach den zu unterscheidenden Klassen etikettiert ist. Der Algorithmus versucht, den Klassifikationsfehler auf dieser Datenbank zu minimieren. 2. Beim *nicht-überwachten Lernen* oder *Clustering* werden „natürliche Gruppierungen“ für die Eingabemuster geformt, die nicht etikettiert sind. Der Begriff „natürlich“ wird für das Verfahren implizit oder explizit definiert, so dass verschiedene Clustering-Algorithmen zu verschiedenen Gruppen führen. Oft gibt der Analyst zuvor eine Hypothese über die Anzahl der verschiedenen Cluster vor. Beim *Reinforcement Learning* schließlich wird zunächst das tentative Klassenetikett einer Trainingsinstanz vom System berechnet. Anschließend wird dann das tatsächliche Etikett dazu genutzt, um den Klassifizierer zu verbessern. Statt einer Kategorie wird häufig auch nur angegeben, ob die vorgeschlagene Klasse die richtige war oder nicht. Sämtliche der untersuchten Verfahren lernen nach Methode eins bzw. drei. Clustering-Methoden werden in der Regel immer dann eingesetzt, wenn noch sehr wenig über die Struktur des Klassifikationsproblems bekannt ist. Da hier jedoch bereits Hypothesen herausgearbeitet werden konnten, ist das überwachte Lernen die geeignetere Methode, um dieses Wissen bei der Konfiguration des Klassifizierers mit einfließen zu lassen. Reinforcement ist in diesem Sinne auch dem überwachten Lernen zuzuordnen, da ebenfalls von etikettierten Trainingsinstanzen Gebrauch gemacht wird, um den Klassifikationsfehler zu verringern.

8.1 Hidden-Markov-Models

Der Vergleich der Klassifikationsverfahren beginnt mit einer Methode, die in AGENDER nicht zum Einsatz gekommen ist, nämlich den *Hidden-Markov-Models* (HMMs). Der Grund, weshalb sie an dieser Stelle dennoch Erwähnung finden ist der, dass HMMs im Bereich der Spracherkennung als Klassifikationsverfahren weit verbreitet sind, und daher die Frage berechtigt ist, warum sie in der verwandten Domäne der Sprecherklassifikation nicht ebenfalls eingesetzt werden. Um eine Antwort auf diese Frage geben zu können, ist eine kurze Betrachtung der Funktionsweise von Hidden-Markov-Models notwendig.

Betrachten wir eine Folge von Zuständen ω an aufeinander folgenden Zeitpunkten t: $\omega^T = \omega(1), \omega(2), \ldots, \omega(T)$. Jeder dieser Zustände kann mehrere Male oder auch gar nicht erreicht

werden. Die Übergänge zwischen den Zuständen sind mit Übergangswahrscheinlichkeiten $a_{ij} = P(\omega_j(t+1)|\omega_i(t))$ versehen. Die Gesamtwahrscheinlichkeit, dass ein Modell θ eine Folge von Zuständen $\omega^6 = \omega_1, \omega_4, \omega_2, \omega_2, \omega_1, \omega_4$ erzeugt, ist $P(\omega^T|\theta) = a_{14}a_{42}a_{22}a_{21}a_{14}$.

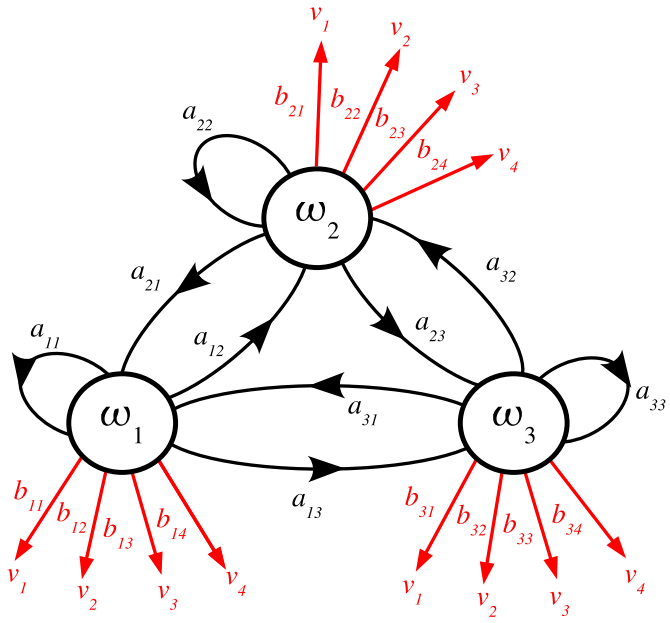


Abbildung 8.1: Ein Hidden-Markov-Model mit drei Zuständen nach Duda et al. (2000, S. 129).

Solche Modelle werden *Markov-Models* genannt (vgl. Duda et al., 2000, S. 129). Ein solches Markov-Model für das Wort „cat“ hätte die Zustände /k/, /a/ und /t/ mit entsprechenden Übergangswahrscheinlichkeiten zwischen den Phonemen. Der Spracherkenner hat jedoch keinen direkten Zugriff auf die Phoneme, sondern lediglich auf bestimmte Eigenschaften des Schalls, der aufgezeichnet wurde. Das Markov-Model muss also erweitert werden, so dass zwischen *sichtbaren* und *verborgenen* Zuständen unterschieden werden kann. Gegeben sei eine Folge von sichtbaren Zuständen (Symbolen) $\mathbf{V}^T = v(1), v(2), \ldots, v(n)$. In jedem Zustand ω gibt es eine Wahrscheinlichkeit $b_{jk} = P(v_k(t)|\omega_j(t))$, dass ein bestimmter sichtbarer Zustand $v_k(t)$ in ω_j gesehen wird. Modelle dieser Art werden *Hidden-Markov-Models* (HMMs) genannt.

Netze in der Art, wie das in Abbildung 8.1 dargestellte, werden *endliche Automaten* genannt – HMMs heißen sie dann, wenn sie mit Übergangswahrscheinlichkeiten versehen sind. Sie sind streng kausal: Die Wahrscheinlichkeit eines Zustands hängt ausschließlich von dem vorangehenden Zustand ab. Einen Zustand ω_0, der nicht wieder verlassen werden kann, nennt man *Endzustand* ($a_{00} = 1$). HMMs verlangen, dass nach jedem Schritt $t \rightarrow t+1$ ein Übergang erfolgt, auch wenn der Übergang auf denselben Zustand erfolgt. Außerdem muss bei jedem Schritt ein Symbol gelesen werden.

Man unterscheidet dreierlei Arten von Problemen, die mit HMMs gelöst werden können: (1) Das Evaluationsproblem: Gegeben sei ein HMM mit den Übergangswahrscheinlichkeiten a_{ij} und b_{jk}. Bestimme die Wahrscheinlichkeit, dass eine bestimmte Folge von Symbolen $\mathbf{V}^T$ von dem Modell erzeugt wurde. (2) Das Dekodierungsproblem: Gegeben sei ein HMM und eine Folge von

Symbolen $\mathbf{V}^T$. Bestimme diejenige Folge von versteckten Zuständen ω^T, die am wahrscheinlichsten zu der Beobachtung der Symbole geführt hat. (3) Das Lernproblem: Gegeben sei die Struktur eines Modells, d. h. die Anzahl der Zustände und Symbole, nicht jedoch die Wahrscheinlichkeiten a_{ij} und b_{jk}. Gegeben sei außerdem eine Menge von Trainingsbeobachtungen. Bestimme die fehlenden Parameter.

Hidden-Markov-Models (HMMs) kommen in erster Linie in Domänen zum Einsatz, die eine inhärente Temporalität besitzen, d. h. bei denen ein Zustand zum Zeitpunkt t direkt beeinflusst wird durch den Zustand bei $t-1$. Spracherkennung oder Gestenerkennung sind Beispiele für solche Domänen. Es gibt allerdings auch eine Reihe von Arbeiten zur Sprachklassifikation auf der Basis von HMMs. Iurgel, Werner und Gerhard (2003) z. B. verwenden HMMs zur automatischen Segmentierung von Fernsehnachrichten in Abschnitte mit Sprache und Abschnitte ohne Sprache. Als zugrunde liegende Merkmale verwenden sie die spektralen Eigenschaften und die Pausenrate sowie eine Reihe von Merkmalen, die von der Pitch-Kontur abgeleitet werden (vgl. 2.2.2). Iurgel et al. trainierten HMMs für die Klassen SPRACHE, MUSIK und (andere) NICHT-SPRACHE, wobei sie zwischen weiblichen und männlichen Sprechern unterschieden. Als Trainingsmaterial für die Modelle wurden Fernsehnachrichten verwendet, deren Segmente zuvor manuell gekennzeichnet wurden. Anschließend testeten Iurgel et al. einen speziell für Nachrichten entwickelten Spracherkenner in zwei Durchgängen, wobei sie beide Male dasselbe Eingabematerial verwendeten, einmal automatisch segmentiert und einmal manuell. Die Ergebnisse zeigten im Fall der automatischen Segmentierung eine leicht erhöhte Fehlerrate (von 32.15 auf 37.02 %).

Bou-Ghazale und Hansen (1996) stellen einen Ansatz vor, bei dem HMMs ebenfalls nicht ausschließlich zur Spracherkennung, sondern darüber hinaus zur Unterscheidung von *unmarkierten* (neutralen) und *markierten* Segmenten der Sprache verwendet werden. Sie untersuchen verschiedene Arten der Markierung, z. B. LAUT (Sprachsegmente, die durch Anheben der Lautstärke hervorgehoben wurden) und ÄRGER (Sprachsegmente, die aufgrund des emotional angeregten Zustands des Sprechers verändert wurden). Eines der wesentlichen Merkmale stellt auch in diesem Fall die Pitch-Kontur dar. Zum Training wurde ein Text von verschiedenen Sprechern in der neutralen und der markierten Variante gelesen. Bou-Ghazale und Hansen verwenden die fertigen Modelle, um neutrale Sprache so zu modifizieren, als sei der Sprecher verärgert. In einem Experiment ließen sie den „synthetischen Ärger“ von Versuchspersonen bewerten mit dem Resultat, dass 97 von 112 modifizierten Äußerungen als „mehr verärgert“ wahrgenommen wurden als die Originale.

Orio und SistiSette (2003) verwenden HMMs in einem so genannten *Query-by-Humming-System*, bei dem die Benutzer eine Sequenz aus einem Musikstück summen (oder singen), und das System versucht, den passenden Titel zu finden (z. B. bei einem Online-Plattenladen). Die Kern-Technologie solcher Systeme wird *Pitch-Tracking* genannt, wobei die Anfrage in eine Folge von Pitch-Werten übersetzt wird. Bei *Query-by-Humming-Systemen* ist dies deshalb problematisch, da die Anfragen häufig von nicht geübten „Sängern“ gestellt werden und daher mit systematischen oder lokalen Fehltönen zu rechnen ist. Zur Lösung dieses Problems schlagen Orio und SistiSette einen Ansatz auf der Basis eines zweistufigen HMMs vor: Auf der so genannten *Ereignisebene* wird dabei die Melodie als ein Pfad zwischen Tönen (Knoten des Netzes) modelliert. Die Über-

gangswahrscheinlichkeiten basieren auf dem melodischen Intervall zwischen den Tönen. Auf der *Audio-Ebene* dagegen wird die Wahrscheinlichkeit von Noten auf der Basis elementarer Merkmale wie *Anschlag* (engl. attack), *Dehnung* (engl. sustain) und *Pause* (engl. rest) ermittelt. Das System wurde mit 18 Anfragen mit insgesamt 300 Noten getestet, die von ungeübten Benutzern mit einem minderwertigen Mikrofon aufgezeichnet wurden. Die Fehlerrate lag bei etwas über zehn Prozent (falscher Noten).

In AGENDER ist eine solche Temporalität auf der Ebene der Klassifikation jedoch nicht vorhanden. Dabei ist keine grundsätzliche Inkompatibilität mit den Annahmen gegeben, die bei der Modellierung mit HMMs gemacht werden. Die Gründe liegen eher in der Auswahl der Merkmale: 1. Sämtliche Merkmale werden über die gesamte Äußerungseinheit ermittelt. Unter dieser Voraussetzung können HMMs dann eingesetzt werden, wenn Merkmale betrachtet werden, die sich im Verlauf der Äußerung ändern, so dass das Muster der Änderungen eine Kontur ergibt. Derzeit ist dies jedoch nicht der Fall. 2. Die Merkmale, die betrachtet werden, sind keine Symbole, sondern reelle Werte.

8.2 Gaussian Mixture Models

Gaussian Mixture Models (Gauß'sche Mixtur-Modelle, GMMs) sind sehr eng mit dem Bayes'schen Klassifizierer verwandt. Sie gelten als ein probabilistisches Modell für multivariate Wahrscheinlichkeitsdichten, die beliebige Dichten (Gauß'sche, Laplace'sche usw.) repräsentieren können (Reynolds et al., 2000). In der Anwendung wird mithilfe von GMMs die zugrunde liegende klassenspezifische Wahrscheinlichkeitsdichte berechnet, auf Basis derer ein Likelihood-Ratio-Klassifizierer dann ein gegebenes Muster einer Kategorie zuweist. Für einen d-dimensionalen Merkmalsvektor $\mathbf{x}$ ist die Mixtur-Dichte (*mixture density*) definiert als:

$$p(\mathbf{x}|\gamma) = \sum_{j=1}^{M} w_j p_j(\mathbf{x}). \tag{8.1}$$

Die Wahrscheinlichkeitsdichte ist eine lineare Kombination aus M Gauß'schen Wahrscheinlichkeitsdichten (vgl. Gleichung 7.2). Die Mixtur-Gewichte w_j erfüllen darüber hinaus die Bedingung

$$\sum_{j=1}^{M} w_j = 1. \tag{8.2}$$

Zusammengenommen werden die Parameter des Modells durch $\gamma = \{w_j, \mu_j, \Sigma_j\}$ denotiert, wobei $j = 1, \ldots, M$. Auf Basis einer Sammlung von Trainingsproben werden die Parameter mithilfe des iterativen *Expectation-Maximization*-Algorithmus (EM-Algorithmus) bestimmt. Dieser passt die Parameter des GMMs so an, dass eine monotone Verbesserung der *Likelihood* des Modells der beobachteten Merkmalsvektoren erreicht wird. Für die Iterationen k und $k+1$ gilt beispielsweise $p(\mathbf{x}|\gamma_{k+1} > p(\mathbf{x}|\gamma_k)$.

Ein GMM kann als hybride Methode zwischen parametrischen und nichtparametrischen Modellen angesehen werden, da – obwohl grundsätzlich Parameter das Verhalten bestimmen – ein hoher Grad an Freiheit beliebige Wahrscheinlichkeitsdichten zulässt.

8.2.1 Gaussian Mixture Models in verwandten Arbeiten

Gaussian Mixture Models finden eine weite Verbreitung im Bereich der textunabhängigen Sprechererkennung (und -verifikation). Reynolds et al. (2000) beschreiben beispielsweise einen Ansatz zur Sprecherverifikation auf der Basis von GMMs. Eine Besonderheit dieses Systems besteht darin, dass ein gegebenes Sprechermodell mit einem so genannten *Universal Background Model* verglichen wird, welches auf Basis von Sprachproben einer großen Anzahl von Sprechern erzeugt wurde. Diese repräsentieren die bei der Erkennung zu erwartende Population von Sprechern. Das Verifikationssystem, das Reynolds et al. (2000) beschreiben, trägt dementsprechend den Namen *GMM-UBM Verification System*. Das verwendete GMM entspricht dem oben beschriebenen Formalismus, d. h. es handelt sich prinzipiell um einen Likelihood-Ratio-Klassifizierer, bei dem $p(\mathbf{x}|\omega)$ gemäß Gleichung 8.1 berechnet wird. Statt einer vollständigen Kovarianzmatrix verwenden Reynolds et al. allerdings eine diagonale Kovarianzmatrix, da diese weniger aufwendig zu berechnen ist, und die resultierenden Modelle nach Angaben der Autoren in empirischen Tests bessere Ergebnisse erzielen. Die Qualität der Klassifikation wird in Form eines *Detection Error Tradeoffs* (DET) angegeben, welches eine Art umgekehrtes *Precision-Recall*-Diagramm darstellt: Die (negative) Genauigkeit entspricht der Wahrscheinlichkeit, dass der Zielsprecher nicht erkannt wird (*Miss Probability*), und die (negative) Vollständigkeit entspricht der Wahrscheinlichkeit, dass ein anderer Sprecher fälschlicherweise eine Verifikation erhält (*False Alarm Probability*). Im Zusammenhang mit Analysen dieser Art wird als einzelner Wert für das Qualitätsmaß häufig die so genannte *Equal Error Rate* EER angegeben, d. h. der Wert für den *Miss Probability* und *False Alarm Probability* gleich sind. Die EER für dieses Modell liegt bei 10 %.

8.2.2 Gaussian Mixture Models in AGENDER

In AGENDER wurde ein GMM mit Diskriminantenfunktionen gemäß Gleichung 8.1 verwendet. Anders als bei Reynolds et al. (2000) werden bei dieser Variante allerdings die Gewichte nicht mithilfe des EM-Algorithmus gelernt, sondern auf Basis einer Evaluation jedes einzelnen Merkmals ermittelt. Bei diesem Verfahren wird zunächst ein GMM mit $\omega_j = 1$ für $j = 1, \ldots, M$ gelernt und kreuzvalidiert, wobei nicht nur das Ergebnis des Klassifizierers als Ganzes geprüft wird, sondern darüber hinaus die Entscheidungen, die auf Basis jedes einzelnen Merkmals getroffen wurden. Die auf diese Weise ermittelten Genauigkeiten werden in einem zweiten Schritt als Basis für die Gewichtung herangezogen.

Die Klassifikationsgenauigkeit des GMMs bezüglich der Kontrollgruppierung (Gruppierung 0) wird in der folgenden Konfusionsmatrix dargestellt. Bei dieser Gruppierung werden alle acht Klassen einzeln trainiert, wobei sämtliche der in Tabelle 7.5 aufgeführten Merkmale verwendet werden. Konfusionsmatrizen stellen ein geeignetes Mittel zur Darstellung der Performanz von Polychotomisierern dar, da sie nicht nur die globale Genauigkeit in Form von *True Positive Rates*

ausdrücken, sondern darüber hinaus deutlich machen, welche Klassen im Fehlerfall statt der richtigen gewählt wurden. Hierfür werden in der linken Spalte untereinander die korrekten Klassen aufgeführt; die Zeilen beinhalten die Ergebnisse des Klassifizierers. Dementsprechend stellt die Diagonale, die hier hervorgehoben wurde, diejenigen Fälle dar, in denen der Klassifizierer eine korrekte Entscheidung getroffen hat. Die Angabe erfolgt in Prozent.

8.1	Gruppierung 0					Gesamtgenauigkeit 50.36 %		
	Kw	Km	Jw	Jm	Ew	Em	Sw	Sm
Kw	73	5.15	14.19	3.73	2.85	0.05	0.98	0.05
Km	58.22	8.84	14.19	10.11	5.15	0.05	3.39	0.05
Jw	52.53	4.57	24.01	8.69	5.99	0.05	4.17	0
Jm	24.35	7.61	7.85	21.6	11.93	3.53	10.36	12.76
Ew	0.83	0.34	0.25	1.47	83.46	7.07	3.04	3.53
Em	0	0.2	0	0.54	10.26	83.06	1.57	4.37
Sw	1.67	2.95	1.52	8.89	19.59	3.04	41.24	21.11
Sm	0.05	0.59	0	5.11	7.76	8.35	10.46	67.7

Die Klassifikationsgenauigkeit bezüglich der weiblichen Kinder (Kw) beträgt 73%, was als relativ hoch eingeschätzt werden kann, wenn man bedenkt, dass das Zufallsniveau (ZN) mit acht Klassen bei 12.5 % liegt. Mit 14.19 % wurden die meisten der falsch klassifizierten Instanzen als weibliche Jugendliche (Jw) eingeschätzt; in 5.15 % der Fehlerfälle war das Ergebnis männliche Kinder (Km). Beides ist nicht überraschend, da diese als die benachbarten Klassen angesehen werden können. Entsprechend treten je „entfernter“ die Klassen liegen, desto seltener Verwechselungen auf. Die Tatsache, dass ein sehr hoher Prozentsatz der männlichen Kinder (Km) fälschlicherweise als Kw klassifiziert wurde (58.22 %), entspricht den Erwartungen: Die beiden Klassen sind nur bedingt voneinander unterscheidbar, wie aus den Korpusanalysen in Teil I hervorgeht. Allerdings ist auffällig, dass die männlichen Kinder mit 10.11 % häufiger mit den männlichen Jugendlichen verwechselt werden, als die weiblichen Kinder (3.73 %), und offensichtlich auch näher an den Klassen Ew (jüngere, weibliche Erwachsene) und Sw (Seniorinnen) liegen. Zusammengenommen beträgt die True Positive Rate für die Kinder 72.61 % (ZN = 14.3 %).

Bei der Klasse der Jugendlichen liegt die Genauigkeit deutlich darunter: Jw verzeichnet eine True Positive Rate von 24.01 %, und die von Jm liegt bei 21.6 %. Sowohl die männlichen als auch die weiblichen Jugendlichen werden am häufigsten mit Kw verwechselt (in 52.53 % bzw. 24.35 % der Fälle). Allerdings ist zu erkennen, dass auch im Fall von Jm wie bereits bei Km eine Tendenz zur Verwechselung mit Stimmen erwachsener Frauen (Ew und Sw) besteht.

Die jüngeren Erwachsenen (Ew und Em) verzeichnen mit 83.46 % bzw. 83.06 % die höchsten True Positive Rates. In den Fällen, in denen Verwechselungen auftreten, geschieht dies zumeist

innerhalb dieser beiden Klassen: Die jüngeren, weiblichen Erwachsenen werden in 7.07 % der Fälle mit jüngeren Männern verwechselt; die umgekehrte Situation tritt in 10.26 % der Fälle auf.

Was die Senioren betrifft, ist ein deutlicher Unterschied zwischen den Geschlechtern zu erkennen: Während Sm eine verhältnismäßig hohe True Positive Rate von 67.7 % verzeichnet, beträgt diese bei Sw lediglich 41.24 %. Hier findet eine häufige Verwechselung mit den jüngeren Frauen statt, was durchaus den Erwartungen entspricht, da die Symptome der Stimmalterung bei Männern im Allgemeinen deutlicher sind als bei Frauen (vgl. Kapitel 3.4).

8.2	Gruppierung 1		Gesamtg. 69.06 %	
	KwKmJw	Ew	Sw	EmSm
KwKmJw	80.36	15.32	3.98	0.34
Ew	14.14	65.44	10.9	9.52
Sw	2.99	25.82	45.61	25.58
EmSm	0	3.49	11.68	84.83

Dieses Bild spiegelt sich auch in den Ergebnissen bezüglich Gruppierung 1 wider, bei der die beiden Klassen der erwachsenen Frauen (Ew und Sw) getrennt betrachtet werden, während die Kinder und Jugendlichen (Kw,Km,Jw und Jm) sowie die Männer (Em und Sm) zu jeweils einer Gruppe zusammengefasst werden. Erwartungsgemäß fallen die True Positive Rates der beiden Gruppen mit 80.36 % bzw. 84.83 % vergleichsweise hoch aus. Auffällig ist jedoch, dass die Klassifikationsgenauigkeit bezüglich Ew mit 65.44 % deutlich geringer ist als bei Gruppierung 0. Darüber hinaus wird die Klasse Sw häufiger fälschlicherweise als Ew eingeschätzt. Dieser Rückgang der Klassifikationsgenauigkeit ist möglicherweise darin zu begründen, dass eine Auswahl von Merkmalen getroffen wurde, während bei Gruppierung 0 alle Merkmale verwendet worden sind. Allerdings ist die Klassifikation dadurch mit einem geringeren Aufwand verbunden, da weniger Merkmale extrahiert werden müssen und auch die Entscheidungsfunktionen weniger komplex sind. Die Gesamtperformanz des GMMs bezüglich Gruppierung 1 liegt bei 69.06 %, dem 2.76-fachen ZN. In Fall von Gruppierung 0 liegt die Gesamtperformanz bei 50.36 %, was dem 4-fachen ZN entspricht.

8.3	Gruppierung 2		Gesamtg. 66.9 %	
	KwKmJw	JmEw	SwSm	Em
KwKmJw	83.9	12.96	2.99	0.15
JmEw	28.92	45.7	13.7	11.68
SwSm	1.72	17.62	52.28	28.38
Em	0	5.11	9.18	85.71

Bei Gruppierung 2 ist die Genauigkeit in zwei Fällen recht hoch, und zwar bei der Gruppe KwKmJw mit 83.9 % und bei der Klasse Em mit 85.71 %, während sie in den beiden anderen Fällen deutlich darunter liegt: Die Gruppen JwEw und SwSm erreichen eine True Positive Rate von 45.7 % bzw. 52.28 %. Die Gesamtperformanz des GMMs bezüglich dieser Gruppierung beträgt 66,9 %. Dies entspricht dem 2.68-fachen ZN, welches wie bei Gruppierung 1 bei 25 % liegt.

8.4	Gruppierung 3			
	Stimme (81.97 %)		Sprechverhalten (58.46 %)	
	KwJwEwEm	SwSm	KwJwEwEm	SwSm
KwJwEwEm	84.98	15.02	90.75	9.25
SwSm	21.04	78.96	73.83	26.17

Die Klassifikationsgenauigkeit des GMMs bezüglich Gruppierung 3 ist in erster Linie deshalb interessant, da sowohl ein Modell auf Basis von Merkmalen trainiert wurde, welche die Charakteristika der Stimme betreffen (*Stimmmerkmalsmodell*), als auch ein Modell mit Merkmalen des Sprechverhaltens (*Sprechverhaltensmodell*) (vgl. Tabelle 7.5). Die Gesamtperformanz des zuerst genannten Modells beträgt 81.97 %, was dem 1.6-fachen ZN entspricht. Die Performanz des letzteren liegt mit 58.46 % (1.17-mal ZN) weit darunter. Es ist deutlich zu erkennen, dass diese geringe Klassifikationsgenauigkeit darauf zurückzuführen ist, dass ein hoher Prozentsatz der älteren Sprecher zu der falschen Gruppe zugeordnet worden ist. Daher kann dieses Modell kaum als Basis einer sinnvollen Unterscheidung herangezogen werden. Das gilt allerdings nicht für das Stimmmerkmalsmodell, bei dem die True Positive Rates beider Gruppen zufriedenstellend hoch sind.

8.5	Gruppierung 4	Gesamtg. 76.73 %	
	KwKmJwJmEw	SwSm	Em
KwKmJwJmEw	87.53	9.62	2.85
SwSm	18.02	55.82	26.17
Em	2.65	10.51	86.84

Die Gesamtgenauigkeit des GMMs bezüglich Gruppierung 4 liegt mit 76.73 % beim 2.3-fachen ZN. Das Modell ist in der Lage, die Klasse der jüngeren, erwachsenen Männern mit einer Trefferquote von 86.84 % recht zuverlässig zu identifizieren. Die Gruppe der Kinder, Jugendlichen und jüngeren Frauen wird ebenfalls mit einer vergleichsweise hohen Genauigkeit von 87.53 % klassifiziert. Die True Positive Rate bei den Senioren (SwSm) liegt zwar deutlich darunter, entspricht mit 55.82 % jedoch immerhin noch dem 1.69-fachen ZN.

8.6	Gruppierung 5			
	Stimme (85.59 %)		Sprechverhalten (85.13 %)	
	KwKmJwJmSwSm	EwEm	KwKmJwJmSwSm	EwEm
KwKmJwJmSwSm	83.65	16.35	79.58	20.42
EwEm	12.47	87.53	9.33	90.67

Bezüglich Gruppierung 5 ist der Performanzunterschied zwischen dem Stimmmerkmalsmodell und Sprechverhaltensmodell weniger groß, als es bei Gruppierung 3 der Fall ist: Das zuerst genannte erreicht eine Gesamtgenauigkeit von 85.59 % (1.72 mal ZN) und das zuletzt genannte eine Gesamtgenauigkeit von 85,13 % (1.7 mal ZN). Die jüngeren, erwachsenen Sprecher können demnach mithilfe beider Modelle mit zufriedenstellender Zuverlässigkeit von den übrigen Altersklassen unterschieden werden. Es gilt zu überprüfen, wie sich die True Positive Rates unter Berücksichtigung verschiedener Kontexte verändern. Die Hypothese besagt, dass besonders in Situationen mit lauten Hintergrundgeräuschen das Stimmmerkmalsmodell stärkere Performanzeinbußen zu verzeichnen hat als das Modell auf Basis des Sprechverhaltens (vgl. Abschnitt 8.8).

In der nachfolgenden Tabelle wird schließlich die Genauigkeit des GMMs bezüglich der reinen Geschlechtsklassifikation dargestellt. Das Modell wurde ausschließlich anhand von Daten jüngerer, erwachsener Sprecher trainiert und getestet. Die erreichte Gesamtgenauigkeit beträgt 91.98 %. Die True Positive Rate der Klasse Ew liegt dabei mit 89.89 % deutlich unter derjenigen der Klasse Em mit 94.07 %. Abbildung 8.2 stellt eine Übersicht der Gesamtperformanz der GMMs bezüglich der verschiedenen Gruppierungen dar, wobei die roten Balken das jeweilige Zufallsniveau markieren.

8.7	Geschlecht	Gesamtg. 91.98 %
	Ew	Em
Ew	89.89	10.11
Em	5.93	94.07

8.3 Naive-Bayes

Die Methode *Naive-Bayes* (NB) ist ebenfalls unmittelbar von den Bayes'schen Klassifizierern abgeleitet. Nach Witten und Frank (1999, S. 88) besticht sie durch ihre Einfachheit und ihre klare Semantik bezüglich des Lernens, der Repräsentation und der Benutzung von probabilistischem

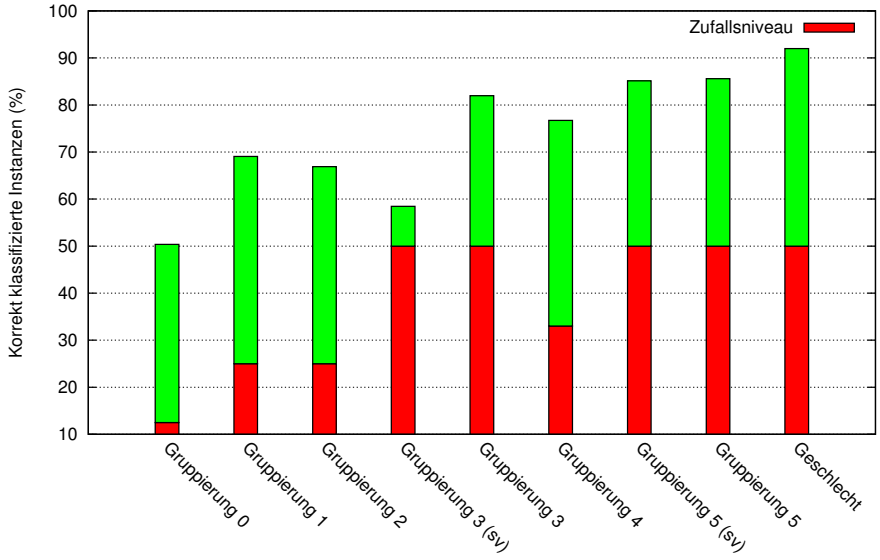


Abbildung 8.2: Übersicht über die Gesamtperformanz der GMMs bezüglich der verschiedenen Gruppierungen. Die roten Balken markieren das jeweilige Zufallsniveau.

Wissen. Allerdings geht sie „naiverweise“ von der Annahme aus, dass alle Merkmale gleich wichtig und statistisch unabhängig voneinander sind, was für die AGENDER Trainingsdatenbank nicht zutrifft. Bei der Berechnung der *Likelihood* aus der klassenspezifischen Wahrscheinlichkeitsdichte $p(\mathbf{x}|\omega_i)$ werden die Wahrscheinlichkeiten der Elemente des Merkmalsvektors $\mathbf{x}$ miteinander multipliziert. Wenn einige dieser Elemente statistisch nicht unabhängig voneinander sind, vergrößert sich ihr Einfluss durch die Multiplikation auf unangemessene Art und Weise. Witten und Frank relativieren jedoch dieser Kritik, in dem sie darauf hinweisen, dass die Methode (wie viele einfache Klassifikationsmethoden) trotz dieser „Naivität“ in der Praxis oftmals erstaunlich gut funktioniert (vgl. ebd.).

8.3.1 Naive-Bayes in verwandten Arbeiten

Zervas et al. (2004) verwenden unter anderem NBs, um Pitch-Akzente für die Sprachsynthese im Voraus zu bestimmen. Dabei geht es darum, die richtige Markierung in der Beschreibung der zu synthetisierenden Äußerung zu setzen, die dann von einem Text-to-Speech-System in Sprache umgewandelt wird. Die korrekte Auswahl von Pitch-Akzenten sollte aus einem Korpus gelernt werden und zu einer besseren Verständlichkeit und zu einem natürlicheren Klang des Ergebnisses führen. Dazu untersuchten Zervas et al. sowohl einen allgemeinen Korpus als auch domänenspezifische Korpora, wie beispielsweise einen *Museumskorpus*, der Äußerungen eines professionellen Sprechers mit typischen Museumstouren enthielt.

Die Merkmale für die Klassifikationsaufgabe waren: Wortart, flache syntaktische Informationen, Anzahl der Silben pro Wort, Unterbrechungsindizes, Grenztöne, Anzahl der Wörter seit der letzten Hauptunterbrechung und Anzahl der Wörter bis zur nächsten Hauptunterbrechung. Hier wird bereits ein Unterschied zwischen Klassifikationsproblemen auf der Spracheingabeseite, zu denen auch die Sprecherklassifikation gehört, und denjenigen auf der Sprachausgabeseite deutlich: Bei Text-to-Speech können auch Merkmale betrachtet werden, die sozusagen „in der Zukunft" liegen, wie das zuletzt genannte Merkmal.

In einer ersten Studie wurden die Klassifizierer mit Daten aus der Domäne getestet, mit der sie trainiert wurden (domänengleich). Beim zweiten Versuch wurden sie domänengekreuzt, d. h. mit der jeweils anderen Domäne getestet, als sie trainiert worden waren. Auf Basis einer zehnfachen Kreuzvalidierung fanden Zervas et al. heraus, dass der Naive-Bayes-Klassifizierer im Vergleich zu anderen getesteten Verfahren – u.a. CART (vgl. 8.5) – besonders bei fehlenden Daten und bei seltener vorkommenden Pitch-Akzenten bessere Ergebnisse erzielt. Der NB-Klassifizierer zeichnet sich darüber hinaus besonders dadurch aus, dass er sehr stabil beim Test zwischen den verschiedenen Domänen ist. Was die Gesamtgenauigkeit über alle Pitch-Akzente und Test-Konfigurationen betrifft, zeigt jedoch der CART-Klassifikator mit 72 % Gesamtgenauigkeit Vorteile gegenüber den NBs mit 65 %.

Rienks, Poppe und Poel (2005) verglichen die Naive-Bayes-Methode sowie die in Abschnitt 8.7 besprochenen *Künstlichen Neuronalen Netze* (engl. *Artificial Neural Networks*, ANNs) direkt mit der menschlichen Klassifikationsfähigkeit. Die Klassifikationsaufgabe klingt zunächst etwas ungewöhnlich: In einem Meeting soll auf der Basis von Kopfbewegungen herausgefunden werden, wer der Sprecher ist. Hintergrund dieser Untersuchung ist die Forschung an so genannten *smart meeting rooms*, die besondere Dienstleistungen für eine Besprechung anbieten sollen. Dazu gehört das An- und Abschalten von Mikrofonen für die jeweiligen Sprecher, die Erstellung einer automatischen Mitschrift und sogar eine maschinell gesteuerte Leitung der Sitzung, bei welcher das System die Aufgabe eines *session chairs* übernimmt.

In einem solchen System dienen Kameras als eine Art von Eingabesensoren. Die spezielle Untersuchung diente dem Zweck, Wissen über das menschliche Blickverhalten auf Basis von Daten über die horizontalen Kopfbewegungen zu ermitteln. Um Datenmaterial für das Training zu erhalten, zeichneten die Autoren drei Besprechungen von insgesamt 21 Minuten Länge auf. Neben dem Video-Signal wurden mit speziellen, am Kopf der Versuchspersonen fixierten Sensoren die Kopfbewegungen aufgezeichnet.

Ähnlich wie bei Zervas et al. (2004) wurden auch hier sowohl domänengleiche als auch domänengekreuzte Evaluationsserien durchgeführt, wobei die jeweilige Besprechung als Domäne galt. In der ersten Serie wurden sowohl Trainings- als auch Testdaten aus einer Besprechung entnommen. In Serie zwei wurden Training und Test auf Basis von Daten aus allen Besprechungen durchgeführt. In Serie drei wurde mit zwei Besprechungen trainiert und mit der Dritten getestet, und in Serie vier wurde schließlich mit einer Besprechung trainiert und mit den beiden anderen getestet. In Serie eins (einzelne Besprechung) erzielten die Klassifizierer gute Ergebnisse: Die NBs erreichten eine Genauigkeit von 83.13 %, die Künstlichen Neuronalen Netze 78,43 %. In Serie

zwei ging die Genauigkeit bereits deutlich zurück: NB 69.4 %, ANN 62.9 %. In Serie drei und vier schließlich konnte nur noch eine sehr geringe Genauigkeit erzielt werden: NB 39.53% / 38.16%, ANN 45.33% / 40.27%.

Es ist festzustellen, dass die NBs zwar in den ersten beiden Serien bessere Ergebnisse erzielt, das Neuronale Netz jedoch bei den offensichtlich sehr schwierigen Aufgaben drei und vier etwas im Vorteil ist. Rienks et al. führen den deutlichen Rückgang der Genauigkeit bei den domänengekreuzten Untersuchungen darauf zurück, dass erstens bei den Besprechungen verschiedene Themen behandelt werden und daher Unterschiede im Verhalten der Teilnehmer zu erwarten waren, und zweitens das Blickverhalten stark auf individuellen Eigenarten basiert und dass daher eine Verallgemeinerung sehr schwierig ist. Sie kündigen Untersuchungen unter Hinzunahme weiterer Merkmale an.

Ein interessanter Aspekt bei dieser Arbeit ist der direkte Vergleich der maschinellen mit der menschlichen Klassifikationsleistung. Um zu testen, wie gut Menschen in der Lage sind, den Sprecher auf Basis der zum Training verwendeten Merkmale vorherzusagen, konnten Rienks et al. ihren Versuchspersonen jedoch nicht die Videoaufzeichnungen präsentieren, da diese dann ihre Entscheidungen auch auf der Basis von anderen, nicht betrachteten Merkmalen hätten treffen können. Ebenso wenig waren die Merkmalsvektoren an sich als Stimuli geeignet, da Menschen bekanntermaßen bei der Verarbeitung von Zahlenkolonnen einen erheblichen Nachteil gegenüber Maschinen haben. Rienks et al. lösten diese Problem auf elegante Art und Weise: Sie präsentierten die jeweiligen Sitzungen als Computersimulation unter Verwendung von virtuellen Charakteren. Dadurch hatten sie eine völlige Kontrolle über die Eigenschaften der „Teilnehmer“ und konnten somit alle bis auf die gewünschten Hinweise (nämlich die horizontalen Kopfbewegungen) weglassen. Die Charaktere sahen z. B. alle gleich aus und waren zur Identifikation mit Nummern versehen. Diese Art der Untersuchung ist dann geeignet, wenn es nicht darum geht, das Merkmalset zu optimieren, sondern ausschließlich darum, die Leistung des Klassifizierers zu quantifizieren. Die Ergebnisse dieses Tests zeigen, dass menschliche Versuchspersonen mit ihren Einschätzungen mit ungefähr 38 % deutlich hinter der Genauigkeit der Klassifizierer zurückbleiben.

Huang und Hsu (2002) wenden auf das Problem der Sprecheridentifikation keinen kanonischen Naive-Bayes-Klassifizierer, sondern eine von den Autoren selbst entwickelte Variante an. Sie nennen dieses Variante *Homologous Naive Bayes* (HNB), um anzudeuten, dass es sich um einen Klassifikator handelt, der auf *homologe Mengen* von Eingabevektoren angewendet wird. Homolog ist eine Menge von Vektoren dann, wenn alle Elemente zu derselben nicht bekannten Klasse gehören. Die Sprecheridentifikation ist für Huang und Hsu ein Beispielproblem, bei dem derartige Merkmalsmengen vorkommen. Tatsächlich extrahierten sie in ihrer Studie akustische Merkmale zur Sprecheridentifikation mit einem Hanning-Fenster von 30 ms. Daher lag pro Äußerung nicht nur ein Merkmalsvektor vor, sondern eine Serie, die zum selben Sprecher gehört. Die Aufgabe des HNB-Klassifizierers ist es, die Elemente der Merkmalsmenge nicht unabhängig voneinander zu klassifizieren, sondern die homologe Eigenschaft zu berücksichtigen.

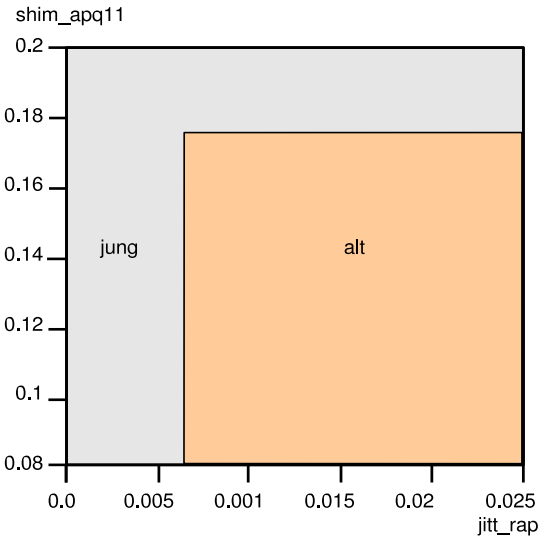


Abbildung 8.3: Entscheidungsregionen eines Naive-Bayes-Klassifizierers für das Sprecheralter auf Basis der beiden Merkmale jitt_rap und shim_apq11.

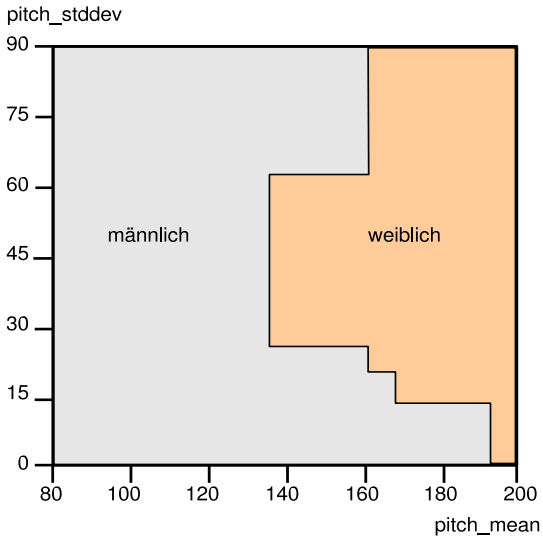


Abbildung 8.4: Entscheidungsregionen eines Naive-Bayes-Klassifizierers für das Sprechergeschlecht auf Basis der beiden Merkmale pitch_mean und pitch_stddev.

In einer Studie verglichen die Autoren die Performanz des HNB-Klassifizierers mit der eines Gaussian-Mixture-Models (GMMs). Sie betrachteten dabei sowohl die mittlere Genauigkeit als auch die mittlere CPU-Zeit für verschiedene Größen von homologen Mengen. Eine Analyse der dargestellten Tabellen zeigt, dass beide Verfahren in dem Test eine Genauigkeit von 99 % erreichen. Die HNBs benötigen mehr Vektoren, um die gleiche Genauigkeit wie die GMMs zu erreichen, sind jedoch schneller und nehmen weniger Rechenleistung in Anspruch.

8.3.2 Naive-Bayes in AGENDER

Für die Sprecherklassifikation in AGENDER wird eine NB-Implementierung aus dem WEKA-Paket verwendet (vgl. Witten und Frank, 1999). Ausgehend von einer univariaten Gauß'schen Wahrscheinlichkeitsdichte p für alle n Elemente des Merkmalsvektors $\mathbf{x}$, wird bei dieser Implementierung die Likelihood von ω_i berechnet durch

$$L(\omega_i|\mathbf{x}) = \prod_{i=1}^{n} p(x_i) \tag{8.3}$$

und die Wahrscheinlichkeit von ω_i durch

$$P(\omega_i|\mathbf{x}) = \frac{L(\omega_i)}{L(\omega_i) + L(\omega_j)}. \tag{8.4}$$

In Abbildung 8.3 wird eine $jitt\_rap$ / $shim\_apq11$-Projektion der Entscheidungsgrenze eines NBs für zwei Altersklassen dargestellt, in Abbildung 8.4 eine $pitch\_stddev$ / $pitch\_mean$-Projektion der Entscheidungsgrenze eines NBs für das Sprechergeschlecht. Es ist deutlich zu erkennen, dass die Methode zu einfachen, treppenförmigen Entscheidungsgrenzen tendiert.

Die folgende Tabelle stellt die Konfusionsmatrix eines NBs bezüglich Gruppierung 0 dar. Die Gesamtgenauigkeit liegt mit 50.39 % auf annähernd demselben Niveau wie die des GMMs, und auch die einzelnen True Positive Rates unterscheiden sich zwischen den beiden Klassifizierern nur geringfügig.

8.8	Gruppierung 0					Gesamtgenauigkeit 50.39 %		
	Kw	Km	Jw	Jm	Ew	Em	Sw	Sm
Kw	73.83	5.01	13.55	3.88	2.6	0.05	1.03	0.05
Km	58.32	8.98	13.75	10.26	5.15	0.05	3.49	0
Jw	52.68	4.32	23.76	8.89	6.04	0.1	4.22	0
Jm	24.15	7.51	8.05	21.4	12.13	3.63	10.16	12.96
Ew	0.88	0.39	0.2	1.47	83.55	7.07	2.9	3.53
Em	0	0.25	0	0.44	10.26	83.01	1.62	4.42
Sw	1.72	3.24	1.67	8.89	19.15	3.04	41.24	21.06
Sm	0.05	0.49	0	5.15	7.95	8.3	10.7	67.35

Die Performanz bezüglich der übrigen Gruppierungen lässt ebenfalls nur einen marginalen Unterschied im Vergleich zum GMM erkennen, weshalb an dieser Stelle auf eine eingehende Interpretation verzichtet werden kann. Abbildung 8.5 stellt eine Übersicht der Gesamtperformanz der NBs bezüglich der verschiedenen Gruppierungen dar, wobei die roten Balken das jeweilige Zufallsniveau markieren.

8.9	Gruppierung 1			Gesamtg. 69.01 %
	KwKmJw	Ew	Sw	EmSm
KwKmJw	80.41	15.22	4.03	0.34
Ew	14.19	65.19	11.09	9.52
Sw	3.04	25.77	45.51	25.68
EmSm	0	3.44	11.63	84.93

8.10	Gruppierung 2		Gesamtg. 66.88 %	
	KwKmJw	JmEw	SwSm	Em
KwKmJw	84.05	12.76	2.99	0.2
JmEw	28.82	45.85	13.79	11.54
SwSm	1.62	17.87	51.94	28.57
Em	0	5.11	9.23	85.67

8.11	Gruppierung 3			
	Stimme (82.01 %)		Sprechverhalten (58.56 %)	
	KwJwEwEm	SwSm	KwJwEwEm	SwSm
KwJwEwEm	84.88	15.12	90.87	9.13
SwSm	20.86	79.14	73.78	26.22

8.12	Gruppierung 4		Gesamtg. 76.70 %
	KwKmJwJmEw	SwSm	Em
KwKmJwJmEw	87.53	9.72	2.75
SwSm	18.07	55.62	26.31
Em	2.75	10.31	86.94

8.13	Gruppierung 5			
	Stimme (85.58 %)		Sprechverhalten (85.19 %)	
	KwKmJwJmSwSm	EwEm	KwKmJwJmSwSm	EwEm
KwKmJwJmSwSm	83.7	16.3	79.75	20.25
EwEm	12.54	87.46	9.38	90.62

8.14	Geschlecht	Gesamtg. 92.01 %
	Ew	Em
Ew	89.99	10.01
Em	5.97	94.03

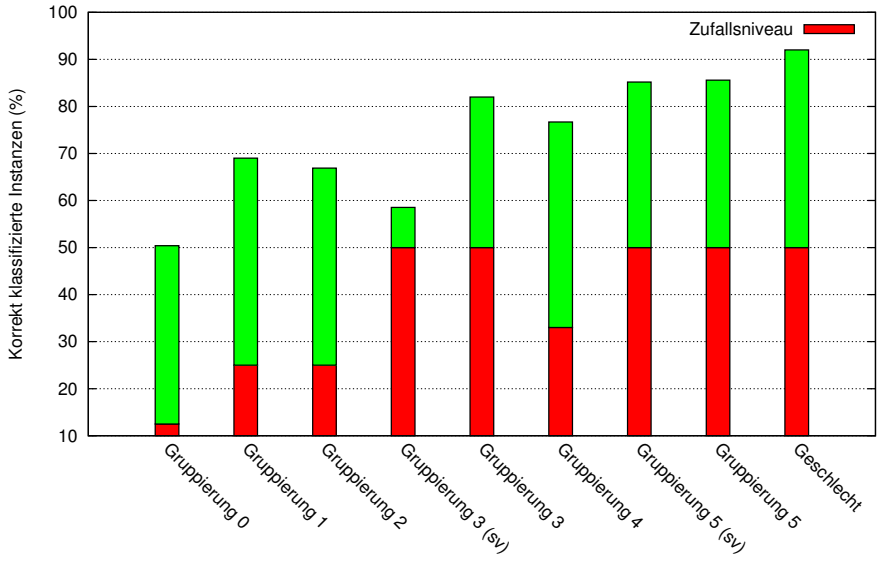


Abbildung 8.5: Übersicht über die Gesamtperformanz der NBs bezüglich der verschiedenen Gruppierungen. Die roten Balken markieren das jeweilige Zufallsniveau.

8.4 k-Nearest-Neighbor

Die bisher betrachteten Methoden haben gemein, dass sie die zugrunde liegenden Wahrscheinlichkeitsdichten als bekannt voraussetzen. Nach Duda et al. (2000, S. 161) ist es jedoch in der Praxis nicht in jedem Fall gegeben, dass hochdimensionale Wahrscheinlichkeitsdichten akkurat durch das Produkt von eindimensionalen Verteilungen repräsentiert werden, wie es bei GMMs und NBs der Fall ist. Die *Nearest-Neighbor*-Regel gehört zu den so genannten *nicht-parametrischen* Methoden, welche nicht von einer bekannten Wahrscheinlichkeitsdichte ausgehen. Die Nearest-Neighbor-Regel überspringt vielmehr die Einschätzung der Parameter und geht direkt zu der Entscheidungsfunktion über: Sei $D^n = \{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$ eine Menge von n etikettierten Prototypen. $\mathbf{x}' \epsilon D^n$ sei der Prototyp, der am nächsten zu einem Testpunkt $\mathbf{x}$ gelegen ist. Dann heißt die Nearest-Neighbor-Regel zur Klassifikation von $\mathbf{x}$: Weise $\mathbf{x}$ die Kategorie zu, die dem Etikett von $\mathbf{x}'$ entspricht.

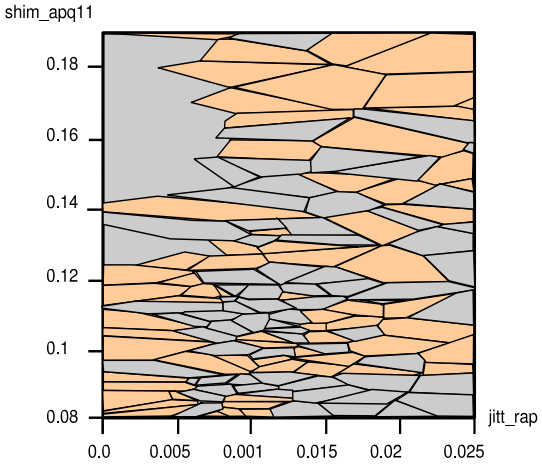


Abbildung 8.6: Voronoi-Mosaik auf Basis einer $jitt\_rap$ / $shim\_apq11$-Projektion des Merkmalsraums aus der Altersklassifikation.

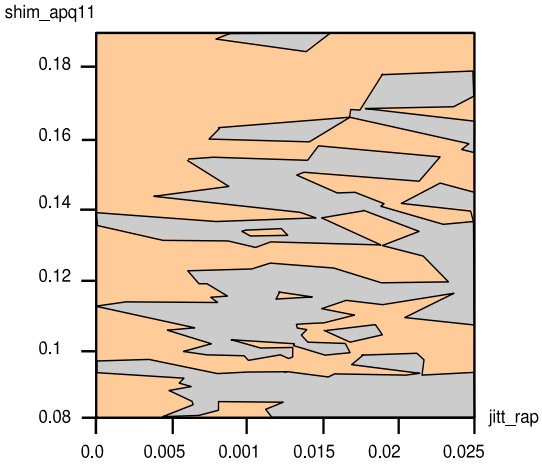


Abbildung 8.7: Editiertes Voronoi-Mosaik.

Die Nearest-Neighbor-Methode wurde bereits in den 50er Jahren für die Statistik entwickelt und seit den 60er Jahren als Klassifikationsalgorithmus eingesetzt (Witten und Frank, 1999). Instanzbasierte Algorithmen wie der Nearest-Neighbor-Algorithmus bringen einen verhältnismäßig geringen Trainingsaufwand mit sich, da sie keine Abstraktionsleistung verrichten müssen. Dafür ist der Speicherbedarf sowie der Aufwand bei der Klassifikation neuer Instanzen größer, vor allem dann, wenn eine inkrementelle Erweiterung der Trainingsdatenbank erfolgen soll. Aha, Kibler und Albert (1991) formulieren eine Reihe von Problemen, die verschiedene Derivate der Nearest-Neighbor-Methode betreffen: 1. Sie seien komputationell aufwendig, weil sie sämtliche Trainingsdaten vorhalten. 2. Sie seien intolerant gegenüber fehlenden Daten und Rauschen. 3. Sie seien intolerant gegenüber irrelevanten Attributen. 4. Ihre Qualität sei in hohem Maße abhängig von der Wahl der Distanzfunktion. 5. Sie erklärten nicht die Struktur der Daten.

Kritikpunkt Nummer eins ist insofern gerechtfertigt, als die Nearest-Neighbor-Methode ein *instanzbasiertes Lernverfahren* ist, das nicht wie andere Klassifikationsmethoden die Trainingsdaten in irgendeiner Form in eine abstraktere Repräsentation (z. B. Regeln) überführt. Die Repräsentation eines Nearest-Neighbor-Modells erfolgt mithilfe eines so genannten *Voronoi-Mosaiks*, wie es in Abbildung 8.6 dargestellt wird (vgl. Duda et al., 2000, S. 178). Die Abbildung zeigt einen Ausschnitt des Merkmalsraums zur Sprecherklassifikation mit zwei Altersklassen aus AGENDER mit der Projektion $jitt\_rap$ / $shim\_apq11$. Die Abschnitte, in denen einzelne Instanzen die „nächsten Nachbarn“ darstellen, sind als rote Flächen für die Klasse ALT bzw. grüne Flächen für die Klasse JUNG erkennbar. Die Menge der Datenpunkte wurde für dieses Beispiel auf einhundert zufällig ausgewählte Instanzen beschränkt, um sichtbare Flächen zu erhalten.

Das Voronoi-Mosaik in Abbildung 8.6 lässt unschwer erkennen, dass die Grenzen zwischen gleichfarbigen Regionen für die Klassifikation nicht relevant sind. Die Komplexität und damit der Speicherplatzbedarf kann durch die so genannte *Editing*-Methode oder *Pruning*-Methode reduziert werden. Dabei werden diejenigen Prototypen entfernt, die von Prototypen derselben Kategorie umgeben sind. Das daraus resultierende Mosaik wird in Abbildung 8.7 dargestellt und repräsentiert die Entscheidungsregionen der Nearest-Neighbor-Regel.

Die Anfälligkeit der Nearest-Neighbor-Methode gegenüber verrauschten Daten kann dadurch verringert werden, dass die Klassifikation nicht nur auf der Basis von einem nächsten Nachbarn erfolgt, sondern auf Basis mehrerer. Ist dies der Fall, spricht man von einer *k-Nearest-Neighbor-Regel* (KNN). Sie weist $\mathbf{x}$ einer Klasse zu auf der Basis der am häufigsten vorgefundenen Klassen unter k nächsten Nachbarn. In anderen Worten: Es wird eine „Abstimmung“ unter einer festen Anzahl von nächsten Nachbarn vorgenommen, wobei durch die Wahl eines ungeraden Wertes für k ein Gleichstand vermieden wird. Die Anfrage startet an dem Testpunkt $\mathbf{x}$ und wächst zu einer sphärischen Region an, bis diese k Trainingsdaten enthält. Der Testpunkt wird dann auf Basis einer Mehrheit der Trainingsdaten kategorisiert. Je größer k ist, desto verlässlicher die Schätzung. Andererseits ist es wünschenswert, dass alle k nächsten Nachbarn $\mathbf{x}'$ möglichst nahe an $\mathbf{x}$ liegen. Das zwingt uns dazu, einen Kompromiss zu wählen, so dass k ein kleiner Teil von n ist (vgl. Duda et al., 2000, S. 184).

Was die Anfälligkeit gegenüber irrelevanten oder auch stark korrelierenden Merkmalen betrifft, ist die Nearest-Neighbor-Methode ebenso „naiv“ wie Naive Bayes. Des Weiteren hängt die Qualität des Klassifikators tatsächlich stark von der Wahl der Distanzfunktion oder Metrik ab (Kritikpunkt Nummer vier). Eine Metrik ist eine Funktion, die uns eine verallgemeinerte skalare Distanz zwischen zwei Mustern gibt.

$$D(\mathbf{a}, \mathbf{b}) = \left(\sum_{k=1}^{d} (a_k - b_k)^2 \right)^{\frac{1}{2}} \tag{8.5}$$

Häufig wird die *Euklidische Distanz* (Gleichung 8.5) verwendet, die jedoch den folgenden Nachteil besitzt: Wenn alle Koordinaten im Merkmalsraum mit einer beliebigen Konstante multipliziert werden, dann sind die Beziehungen bezüglich der Euklidischen Distanz möglicherweise sehr von denen des Originals verschieden. Das Problem betrifft auch unterschiedlich skalierte Trainingsdaten. Hätte eines der Merkmale in Abbildung 8.6 bzw. 8.7 eine kleinere Skala als das andere, würde es praktisch nichts zu der Distanz beitragen. Der unerwünschte Effekt kann jedoch durch eine zuvorige Normierung aufgehoben werden (vgl. Abschnitt 4.3).

8.4.1 k-Nearest-Neighbor in verwandten Arbeiten

Kakumanu et al. (2001) verwenden die Nearest-Neighbor-Methode zur Erzeugung realistisch animierter Gesichter, die die Lippen synchron zu einer Sprachausgabe bewegen und zusätzlich Emotionen mit der passenden Mimik ausdrücken. Dabei müssen die *spatio-temporalen Beziehun-*

gen zwischen Sprachakustik und Gesichtsanimation modelliert werden. Die Nearest-Neighbor-Modelle werden dann dazu verwendet, neue artikulatorische Trajektorien vorherzusagen.

Die Lippenbewegungen werden anhand von Videoaufzeichnungen gelernt, wobei auf der sub-phonetischen Ebene akustische Merkmale der Sprache (z. B. Spektrale Energie, F0) auf so genannte *orale Gesichtsbewegungen* („orofacial movements“) abgebildet werden. Eine Evaluation mit 25 Testsätzen ergab eine gute Performanz des Systems: Der mittlere quadratische Fehler betrug durchschnittlich 0.28 bei einem durchschnittlichen Korrelationskoeffizienten zwischen vorhergesagter und tatsächlicher Trajektorie von 0.77. Kakumanu et al. weisen in ihrer Konklusion allerdings darauf hin, es sei zu erwarten, dass die Nearest-Neighbor-Methode sich bei anwachsender Datenbasis als nicht länger praktikabel herausstellen wird (was die Suchzeiten und den Speicherbedarf betrifft). Sie kündigen für diesen Fall die Untersuchung alternativer Modelle an.

Peskin et al. (2003) untersuchen die Möglichkeit, prosodische Sprachmerkmale als Basis für ein Sprechererkennungssystem zu verwenden. Entgegen den traditionellen Arbeiten in dieser Domäne, die eher auf Merkmalen niedriger Ebene basieren (vgl. Abschnitt 1.5.1), verwenden Peskin et al. Merkmale wie Wortdauer, Pausenfrequenz, verschiedene Statistiken zu Pitch und die Pitchdynamik. Als Klassifikator dient ein KNN mit einer symmetrischen Kullback-Leibler-Metrik. Als Wert für k wählten Peskin et al. $k = 3$. Weitere Werte für k wurden ebenfalls getestet, ohne dass jedoch eine Verbesserung festgestellt werden konnte. Die Modelle zeigten bei einzelnen Merkmalen eine EER (equal error rate) von etwa 15%. Durch die Kombination von Merkmalen konnte die Fehlerrate auf etwa 8% reduziert werden.

Ein k-Nearest-Neighbor-Klassifizierer wird bei Lisetti, Nasoz, LeRouge, Oyzer und Alvarez (2003) angewendet, um Emotionen wie Ärger, Angst, Traurigkeit und Frustration zu erkennen. Der Erkenner soll in ein so genanntes *tele-home health care system* integriert werden, d. h. ein System, bei dem Patienten online medizinische Beratung erhalten. In einem Pilot-Experiment statteten sie zehn Versuchspersonen mit einem tragbaren Computer aus, der die folgenden physiologischen Daten aufzeichnete: Galvanischer Hautleitwert, Hauttemperatur, Wärmefluss und Bewegung. Jede Versuchsperson durchlief fünf verschiedene Bedingungen: eine Bedingung für jede der oben genannten Emotionen sowie eine neutrale Bedingung (Baseline). Die gesammelten Daten wurden normiert, indem die Merkmale mit dem jeweiligen Mittelwert einer Versuchsperson in der neutralen Bedingung in Beziehung gesetzt wurden. Lisetti et al. fassen die Ergebnisse wie folgt zusammen: Der Nearest-Neighbor-Klassifizierer war in der Lage, Ärger und Angst in 80 % der Fälle korrekt zu erkennen, Traurigkeit in 90 % der Fälle und Frustration in 70 % der Fälle. Aus der Beschreibung des Experimentes ist allerdings nicht ersichtlich, wie die verschiedenen Emotionen bei den Versuchspersonen ausgelöst worden sind. Daher sind diese Ergebnisse nur schwer interpretierbar. Lisetti et al. weisen jedoch explizit darauf hin, dass es sich um eine vorläufige Studie handelt und kündigen weiterführende Untersuchungen an. Die Möglichkeit der Klassifikation von physiologischen Daten im Hinblick auf affektive und kognitive Zustände des Benutzers wurde auch im Rahmen dieser Arbeit (am Rande) untersucht, um die Erweiterbarkeit der verwandten Methoden zu eruieren.

8.4.2 k-Nearest-Neighbor in AGENDER

In AGENDER wird eine KNN-Implementierung aus dem WEKA-Paket (vgl. Witten und Frank, 1999) getestet. Die Modelle wurden mit $k = 3$ trainiert.

8.15	Gruppierung 0					Gesamtgenauigkeit 60.02 %		
	Kw	Km	Jw	Jm	Ew	Em	Sw	Sm
Kw	57.93	14.24	23.61	3.49	0.44	0	0.2	0.1
Km	38.59	26.71	21.45	9.38	1.37	0.1	2.21	0.2
Jw	39.52	12.32	39.18	6.48	1.77	0	0.59	0.15
Jm	17.08	10.11	16.2	44.33	3.58	1.33	4.27	3.09
Ew	1.03	0.2	0.69	0.29	89.05	4.17	3.34	1.23
Em	0	0.05	0	0.88	5.01	91.41	0.44	2.21
Sw	0.79	0.64	1.03	1.87	7.66	1.72	73.34	12.96
Sm	0	0.05	0	2.31	4.12	2.9	8.44	82.18

Während sich die GMMs und die NBs nur unwesentlich voneinander unterscheiden, zeichnen sich beim KNN bereits für Gruppierung 0 deutliche Unterschiede ab: Die Gesamtgenauigkeit liegt mit 60.02 % höher als die der beiden zuvor beschriebenen Klassifikationsmethoden. Die Genauigkeit bei der Klasse Kw ist zwar geringer, allerdings zugunsten einer Verbesserung bei Km. Zusammengenommen erreicht das KNN-Modell bei den Kindern eine True Positive Rate von 71.76 % (ZN = 14.3), was annähernd dasselbe Niveau wie das des GMMs bzw. NBs ist. Die Genauigkeit für die übrigen Klassen liegt jedoch in allen Fällen deutlich darüber, so dass das KNN bei der Klasse Ew sogar mit einer Trefferquote von 91.41 % die 90-Prozent-Marke überschritten hat. Die Werte im Bereich von 40 %, welche die beiden Vorgänger bei den weiblichen Senioren zu verzeichnen haben, konnten ebenfalls stark verbessert werden, so dass er hier bei zufriedenstellenden 74.34 % liegt.

8.16	Gruppierung 1			Gesamtg. 69.84 %
	KwKmJw	Ew	Sw	EmSm
KwKmJw	80.81	15.02	3.53	0.64
Ew	15.22	58.66	17.33	8.79
Sw	5.5	15.91	59.16	19.44
EmSm	0	3.44	15.81	80.76

Die Gesamtperformanz des KNNs bezüglich Gruppierung 1 ist zwar mit 69.06 % gegenüber derjenigen des GMMs mit 69.84 % nur geringfügig größer, aber die True Positive Rates der einzelnen Klassen sind ausgeglichener. Vor allen Dingen für die Klasse Sw verzeichnet das KNN deutlich bessere Ergebnisse.

8.17	Gruppierung 2		Gesamtg. 70.59 %	
	KwKmJw	JmEw	SwSm	Em
KwKmJw	77.42	18.36	4.07	0.15
JmEw	21.06	48.9	22.19	7.85
SwSm	1.42	12.03	76.49	10.06
Em	0	3.88	16.54	79.58

Ein ähnliches Bild zeigt sich auch bezüglich Gruppierung 2. Die Differenz der Gesamtgenauigkeit ist zwar höher (70.59 % beim KNN und 66.9 % beim GMM), aber dennoch nicht so hoch, wie Gruppierung 0 es hätte erwarten lassen. Dafür ist die Konfusionsmatrix ausgeglichener: Während beim GMM die beiden Zellen oben links und unten rechts vergleichsweise hohe Werte und die mittleren Zellen der hervorgehobenen Diagonale geringe Werte beinhalten, ist diese Unausgeglichenheit hier nicht gegeben. Lediglich die Klasse der Jugendlichen verzeichnet noch eine mit einigem Abstand geringere True Positive Rate.

8.18	Gruppierung 3			
	Stimme (95.4 %)		Sprechverhalten (71 %)	
	KwJwEwEm	SwSm	KwJwEwEm	SwSm
KwJwEwEm	94.08	5.92	74.96	25.04
SwSm	3.92	96.71	32.97	67.03

Ausgeglichener sind auch die Leistungen der beiden Modelle bezüglich Gruppierung 3. Das Stimmmerkmalsmodell ist zwar mit 95.4 % immer noch deutlich besser als das Sprechverhaltensmodell, dessen True Positive Rate der Klasse SwSm ist jedoch deutlich besser, als es bei den bisher besprochenen Methoden zu beobachten ist. Daher kann das KNN-Sprechverhaltensmodell durchaus zur Unterscheidung der Senioren von den übrigen Altersklassen herangezogen werden.

8.19	Gruppierung 4	Gesamtg. 87.07 %	
	KwKmJwJmEw	SwSm	Em
KwKmJwJmEw	86.7	11.59	1.72
SwSm	5.84	85.57	8.59
Em	0.54	10.51	88.95

Bezüglich Gruppierung 4 erreicht das KNN eine Gesamtgenauigkeit von 87.07 %, was rund 10 % über der des GMMs bzw. NBs liegt. Diese Verbesserung rührt vor allen Dingen von einer wesentlich höheren True Positive Rate der Klasse SwSm, die bei den beiden zuvor diskutierten Methoden jeweils bei etwa 55 % liegt.

8.20	Gruppierung 5			
	Stimme (91.52 %)		Sprechverhalten (85.79 %)	
	KwKmJwJmSwSm	EwEm	KwKmJwJmSwSm	EwEm
KwKmJwJmSwSm	89.91	10.09	86.45	13.55
EwEm	6.97	93.13	14.87	85.13

Da es bezüglich Gruppierung 5 weder im Fall des Stimmmerkmalsmodells noch im Fall des Modells auf Basis des Sprechverhaltens eine starke Unausgeglichenheit in den True Positive Rates gibt, ist der Performanzvorteil des KNNs im Vergleich zu diesen Methoden weniger groß. Die Gesamtgenauigkeit des zuletzt genannten Modells liegt mit 85.79 % auf demselben Niveau wie das des GMMs (85.13 %). Das Stimmmerkmalsmodell verzeichnet allerdings mit 91.52 % eine bessere Trefferquote – beim GMM liegt sie bei 85.59 %.

8.21	Geschlecht	Gesamtg. 92.30 %
	Ew	Em
Ew	90.9	9.1
Em	6.29	93.71

Auch die separat evaluierte Geschlechtsklassifikation mit den beiden Klassen Ew und Em liegt mit 92.30 % auf einem ähnlichen Niveau wie die beiden Vorgängermethoden. Die Leistungen der KNNs bezüglich der verschiedenen Gruppierungen werden in Abbildung 8.8 zusammenfassend dargestellt, wobei die roten Balken das jeweilige Zufallsniveau kennzeichnen.

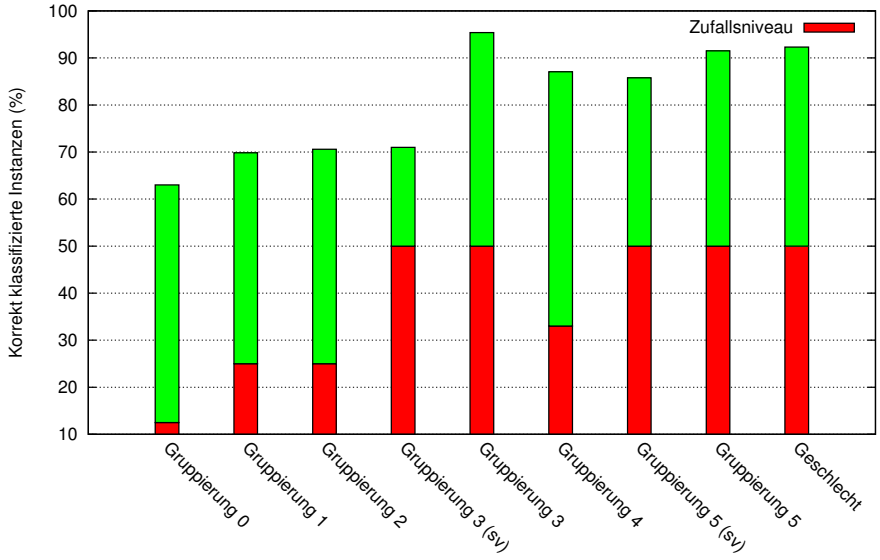


Abbildung 8.8: Übersicht über die Gesamtperformanz der KNNs bezüglich der verschiedenen Gruppierungen. Die roten Balken markieren das jeweilige Zufallsniveau.

8.5 Entscheidungsbäume

Die meisten praktischen Methoden zur Mustererkennung beziehen sich auf Probleme, bei denen die Attribute reelle Werte annehmen können, und bei denen eine bestimmte Metrik angewendet werden kann. Es gibt jedoch auch Probleme, bei denen die Attribute nominelle Werte haben, z. B. diskrete Beschreibungen ohne irgendeinen Ähnlichkeitsbegriff. In diesem Fall ist nicht von einem Merkmalsvektor die Rede, sondern von einer *Liste von Attributen*. Für solche diskreten Probleme sind Entscheidungsbäume besonders geeignet (vgl. Duda et al., 2000, S. 394). Entscheidungsbäume können jedoch auch auf reellwertige Attribute angewendet werden, wobei sie sich gegenüber anderen Methoden vor allem durch Performanzvorteile auszeichnen. Nach Duda et al. (2000, S. 395) handelt es sich um eine intuitive Klassifikationsmethode: Ein Muster wird anhand einer Folge von Fragen klassifiziert, wobei es von der Antwort der einen Frage abhängt, welche Frage als nächstes gestellt wird. Dies ist im Fall von nominellen Daten besonders nützlich, da die Antworten „ja/nein“, „wahr/falsch“ oder „Wert(Attribut) Element/nicht Element einer Menge von Werten“ sein können. Eine solche Folge von Fragen wird gängigerweise als Baum repräsentiert, wobei die Klassifikation eines gegebenen Musters beim Wurzelknoten beginnt (vgl. Abbildung 8.9). Die verschiedenen Kanten, die von dem Wurzelknoten ausgehen, korrespondieren mit den verschiedenen Werten, die dieses Merkmal annehmen kann. Auf die gleiche Weise wird bei den nachfolgenden Knoten verfahren, bis ein Blatt des Baumes erreicht wird, das mit einem Kategorie-Etikett versehen ist.

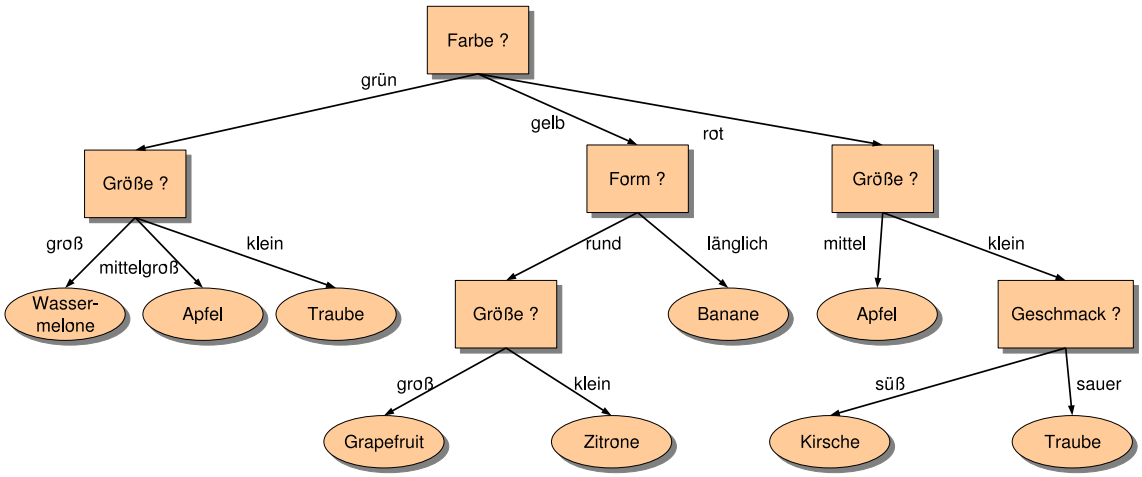


Abbildung 8.9: Einfaches Beispiel für einen Entscheidungsbaum: Klassifikation von Obst anhand von nominellen Merkmalen nach Duda et al. (2000, S. 395).

Der einfache Entscheidungsbaum in Abbildung 8.9 macht einen Vorteil dieser Methode deutlich, nämlich ihre Interpretierbarkeit: Erstens kann die Klassifikation eines bestimmten Musters einfach erklärt werden, indem sie als Konjunktion der Entscheidungen entlang der Kanten des Baumes ausgedrückt wird und zweitens können *Beschreibungen* der Kategorien selbst erstellt werden, indem der Pfad in einen logischen Ausdruck überführt wird. Ein weiterer Vorteil von Entscheidungsbäumen besteht darin, dass die Klassifikation sehr schnell erfolgt, da sie auf einer Folge von sehr einfachen Abfragen basiert (vgl. Duda et al., 2000, S. 396). Entscheidungsbäume stellen einen Spezialfall einer allgemeineren Methodologie zur Erstellung von Bäumen dar, die CART (classification and regression trees) genannt wird. Der CART-Ansatz beinhaltet sechs Fragen bezüglich des Entwurfs eines Entscheidungsbaumes: 1. Wie viel Teilungen sollte es an einem Knoten geben? 2. Welche Eigenschaft sollte an welchem Knoten getestet werden? 3. Unter welchen Bedingungen sollte ein Knoten als Blatt (Endknoten) deklariert werden? 4. Welches sind die Stellen, an denen der Baum „beschnitten“ werden kann? 5. Nach welchen Kriterien sollte die Klasse ausgewählt werden, wenn ein Blatt kein eindeutiges Ergebnis zulässt? 6. Wie soll mit fehlenden Daten umgegangen werden (vgl. Duda et al., 2000, S. 397)?

Die Anzahl der Teilungen wird von dem Designer festgelegt und kann innerhalb eines Baumes variieren, wie es auch in dem Beispiel aus Abbildung 8.9 der Fall ist. Die Anzahl der Verbindungen, die von einem Knoten ausgehen, wird *Verzweigungsfaktor* B (abgeleitet von dem englischen Begriff *branching factor*) genannt. Jeder beliebige Baum kann durch Einführung von zusätzlichen Knoten in einen *binären Baum* ($B = 2$) überführt werden. Binäre Bäume haben gute Eigenschaften, was die Trainings- und Klassifikationszeiten betrifft (vgl. ebd.). Das fundamentale Prinzip bei der Erzeugung von Bäumen ist Einfachheit: Es werden Entscheidungen bevorzugt, die zu einem einfachen, kompakten Baum mit möglichst wenigen Knoten führen. Der Grad der Durchmischung an einem Knoten N wird *Unreinheit* (*impurity*) $i(N)$ genannt. Es gilt $i(N) = 1$, wenn alle Muster,

die N erreichen, zu unterschiedlichen Kategorien gehören, und $i(N) = 0$, falls es nur eine Kategorie gibt. Beim Entwurf des Baumes wird für jeden Knoten eine Eigenschaft gesucht, für welche die Unreinheit $i(N)$ der unmittelbar darauf folgenden Knoten N so gering wie möglich ist. Dabei kommen in der Regel *Gradientenabstiegsverfahren* zum Einsatz.

Zu der Frage, wann die Teilung des Baumes gestoppt werden sollte, ergibt sich nach Duda et al. (2000, S. 402) die folgende Überlegung: Wenn eine Teilung erfolgt, bis jedes Blatt die geringstmögliche Unreinheit aufweist, wird das Ergebnis überangepasst (overfitted) sein. In dem extremen Fall wird jede Trainingsinstanz einem Blatt entsprechen, was keine angemessene Generalisierung darstellt. Wenn die Teilung auf der anderen Seite zu früh gestoppt wird, ist der Fehler auf den Trainingsdaten nicht ausreichend gering. Eine Möglichkeit zur Lösung dieses Problems besteht darin, dass die zu erwartende Fehlerrate eines vollständigen mit der eines verkürzten Baumes auf Basis eines Testsets geschätzt wird. Diese Methode wird *reduced error pruning* genannt (vgl. Witten und Frank, 1999, S. 164). Sie hat den Nachteil, dass der Baum zunächst vollständig aufgebaut werden muss. Eine weitere Möglichkeit besteht darin, die Teilung solange fortzuführen, bis die maximale Verringerung der Unreinheit, die durch den nächsten Zerteilungsschritt erreicht werden kann, nicht größer als ein bestimmter Schwellenwert β ist. Bei Anwendung dieser Methode ist es möglich, dass die Blätter des Baumes auf unterschiedlichen Ebenen liegen. Bei dieser Art von Bäumen, den so genannten *unbalancierten* Bäumen, werden für einfache Entscheidungen weniger Schritte benötigt als für komplexere. Der Nachteil besteht darin, dass der optimale Wert für β nur schwer eingeschätzt werden kann (vgl. Duda et al., 2000, S. 403).

Die einfachste Regel der Zuweisung von Klassen zu bestimmten Endknoten ist die folgende: Wenn alle Muster zur selben Kategorie gehören, wird diese Kategorie gewählt. Wenn die Endknoten eine positive Unreinheit aufweisen, wird diejenige Kategorie ausgewählt, zu der die meisten Muster gehören. Zu den bekanntesten Entscheidungsbaum-Methoden gehört der *ID3-Algorithmus*, der seinen Namen dem Umstand verdankt, dritter in der Reihe der so genannten „interaktiven Dichotomisierer“ zu sein. Da der Algorithmus zur Verarbeitung von nominellen Eingaben entwickelt wurde, werden reellwertige Attribute zunächst zu Intervallen diskretisiert. Der Verzweigungsfaktor entspricht dann der Anzahl der diskreten Werte, die das entsprechende Attribut annehmen kann. Die Tiefe eines *ID3*-Baumes wird bestimmt durch die Anzahl der Eingabewerte: Der Algorithmus fährt mit der Teilung fort, bis alle Knoten rein sind oder es keine Variablen mehr gibt, auf deren Basis verzweigt werden könnte (vgl. Duda et al., 2000, S. 411). Der C 4.5-Algorithmus ist der Nachfolger des *ID3* und gehört zu den am meisten verbreiteten Baum-Methoden . Reellwertige Attribute werden wie in CART behandelt. Für nominelle Daten wird ein Verzweigungsfaktor von $B > 2$ gewählt. Das Problem der Tiefe der Teilung löst der Algorithmus mithilfe einer Heuristik, die auf der statistischen Signifikanz der Verzweigungen basiert (vgl. Quinlan, 1993). Witten und Frank (1999, S. 164) bezeichnen die statistische Grundlage der Heuristik als „ad-hoc“, räumen jedoch ein, dass sie in der Praxis gut funktioniert.

8.5.1 Entscheidungsbäume in verwandten Arbeiten

Jande (2004) verwendet einen C 4.5 Entscheidungsbaum zur Bestimmung von Aussprachevariationen auf Basis von verschiedenen linguistischen Parametern. Hintergrund dieses Klassifikationsproblems ist die Tatsache, dass die Art der Aussprache von Wörtern stark von dem Kontext abhängt, in dem sie vorkommen. Bei einem Text-to-Speech-System kann ein allgemeines Modell dieser Variationen die Natürlichkeit der Ausgabe erhöhen und sogar verschiedene Sprechstile simulieren. Als Trainingsdatenbank verwendet Jande ein Korpus mit etwa einhundert Minuten Spontansprache mit den folgenden Annotationen: Manuelle Transkription, prosodische Grenzen, fokaler Stress, Zögerungen, Wortfragmente und Geschlecht des Sprechers. Hinzu kamen eine Reihe automatisch extrahierter und nicht näher beschriebener phonologischer Merkmale.

Als Grund für die Wahl eines C 4.5 Entscheidungsbaumes führt Jande die Möglichkeit an, das Modell in interpretierbare logische Regeln übersetzen zu können. In der Studie wurden verschiedene Implementierungen verglichen, unter anderem auch die in dieser Arbeit verwendete WEKA-Implementierung der Universität von Waikato (vgl. Abschnitt 11.1.1). Der Vergleich ergab, dass die Implementierungen zwar unterschiedliche Bäume erzeugen, aber ähnlich hohe Vorhersage-Genauigkeiten erreichen (um die 40 % Fehlerrate).

Schötz (2004b) untersucht verschiedene prosodische Merkmale zur automatischen Erkennung des Sprecheralters unter Verwendung von CARTs. Die Ergebnisse dieser Studie wurden bereits in Abschnitt 1.6.2 beschrieben. Neben dem Argument der Interpretierbarkeit, gibt Schötz als Grund für die Wahl eines Entscheidungsbaumes an, dass die Ergebnisse auf diese Art und Weise unmittelbar mit denen von im Vorfeld durchgeführten Perzeptionsexperimenten vergleichbar sind. Eine der Haupt-Fragestellungen dieser Untersuchung war nämlich, ob der automatische Klassifizierer dieselben Merkmale bevorzugen würde, die sich auch bei den Perzeptionsexperimenten als die wichtigsten herausgestellt hatten. Entscheidungsbäume lassen eine direkte Aussage über die Wichtigkeit von Merkmalen zu, da diese früher als Basis für die Unterteilung der Trainingsdaten herangezogen werden als weniger wichtige. Was die Performanz des Entscheidungsbaumes betrifft, so sind die Ergebnisse nicht direkt mit denen in dieser Arbeit vergleichbar, da statt Altersklassen das (kontinuierliche) Alter betrachtet wurde, und somit der Mittlere Fehler anstelle der *True Positive Rates* genannt wird. Dieser liegt für weibliche Sprecher bei durchschnittlich 16.46 Jahren und für männliche Sprecher bei durchschnittlich 16.41 Jahren. In einer weiteren, bisher unveröffentlichten Untersuchung hat Schötz jedoch auch binäre Altersklassen untersucht (ebenfalls mit Entscheidungsbäumen). Diese erreichten eine Genauigkeit von 72.14 % (Altersklasse) und 92.62 % (Geschlecht).

8.5.2 Entscheidungsbäume in AGENDER

In AGENDER wird die C 4.5-Implementation aus dem WEKA-Paket verwendet (Witten und Frank, 1999, S. 159). Der Konfidenzschwellenwert für das *Pruning* liegt bei 0.25. Die minimale Anzahl der Instanzen pro Blatt beträgt zwei. Die Abbildungen 8.10 und 8.11 zeigen die treppenförmigen Entscheidungsgrenzen, die bei Entscheidungsbäumen üblich sind.

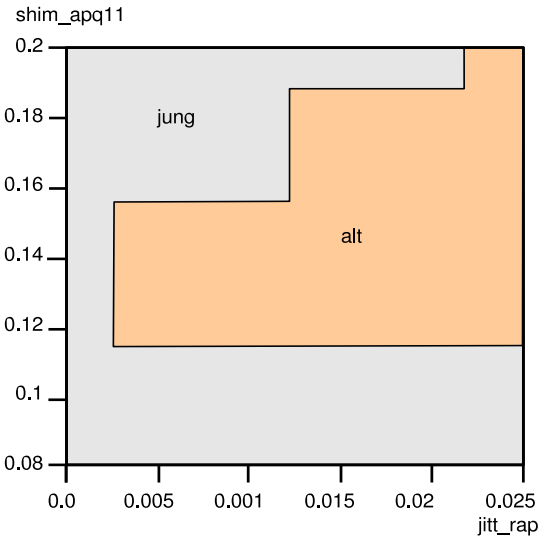


Abbildung 8.10: Entscheidungsregionen eines CART-Klassifizierers für das Sprecheralter auf Basis der beiden Merkmale jitt_ppq und shim_apq11.

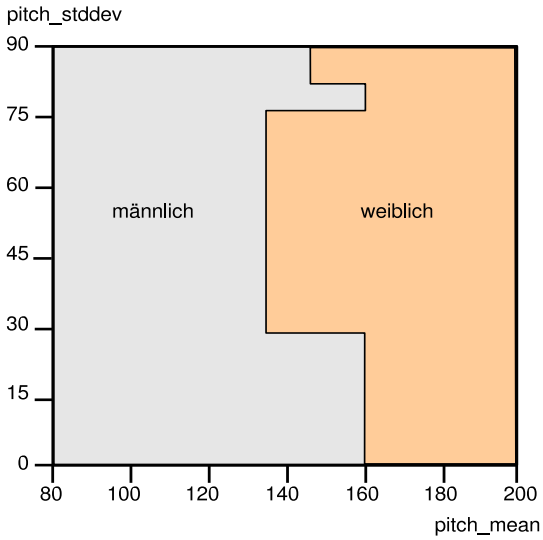


Abbildung 8.11: Entscheidungsregionen eines CART-Klassifizierers für das Sprechergeschlecht auf Basis der beiden Merkmale p_mean und p_stddev.

8.22	Gruppierung 0				Gesamtgenauigkeit 57.61 %			
	Kw	Km	Jw	Jm	Ew	Em	Sw	Sm
Kw	40.35	22.58	26.36	9.43	0.59	0.05	0.64	0
Km	24.15	32.65	22.88	16.3	0.93	0.2	2.75	0.15
Jw	26.71	22.09	35.84	12.47	1.47	0.1	1.13	0.2
Jm	9.38	15.86	11.83	43.54	2.21	2.36	9.18	5.65
Ew	0.69	1.03	1.57	1.72	81.3	3.83	6.82	3.04
Em	0	0.05	0.05	1.91	3.68	86.5	2.55	5.25
Sw	0.44	2.9	1.13	7.66	5.6	2.8	65.59	13.89
Sm	0.1	0.25	0.05	4.86	2.7	4.32	12.62	75.11

Bezüglich Gruppierung 0 ist die Gesamtgenauigkeit des C 4.5 mit 57.61 % zwischen denjenigen der parametrischen Methoden GMM und NB (50.36 % bzw. 50.39 %) und derjenigen des instanzbasierten KNN (60.02 %) angesiedelt. Die Verteilung der True Positive Rates ist mit einer Standardabweichung von 21.92 % noch etwas ausgeglichener als beim KNN mit 24.56 %. Zum Vergleich: Die Standardabweichung der True Positive Rates beim GMM beträgt 30 %.

8.23	Gruppierung 1		Gesamtg. 70.36 %	
	KwKmJw	Ew	Sw	EmSm
KwKmJw	81.49	14.04	4.03	0.44
Ew	14.04	61.07	16.69	8.2
Sw	5.06	18.31	59.99	16.64
EmSm	0.05	4.47	16.59	78.89

Die Perfomanz des C 4.5 bezüglich Gruppierung 1 ist mit 70.36 % sogar höher als diejenige des KNNs (69.84 %). Die Standardabweichung der True Positive Rates beträgt hier nur 11 %, was auf eine verhältnismäßig ausgeglichene Genauigkeit hindeutet. Beim KNN liegt die Standardabweichung bei 12.93 %.

8.24	Gruppierung 2		Gesamtg. 70.44 %	
	KwKmJw	JmEw	SwSm	Em
KwKmJw	84.24	13.79	1.72	0.25
JmEw	23.86	47.77	18.7	9.67
SwSm	2.75	13.94	69.86	13.45
Em	0	6.43	13.65	79.92

Bezüglich Gruppierung 2 ist die Gesamtgenauigkeit annähernd auf demselben Niveau wie diejenige des KNNs. Allerdings bestätigt sich in diesem Fall nicht der bisherige Vorzug des C 4.5, nämlich die relativ hohe Ausgeglichenheit der True Positive Rates: Die Standardabweichung ist mit 16.27 % etwas höher angesiedelt als beim KNN (15.78 %).

8.25	Gruppierung 3			
	Stimme (92.3 %)		Sprechverhalten (74.27 %)	
	KwJwEwEm	SwSm	KwJwEwEm	SwSm
KwJwEwEm	91.78	8.22	82.72	17.28
SwSm	7.19	92.81	34.19	65.81

Das C 4.5-Stimmmerkmalsmodell bleibt bezüglich seiner Gesamtgenauigkeit mit einem Ergebnis von 92.3 % hinter der des KNNs zurück (95.4 %). Dafür ist die Performanz des Modells

auf Basis des Sprechverhaltens mit 74.27 % höher, als es beim KNN der Fall ist (71 %), wobei dies z.T. auf Kosten der Ausgeglichenheit geht: Die Standardabweichung liegt hier bei 11.96 %, während sie für das KNN lediglich 5.61 % beträgt.

8.26	Gruppierung 4	Gesamtg. 88.08 %	
	KwKmJwJmEw	SwSm	Em
KwKmJwJmEw	88.27	9.67	2.06
SwSm	8.2	79.04	12.76
Em	1.47	11.54	86.99

Der Unterschied zwischen C 4.5 und KNN bezüglich Gruppierung 4 ist mit Vorsicht zu interpretieren: Während die Gesamtgenauigkeit des C 4.5 mit 88.08 % etwas höher ist als diejenige des KNNs mit 87.07 %, ist die Ausgeglichenheit geringer: C 4.5 verzeichnet eine Standardabweichung von 8.27 %, KNN eine Standardabweichung von nur 1.72 %.

8.27	Gruppierung 5			
	Stimme (89.06 %)		Sprechverhalten (87.5 %)	
	KwKmJwJmSwSm	EwEm	KwKmJwJmSwSm	EwEm
KwKmJwJmSwSm	89.4	10.6	90.43	9.57
EwEm	11.29	88.71	15.44	84.56

Beim C 4.5 ist bezüglich Gruppierung 5 erstmals ein Modell auf Basis des Sprechverhaltens annähernd so genau wie ein Stimmmerkmalsmodell. Gegenüber dem KNN mit 91.52 % hat das Stimmmerkmalsmodell des C 4.5 mit 89.06 % Genauigkeitseinbußen zu verzeichnen, während das Sprechverhaltensmodell mit 87.5 % im Vergleich zum KNN mit 85.79 % etwas an Performanz zulegt. Allerdings geht dies auch in diesem Fall zu Lasten der Ausgeglichenheit, denn die Standardabweichung beim C 4.5 beträgt 4.15 %, während sie beim KNN bei 0.93 % angesiedelt ist.

8.28	Geschlecht	Gesamtg. 92.86 %
	Ew	Em
Ew	91	9
Em	5.28	94.72

Bezüglich der separaten Evaluation der Geschlechtsklassifikation kann das C 4.5-Modell mit 92.96 % gegenüber dem instanzbasierten KNN mit 92.30 % nochmals einen leichten Performanzvorteil verzeichnen. In Abbildung 8.12 wird die Genauigkeit der Entscheidungsbäume bezüglich der verschiedenen Gruppierungen noch einmal in der Übersicht dargestellt. Die roten Balken markieren wie gewohnt das jeweilige Zufallsniveau.

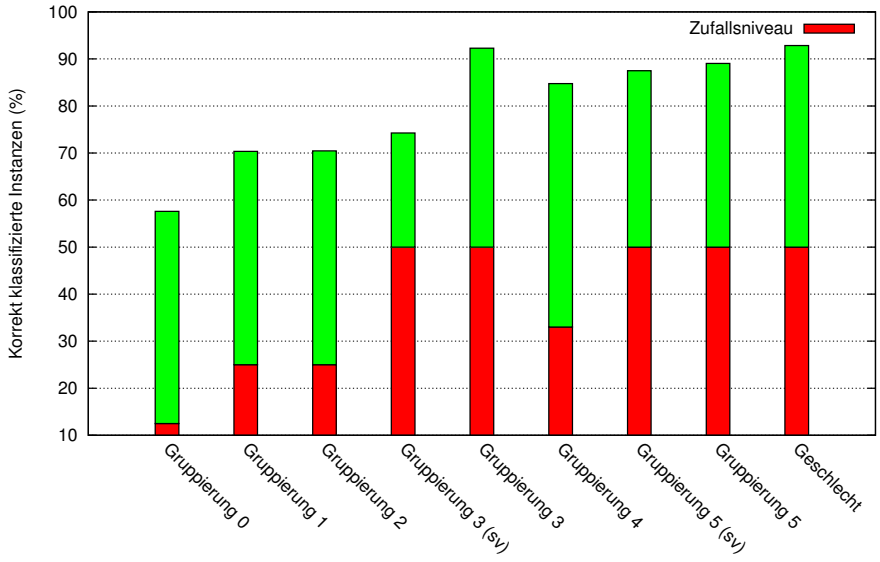


Abbildung 8.12: Übersicht über die Gesamtperformanz der C 4.5 Entscheidungsbäume bezüglich der verschiedenen Gruppierungen. Die roten Balken markieren das jeweilige Zufallsniveau.

8.6 Support-Vector-Machines

Während die parametrischen Klassifikationsmethoden von einer bekannten Wahrscheinlichkeitsdichte ausgehen und deren Parameter mithilfe der Trainingsdaten schätzen, nehmen *Support-Vector-Machines* die Form der Diskriminantenfunktionen als linear an und schätzen die Parameter des Klassifizierers. Sie werden daher zu den nicht-parametrischen Methoden gezählt (vgl. Duda et al., 2000, S. 216). . Eine Diskriminantenfunktion, welche eine lineare Kombination der Komponenten von $\mathbf{x}$ darstellt, kann beschrieben werden durch

$$g(\mathbf{x}) = \mathbf{w}^t\mathbf{x} + w_0, \tag{8.6}$$

wobei $\mathbf{w}$ den *Gewichtungsvektor*, $\mathbf{w}^t\mathbf{x}$ das innere Produkt desselben mit dem Merkmalsvektor $\mathbf{x}$, und w_0 ein *Schwellenwertgewicht* repräsentiert. Die Entscheidungsregel für den zweikategorialen Fall ist: Entscheide ω_1, wenn $g(\mathbf{x}) > 0$, und ω_2, wenn $g(\mathbf{x}) < 0$. Somit wird $\mathbf{x}$ die Kategorie ω_1 zugewiesen, wenn das innere Produkt den Schwellenwert $-\omega_0$ übersteigt, und ansonsten ω_2.

Die Entscheidungsgrenze wird durch $g(\mathbf{x}) = 0$ gebildet. Im multivariaten Fall handelt es sich um eine Entscheidungsoberfläche, eine *Hyperebene* (vgl. Duda et al., 2000, S. 217). Abbildung 8.13 stellt einen einfachen linearen Klassifizierer mit d Eingabeeinheiten (Merkmalsvektor) dar. Die Schwellenwert-Einheit gibt immer den konstanten Wert 1.0 aus. Jeder der Eingabewerte x_i wird mit seinem Gewicht w_i multipliziert, so dass an der Ausgabeeinheit $\sum w_i x_i$ anliegt. Diese gibt in dem Fall $+1$ aus, falls $\mathbf{w}^t\mathbf{x} + \omega_0 > 0$, und ansonsten -1.

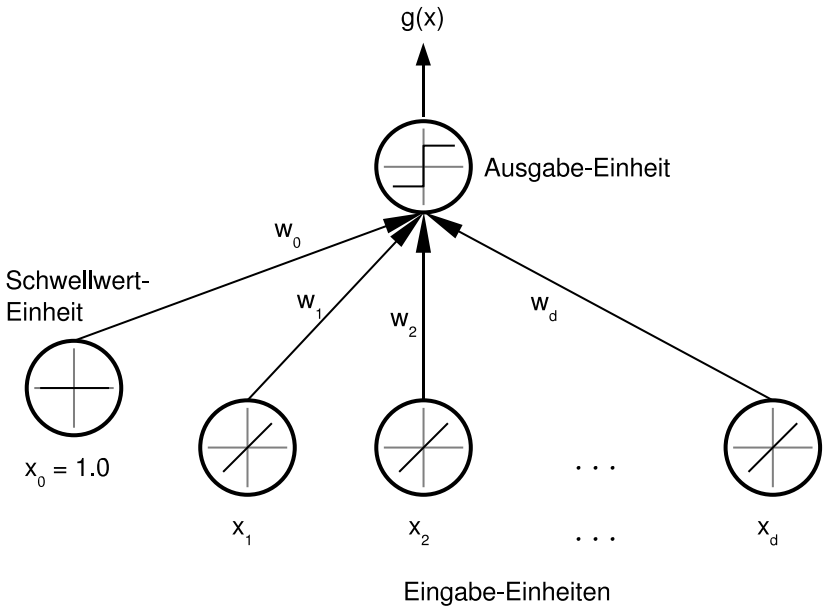


Abbildung 8.13: Ein einfacher linearer Klassifizierer nach Duda et al. (2000, S. 216).

Die trennende Hyperebene H teilt den Merkmalsraum in zwei Halbräume auf: die Entscheidungsregion R_1 für ω_1 und die Entscheidungsregion R_2 für ω_2. Häufig wird die Ausdrucksweise verwandt, alle $\mathbf{x}$ in R_1 seien auf der *positiven Seite* von H und alle $\mathbf{x}$ in R_2 seien auf der *negativen Seite*. Die Lage von H wird bestimmt durch das Schwellenwertgewicht w_0 und die Neigung durch den Gewichtungsvektor $\mathbf{w}$ (vgl. ebd.).

Der positive bzw. negative Wert von $g(\mathbf{x})$ stellt ein algebraisches Maß für die Distanz von $\mathbf{x}$ zu H dar, und zwar durch

$$r = \frac{g(\mathbf{x})}{\|\mathbf{w}\|}, \tag{8.7}$$

wobei $\|\mathbf{w}\|$ die *Euklidische Norm* von $\mathbf{w}$ denotiert, also $\sqrt{w^t w}$. Nach Duda et al. (2000, S. 222)

kann Gleichung 8.6 auch ausgedrückt werden durch

$$g(\mathbf{x}) = w_0 + \sum_{i=1}^{d} w_i x_i. \tag{8.8}$$

Bei linearen Diskriminantenfunktionen gilt:

$$g(\mathbf{x}) = w_0 + \sum_{i=1}^{d} w_i x_i = \sum_{i=0}^{d} w_i x_i, \tag{8.9}$$

wobei gilt $x_0 = 1$. Somit kann der *erweiterte Merkmalsvektor* $\mathbf{y}$ beschrieben werden durch Gleichung 8.10 und analog dazu der *erweiterte Gewichtungsvektor* $\mathbf{a}$ durch Gleichung 8.11.

$$\mathbf{y} = \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_d \end{bmatrix} = \begin{bmatrix} 1 \\ \mathbf{x} \end{bmatrix}. \tag{8.10}$$

$$\mathbf{a} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{bmatrix} = \begin{bmatrix} w_0 \\ \mathbf{w} \end{bmatrix}. \tag{8.11}$$

Die Diskriminantenfunktion $g(\mathbf{x})$ kann somit in der Form $\mathbf{a}^t\mathbf{y}$ geschrieben werden.

Angenommen, in der Trainingsdatenbank ist eine Menge von Proben $\mathbf{y}_1, \ldots, \mathbf{y}_n$ enthalten, von denen einige mit ω_1 etikettiert sind und einige mit ω_2. Diese Proben sollen benutzt werden, um die Gewichte $\mathbf{a}$ in der linearen Diskriminantenfunktion $g(\mathbf{x}) = \mathbf{a}^t\mathbf{y}$ zu bestimmen. Wenn es ein $\mathbf{a}$ gibt, für das alle Trainingsdaten korrekt klassifiziert werden, heißen diese *linear trennbar* (vgl. Duda et al., 2000, S. 223f). In dem zweikategorialen Fall können alle Proben y_i mit den Etiketten ω_2 durch $-y_i$ ersetzt werden. Gesucht wird dann nach einem Gewichtungsvektor $\mathbf{a}$, so dass gilt: $\mathbf{a}^t y_i > 0$ für alle Trainingsdaten. Ein solcher Vektor wird *Trennvektor* oder *Lösungsvektor* genannt. Er maximiert in der Regel die minimale Distanz der Proben zur Hyperebene H, wobei gilt: $\mathbf{a}^t y_i \geq b$ für alle i. Die positive Konstante b wird *Rand* genannt (vgl. ebd.).
Support-Vector-Machines (SVMs) basieren auf den Grundideen linearer Klassifizierer mit Rändern, überführen jedoch die Trainingsdaten zuvor in einen höherdimensionalen Raum. Dabei wird von der Annahme ausgegangen, dass mit einer geeigneten nichtlinearen Abbildung $\gamma(\cdot)$ in eine genügend hohe Dimension die Trainingsdaten zweier Kategorien immer durch eine Hyperebene getrennt werden können (vgl. Duda et al., 2000, S. 262).

SVMs suchen die optimale Hyperebene H, welche diejenige mit dem größten Rand b ist. Die Unterstützungsvektoren (*support vectors*) sind diejenigen (transformierten) Muster, die den Rand und damit H bestimmen (vgl. 8.14). Wenn die Unterstützungsvektoren gegeben sind, können sämtliche anderen Trainingsdaten aus dem Modell gelöscht werden, ohne dass sich die Position und Orientierung von H ändert (Witten und Frank, 1999, S. 190). Sie bilden gleichzeitig die Muster, die am schwierigsten zu klassifizieren sind (vgl. Duda et al., 2000, S. 262).

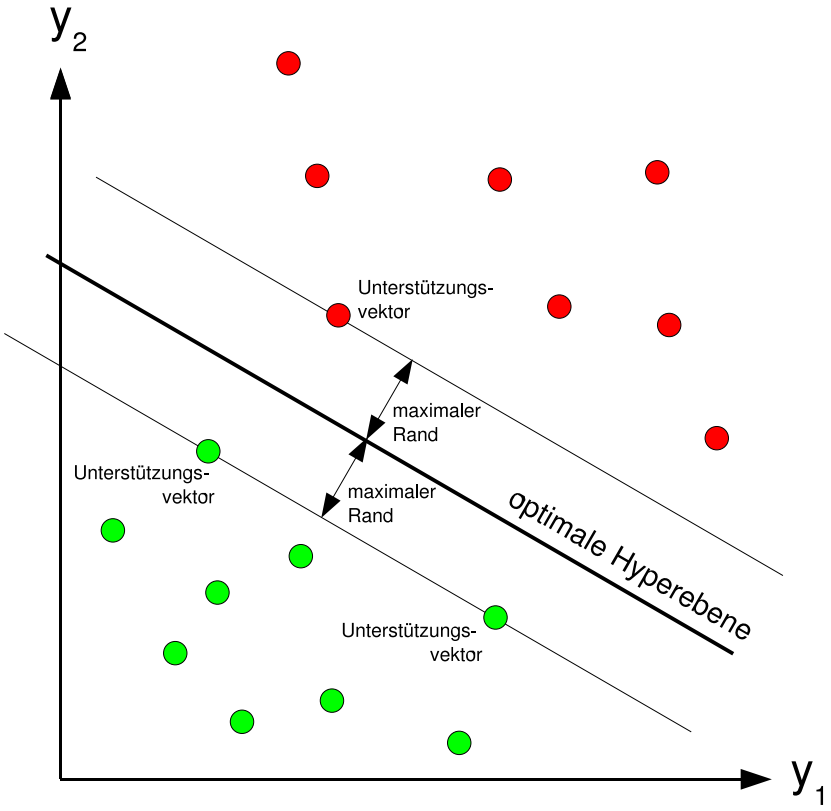


Abbildung 8.14: Entscheidungsgrenze, Rand und Unterstützungsvektoren einer Support-Vector-Machine nach Duda et al. (2000, S. 262).

Das exklusive ODER (XOR) stellt das einfachste Problem dar, das nicht mit einer linearen Diskriminantenfunktion gelöst werden kann. Dem SVM-Ansatz folgend, werden die Merkmale daher in einem Vorverarbeitungsschritt in eine höhere Dimension abgebildet, in der sie dann linear trennbar sind. Duda et al. (2000, S. 264) verwenden für dieses Beispiel die Transformationsfunktionen $\Gamma = \{1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1x_2, x_1^2, x_2^2\}$. Abbildung 8.15 (links) stellt den ursprünglichen Merkmalsraum des Problems dar: Die roten Punkte gehören zur Kategorie ω_1 und die grünen Punkte zu ω_2. Die vier Trainingspunkte werden mithilfe von Γ in einen sechsdimensionalen Raum abgebildet, so dass sie durch die Diskriminantenfunktion $g(\mathbf{x}) = x_1x_2$ trennbar sind. In Abbildung 8.15 (rechts) wird eine zweidimensionale Projektion dieses Raums dargestellt. Aufgrund der starken

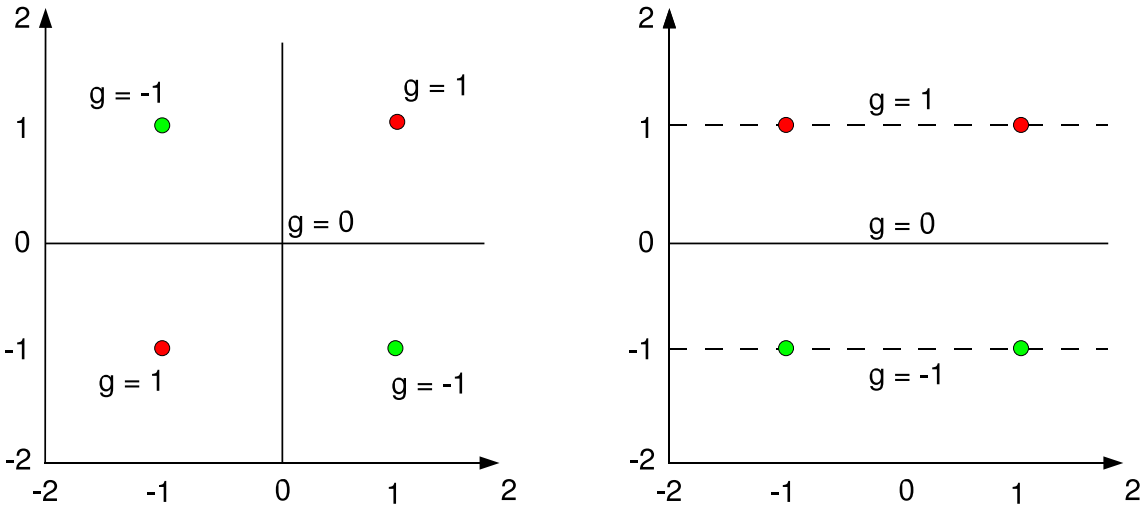


Abbildung 8.15: Links: ursprünglicher Merkmalsraum des XOR-Problems. Rechts: Projektion des in einen sechsdimensionalen Raum überführten Merkmalsraums. x-Achse: $\sqrt{2}x_1$, y-Achse: $\sqrt{2}x_1x_2$. Die Entscheidungsgrenze ist nunmehr linear (vgl. Duda et al., 2000, S. 264).

Symmetrie des Problems sind alle vier Merkmale Unterstützungsvektoren (vgl. Duda et al., 2000, S. 264 f).

Aus diesem Beispiel ist deutlich geworden, dass ein zentraler Schritt bei der Konstruktion einer SVM die Wahl einer geeigneten Menge Γ ist. Sie ist oftmals abhängig von der Domänenkenntnis des Designers. Ansonsten werden häufig polynominale, Gauß'sche oder andere elementare Funktionen gewählt. Witten und Frank (1999, S. 188 f) geben ein Beispiel an, bei dem die ursprüngliche Menge von Attributen durch eine n-fache Faktorisierung transformiert wird. Für zwei Attribute und $n = 3$ wäre dies

$$\mathbf{y} = w_1x_1^3 + w_2x_1^2x_2 + w_3x_1x_2^2 + w_4x_2^3. \tag{8.12}$$

Die Dimensionalität des abgebildeten Raums kann beliebig hoch sein, wird jedoch in der Praxis durch komputationelle Ressourcen beschränkt. Für eine Transformation von ursprünglich zehn Merkmalen mit $n = 5$ müsste der Lernalgorithmus über 2000 Koeffizienten bestimmen (vgl. ebd.).

Ein Vorteil des SVM-Ansatzes besteht darin, dass er im Allgemeinen weniger anfällig für Overfitting-Probleme ist als andere Methoden. Nach Witten und Frank (1999, S. 191) entstehen diese immer dann, wenn die Modelle instabil sind, d. h. sich die Entscheidungsgrenzen mit der Veränderung weniger Instanzen verschieben. Die Hyperebene mit dem größten Rand bleibt jedoch relativ stabil, denn sie ändert sich nur dann, wenn Unterstützungsvektoren hinzukommen oder gelöscht werden. Das gilt auch für einen hochdimensionalen Raum, der durch eine nichtlineare Transformation gespannt wird. Die Unterstützungsvektoren sind globale Repräsentanten der

gesamten Trainingsdatenbank. Es gibt in der Regel nur wenige von ihnen, was eine geringe Flexibilität und damit eine geringere Gefahr von Overfitting bedeutet (vgl. Witten und Frank, 1999, S. 191 f).

8.6.1 Support-Vector-Machines in verwandten Arbeiten

Yacoub, Simske, Lin und Burns (2003) verglichen SVMs und *Künstliche Neuronale Netze* (artificial neural networks, ANNs) für das Problem der Erkennung von Emotionen in so genannten *Voice-Response*-Systemen (vgl. Kapitel 1.1). Zu diesem Zweck führten sie eine Serie von Experimenten mit Sprachproben durch, die von acht Schauspielern gesprochen wurden. In dem ersten Experiment wurden die Daten von sieben Schauspielern für das Training verwendet und diejenigen des achten Schauspielers für den Test. In der zweiten Serie war das Verhältnis sechs zu zwei. Die Evaluation wurde mithilfe einer zehnfachen Kreuzvalidierung durchgeführt. Yacoub et al. berichten, dass für die Erkennung von Ärger die ANNs bessere Ergebnisse erbrachten, wenn es genügend Trainingsdaten gab: in der 7:1 Bedingung erreichten sie 94 % und die SVMs 90.90 %. Bei einem Durchlauf mit einem reduzierten Merkmalssatz, bei dem nur die relevantesten Merkmale verwendet wurden, erzielten die ANNs 91.00 % und die SVMs 90.91 % (jeweils 7:1 Bedingung).

Campbell et al. (2003) verwenden SVMs zur Sprecheridentifikation. Der Ansatz unterscheidet sich von den klassischen Sprecheridentifikationssystemen dadurch, dass statt spektraler Eigenschaften symbolische Merkmale, nämlich *N-Gramme* von Phonemen verwendet werden. Nach Angabe der Autoren geht dies einher mit einer allgemeinen Entwicklung in diesem Bereich, Informationen höherer Ebene mit in den Entscheidungsprozess einfließen zu lassen, und damit die Erkennungsleistung gegenüber klassischen, rein akustischen Systemen zu verbessern (vgl. Abschnitt 1.5.1). Ein weiterer interessanter Aspekt bei Campbell et al. (2003) besteht darin, dass bei der Erkennung der Phonemfolgen verschiedene Module verwendet werden, die jeweils mit einer anderen Sprache trainiert wurden, und zwar mit Englisch, Deutsch, Japanisch, Mandarin und Spanisch. Die Gründe für diese Vorgehensweise sind nach Campbell et al. erstens, dass das fertige System für mehrere Sprachen eingesetzt werden kann, und zweitens, dass die Phonemfolgen einer Sprache, die nicht die Trainingssprache des Modells ist, komplementäre Informationen darstellen, und dadurch die Erkennungsleistung gesteigert werden kann.

Als Eingabe für die SVMs dient ein Merkmalsvektor v, der aus *N-Gramm*-Wahrscheinlichkeiten von Phonemen, die aus dem Trainingsmaterial, der so genannten *conversation site*, ermittelt wurden. Die *N-Gramm*-Wahrscheinlichkeiten für einen gegebenen Sprecher werden mit denjenigen der *conversation site* in Beziehung gesetzt. Bei diesem Vorgang, dem so genannten *term weighting*, wird ein sprecherspezifischer Wert ermittelt, anhand dessen dieser später wieder erkannt werden kann. Campbell et al. (2003) beziffern die Leistung der SVM auf eine EER (*Equal Error Rate*, vgl. 8.2.1) von 3.5 %, was nach ihren Angaben einer Verbesserung von 60 % gegenüber klassischen Systemen auf Basis der *Log-Likelihood-Ratio* entspricht.

8.6.2 Support-Vector-Machines in AGENDER

Da SVMs in der verwendeten WEKA-Version nur als Dichotomisierer implementiert wurden, können lediglich Vergleiche bezüglich Gruppierung 3 und 5 sowie bezüglich der Geschlechtsklassifikation angestellt werden.

8.29	Gruppierung 3			
	Stimme (89.77 %)		Sprechverhalten (67.08 %)	
	KwJwEwEm	SwSm	KwJwEwEm	SwSm
KwJwEwEm	87.6	12.4	61.63	38.17
SwSm	8.05	91.95	27.47	72.53

Die Gesamtgenauigkeit des Stimmmerkmalsmodells ist mit 89.77 % geringer als die der beiden Vorgängermethoden KNN und C 4.5 (95.4 % bzw. 92.3 %), jedoch höher als die der parametrischen Methoden GMM und NB (81.97 % bzw. 82.01 %). Was die Qualität des Sprechverhaltensmodells betrifft, kann die SVM an derselben Position eingereiht werden: Die Trefferquote liegt mit 67.08 % höher als bei den parametrischen Methoden (58.46% bzw. 58.56 %), jedoch niedriger als bei KNN (71 %) und C 4.5 (27.27 %). Die Ausgeglichenheit des zuletzt genannten Modells ist mit einer Standardabweichung von lediglich 7.71 % vergleichsweise hoch. Bei den C 4.5 Entscheidungsbäumen liegt dieser Wert bei 11.96 %.

8.30	Gruppierung 5			
	Stimme (88.26 %)		Sprechverhalten (86.93 %)	
	KwKmJwJmSwSm	EwEm	KwKmJwJmSwSm	EwEm
KwKmJwJmSwSm	91.34	8.66	87.21	12.79
EwEm	14.83	85.17	13.35	86.65

Während die Performanz des Stimmmerkmalsmodells bezüglich Gruppierung 5 mit 88.26 % ebenfalls an besagter mittlerer Position zwischen GMM (85.95 %) / NB (82.01 %) und KNN (91.52 %) / C 4.5 (89.06 %) liegt, zeigt das korrespondierende Sprechverhaltensmodell mit 86.93 % die bisher beste Leistung. Die Gesamtgenauigkeit ist beim C 4.5 mit 87.5 % zwar höher, aber die SVM zeichnet sich dafür durch eine bessere Ausgeglichenheit aus: Die Standardabweichung liegt bei 0.4 %, während sie beim C 4.5 bei 4.15 % angesiedelt ist.

8.31	Geschlecht	Gesamtg. 93.27 %
	Ew	Em
Ew	90.08	9.92
Em	3.54	96.46

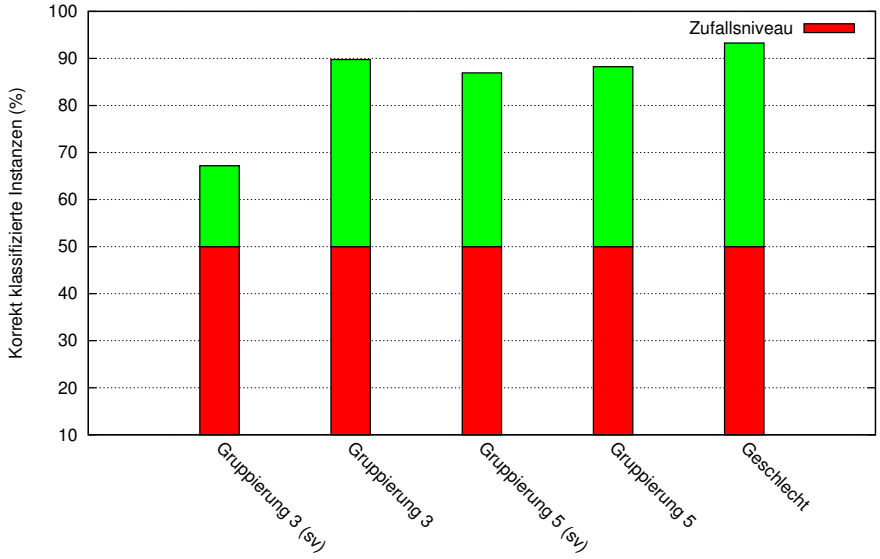


Abbildung 8.16: Übersicht über die Gesamtperformanz der SVMs bezüglich der verschiedenen Gruppierungen. Die roten Balken markieren das jeweilige Zufallsniveau.

Bezüglicher der separaten Evaluation der Geschlechtsklassifikation erreicht die SVM mit 93.27 % die bisher höchste Genauigkeit. Das Niveau der zweitbesten Methode C 4.5 liegt bei 92.86 %. Die Leistungen der SVMs bezüglich der verschiedenen Gruppierungen werden in Abbildung 8.16 zusammenfassend dargestellt, wobei die roten Balken das jeweilige Zufallsniveau kennzeichnen.

8.7 Neuronale Netze

Netzwerke von künstlichen, neuronengleichen Elementen werden seit den 60er Jahren von vielen Wissenschaftlern auf dem Gebiet der adaptiven Signalverarbeitung und Mustererkennung eingesetzt (vgl. Morgan, 1990). Die so genannten *Künstlichen Neuronalen Netze* (*Artificial Neural Networks*, ANNs) stellen eine flexible heuristische Technik zur Mustererkennung mit komplexen Modellen dar. Die konzeptuelle und algorithmische Einfachheit der Rückwärtspropagierung – eines der am meisten genutzten Methoden zum Trainieren von Neuronalen Netzen – zusammen

mit dem offenkundigen Erfolg in einer Vielzahl von anwendungsnahen Problemen machen diese Technik zu einer der bedeutendsten Methoden der Mustererkennung (vgl. Duda et al., 2000, S. 283ff). Beim Entwurf der Netzwerktopologie, die eine wichtige Rolle für die Klassifikation mit Neuronalen Netzen spielt, kann häufig Domänenwissen einfließen, z. B. bei der Wahl der Anzahl von *versteckten Ebenen* (*hidden layers*). Eines der hauptsächlichen Probleme stellt die Regulierung der Netzwerkkomplexität dar: Während die Anzahl der Ein- und Ausgabeknoten durch den Merkmalsraum bzw. die Anzahl der Kategorien bestimmt wird, ist die Gesamtzahl der Gewichte oder Parameter des Netzwerkes zunächst frei. Das Problem besteht darin, dass einerseits eine zu große Anzahl freier Parameter zu einer geringen Generalisierung führt und andererseits bei einer zu geringen Anzahl die Trainingsdaten nicht adäquat gelernt werden. Es ist sehr wichtig zu beachten, dass Neuronale Netze den Designer nicht davon entbinden, sich mit den Daten und der Problemdomäne vertraut zu machen (vgl. ebd.).

Es gibt eine Vielzahl verschiedener Arten von Künstlichen Neuronalen Netze: McCulloch-Pitts-Netze, Kohonen-Netze, Hopfield-Netze und Perzeptron-Netze, um nur einige zu nennen. Der in der vorliegenden Arbeit verwendete Begriff der Neuronalen Netze beschränkt sich jedoch auf den weit verbreiteten, zuletzt genannten Typ: Perzeptron-Netze.

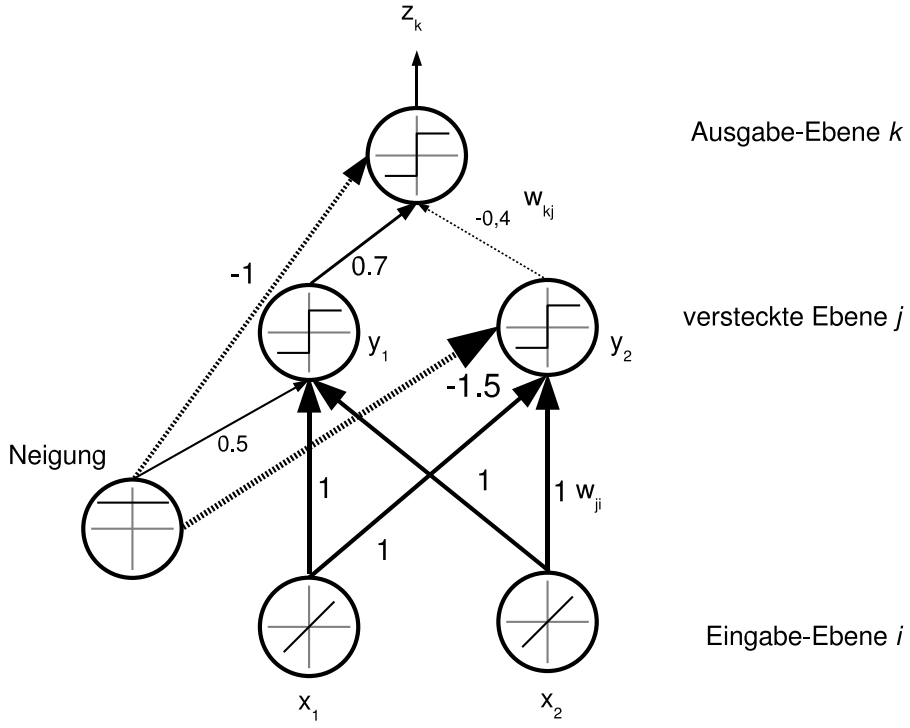


Abbildung 8.17: Einfaches Neuronales Netzwerk mit drei Ebenen zur Lösung des XOR-Problems nach Duda et al. (2000).

Abbildung 8.17 stellt ein einfaches Perzeptron-Netzwerk zur Lösung des aus Abschnitt 8.6 bekannten Problems des exklusiven ORDER (XOR) dar. In diesem Fall handelt es sich um ein sogenanntes *mehrlagiges Perzeptron* (*multilayer perceptron*, MLP). MLPs zeichnen sich dadurch

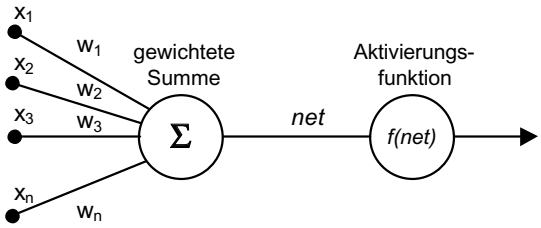


Abbildung 8.18: Aufbau eines „Neurons“.

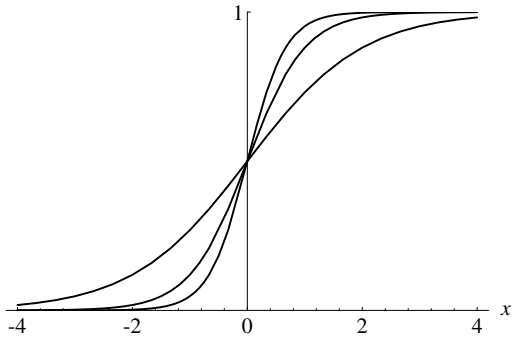


Abbildung 8.19: Sigmoid-Funktion mit c = 1, c = 2 und c = 3 nach Rojas (1996).

aus, dass sie über mehre Neuronenebenen verfügen: eine *Eingabeebene*, eine oder mehrere *versteckte Ebenen* (in dem Fall ist es eine) und eine *Ausgabeebene*. Die Eingabeebene (*input layer*) repräsentiert den zweidimensionalen Merkmalsraum. Daneben gibt es eine einzelne *Neigungseinheit*, die mit allen Einheiten außer den Eingabeeinheiten verbunden ist. Positive („verstärkende“) Verbindungen werde mithilfe von durchgezogenen Linien markiert und negative („hemmende“) Verbindungen mit gestrichelten Linien.

Die Ausgabe einer Einheit wird auf Basis zweier Funktionen berechnet, der *Netzfunktion* (*net*) und der Aktivierungsfunktion (*f(net)*). Wie in Abbildung 8.18 dargestellt wird, stellt *net* eine gewichtete Summe der Eingaben des Knotens dar. In dem Beispiel in Abbildung 8.17 entspricht *f(net)* einer einfachen Schwellenwert- oder Schritt-Funktion (*sign*):

$$f(net) = Sgn(net) \equiv \begin{cases} 1 & \text{falls} \quad net \geq 0 \\ -1 & \text{falls} \quad net < 0. \end{cases} \tag{8.13}$$

Da für die Rückwärtspropagierung jedoch Aktivierungsfunktionen benötigt werden, die differenzierbar und stetig sind, wird in der Praxis häufig die *Sigmoid*-Funktion verwendet, deren allgemeine Form durch

$$S_c(x) = \frac{1}{1 + e^{cx}} \tag{8.14}$$

definiert wird (vgl. Rojas, 1996, S. 150). Die Form der *Sigmoid*-Kurve ist abhängig von der Konstanten c (vgl. Abbildung 8.19). Je höher der Wert für c desto ähnlicher ist die Funktion der Schritt-Funktion. Es ist durchaus möglich, dass die Aktivierungsfunktionen der Ausgabeeinheiten andere sind als die der versteckten Einheiten. Theoretisch kann sogar jede einzelne Einheit ihre eigene Aktivierungsfunktion besitzen.

Eine Konfiguration von Verbindungen, wie das Netzwerk in Abbildung 8.17 sie aufweist, wird als vollständig verbundene 2-2-1-Topologie bezeichnet (vgl. Duda et al., 2000, S. 285). Wie bei den linearen Klassifizierern (vgl. 8.6) wird auch hier aus Gründen der Einfachheit der Eingabe-Vektor um den Wert $x_0 = 1$ für die Neigungseinheit und der Gewichtungsvektor um den Wert w_0 für dessen Gewicht erweitert. Somit kann die *net*-Funktion wie folgt beschrieben werden (vgl. Duda et al., 2000, S. 285):

$$net_j = \sum_{i=1}^{d} x_i w_{ji} + w_{j0} = \sum_{i=0}^{d} x_i w_{ji} = \mathbf{w}_j^t \mathbf{x}, \tag{8.15}$$

wobei i die Eingabeebene indiziert und j die versteckte Ebene. Die Gewichte von der Eingabeebene zur versteckten Ebene werden dementsprechend mit w_{ji} denotiert. In Analogie zur Neurobiologie werden die Verbindungen häufig „Synapsen“ genannt und die entsprechenden Gewichte „synaptische Gewichte“. Die Ausgabeeinheiten berechnen net_k analog (vgl. ebd.):

$$net_k = \sum_{j=1}^{n_h} y_j w_{kj} + w_{k0} = \sum_{i=0}^{n_h} y_j w_{kj} = \mathbf{w}_k^t \mathbf{y}, \tag{8.16}$$

wobei k die Ausgabeebene indiziert und n_h die Anzahl der versteckte Einheiten denotiert. Es kann sehr einfach gezeigt werden, dass das Neuronale Netz aus Abbildung 8.17 tatsächlich das XOR-Problem löst: Die versteckte Einheit y_1 gibt 1 aus, falls $x_1 + x_2 + 0.5 \geq 0$, und ansonsten -1. Entsprechend gibt die versteckte Einheit y_2 -1 aus, falls $x_1 + x_2 - 1.5 \geq 0$, und ansonsten 1. Die Ausgabe-Einheit schließlich gibt 1 aus, wenn $y_1 = 1$ und $y_2 = 1$. Tabelle 8.32 stellt eine Wahrheitstabelle für dieses Problem dar, welche derjenigen des XOR entspricht.

x_1	x_2	y_1	y_2	k
-1	-1	-1	1	-1
-1	1	1	1	1
-1	1	1	1	1
1	1	1	-1	-1

Tabelle 8.32: Wahrheitstabelle für das XOR-Problem.

Neuronale Netze mit versteckten Ebenen heißen nichtlinear, und tatsächlich kann gezeigt werden, dass beliebige Funktionen implementiert werden können – gegeben eine ausreichend große Anzahl versteckter Einheiten, passender nichtlinearer Aktivierungsfunktionen und Gewichten (vgl. Duda et al., 2000, S. 287).

Rückwärtspropagierung ist eine der einfachsten und allgemeinsten Methoden zum Trainieren eines Neuronalen Netzes. Der Algorithmus dient dazu, herauszufinden, in welcher Beziehung die

Ausgabe (und somit der Klassifikationsfehler) mit den Gewichten zwischen der versteckten Ebene und der Ausgabe-Ebene steht (vgl. Duda et al., 2000, S. 289).

Gegeben sei ein Neuronales Netz mit n Eingabeeinheiten, m Ausgabeeinheiten und einer beliebigen Anzahl von versteckten Einheiten. Gegeben sei darüber hinaus ein Trainingssatz $\{(\mathbf{x}_1, \mathbf{t}_1), \ldots, (\mathbf{x}_p, \mathbf{t}_p)\}$ bestehend aus p geordneten Paaren von n- bzw. c-dimensionalen Vektoren (Trainingsmuster mit n Merkmalen und Zielmuster mit c Kategorien). Die Aktivierungsfunktionen sämtlicher Einheiten seien stetig und differenzierbar. Die initialen Gewichte der Verbindungen werden zufällig ausgewählt. Wenn nun ein Trainingsmuster $\mathbf{x}_i$ aus dem Trainingssatz an das Netzwerk angelegt wird, erzeugt dieses eine Ausgabe $\mathbf{z}_i$, die sich im Allgemeinen von dem Zielmuster $\mathbf{t}_i$ unterscheidet. Die Aufgabe des Lernalgorithmus besteht darin, $\mathbf{z}_i$ und $\mathbf{t}_i$ identisch zu machen, d. h. die Fehlerfunktion E zu minimieren (vgl. Rojas, 1996, S. 154). Diese wird definiert als

$$E = \frac{1}{2} \sum_{i=1}^{p} \|\mathbf{z}_i - \mathbf{t}_i\|^2. \tag{8.17}$$

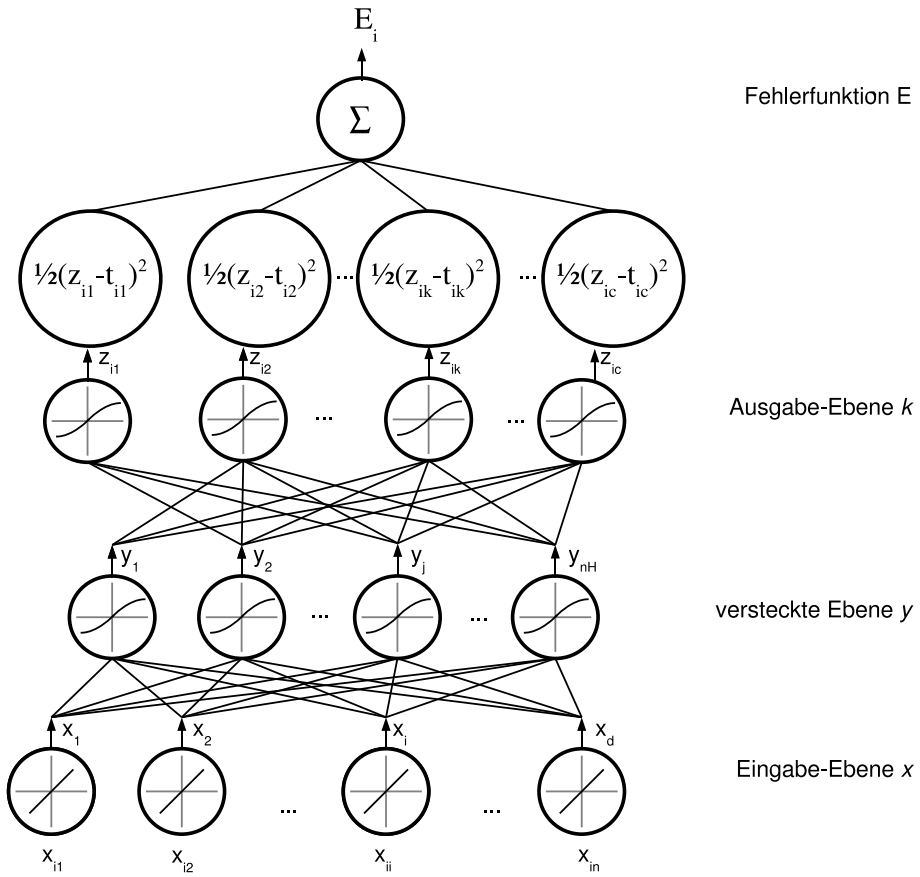


Abbildung 8.20: Erweiterung eines Neuronalen Netzes zur automatischen Berechnung der Fehlerfunktion nach Duda et al. (2000, S. 290).

Zur Korrektur der initialen Gewichte wird der *Gradient* der Fehlerfunktion rekursiv berechnet.

Der erste Schritt besteht darin, das Netz so zu erweitern, dass die Fehlerfunktion automatisch berechnet wird (vgl. Abbildung 8.20). Die Summe aller einzelnen Fehler $E_1, \ldots, E_p$ entspricht der Fehlerfunktion E. Da E ausschließlich durch die Komposition der Aktivierungsfunktionen der Einheiten berechnet wird, ist sie eine stetige und differenzierbare Funktion der Gewichte $w_1, w_2, \ldots, w_\ell$ des gesamten Netzes, und kann daher durch ein iteratives Gradientenabstiegsverfahren minimiert werden (vgl. Rojas, 1996, S. 155).

Nach Duda et al. (2000, S. 293) können beim Training Neuronaler Netze verschiedene *Trainingsprotokolle* angewendet werden: *stochastisches Lernen* und *Stapel-Lernen*. Beim stochastischen Lernen werden die Muster aus der Trainingsdatenbank zufällig ausgewählt und die Gewichte des Netzwerkes für jedes Muster separat angepasst. Die Methode heißt deshalb stochastisch, weil die Trainingsdaten wie eine zufällige Variable angesehen werden können (vgl. ebd.). Beim *Stapel-Lernen* werden zunächst alle Trainingsmuster an das Netz angelegt und die Veränderungen, die sie bei den Gewichten verursachen, summiert. Erst dann erfolgt eine Aktualisierung. Da bei dieser Methode alle Trainingsmuster betrachtet werden, entfällt die zufällige Auswahl (vgl. Duda et al., 2000, S. 295). Beide Verfahren enden dann, wenn ein *Stopp-Kriterium* erfüllt ist, d. h. zum Beispiel wenn die Veränderung der Fehlerfunktion durch die aktuelle Iteration kleiner als ein bestimmter Schwellenwert ist.

Die Ausdrucksmächtigkeit des Neuronalen Netzes und die Komplexität der Entscheidungsgrenze hängt unmittelbar von der Anzahl der versteckten Einheiten (n_H) ab. Während jedoch die Anzahl der Ein- und Ausgabeeinheiten von der Dimensionalität der Muster bzw. der Anzahl der Kategorien abhängt, gibt es in Bezug auf n_H keine solch objektiven Kriterien (vgl. Duda et al., 2000, S. 310). Wenn die Muster gut unterscheidbar oder linear trennbar sind, werden nur wenige versteckte Einheiten benötigt.

8.7.1 Neuronale Netze in verwandten Arbeiten

Konig, Morgan und Chandra (1991) verwenden ANNs zur Einschätzung des Sprechergeschlechts, unter Berücksichtigung dessen die Fehlerrate eines Spracherkenners, wie in Abschnitt 1.3 beschrieben, verringert werden soll. Das Netzwerk für die Klassifikation basiert auf den Sprachproben von 109 Sprechern aus dem Korpus mit der Bezeichnung *DARPA Resource Management*, wobei insgesamt 3 511 Äußerungen zum Training und 479 Äußerungen zum Test verwendet wurden. Über den Anteil von Frauen bzw. Männern wird keine Angabe gemacht. Das Merkmalset besteht aus 12 *mel-cepstral*-Merkmalen und 14 Derivaten der Energie, die über ein Zeitfenster von 20 ms berechnet wurden (vgl. Abschnitt 2.2.3). Das neuronale Netz verfügt demnach über 26 Eingabeeinheiten und zwei Ausgabeeinheiten (eine für jedes Geschlecht). Es wurden mehrere ANNs mit jeweils einer versteckten Ebene, aber einer unterschiedlichen Anzahl versteckter Einheiten, nämlich 16 und 32, trainiert. Darüber hinaus wurde die Größe des betrachteten Sprachabschnittes zwischen 500 ms und 900 ms variiert. Die Genauigkeit der Netze mit 16 versteckten Einheiten liegt in beiden Fällen bei 74.9 %, wohingegen die der Netze mit 32 Einheiten zwischen 71.4 % (500 ms) und 72 % (900 ms) schwankt. Die Performanz ist demnach

bei einer geringen Anzahl versteckter Einheiten überraschenderweise besser. Konig et al. geben an, dass auf der Äußerungsebene eine wesentlich zuverlässigere Erkennung möglich sei. Sie beziffern die erreichte Genauigkeit in dem Fall auf durchschnittlich 93.55 %.

Yu, Chang, Xu und Heung-Shum (2001) vergleichen die Erkennungsleistung von ANNs mit derjenigen verschiedener anderer Klassifikationsverfahren, darunter KNN und SVM, auf dem Gebiet der Emotionserkennung. Die Modelle basieren auf Merkmalen wie der Artikulationsgeschwindigkeit, des mittleren Abstands zwischen stimmhaften Regionen, der mittleren Grundfrequenz und der Anzahl der Anstiege bzw. Abfälle der Grundfrequenz. Neben der neutralen Bedingung untersuchten Yu et al. die emotionalen Zustände Ärger, Freude und Traurigkeit. Die Methode ANN erreichte eine durchschnittliche Genauigkeit von 41.45 % bei einem Zufallsniveau von 25 %. Die Performanz von KNN war mit durchschnittlich 50.89 % jedoch bereits deutlich besser und, SVM erreichte sogar 74.28 %. Der Interpretation der Autoren zufolge besteht der Vorteil von SVMs gegenüber ANNs darin, dass sie auch dann gute Erkennungsleistungen zeigen, wenn die Relevanz der einzelnen Merkmale nicht bekannt ist.

8.7.2 Neuronale Netze in AGENDER

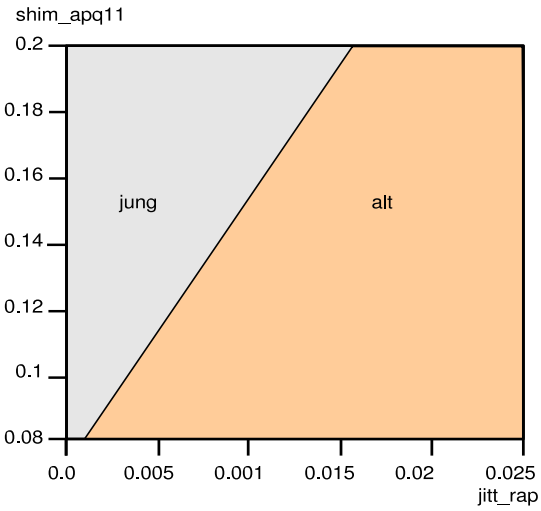


Abbildung 8.21: Entscheidungsregionen eines Neuronalen Netzes für das Sprecheralter auf Basis der beiden Merkmale jitt_rap und shim_apq11.

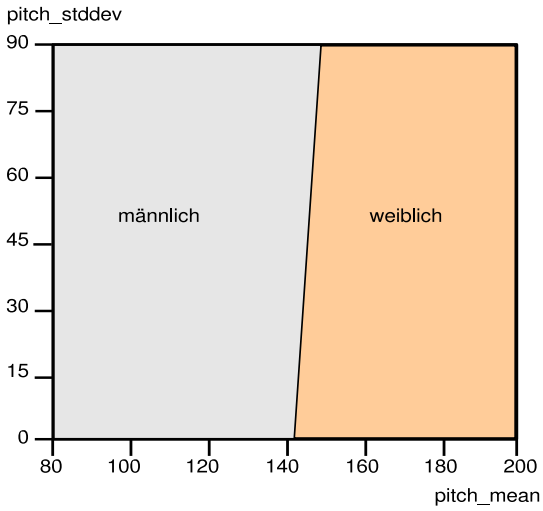


Abbildung 8.22: Entscheidungsregionen eines Neuronalen Netzes für das Sprechergeschlecht auf Basis der beiden Merkmale p_mean und p_stddev.

In AGENDER wird eine ANN-Implementierung aus dem WEKA-Paket verwendet (vgl. Witten und Frank, 1999). Es handelt sich dabei um ein *multilayer Perceptron* (MLP) mit einer versteckten Ebene. Sämtliche Aktivierungsfunktionen sind sigmoid. Die Anzahl der versteckten Einheiten

entspricht

$$\frac{\text{Anzahl der Eingabeeinheiten} + \text{Anzahl der Ausgabeeinheiten} + 1}{2}. \tag{8.18}$$

Das Netz für Gruppierung 0 besitzt demnach 13 versteckte Einheiten, das Netz für Gruppierung 1 besitzt vier usw. In Abbildung 8.21 wird eine $jitt\_rap$ / $shim\_apq11$-Projektion der Entscheidungsgrenze eines ANNs für das Problem der Alterseinschätzung mit zwei Klassen dargestellt. Abbildung 8.22 stellt eine $pitch\_mean$ / $pitch\_stddev$-Projektion der Entscheidungsgrenze eines ANNs für das Problem der Geschlechtsklassifikation dar.

8.33	Gruppierung 0					Gesamtgenauigkeit 63.50 %		
	Kw	Km	Jw	Jm	Ew	Em	Sw	Sm
Kw	76.09	4.07	13.6	5.06	0.54	0.05	0.44	0.15
Km	54.25	12.37	12.52	15.51	1.13	0.25	3.78	0.2
Jw	54.15	2.41	27.44	13.16	1.28	0.1	1.37	0.1
Jm	20.08	3.98	6.33	59.25	1.03	1.13	4.96	3.24
Ew	0.25	0	0.2	0.54	84.73	3.44	6.92	3.93
Em	0	0	0	0.74	3.53	87.87	1.57	6.28
Sw	0.59	1.13	0.15	2.5	3.78	0.93	77.07	13.84
Sm	0	0.05	0	1.67	1.18	1.47	12.47	83.16

Bezüglich Gruppierung 0 verzeichnet das ANN mit 63.50 % die höchste Gesamtgenauigkeit aller verglichenen Klassifikationsmethoden. Das Niveau der Zweitbesten, SVM, ist bei 60.02 % angesiedelt. Dabei ist zu beachten, dass, im Gegensatz zu SVM und C 4.5, die True Positive Rate der Klasse männlichen Kinder (Km) in diesem Fall vergleichsweise gering ist. Wie bei den parametrischen Methoden, wird diese am häufigsten mit der Klasse Kw verwechselt, was damit begründet werden kann, dass die Stimmen von Kindern im vorpubertären Alter wenig geschlechtsspezifische Merkmale aufweisen.

8.34	Gruppierung 1			Gesamtg. 71.56 %
	KwKmJw	Ew	Sw	EmSm
KwKmJw	80.9	16.25	2.41	0.44
Ew	11.39	66.37	12.42	9.82
Sw	4.52	21.01	58.91	15.56
EmSm	0	2.16	17.77	80.07

Auch bezüglich Gruppierung 1 zeigt das ANN mit 71.56 % die beste Performanz, wobei die Unterschiede zwischen den einzelnen Methoden in dem Fall relativ gering sind: Die geringste Genauigkeit weist die Methode NB auf mit 69.01 %; C 4.5 ist mit 70.36 % die zweitbeste Methode. Allerdings weist das ANN auch in Bezug auf die Ausgeglichenheit der True Positive Rates einen leichten Vorteil auf: Die Standardabweichung liegt bei 10.74 %, während sie im Fall des C 4.5 bei 11 % angesiedelt ist.

8.35	Gruppierung 2		Gesamtg. 74.78 %	
	KwKmJw	JmEw	SwSm	Em
KwKmJw	83.6	14.73	1.37	0.29
JmEw	19.49	51.94	19.29	9.28
SwSm	2.85	9.82	78.5	8.84
Em	0	1.57	13.4	85.03

Einen deutlichen Vorsprung hat das ANN bezüglich Gruppierung 2: Die Gesamtgenauigkeit ist mit 74.98 % die beste vor dem zweitplatzierten KNN mit 70.59 %. Die Standardabweichung ist mit 15.47 % annähernd auf dem Niveau des KNNs (15.78 %).

8.36	Gruppierung 3			
	Stimme (94.61 %)		Sprechverhalten (67.37 %)	
	KwJwEwEm	SwSm	KwJwEwEm	SwSm
KwJwEwEm	92.24	7.76	59.65	40.35
SwSm	3.02	96.98	24.91	75.09

Die Gesamtgenauigkeit des ANN Stimmmerkmalsmodells bezüglich Gruppierung 3 fällt bleibt mit 94.61 % hinter der des KNNs mit 95.4 % zurück. Was das Sprechverhaltensmodell angeht, erreicht das ANN mit 67.37 % sogar nur Platz drei: sowohl KNN als auch C 4.5 verzeichnen mit 71 % bzw. 74.27 % bessere Ergebnisse.

8.37	Gruppierung 4	Gesamtg. 87.56 %	
	KwKmJwJmEw	SwSm	Em
KwKmJwJmEw	89.59	8.84	1.57
SwSm	7.56	85.76	6.68
Em	0.39	12.27	87.33

Auch bezüglich Gruppierung 4 ist das ANN mit 87.56 % nicht das Modell mit der höchsten Genauigkeit, da C 4.5 in dem Fall mit 88.08 % besser ist. Dabei ist jedoch zu beachten, dass das ANN ausgeglichenere True Positive Rates aufweist: Die Standardabweichung liegt hier bei 1.93 %, während sie beim C 4.5 bei 8.27 % angesiedelt ist.

8.38

	Gruppierung 5			
	Stimme (90.8 %)		Sprechverhalten (87.08 %)	
	KwKmJwJmSwSm	EwEm	KwKmJwJmSwSm	EwEm
KwKmJwJmSwSm	90.18	9.82	88.71	11.29
EwEm	8.59	91.41	14.56	85.44

Das ANN-Stimmmerkmalsmodell bezüglich Gruppierung 5 belegt mit einer Gesamtgenauigkeit von 90.8 % hinter dem KNN mit 91.52 % den zweiten Platz. Das drittbeste Modell ist der C 4.5 Entscheidungsbaum, der eine Genauigkeit von 80.06 % erreicht. Das Sprechverhaltensmodell liegt mit 87.08 % auf annähernd demselben Niveau wie das bestplatzierte C 4.5 (87.5 %), wobei die Ausgeglichenheit beim ANN höher ist: Die Standardabweichung liegt bei 2.31 %, während sie beim C 4.5 4.15 % beträgt ist.

8.39

	Geschlecht	Gesamtg. 93.14 %
	Ew	Em
Ew	90.63	9.37
Em	4.36	95.64

Was die Geschlechtsklassifikation betrifft, bleibt die Performanz des ANNs mit 93.14 % geringfügig hinter derjenigen der SVMs mit 93.27 % zurück. In Abbildung 8.23 wird die Genauigkeit der ANNs bezüglich der verschiedenen Gruppierungen nochmals in der Übersicht dargestellt. Die roten Balken markieren das jeweilige Zufallsniveau.

8.8 Kontextklassifikation

Die nachfolgende Tabelle stellt die Genauigkeit eines C 4.5 Entscheidungsbaumes angewandt auf das Problem der Kontextklassifikation dar. Die Gesamtgenauigkeit beträgt 93.77 %, wobei die True Positive Rate der Klasse QUIET mit 89.54 % am geringsten und die der Klasse NOISY mit 97.51 % am höchsten ist.

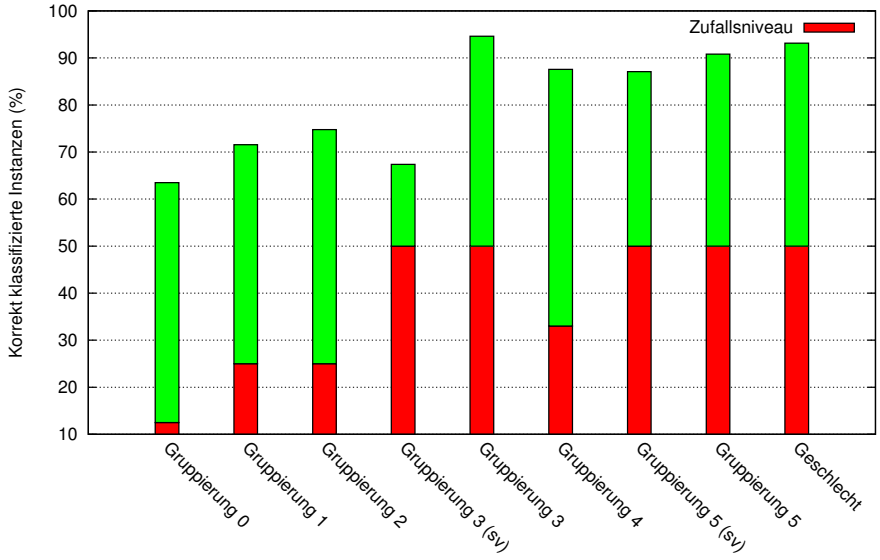


Abbildung 8.23: Übersicht über die Gesamtperformanz der ANNs bezüglich der verschiedenen Gruppierungen. Die roten Balken markieren das jeweilige Zufallsniveau.

8.40	Kontext	Gesamtg. 93.77 %	
	quiet	voicy	noisy
quiet	89.54	9.39	1.07
voicy	4.22	94.27	1.51
noisy	0.56	1.94	97.51

Es verbleibt zu prüfen, ob sich die Hypothese bestätigt, dass durch eine Hinzunahme eines Sprechkontextes die Vorteile der Stimmmerkmalsmodelle geringer werden, und sich evtl. die Verhältnisse sogar zugunsten der Sprechverhaltensmodelle verschieben. Dazu wurde eine Kreuzvalidierung der beiden C 4.5-Modelle bezüglich Gruppierung 5 auf Basis der künstlich erzeugten Kontext-Korpora durchgeführt. Die C 4.5 Methode wurde deshalb ausgewählt, da in diesem Fall das Sprechverhaltensmodell im Vergleich zu dem Stimmmerkmalsmodell das beste Ergebnis erzielte (vgl. Abschnitt 8.5.2).

Die Ergebnisse dieser Untersuchung werden in den folgenden drei Konfusionsmatrizen dargestellt. Die erste entspricht der in Abschnitt 8.5.2 dargestellten Tabelle, die auf Basis des Kontextes QUIET (neutrale Bedingung) ermittelt wurde. Die Grundlage der darauf folgenden Tabelle bilden

die Kontexte NOISY bzw. VOICY.

Insgesamt geht die Klassifikationsgenauigkeit bei Vorhandensein eines Kontextes zurück, und zwar stärker, wenn es sich um einen stimmenähnlichen (VOICY) Kontext handelt. Dieses Ergebnis ist nicht überraschend, da eine stärkere Beeinflussung der zugrunde liegenden Merkmale durch einen harmonischen Hintergrund zu erwarten war. Was die Unterschiede zwischen Stimmmerkmalsmodell und Sprechverhaltensmodell betrifft, ist zu beobachten, dass im Fall des lauten Kontextes (NOISY) die Performanz des ersteren stärker nachlässt als die des zweiten. Zumindest für diesen Fall wird demnach die oben genannte Hypothese bestätigt.

8.41	Gruppierung 5 (ruhiger Kontext)			
	Stimme (89.06 %)		Sprechverhalten (87.5 %)	
	KwKmJwJmSwSm	EwEm	KwKmJwJmSwSm	EwEm
KwKmJwJmSwSm	89.4	10.6	90.43	9.57
EwEm	11.29	88.71	15.44	84.56

8.42	Gruppierung 5 (lauter Kontext)			
	Stimme (79.38 %)		Sprechverhalten (83.7 %)	
	KwKmJwJmSwSm	EwEm	KwKmJwJmSwSm	EwEm
KwKmJwJmSwSm	81.36	18.64	86.78	13.22
EwEm	22.6	77.4	19.38	80.62

8.43	Gruppierung 5 (stimmenähnlicher Kontext)			
	Stimme (78.56 %)		Sprechverhalten (78.98 %)	
	KwKmJwJmSwSm	EwEm	KwKmJwJmSwSm	EwEm
KwKmJwJmSwSm	80.34	19.66	79.38	20.62
EwEm	23.22	76.78	21.41	78.59

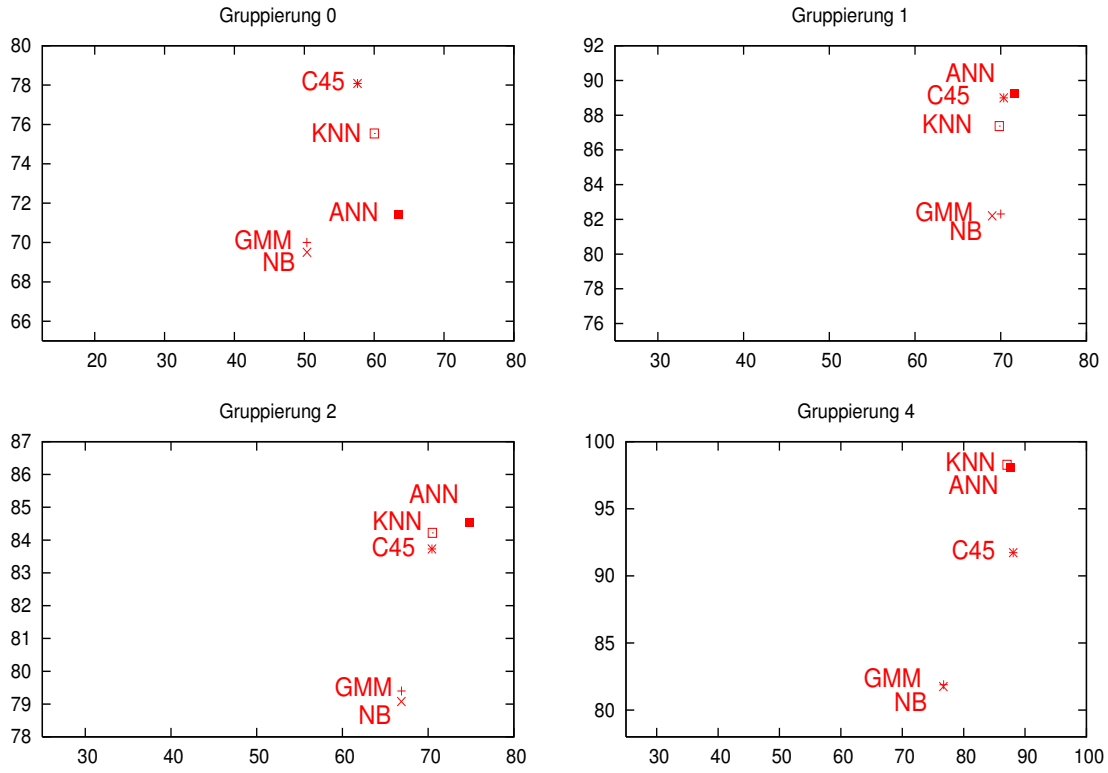


Abbildung 8.24: Übersicht über die Evaluationsergebnisse der Klassifikationsmethoden bezüglich der verschiedenen Gruppierungen (Teil 1). Die x-Achse entspricht der Klassifikationsgenauigkeit in %, wobei die Nullstelle das jeweilige Zufallsniveau markiert. Die y-Achse entspricht der Ausgeglichenheit der einzelnen True Positive Rates (100-Standardabweichung).

8.9 Übersicht über die Evaluationsergebnisse

In den Abbildungen 8.24 und 8.25 wird eine Übersicht über die Performanz der Klassifikationsmethoden bezüglich der einzelnen Gruppierungen dargestellt. Die x-Achse entspricht der Gesamtgenauigkeit in Prozent, wobei die Nullstelle das jeweilige Zufallsniveau markiert. Die y-Achse stellt die Ausgeglichenheit in Form des Maßes 100 - Standardabweichung dar.

Insgesamt sind die Ergebnisse äußerst zufriedenstellend: Bei allen Gruppierungen konnten Klassifikationsgenauigkeiten erreicht werden, die deutlich über das Zufallsniveau hinausgehen. Insbesondere die hohen True Positive Rates, die mithilfe der meisten Methoden bezüglich der Kontrollgruppierung 0 erreicht wurden, belegen eine prinzipielle Diskriminierbarkeit der Sprecherklassen auf Basis der untersuchten Merkmale. Bei der Unterscheidung der Senioren von den übrigen Altersklassen (Gruppierung 3) erreichte das bestplatzierte Modell – das KNN-

Stimmmerkmalsmodell – eine Trefferquote von 95.4 %. Was die Geschlechtserkennung betrifft, konnte mit einer SVM eine Genauigkeit von 93.27 % erreicht werden.

Gemessen an der hohen Gesamtperformanz bleiben die Leistungen der parametrischen Methoden hinter den Erwartungen zurück. Dabei ist besonders auffällig, dass GMM in allen Fällen annähernd dieselben bzw. nur geringfügig bessere Ergebnisse aufweist als NB. Möglicherweise ist dies auf die Tatsache zurückzuführen, dass die Gewichte nicht mithilfe des EM-Algorithmus gelernt, sondern durch eine Kreuzvalidierung ermittelt wurden. Allerdings entspricht es den Erwartungen, dass die Performanz der parametrischen Methoden bezüglich Gruppierung 0 am geringsten ist. Die Hypothese lautete, dass diese Methoden den größten Nutzen aus einer Top-Down-Selektion der Merkmale ziehen würden. Dies wird bestätigt durch die Ergebnisse von KNN und ANN, die zwar in fast allen Gruppierungen die vorderen Plätze einnehmen, jedoch besonders in Gruppierung 0 einen deutlichen Vorsprung aufweisen.

Die Genauigkeit der Stimmmerkmalsmodelle ist wie erwartet höher als die der entsprechenden Sprechverhaltensmodelle. Während die Performanz bezüglich Gruppierung 5 jedoch zumindest auf einem ähnlich hohen Niveau bleibt, sind bezüglich Gruppierung 3 gravierende Performanzeinbußen zu verzeichnen. Darüber hinaus ist festzustellen, dass in beiden Fällen jeweils der C 4.5 Entscheidungsbaum die besten Leistungen zeigt. Unter Hinzunahme eines Sprechkontextes geht die Klassifikationsperformanz insgesamt zurück. Im Fall von NOISY verzeichnet jedoch das Sprechverhaltensmodell einen leichten Vorteil gegenüber dem Stimmmerkmalsmodell.

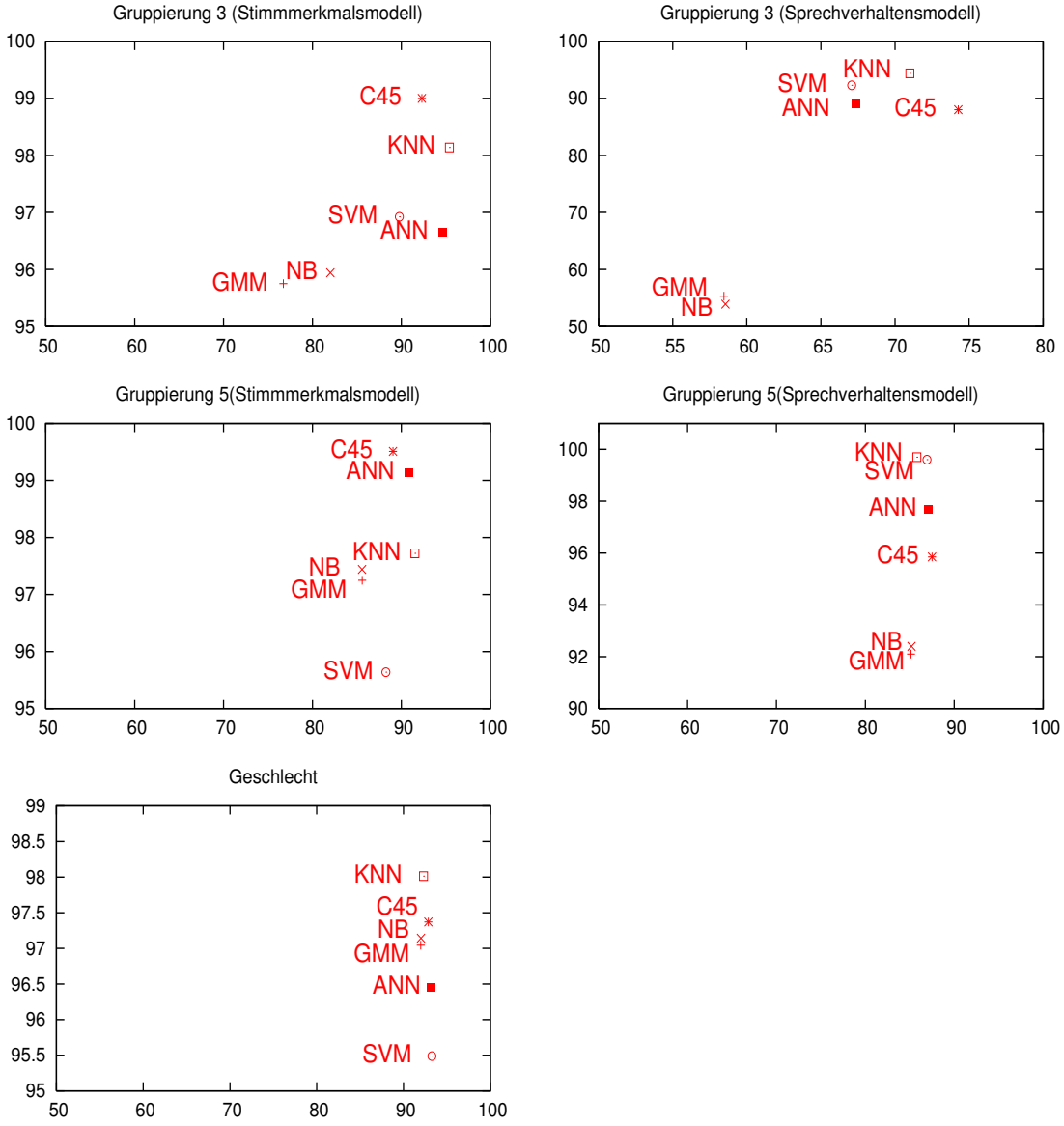


Abbildung 8.25: Übersicht über die Evaluationsergebnisse der Klassifikationsmethoden bezüglich der verschiedenen Gruppierungen (Teil zwei). Die x-Achse entspricht der Klassifikationsgenauigkeit in %, wobei die Nullstelle das jeweilige Zufallsniveau markiert. Die y-Achse entspricht der Ausgeglichenheit der einzelnen True Positive Rates (100-Standardabweichung).

9 Einbeziehung von Top-Down-Wissen und Behandlung der inhärenten Unsicherheit mithilfe Dynamischer Bayes'scher Netze (Zweite Ebene)

Nach Duda et al. (2000, S. 13) ist die Nachverarbeitung (*Post-Processing*) ein wesentlicher Bestandteil eines Mustererkennungssystems. Ihre Notwendigkeit ergibt sich daraus, dass auf Basis der Klassifikationsergebnisse in der Regel Handlungen ausgelöst oder empfohlen werden, was eine Reihe von Problemen mit sich bringt. Erstens gibt es unter Umständen mehrere Klassifizierer, deren Ergebnisse kombiniert werden müssen. Zweitens ist es oftmals sinnvoll, Expertenwissen – oder allgemeiner: *Top-Down-Wissen* – in den Entscheidungsprozess mit einfließen zu lassen. Dieser Begriff beschreibt diejenigen Informationen, die nicht aus den (Sensor-)Daten abgeleitet werden (*Bottom-Up-Wissen*), sondern sich aus der Kenntnis von kausalen Zusammenhängen innerhalb der Domäne ergeben. Zum Top-Down-Wissen gehört der *Kontext*, d. h. eingabeunabhängige Informationen.

Das Problem der Fusion mehrerer Klassifikationsergebnisse umfasst in AGENDER einen statischen und einen dynamischen Aspekt: Der statische Aspekt betrifft die Kombination mehrerer Ergebnisse derselben Äußerungen, die dadurch zustande kommen, dass es verschiedene Gruppierungen von Klassen mit jeweils anderen Merkmalsets gibt. Der dynamische Aspekt besteht darin, dass unter Umständen mehrere Äußerungen desselben Sprechers betrachtet werden. Zur Lösung sowohl des statistischen als auch des dynamischen Aspektes wird in AGENDER ein *Dynamisches Bayes'sches Netz* (DBN) angewendet. Der Vorteil von DBNs besteht darin, dass sie darüber hinaus einen Mechanismus darstellen, der geeignet ist, um die der Klassifikation inhärente Unsicherheit explizit zu modellieren und dabei Top-Down-Wissen zu berücksichtigen. Die Nachverarbeitungsphase wird in der vorliegenden Arbeit *Zweite Ebene* (*Second Layer*) genannt, wobei die *Erste Ebene* (*First Layer*) durch die Klassifikation einschließlich der Merkmalsextraktion gebildet wird.

9.1 Grundlagen Dynamischer Bayes'scher Netze

Ein Domänenexperte, wie z. B. ein Formel-1-Ingenieur, ist in der Lage, die kausalen Zusammenhänge zu bestimmen, die zwischen Konzepten (Variablen) innerhalb der Domäne bestehen. Der

Ingenieur hat Zugriff auf die Zustände einer Reihe von Sensoren am Rennfahrzeug, wie Motortemperatur, Bremsflüssigkeitsdruck, Reifendruck und Spannungen im elektrischen System – wobei die Arbeit mit einem Formel-1-Rennwagen zweifellos bei Weitem perfidere Messungen erfordert als diese. Bei der Interpretation der Messungen berücksichtigt er beispielsweise, dass zwischen der Temperatur des Motors und der Öltemperatur ein kausaler Zusammenhang besteht, während dies zwischen dem Öldruck im Motor und dem Luftdruck in den Reifen nicht der Fall ist. Darüber hinaus macht er von dem Wissen Gebrauch, dass einige Variablen von mehreren anderen beeinflusst werden: Die Temperatur des Kühlers ist z. B. abhängig von der Motortemperatur und von der Fahrgeschwindigkeit, da bei höheren Geschwindigkeiten die herbeigeführte Luft eine stärkere Kühlwirkung entwickelt.

Derartig strukturelles Wissen über eine Domäne wird als *theoriegesteuertes Wissen* oder *Top-Down-Wissen* bezeichnet und damit von dem *datengesteuerten Wissen* (*Bottom-Up-Wissen*) abgegrenzt, welches durch die in Kapitel 7 beschriebenen Prozesse der Mustererkennung abgeleitet wird.

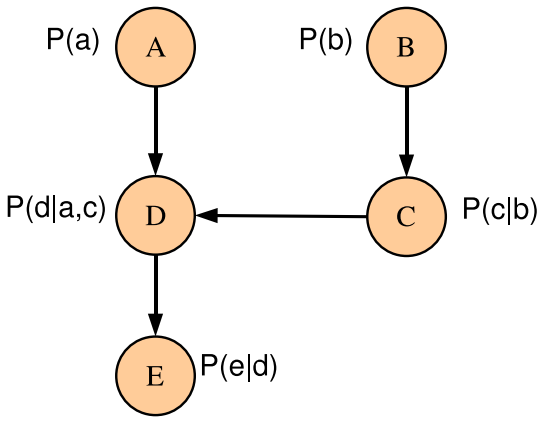


Abbildung 9.1: Beispiel für ein Bayes'sches Netz nach Duda et al. (2000, S. 57).

Zur Modellierung der kausalen Zusammenhängen, die in einer Domäne bestehen, werden häufig *Bayes'sche Netze* (BNs) verwendet (vgl. Duda et al., 2000, S. 56). Die Knoten eines BN repräsentieren Systemvariablen, die diskrete Werte annehmen können. Abbildung 9.1 stellt ein solches Netz dar: Die Knoten werden mit **A**, **B** usw. und die dazugehörenden Variablen mit den entsprechenden Kleinbuchstaben bezeichnet. Die Variablen können eine diskrete Anzahl von Zuständen annehmen, z. B. a_1 und a_2, und für jeden der Zustände gilt eine reellwertige Wahrscheinlichkeit, z. B. $P(a_1) = 0.739$, $P(a_2) = 0.261$. Die Kanten eines Bayes'schen Netzes sind gerichtet und verbinden jeweils zwei Knoten miteinander. Sie repräsentieren den kausalen Einfluss eines Knotens auf einen anderen. In Abbildung 9.1 wird **D** von **A** direkt und von **B** indirekt durch **C** beeinflusst. Bezüglich eines einzelnen Knotens bezeichnen wir die Menge der direkten Vorgänger als *Elternknoten* und die Menge der direkten Nachfolger als *Kinderknoten*. In Abbildung 9.1 sind **A** und **C** die Eltern und **E** ein Kind von **D**. In einem BN kann die Wahrscheinlichkeit aller Konfigurationen von Variablen durch direkte Anwendung der Bayes'schen Formel (9.1) bestimmt

werden.

$$P(a_i|x) = \frac{p(x|a_i)P(a_i)}{p(x)} \tag{9.1}$$

Die Bayes'sche Formel drückt aus, dass durch Beobachtung des Wertes von x die A-priori-Wahrscheinlichkeit $P(a_i)$ in eine A-posteriori-Wahrscheinlichkeit $P(a_i|x)$ umgewandelt werden kann. Diese drückt die Wahrscheinlichkeit aus, dass die Variable a sich in dem Zustand i befindet unter der Bedingung, dass der Wert von x festgestellt wurde (vgl. Duda et al., 2000, S. 22). $P(a_i|x)$ wird häufig *Likelihood* von a_i genannt.

Um die Likelihood der Knoten zu berechnen, werden die *Tabellen der bedingten Wahrscheinlichkeiten* (*conditional probability tables,* CPT*s*) benötigt, die Wahrscheinlichkeiten aller Zustände einer Variablen für alle Bedingungen angeben. Die Bedingungen werden durch die Zustände der Variablen an den Elternknoten repräsentiert. Die Summe einer CPT-Spalte ist gleich eins. Hat ein Knoten keine Elternknoten, enthält die CPT lediglich die A-priori-Wahrscheinlichkeiten der Variablen.

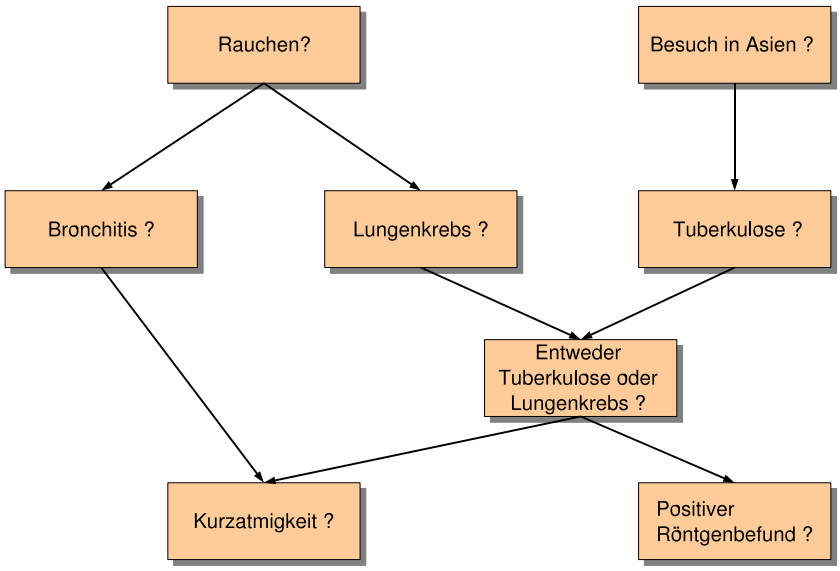


Abbildung 9.2: Asien-Netzwerk.

Abbildung 9.2 stellt ein klassisches Beispiel eines Bayes'schen Netzes nach Cowell, Dawid, Lauritzen und Spiegelhalter (1999, S. 20) dar. Die dazugehörige CPT entspricht Tabelle 9.1. Das Netz drückt den folgenden medizinische Sachverhalt aus: KURZATMIGKEIT kann durch TUBERKULOSE, LUNGENKREBS, BRONCHITIS, eine Kombination der genannten Krankheiten oder andere Ursachen ausgelöst werden. Hat der Patient kürzlich eine Reise nach Asien unternommen, steigt die Wahrscheinlichkeit der TUBERKULOSE als Ursache (in einer aktuelleren Variante des Beispiels wäre vermutlich SARS an die Stelle der Tuberkulose getreten), während RAUCHEN

als Risiko-Faktor für LUNGENKREBS und BRONCHITIS bekannt ist. Eine einfache RÖNTGEN-Untersuchung oder die Beobachtung von KURZATMIGKEIT tragen nicht zur Unterscheidung zwischen LUNGENKREBS und TUBERKULOSE bei. Wenn wir erfahren, dass der Patient RAUCHER ist, werden die Wahrscheinlichkeiten von LUNGENKREBS und BRONCHITIS angepasst (erhöht). Die Wahrscheinlichkeit von TUBERKULOSE bleibt allerdings unverändert, d. h. TUBERKULOSE und RAUCHEN sind bedingt unabhängig voneinander. Ein positiver RÖNTGEN-Befund verändert die Wahrscheinlichkeit von TUBERKULOSE und LUNGENKREBS, nicht jedoch die von BRONCHITIS (d. h. BRONCHITIS ist bedingt unabhängig von RÖNTGEN gegeben RAUCHEN). Falls wir jedoch zusätzlich wüssten, dass der Patient unter KURZATMIGKEIT leidet, hätte das RÖNTGEN-Ergebnis auch die Wahrscheinlichkeit von BRONCHITIS beeinflusst (BRONCHITIS ist nicht bedingt unabhängig von RÖNTGEN gegeben RAUCHEN und KURZATMIGKEIT).

A:	$p(a)$	=	0.01	L:	$p(l\|s)$	=	0.1
					$p(l\|\bar{s})$	=	0.01
B:	$p(b\|s)$	=	0.6	S:	$p(s)$	=	0.5
	$p(b\|\bar{s})$	=	0.3				
D:	$p(d\|b,e)$	=	0.9	T:	$p(t\|a)$	=	0.05
	$p(d\|\bar{b},e)$	=	0.7	T:	$p(t\|\bar{a})$	=	0.01
	$p(d\|b,\bar{e})$	=	0.8				
	$p(d\|\bar{b},\bar{e})$	=	0.1				
E:	$p(e\|l,t)$	=	1	X:	$p(x\|e)$	=	0.98
	$p(e\|\bar{l},t)$	=	1		$p(x\|\bar{e})$	=	0.05
	$p(e\|l,\bar{t}$	=	1				
	$p(e\|\bar{l},\bar{t}$	=	0				

Tabelle 9.1: CPT (conditional probability table) des „Asien“-Beispiels aus Abbildung 9.2.

Dieses Beispiel verdeutlicht, dass bereits ein Bayes'sches Netz mit wenigen Knoten einen durchaus komplexen Zusammenhang modellieren kann. Tatsächlich erfreuen sich Bayes'sche Netze vor allem als Inferenzmechanismus in den so genannten *Expertensystemen* einer weiten Verbreitung. Jensen (1997, S. 4–6) führt einige Beispiele für Expertensysteme auf der Basis von Bayes'schen Netzen aus den unterschiedlichsten Bereichen auf: ein System zur Eindämmung von Mehltau bei Winterweizen (Landwirtschaft), Diagnose-Assistenten für neuromuskuläre Krankheiten (Medizin), ein Wettervorhersage-System und ein System zur Interpretation von Gedichten.

Es stellt sich die Frage nach der Herkunft der bedingten Wahrscheinlichkeiten. Nach Russel und Norvig (1995, S. 430) gibt es drei konkurrierende Sichtweisen: Die *Frequentivisten* (von engl.

frequent=häufig) ermitteln die Zahlen ausschließlich mithilfe von Experimenten: Wenn einhundert Personen untersucht werden und zehn davon besitzen eine bestimmte Eigenschaft, dann kann diese Eigenschaft mit einer Wahrscheinlichkeit von 0.1 angenommen werden. Die Schwierigkeit besteht darin, dafür zu sorgen, dass diese Feststellungen verlässlich sind. Die Sichtweise der *Objektivisten* ist die, dass die Wahrscheinlichkeiten reelle Aspekte der Welt sind, statt nur eine Beschreibung der Annahmen des Betrachters. In diesem Sinne stellen die Messungen der Frequentivisten Versuche dar, diese reellen Wahrscheinlichkeitswerte zu messen. Die *Subjektivisten* schließlich beschreiben die Wahrscheinlichkeiten ohne jegliche externe physikalische Referenz. Stattdessen legt ein Experte die Zahlen auf Basis seiner eigenen Einschätzung fest.

Die Art und Weise, in der die CPTs für die Zweite Ebene des AGENDER-Ansatzes ermittelt werden, entspricht teilweise der frequentivistischen Sichtweise und teilweise der subjektivistischen. Zur Modellierung der klassifikationsinhärenten Unsicherheit beispielsweise werden Wahrscheinlichkeiten verwendet, die aus der Evaluation des jeweiligen Klassifikators abgeleitet werden (vgl. Abschnitt 9.3.1), was eine statistische Information und somit frequentivistisch ist. Für die Modellierung einiger Top-Down-Aspekte, die sich auf die Progressivität der Inferenzen auswirkt (vgl. Abschnitt 9.3.2), werden die CPT-Einträge dagegen vom Designer eingeschätzt. Allerdings handelt es sich dabei um weniger zentrale Aspekte der Zweiten Ebene.

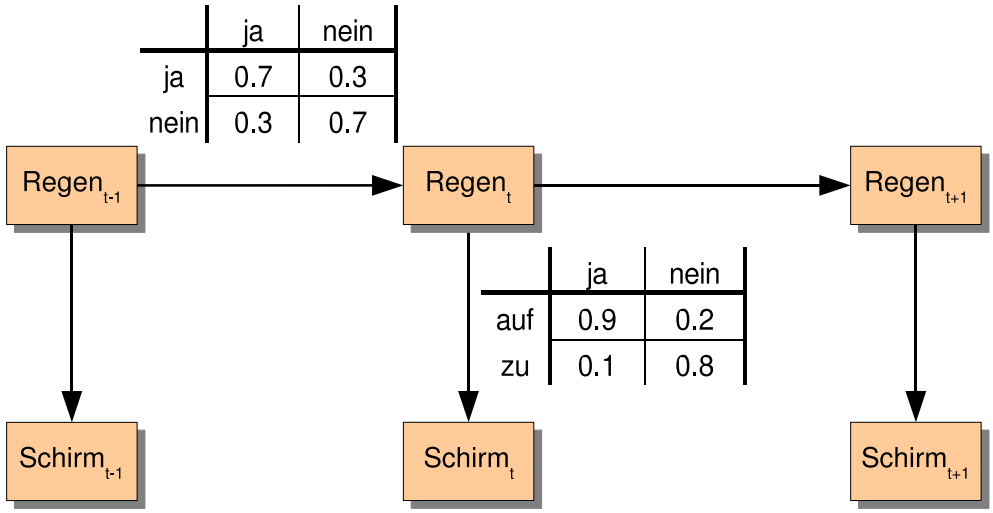


Abbildung 9.3: Beispiel eines Dynamischen Bayes'schen Netzes mit drei Zeitscheiben nach Russel und Norvig (2003).

Abbildung 9.3 stellt ein Beispiel für ein *Dynamisches Bayes'sches Netz* (DBN) dar. Mittels DBNs können kausale *Prozesse* modelliert werden: Der Zusammenhang zwischen der Tatsache, dass es regnet, und dem Erscheinen von Regenschirmen auf der Promenade ist *statisch*. Die Wahrscheinlichkeit, dass es regnet in Abhängigkeit von der Tatsache, ob es bereits am Tag zuvor geregnet hat, beinhaltet jedoch eine zeitliche Komponente und wird deshalb als *dynamisch* bezeichnet. Ein DBN setzt sich aus mehreren BN zusammen, wobei jedes einzelne davon eine so genannte *Zeitscheibe* darstellt. Die Auflösung der Zeitscheiben hängt von der Domäne ab: In dem Beispiel aus

Abbildung 9.3 könnten die Zeitpunkte $t-1, t$ und $t+1$ auch Stunden oder Minuten denotieren. In einem DBN gibt es feste Übergänge zwischen den Zeitscheiben. Sie besitzen dieselben Eigenschaften wie die Kanten eines BN. Wie in Abbildung 9.3 bereits angedeutet, wird der Einfluss einer Zeitscheibe auf die nächste durch die CPTs der Verbindungsknoten repräsentiert.

9.2 Bayes'sche Netze in verwandten Arbeiten

Garg, Pavlovic und Rehg (2000) beschreiben ein System zur audio-visuellen Erkennung von Sprechaktivität, bei dem sensorische Informationen mithilfe eines *Dynamischen Bayes'schen Netzes* kombiniert werden. Bei dem zugrunde liegenden Szenario handelt es sich um ein öffentlich zugängliches Informationsterminal (Kiosk), mit welchem die Benutzer per Sprache interagieren können. Mithilfe einer Kamera (visuelle Eingabe) und eines Mikrofons (auditive Eingabe) soll das System entscheiden, ob zu einem gegebenen Zeitpunkt ein Sprecher eine Anfrage (Äußerung) an das Terminal stellt oder nicht. Bei der Beispielanwendung handelt es sich um ein „Blackjack"-Spiel, bei dem die Spieler (Sprecher) einfache Kommandos an den „Groupier" (System) richten können.

Garg et al. führen Gründe für die Verwendung eines DBN als Inferenzmechanismus auf, die auch die Einführung einer Zweiten Ebene in AGENDER motivieren, nämlich dass Bayes'sche Netze 1. auf elegante Art und Weise eine datengesteuerte Modellierung mit einer – wie sie es nennen – *expertengesteuerten* Modellierung verbinden, und 2. einen guten Mechanismus darstellen, um verschiedene „Sensoren" miteinander zu kombinieren. Die Menge der Sensoren umfasst in diesem Fall: einen Gesichtserkenner, einen Hautfarbenerkenner, einen Gesichtstexturerkenner, einen Lippenbewegungserkenner und einen Stille-Erkenner.

In Abbildung 9.4 wird die Topologie einer Zeitscheibe des verwendeten DBNs dargestellt. Die Knoten auf der untersten Ebene repräsentieren die Sensoren (beobachtbare Knoten), die jeweils zwei Zustände (ja/nein) besitzen. Der Gesamtgraph setzt sich aus einem visuellen Teilgraphen und einem auditiven Teilgraphen zusammen. Durch Rückwärtspropagierung wird die Likelihood für das Vorhandensein eines Sprechers ermittelt. Außer den Sensor-Knoten wird hierfür noch der Knoten KONTEXT instanziiert. Er repräsentiert den Dialog-Kontext, auf Basis dessen die Erwartung einer Äußerung von Seiten des Benutzers bestimmt wird. Beim „Blackjack" wird eine Äußerung genau dann erwartet, wenn der Sprecher an der Reihe ist.

In diesem DBN stellen alle Knoten Verbindungsknoten zur nächsten Zeitscheibe dar, was Garg et al. mit dem fehlenden Wissen um die exakten temporalen Beziehungen in der Domäne begründen. Dagegen profitiert das System von der Tatsache, dass durch die Beschränkung auf das „Blackjack"-Spiel die kontextuellen Rahmenbedingungen wohl definiert sind, wie z. B. die Länge der zulässigen Kommandos. Die Information aus Zeitscheiben, die älter sind als der aktuelle Dialogschritt können daher abgeschnitten werden. Die Klassifikationsgenauigkeit wird von den Autoren auf 85 % beziffert. Allerdings stammt dieser Wert aus einem Experiment, das mit Daten von nur einem Sprecher ausgeführt wurde.

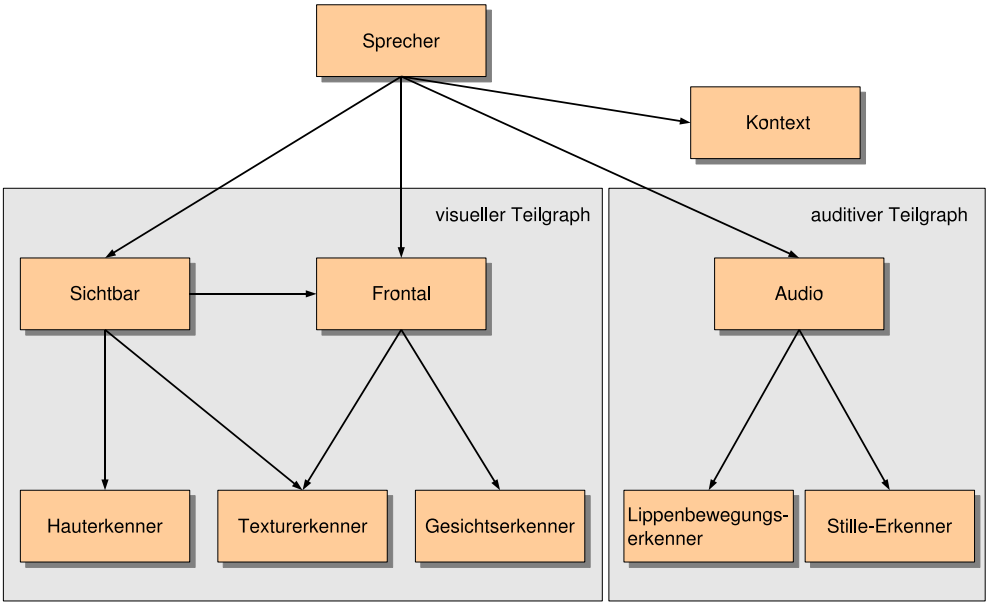


Abbildung 9.4: Zeitscheibe eines Dynamischen Bayes'sches Netzes zur audio-visuellen Erkennung von Sprechaktivität nach Garg et al. (2000).

9.3 Bayes'sche Netze in AGENDER

Mithilfe von DBNs können in AGENDER eine Reihe von Problemen, die im Rahmen eines Mustererkennungsansatzes auftreten, auf elegante Art und Weise gelöst werden. Dazu gehört die explizite Modellierung der klassifikationsinhärenten Unsicherheit, die Einbeziehung von domänenspezifischem Top-Down-Wissen und die Fusion mehrer Klassifikationsergebnisse – sowohl die statische als auch die dynamische Variante.

9.3.1 Explizite Modellierung der klassifikationsinhärenten Unsicherheit

Bei einem nicht-trivialen Klassifikationsproblem ist die Entscheidung des Klassifizierers stets mit Unsicherheit behaftet, d. h. wenn sich der Klassifizierer für eine bestimmte Klasse ω_i entscheidet, kann diese nur mit einer bestimmten Wahrscheinlichkeit $P(\omega_i) < 1$ angenommen werden. Die tatsächliche Unsicherheit ist nicht bekannt, jedoch stellt das Ergebnis der Kreuzvalidierung eine plausible Annäherung dar: Wenn diese beispielsweise eine *True Positive Rate* von 0.92 für ω_i ergeben hat, dann gilt: $P(\omega_i) = 0.92$.

Die Unsicherheit wird nun wie folgt auf der Zweiten Ebene modelliert: Die Entscheidung des Klassifizierers wird als eine Beobachtung angesehen, die Rückschlüsse auf die tatsächlichen Sprecher- oder Kontexteigenschaften zulässt. Der kausale Zusammenhang entspricht dem von Krankheit und Symptom: Es ist die Krankheit, die das Symptom auslöst, obwohl von der Beobachtung des Symptoms auf die Krankheit geschlossen wird. In Abbildung 9.5 wird dieser Zu-

sammenhang anhand eines vereinfachten Beispiels dargestellt.

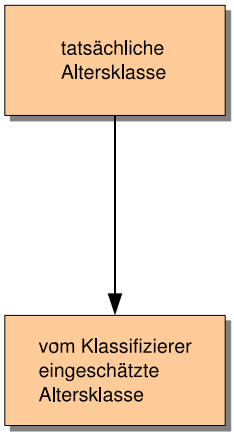
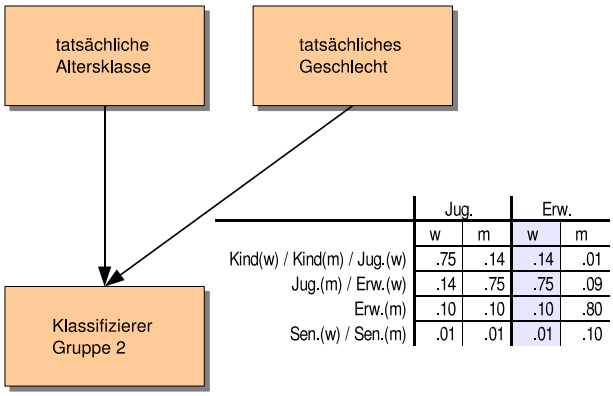


Abbildung 9.5: Kausaler Zusammenhang zwischen tatsächlicher Sprechereigenschaft und Ergebnis des Klassifizierers.

Die Unsicherheit wird dabei durch die CPTs ausgedrückt wie in Abbildung 9.6 anhand des Beispiels eines *Gruppe-2-Klassifizierer* dargestellt wird, der vier Klassen unterscheidet: 1. KWKMJW, 2. JMEW, 3. EM und 4. SWSM. Die dargestellte CPT modelliert eine Situation, in der die Kreuzvalidierung ergeben hat, dass die jüngeren, erwachsenen Test-Sprecherinnen in 75 Prozent der Fälle korrekt klassifiziert worden sind, in 14 Prozent der Fälle fälschlicherweise der Klasse KWKMJW zugeordnet wurde, in 10 Prozent der Fälle der Klasse EM und in einem Prozent der Fälle der Klasse SWSM. Wenn nun der Klassifizierer beispielsweise das Ergebnis JMEW ausgibt, dann werden auch die Fälle der möglichen Falschklassifikation in der Likelihood von ALTERSKLASSE und GESCHLECHT berücksichtigt. Die CPT in Abbildung 9.6 wurde aus Gründen der Übersichtlichkeit verkürzt dargestellt: Es fehlen die Einträge für die Altersklassen KINDER und SENIOREN, die analog zu dem beschriebenen Fall ermittelt werden.



	Jug.		Erw.	
	w	m	w	m
Kind(w) / Kind(m) / Jug.(w)	.75	.14	.14	.01
Jug.(m) / Erw.(w)	.14	.75	.75	.09
Erw.(m)	.10	.10	.10	.80
Sen.(w) / Sen.(m)	.01	.01	.01	.10

Abbildung 9.6: Beispiel für die Modellierung der Unsicherheit eines Klassifizierers durch entsprechende CPT-Einträge.

9.3.2 Einbeziehung von domänenspezifischem Top-Down-Wissen

Wie die Beispiele aus Abschnitt 9.1 deutlich machen, kann Top-Down-Wissen mithilfe von DBNs unmittelbar repräsentiert werden. In dem zweistufigen Mustererkennungsansatz, wie er in der vorliegenden Arbeit vorgeschlagen wird, bei dem DBNs die Zweite Ebene bilden, kann diese Eigenschaft genutzt werden. Zunächst einmal können die A-priori-Wahrscheinlichkeiten der Klassen auf einfache Art und Weise repräsentiert werden, indem diese direkt in die CPTs der entsprechenden Knoten eingetragen werden. Zwei weniger offensichtliche Beispiele werden im Folgenden dargestellt: die *Kosten einer Falschklassifikation* und die *Einbeziehung des Kontextes*.

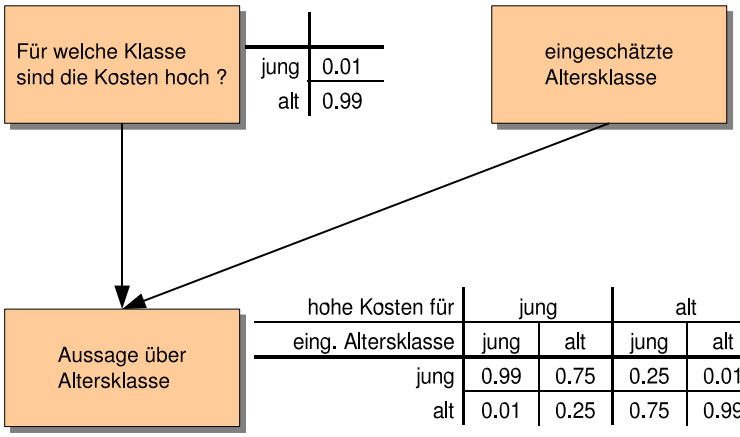
Kosten einer Falschklassifikation

Duda et al. (2000, S. 3) weisen darauf hin, dass die Kosten einer Falschklassifikation oftmals nicht für alle Klassen gleich sind, und illustrieren dies anhand des Fischfabrik-Beispiels: Wenn der Kunde einen Seebarsch kauft und stattdessen einen Lachs in der Packung vorfindet, ist dies weniger dramatisch als umgekehrt. Diese Kostenverteilung ist im höchsten Maße domänenabhängig: Ohne sich auf eine spezifische Anwendung zu beziehen, kann nicht entschieden werden, ob es „teurer" ist, einen älteren Menschen als Jugendlichen einzuschätzen oder umgekehrt. Es ist daher von Vorteil, diese Information explizit bei der Konfiguration des Systems angeben zu können, ohne dass die Modelle neu trainiert werden müssen. In Abbildung 9.7 wird dargestellt, wie variable Kosten auf der Zweiten Ebene berücksichtigt werden können. In diesem Beispiel wurde die Einschätzung des Sprecheralters auf zwei Klassen JUNG und ALT beschränkt. Es wird der Fall dargestellt, bei dem höhere Kosten verursacht werden, wenn ein älterer Sprecher fälschlicherweise als jung eingeschätzt wird, als umgekehrt. Die CPT-Einträge des Knotens AUSSAGE ÜBER ALTERSKLASSE sind so gewählt, dass dann, wenn der Klassifizierer den Sprecher als jung einschätzt, die Aussage konservativer ist, d. h. die Likelihood für JUNG ist geringer (72.2 %). Dagegen ist die Aussage über die Altersklasse im Fall von ALT mit einer Likelihood von 95.4 % progressiver. Die Likelihood-Werte werden durch Anwendung der Bayes'schen Formel (Gleichung 9.1, Seite 183) auf die in Abbildung 9.7 dargestellten Wahrscheinlichkeiten ermittelt.

Einbeziehung des Kontextes

In Abschnitt 1.5.1 wurde ein Ebenemodell der Merkmale vorgeschlagen, das zwischen stimmlichen Charakteristika (z. B. Stimmtonhöhe, Stimmqualität) auf der unteren Ebene und dem Sprechverhalten (z. B. Äußerungsgeschwindigkeit) auf der oberen Ebene unterscheidet. Darüber hinaus wurde das folgende trade-off formuliert: Die Merkmale der niedrigeren Ebene erlauben unter neutralen Bedingungen bessere Vorhersagen der Sprecherklassen, sind jedoch anfälliger gegenüber Veränderungen des (Sprech-)Kontextes. Bei den Merkmalen höherer Ebene dagegen ist es umgekehrt. Als *Kontext* wird dabei diejenige akustische Information bezeichnet, die nicht unmittelbar von der Äußerung des Sprechers abhängt, sondern von der Umgebung (Hintergrundgeräusche).

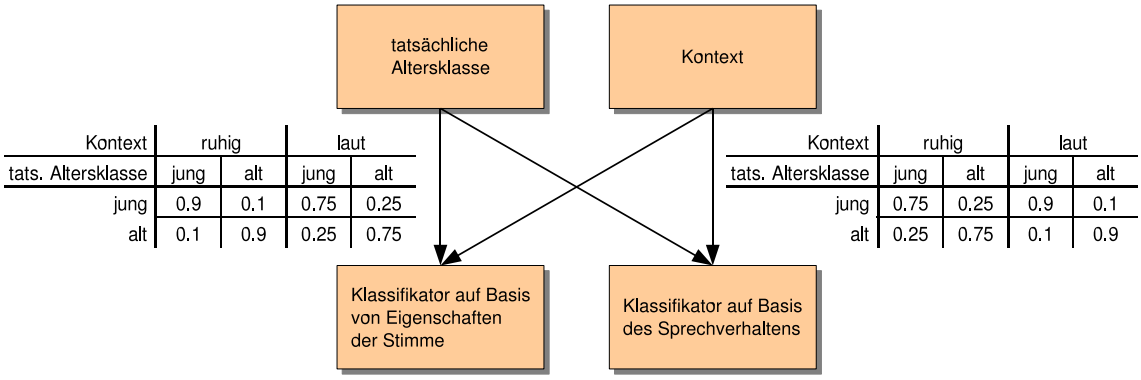
Das Beispiel in Abbildung 9.8 macht deutlich, wie mithilfe der Zweiten Ebene ein solches trade-off gelöst werden kann, wobei vereinfachend von jeweils zwei Sprecher- und Kontextklas-



hohe Kosten für	jung		alt	
eing. Altersklasse	jung	alt	jung	alt
jung	0.99	0.75	0.25	0.01
alt	0.01	0.25	0.75	0.99

Abbildung 9.7: Mehr oder weniger progressive Inferenz aufgrund variabler Kosten einer Falschklassifikation.

sen ausgegangen wird. Die Struktur des Netzwerkes wurde so angepasst, dass die Ergebnisse der Klassifizierer nicht nur von der tatsächlichen Sprechereigenschaft, sondern darüber hinaus vom Kontext abhängig sind. Die CPTs drücken aus, dass bei einem ruhigen Kontext die Stimmmerkmalsmodelle mit einer höheren Wahrscheinlichkeit die richtige Klasse ausgibt. Bei einem lauten Kontext dagegen wird das Ergebnis Sprechverhaltensmodells als verlässlicher eingeschätzt. Je nach Netzwerkstruktur ist es unter Umständen sinnvoll, die CPTs der Modelle so anzupassen, dass die Aussagen über die Altersklasse bei denjenigen Kontexten, bei denen sie weniger verlässlich sind, konservativer zu gestalten, als die Evaluationsergebnisse es zulassen. Dies ist immer dann der Fall, wenn vermieden werden soll, dass sowohl dem Stimmmerkmals- als auch dem Sprechverhaltensmodell ein Gewicht bei der Kombination mit anderen Klassifizierern beigemessen wird. In Kapitel 10.1 wird ein Beispiel aufgeführt, bei dem aus besagtem Grund die betreffenden Spalten der CPTs auf 50:50 gesetzt wurden.



Kontext	ruhig		laut	
tats. Altersklasse	jung	alt	jung	alt
jung	0.9	0.1	0.75	0.25
alt	0.1	0.9	0.25	0.75

Kontext	ruhig		laut	
tats. Altersklasse	jung	alt	jung	alt
jung	0.75	0.25	0.9	0.1
alt	0.25	0.75	0.1	0.9

Abbildung 9.8: Kontextabhängige Gewichtung von Merkmalen.

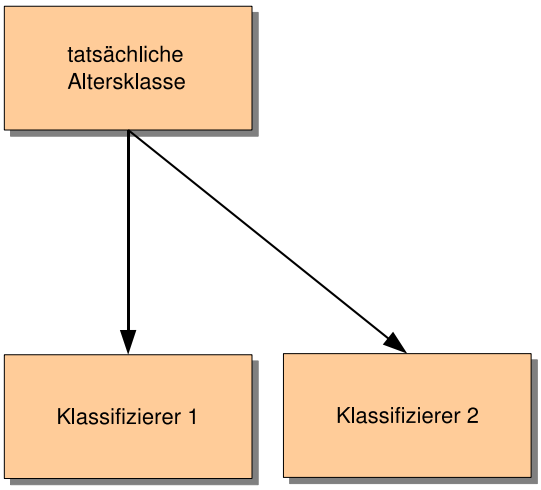


Abbildung 9.9: Fusion mehrerer Klassifikationsergebnisse, die dieselbe Äußerung betreffen (statische Variante).

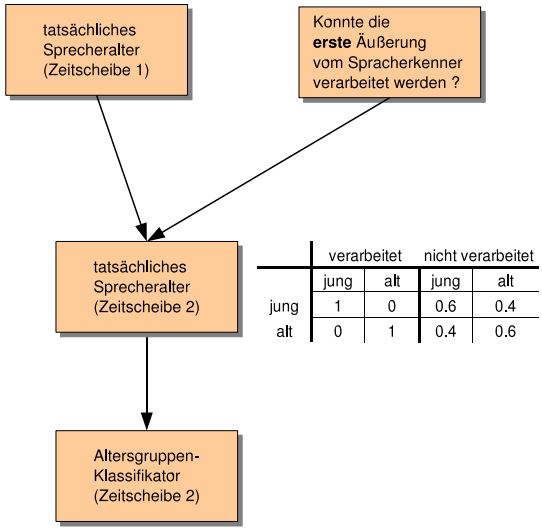


Abbildung 9.10: Ein Beispiel für die variable Kombination von Einschätzungen auf Basis mehrerer Äußerungen.

9.3.3 Fusion mehrerer Klassifikationsergebnisse

Die Fusion mehrerer Klassifikationsergebnisse beschreibt ein Problem, das nach Huang und Hsu (2002) als Klassifikation einer *homologen Menge* bezeichnet werden kann: Sei $\mathbf{X} = \{x_1, \ldots, x_n\}$ eine Menge von Merkmalsvektoren, die zur selben, unbekannte Klasse c gehören. Jeder der Vektoren x_t habe r Merkmale $(x_{t1}, \ldots, x_{tr})$. O denotiert die Tatsache, dass alle Merkmale zur Klasse c gehören.

Nach der Bayes'schen Entscheidungstheorie (siehe Abschnitt 7.2) sollte die Menge der Merkmalsvektoren so klassifiziert werden, dass $p(c|\mathbf{X}, O)$ maximiert wird. Nach Huang und Hsu gibt es zur Lösung dieses Problems drei Alternativen: die *Abstimmungsmethode*, die *Durchschnittsmethode* und die *Maximummethode*. Bei der Abstimmungsmethode werden alle Elemente unabhängig voneinander klassifiziert. Das Ergebnis ist die am Häufigsten vorkommende Klasse. Die Durchschnittsmethode ermittelt $p(c|\mathbf{x_t})$ für alle $\mathbf{x}_t \epsilon \mathbf{X}$ und bildet dann den Durchschnitt. Die Maximummethode ermittelt ebenfalls $p(c|\mathbf{x_t})$ für alle $\mathbf{x}_t \epsilon \mathbf{X}$ und weist $p(c|\mathbf{X})$ den Maximalwert zu. Parris und Carey (1996) verwenden diese Methode, um die Ergebnisse zweier Klassifikatoren zur Geschlechtserkennung zu kombinieren.

In AGENDER ist jedoch zur Lösung dieses Problem kein gesonderter Mechanismus notwendig, da dies ebenfalls mithilfe eines DBN auf der Zweiten Ebene bewerkstelligt werden kann. Die Fusion mehrer Klassifikationsergebnisse, die dieselbe Äußerung betreffen (statische Variante), kann dabei als eine Erweiterung des in Abbildung 9.5 dargestellten Falls angesehen werden. Die Fusion wird dadurch erreicht, dass der Knoten, der die tatsächliche Sprechereigenschaft repräsentiert, mehrere Kinderknoten besitzt (vgl. Abbildung 9.9). Da die CPTs durch die Validierungsergebnisse

bestimmt werden, ist gewährleistet, dass bei der Fusion die Qualität des jeweiligen Klassifizierers berücksichtigt wird.

Die Fusion von Klassifikationsergebnissen, die jeweils eine von mehreren aufeinander folgenden Äußerungen desselben Sprechers betreffen, kann ebenfalls mithilfe eines DBN realisiert werden. Dabei entspricht jede Äußerung einer Zeitscheibe. Die CPTs an den Verbindungsknoten bestimmen, wie stark die Informationen aus den vergangenen Zeitscheiben mit in die Gesamteinschätzung einfließt. In Abbildung 9.10 wird dies anhand eines einfachen Beispiels verdeutlicht: Zwei Zeitscheiben sind über den Knoten ALTERSKLASSE mit den Zuständen JUNG und ALT miteinander verbunden. Durch die Wahl der Übergangswahrscheinlichkeiten, wie sie auf der linken Seite der abgebildeten CPT dargestellt werden, wird ein *1:1-Übergang* hergestellt, bei dem die A-priori-Wahrscheinlichkeit der Zeitscheibe t Likelihood aus $t - 1$ entspricht. In dem Beispiel ist die Übergangswahrscheinlichkeit jedoch von einem weiteren Knoten abhängig, der die beiden Optionen repräsentiert, dass die erste Äußerung von einem Spracherkenner verarbeitet bzw. zurückgewiesen wurde. Hintergrund dieser Konfiguration ist ein Anwendungsszenario, bei dem die Sprecherklassifikation in ein natürlichsprachliches Dialogsystem integriert wurde, und zwar dergestalt, dass die Verarbeitung durch den Spracherkenner im Anschluss an die Klassifikation erfolgt. Wenn eine Äußerung nicht erkannt wurde, weil der Sprecher beispielsweise nichts gesagt, sondern nur gehustet hat, geben viele Spracherkenner ein *No-Match-Signal* zurück. In einem solchen Fall ist davon auszugehen, dass das Ergebnis der Klassifikation weniger verlässlich ist. Dieser Umstand wird durch die Abschwächung der Übergangswahrscheinlichkeiten zum Ausdruck gebracht.

10 Ablaufbeispiel einer Sprecherklassifikation und Zusammenfassung des zweistufigen Ansatzes

10.1 Ablaufbeispiel einer kontextsensitiven Sprecherklassifikation über beide Ebenen

In Abbildung 10.1 wird der Ablauf einer Sprecherklassifikation mit dem AGENDER-Ansatz anhand eines einfachen Beispiels dargestellt. Dieser ist zunächst unabhängig von dem zugrunde liegenden Anwendungsszenario, was durch die Darstellung von Mikrofon, Smartphone und Telefon als alternative Aufnahmegeräte angedeutet wird. Auf die Übertragung der Daten zwischen den einzelnen Systemkomponenten wird in Teil III genauer eingegangen. Die Art und Weise der Vorverarbeitung ist ebenfalls von der Anwendung abhängig: In der hier beschrieben AGENDER-Variante wird in diesem Schritt die Abtastfrequenz auf 8 KHz reduziert. Für telefonbasierte Anwendungen ist eine zusätzliche Anwendung von kanalabhängigen Filtern möglich. Zur Merkmalsextraktion werden dieselben Verfahren angewendet, die bereits als Grundlage der Korpusanalyse dienten: Pitch, Jitter, Shimmer, Harmonicity-to-Noise-Ratio und Intensität werden mithilfe von PRAAT-Funktionen ermittelt, die Artikulationsgeschwindigkeit mithilfe von ENRATE und die Pausen mithilfe von SRSAD (vgl. Kapitel 4).

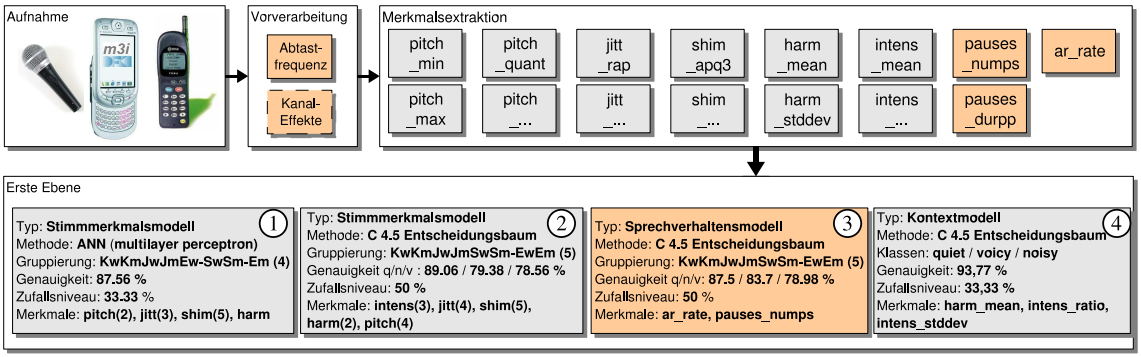


Abbildung 10.1: Beispiel für den Ablauf einer Sprecherklassifikation mit AGENDER bis zur Ersten Ebene.

Aus Gründen der Übersichtlichkeit werden in diesem Beispiel nur zwei Gruppierungen betrachtet, nämlich Gruppierung 4 (KMKMJWJMEW/SWSM/EM) und Gruppierung 5 (KWKM-JWJMSWSM/EWEM). Durch Fusion der Ergebnisse – welche auf der Zweiten Ebene erfolgt (siehe unten) – wird eine Unterscheidung der folgenden Klassen ermöglicht: 1. Kinder und Jugendliche beider Geschlechter (KWKMJWJM), 2. jüngere, erwachsene Frauen (EW), 3. jüngere, erwachsene Männer (EM) und 4. Senioren beider Geschlechter (SWSM). Als Klassifikator wurde für Gruppierung 4 ein ANN (multilayer perceptron) ausgewählt, da dieses die besten Ergebnisse in der Evaluation erzielte. Um das Prinzip der Kontextsensitivität verdeutlichen zu können, wurden für Gruppierung 5 das C 4.5 Stimmmerkmals- und das C 4.5 Sprechverhaltensmodell gewählt – bisher liegen lediglich für diese Methode Genauigkeitsanalysen unter Einwirkung der verschiedenen Kontexte vor. Die Klassifikatoren verwenden die Merkmale, die in Tabelle 7.4 (Seite 127) für die jeweilige Gruppierung angegeben werden.

Die Einheiten auf der Ersten Ebene wurden zur besseren Referenzierung von links nach rechts nummeriert: Klassifizierer eins ist ein Stimmmerkmalsmodell auf Basis eines ANNs (multilayer perceptron) mit einer Gesamtgenauigkeit von 87.56 % (das Zufallsniveau bei dieser Gruppierung liegt bei 33.33 %). Das Modell verwendet zwei Derivate der Grundfrequenz ($pitch\_quant$ und $pitch\_mean$), drei Jitter-Varianten ($jitt\_la$, $jitt\_rap$ und $jitt\_ddp$), fünf Shimmer-Varianten ($shim\_l$, $shim\_ldb$, $shim\_apq3$, $shim\_apq11$ und $shim\_ddp$) und die mittlere Harmonicity-to-Noise-Ratio $harm\_mean$. Das Netz verfügt über eine versteckte Ebene mit acht Einheiten. Wie bei allen in AGENDER verwendeten ANNs sind sämtliche Aktivierungsfunktionen des Netzes sigmoid.

Bei den Klassifizierern zwei bis vier handelt es sich um C 4.5 Entscheidungsbäume mit einem Konfidenzschwellenwert für das Pruning von 0.25 und einer minimalen Anzahl von Instanzen pro Blatt von zwei. Klassifizierer zwei repräsentiert das Stimmmerkmalsmodell für Gruppierung 5 und basiert auf den Merkmalen $intens\_mean$, $intens\_ratio$, $intens\_stddev$, $jitt\_l$, $jitt\_rap$, $jitt\_ddp$, $jitt\_ppq$, $shim\_l$, $shim\_ldb$, $shim\_apq3$, $shim\_apq11$, $shim\_ddp$, $harm\_mean$, $harm\_stddev$, $pitch\_min$, $pitch\_max$, $pitch\_quant$ und $pitch\_mean$. Das Modell weist die folgende, kontextabhängige Performanz auf: Bei einem ruhigem Kontext (QUIET) beträgt die Gesamtgenauigkeit 89.06 %, bei einem lauten Kontext (NOISY) 79.38 % und bei einem stimmenähnlichen Kontext (VOICY) 78.56 %[1]. Da Gruppierung 5 nur zwei Klassen umfasst, liegt das Zufallsniveau in diesem Fall bei 50 %. Klassifizierer drei stellt das entsprechende Sprechverhaltensmodell dar und verwendet die Merkmale ar_rate und $pauses\_numps$. Die kontextabhängige Performanz beträgt 87.5 % (QUIET), 83.7 % (NOISY) und 78.98 % (VOICY). Das Zufallsniveau liegt gemäß der zugrunde liegenden Gruppierung bei 50 %. Klassifizierer vier repräsentiert das Kontextmodell auf Basis der Merkmale $harm\_mean$, $intens\_ratio$ und $intens\_stddev$ und erreichte bei der Kreuzvalidierung eine Gesamtgenauigkeit von 93.77 %. Mit drei Klassen liegt das Zufallsniveau bei 33.33 %.

[1] Aufgrund ihrer Kürze und Prägnanz werden in den folgenden Abbildungen und Tabellen die englischen Bezeichnungen der Kontexte QUIET, NOISY und VOICY verwendet.

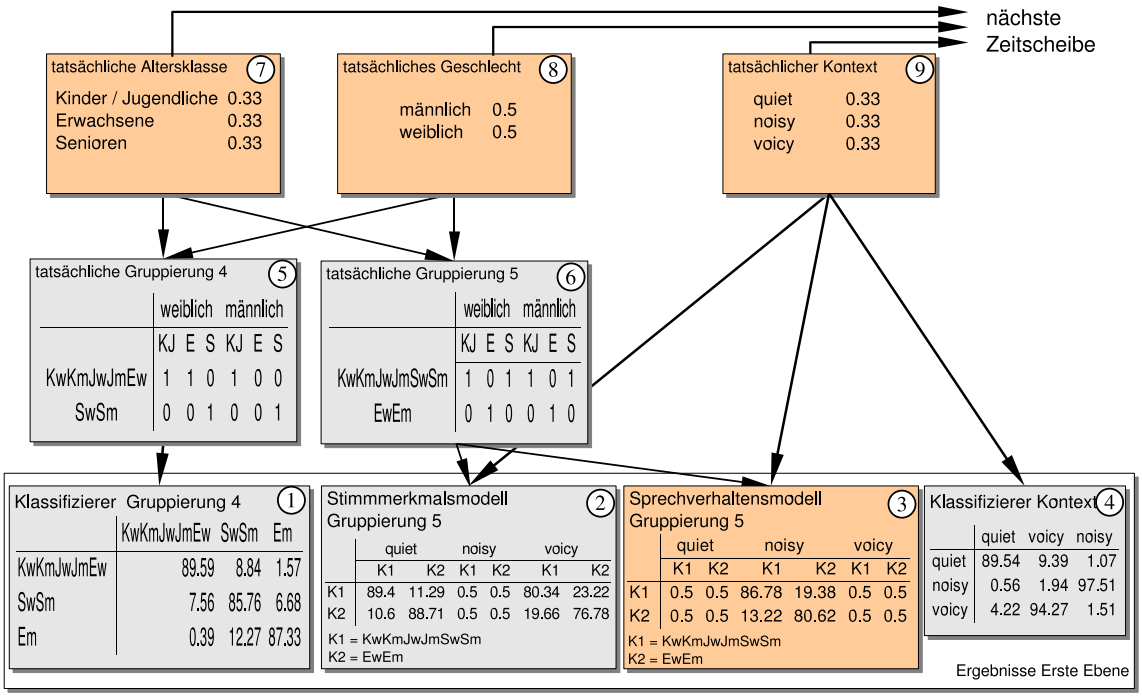


Abbildung 10.2: Beispiel für den Ablauf einer Sprecherklassifikation mit AGENDER: erste Zeitscheibe der Zweiten Ebene.

Nach der Aufzeichnung und Vorverarbeitung der Äußerung, der Extraktion der Merkmalsvektoren und deren Zuweisung zu einer jeweiligen diskreten Klasse, ist die Erste Ebene abgeschlossen. Die Aufgaben der Zweiten Ebene sind in diesem Beispiel die statische und dynamische Fusion der einzelnen Klassifikationsergebnisse, die explizite Repräsentation der Unsicherheit sowie die Einbeziehung des auditiven Kontextes. Die erste Zeitscheibe des entsprechenden Dynamischen Bayes'schen Netzes (DBN) wird in Abbildung 10.2 dargestellt. Man beachte, dass die Richtung der Pfeile nun nicht mehr den zeitlichen Ablauf der Verarbeitung andeuten, sondern – gemäß der Konvention zur Darstellung von DBNs – die Richtung der bestehenden kausalen Zusammenhänge. Zum besseren Verständnis sei noch einmal auf den in Abbildung 9.2 (Seite 183) dargestellten Zusammenhang von Krankheit und Symptom verwiesen: Die Krankheit verursacht das Symptom, was durch die Richtung der Pfeile von oben nach unten angedeutet wird. Was beobachtet wird, ist jedoch das Symptom, so dass die Propagierung des Wissens entgegengesetzt der Pfeilrichtung verläuft.

Analog dazu sind die beobachtbaren Knoten eins bis vier des DBNs in Abbildung 10.2 am unteren Ende des Netzwerkes angeordnet. Sie werden entsprechend der Ergebnisse der jeweiligen Klassifizierer auf der Ersten Ebene instanziiert. Die Knoten fünf und sechs repräsentieren den tatsächlichen Weltzustand bezüglich der jeweiligen Gruppierung. Auf diese Art und Weise kann die Unsicherheit der Klassifizierer, wie in Abschnitt 9.3.1 beschrieben, modelliert werden. Die Verbindung zwischen ihnen und den oberen Knoten sieben und acht dient zur Fusion der Ergebnisse bei gleichzeitiger Disambiguierung der Klassen. Die zuletzt genannten Knoten stellen

Fall	Beobachtungen					Likelihood-Werte				
	1			2		7			8	
	KwKm-JwJm-Ew	SwSm	Em	KwKm-JwJm-SwSm	EwEm	KJ	E	S	w	m
1	100	0	0	100	0	82.93	9.60	7.47	54.63	45.37
2	100	0	0	0	100	27.64	69.87	2.49	83.71	16.29
3	0	100	0	100	0	8.64	1.87	89.49	50.05	49.95
*4	0	100	0	0	100	6.22	29.38	64.40	51.76	49.24
*5	0	0	100	100	0	1.79	46.76	51.45	26.82	73.18
6	0	0	100	0	100	0.17	95.05	4.79	2.89	97.11

Tabelle 10.1: Likelihood-Werte der Knoten sieben und acht unter Berücksichtigung aller möglichen Ergebnisse von Klassifizierer eins und zwei (statische Fusion).

das letztendliche Sprechermodell dar, welches an die Anwendung weitergegeben wird. Von dort aus gehen Verbindungen zur nächsten Zeitscheibe mit 1:1-Übergangswahrscheinlichkeiten. Im Fall des Kontextmodells fehlt die Zwischenebene, da keine Fusion erforderlich ist. Analog zum Sprechermodell dient die Verbindung zwischen Knoten vier und Knoten neun zur Modellierung der Unsicherheit auf Basis der Kreuzvalidierungsergebnisse des Kontextklassifizierers. Darüber hinaus gibt es von Knoten neun ausgehend Verbindungen zu Knoten zwei und Knoten drei. Auf diese Art und Weise wird eine kontextabhängige Gewichtung von Stimmmerkmals- und Sprechverhaltensmodell erreicht, wie sie in Abschnitt 9.3.2 erläutert wurde.

Im Folgenden wird eine Reihe von Tabellen präsentiert, welche die Likelihood-Werte des Bayes'schen Netzes unter Berücksichtigung verschiedener Beobachtungen enthalten. In Tabelle 10.1 wird zunächst der Aspekt der statischen Fusion verdeutlicht, indem alle möglichen Kombinationen von Beobachtungen der Knoten eins und zwei und die daraus abgeleiteten Likelihood-Werte der Knoten sieben und acht dargestellt werden. Bezüglich des Kontextes wird in diesem Beispiel davon ausgegangen, dass dieser den Wert QUIET hat, so dass das Stimmmerkmalsmodell den Vorzug gegenüber dem Sprechverhaltensmodell genießt. Der Übersicht wegen wurde Knoten drei nicht instanziiert, da dessen Einfluss unter den genannten Voraussetzungen marginal ist.

Die erste Zeile stellt den Fall dar, bei dem der Klassifizierer für Gruppierung 4 (Knoten 1) die Klasse KWKMJWJMEW und der Klassifizierer für Gruppierung 5 (Knoten 2) die Klasse KWKM-JWJMSWSM ausgibt. Bei dieser Konstellation ist erwartungsgemäß die Likelihood für die Altersklasse KJ mit 82.93 % am höchsten, während bezüglich des Geschlechts erwartungsgemäß mit

	Beobachtungen					Likelihood-Werte				
	1			2		7			8	
Fall	KwKm-JwJm-Ew	SwSm	Em	KwKm-JwJm-SwSm	EwEm	KJ	E	S	w	m
1	100	0	0	100	0	96.82	2.39	0.79	55.68	44.32
*2	100	0	0	0	100	84.72	14.60	0.69	61.04	38.96
*3	0	100	0	100	0	50.63	2.15	47.22	54.68	45.32
*4	0	100	0	0	100	44.85	13.32	41.83	54.97	45.03
*5	0	0	100	100	0	12.16	56.40	31.44	24.11	75.89
*6	0	0	100	0	100	2.78	90.03	7.19	5.92	94.08

Tabelle 10.2: Likelihood-Werte der Knoten sieben (Altersklasse) und acht (Geschlecht) in der zweiten Zeitscheibe unter der Annahme, dass das Ergebnis der vorangegangenen Zeitscheibe Fall eins aus Tabelle 10.1 ist (dynamische Fusion).

54.63:45.37 nur eine sehr konservative Aussage getroffen wird. Die Tatsache, dass überhaupt ein Unterschied besteht, ist auf die Differenzen der True Positive Rates zurückzuführen, die als Grundlage für die CPTs dienen. Analog dazu führen die Konstellationen von Beobachtungen in den Zeilen zwei, drei und sechs zu den Ergebnissen EW, SWSM und EM. Die mit einem Stern markierten Zeilen stellen diejenigen Fälle dar, in denen die Ergebnisse der Ersten Ebene nicht konsistent sind: In Zeile vier wird ein Widerspruch zwischen SWSM und EWEM beschrieben, der in dem Fall zugunsten der zuerst genannten Klasse aufgelöst wird. In der darauf folgenden Zeile wird ein Widerspruch zwischen EM und KWKMJWJMSWSM dargestellt, was sich hier in einer annähernd ausgeglichenen Likelihood beider widerspiegelt. Dennoch wird aus dieser Konstellation ein informatives Ergebnis abgeleitet: Ein Sprecher männlichen Geschlechts wird als wahrscheinlicher angenommen als ein Sprecher weiblichen Geschlechts.

In Tabelle 10.2 werden die alternativen Sprechermodelle – d. h. Likelihood-Werte der Knoten sieben (Altersklasse) und acht (Geschlecht) – der zweiten Zeitscheibe aufgeführt, und zwar unter der Annahme, das Ergebnis der ersten Zeitscheibe ist KWKMJWJMEW / KWKMJWJMSWSM (Zeile eins in Tabelle 10.1). Sind die Beobachtungen in Zeitscheibe zwei dieselben, führt dies zu einer Erhöhung der Likelihood des bisherigen Sprechermodells. In den mit Stern markierten Fällen trifft dies nicht zu: In Zeile zwei wird beispielsweise die Situation beschrieben, dass der Klassifizierer für Gruppierung 5 EWEM ausgibt, wodurch die Likelihood der Klasse EW ansteigt. Dieses einfache Beispiel verdeutlicht bereits, dass Dynamische Bayes'sche Netze einen eleganten Mechanismus darstellen, um eine Fusion der Ergebnisse mehrerer Klassifizierer zu erreichen.

	Beobachtungen							Likelihood-Werte					
	2		3		4			7			9		
Fall	KwKm-JwJm-SwSm	EwEm	KwKm-JwJm-SwSm	EwEm	Q	N	V	KJ	E	S	Q	N	V
1	100	0	100	0	100	0	0	46.82	6.36	46.82	90.42	8.60	0.99
2	100	0	100	0	0	100	0	45.05	9.89	45.05	4.93	93.85	1.51
3	100	0	100	0	0	0	100	43.73	12.54	43.73	0.61	1.93	97.46
4	100	0	0	100	100	0	0	45.28	9.45	45.28	94.02	4.96	1.02
5	100	0	0	100	0	100	0	15.92	68.16	15.92	7.86	89.44	2.60
6	100	0	0	100	0	0	100	43.37	13.26	43.37	0.62	1.08	98.30
7	0	100	100	0	100	0	0	14.66	70.69	14.66	85.01	13.98	1.01
8	0	100	100	0	0	100	0	43.73	12.54	43.73	2.75	96.27	0.98
9	0	100	100	0	0	0	100	17.74	64.52	17.74	0.56	3.02	96.42
10	0	100	0	100	100	0	0	9.952	80.11	9.95	90.65	8.27	1.08
11	0	100	0	100	0	100	0	12.30	75.41	12.30	4.81	93.48	1.71
12	0	100	0	100	0	0	100	16.81	66.37	16.81	0.56	1.70	97.74

Tabelle 10.3: Likelihood-Werte der Knoten sieben (Altersklasse) und neun (Kontext) unter Berücksichtigung aller möglichen Resultate von Stimmmerkmals- und Sprechverhaltensmodell (Knoten zwei und drei) sowie des Kontextklassifizierers (Knoten vier).

Das gilt sowohl für den statischen Fall, bei dem sich die Ergebnisse auf dieselbe Äußerung beziehen, als auch für den dynamischen Fall, bei dem das Modell sukzessive auf Basis mehrerer Äußerungen desselben Sprechers aufgebaut wird.

Tabelle 10.3 dient zur Verdeutlichung des Aspektes der Kontextsensitivität. Dazu werden die Likelihood-Werte der Knoten sieben (Altersklasse) und neun (Kontext) unter Berücksichtigung aller möglichen Resultate von Stimmmerkmals- und Sprechverhaltensmodell (Knoten zwei und drei) sowie des Kontextklassifizierers (Knoten vier) dargestellt. Knoten eins wird in diesem Fall der geringen Komplexität wegen nicht instanziiert. Aufgrund der fehlenden Disambiguierung betragen die Likelihood-Werte des Knotens acht (Geschlecht) im gesamten Beispiel 50 % und werden daher in der Tabelle ebenfalls nicht dargestellt.

Die Zeilen eins bis drei sowie zehn bis zwölf beschreiben diejenigen Fälle, in denen das

Stimmmerkmalsmodell und das Sprechverhaltensmodell dasselbe Ergebnis haben. Es ist dennoch zu erkennen, dass die Likelihood-Werte des Knotens sieben sich in Abhängigkeit vom Kontext unterscheiden. Dies ist darauf zurückzuführen, dass die Kreuzvalidierungsergebnisse der Modelle je nach Kontext verschieden sind. In den Zeilen eins bis drei nimmt beispielsweise die Likelihood der Klasse E von QUIET bis VOICY kontinuierlich zu. Die Zeilen vier bis neun repräsentieren diejenigen Fälle, in denen es einen Widerspruch zwischen den beiden Modellen gibt. Bei einem lauten Kontext (NOISY) erhält die Aussage des Sprechverhaltensmodells aufgrund dessen besserer Kreuzvalidierungsergebnisse bei der Bestimmung der Likelihood von Knoten sieben ein höheres Gewicht. Bei QUIET und VOICY wird dagegen das Stimmmerkmalsmodell bevorzugt. Die Unterschiede zwischen den beiden zuletzt genannten Fällen sind wiederum auf die verschiedenen Kreuzvalidierungsergebnisse zurückzuführen.

10.2 Zusammenfassung des zweistufigen Ansatzes zur Sprecherklassifikation

Die durch die wechselseitige Abhängigkeit von Sprecheralter und -geschlecht verursachte Komplexität kann verringert werden, indem das aus vier mal zwei Klassen bestehende Problem (vier Altersklassen und zwei Geschlechter) in ein Acht-Klassen-Problem umformuliert wird. Diese werden bezeichnet mit KW, KM, JW, JM, EW, EM, SW und SM, wobei der erste Buchstabe für die Altersklasse steht (Kinder, Jugendliche, Erwachsene und Senioren) und der zweite Buchstabe für das Geschlecht (weiblich und männlich).

Bezüglich der veränderten Problemstellung wurde eine erneute Auswertung der Korpusanalyse-Ergebnisse vorgenommen, wobei festgestellt wurde, dass die Bildung multipler Gruppierungen der Klassen möglich ist, die jeweils von unterschiedlichen Kombinationen von Merkmalen unterstützt werden. Gruppierung 1 wird beispielsweise definiert als KWKMJW-EW-SW-EMSM, d. h. die Kinder und (weiblichen) Jugendlichen bilden eine Gruppe, die jüngeren, erwachsenen Frauen und die Frauen im Seniorenalter werden separat betrachtet, und die jüngeren und älteren Männer bilden wiederum eine Gruppe. Gruppierung 1 wird hauptsächlich von den Maßen $jitt\_ppq$ und $pitch\_mean$ unterstützt, bedingt jedoch auch von $shim\_apq11$ und $pitch\_min$. Insgesamt wurden fünf solcher Gruppierungen gefunden, die zusammen mit der Kontrollgruppierung 0, bei der alle acht Klassen separat betrachtet werden, das Klassifikationsproblem der *Ersten Ebene* definieren. Um einen direkten Vergleich mit Systemen zur Einschätzung des Sprechergeschlechts zu ermöglichen, wird darüber hinaus die Gruppierung EW-EM betrachtet.

Mit dem Begriff Erste Ebene werden in AGENDER diejenigen Phasen der Mustererkennung bezeichnet, welche die Merkmalsextraktion und die Klassifikation betreffen. Die Merkmalsextraktoren wurden auf Basis des in Teil I beschriebenen Korpusanalyse-Verfahrens bestimmt. Bezüglich der Klassifikation wurden die folgenden bekannten Methoden des maschinellen Lernens untersucht: 1. Naive Bayes (NB), 2. Gaussian-Mixture-Models (GMMs), 3. k-Nearest-Neighbor (KNN), 4. C 4.5 Entscheidungsbäume (C 4.5), 5. Support-Vector-Machines (SVMs) und 6. Künstliche Neuronale Netze (Artificial Neural Networks, ANNs). Die ersten beiden werden

zu den parametrischen Methoden gezählt, da sie vollständig durch eine bestimmte Anzahl von Parametern definiert werden. Im Fall von NB sind dies die klassenspezifischen Mittelwerte und Standardabweichungen – im Fall von GMMs kommen noch die Gewichte der einzelnen Merkmale hinzu. Die Methode KNN gehört dagegen zu den instanzbasierten Methoden, da in dem Fall die Zuordnung eines Merkmalsvektors zu einer Klasse nicht anhand eines zuvor generalisierten Modells, sondern anhand der einzelnen Trainingsinstanzen erfolgt. Entscheidungsbäume nehmen insofern eine Sonderstellung ein, als sie ursprünglich für die Klassifikation auf Grundlage von nominellen Merkmalen entwickelt wurden (z. B. ROT/GRÜN oder RUND/ECKIG). Sie gelten dennoch als ebenso geeignet für Klassifikationsprobleme mit reellwertigen Merkmalen, was auch bei dieser Untersuchung bestätigt werden konnte. SVMs stellen eine besondere Form von Klassifikationsverfahren auf Basis linearer Diskriminantenfunktionen dar, bei denen der ursprüngliche Merkmalsraum zunächst in einen höherdimensionalen Raum überführt wird, um auch nichtlineare Probleme lösen zu können. ANNs schließlich sind einem Mechanismus nachempfunden, bei dem man davon ausgeht, dass er bei der menschlichen Informationsverarbeitung angewendet wird. Die Entscheidungen werden durch künstliche Synapsen herbeigeführt, die über mehrere Ebenen miteinander vernetzt sind und deren Aktivierungsfunktionen auf Basis der Trainingsdaten angepasst werden.

Alle genannten Verfahren wurden auf das gegebene Klassifikationsproblem angewendet, um festzustellen, welches am besten geeignet ist. Im Vordergrund stand dabei die globale Klassifikationsgenauigkeit. Die Studie sollte jedoch ebenfalls Aufschluss darüber geben, inwiefern erstens Vorteile aus den Gruppierungen gezogen werden können, und zweitens die Leistungen der Stimmmerkmalsmodelle sich von denjenigen der Sprechverhaltensmodelle unterscheiden. Was den zuerst genannten Aspekt betrifft, ließen die jeweiligen theoretischen Eigenschaften der Verfahren unterschiedliche Hypothesen zu: Es wurde erwartet, dass besonders die parametrischen Methoden von einer Gruppierung der Klassen und einer vorherigen manuellen Selektion der Merkmale profitieren, wohingegen dies bei Verfahren mit hoch entwickelten Trainingsprozessen, wie z. B. ANNs, in geringerem Maße der Fall sein sollte. Bezüglich des Vergleichs der Stimmmerkmals- und Sprechverhaltensmodelle wurde angenommen, dass in einem ruhigen Kontext erstere eine bessere Performanz aufweisen, sich dieser Vorzug jedoch durch einen lauten Kontext relativiert und evtl. umkehrt.

Die Ergebnisse, die erzielt werden konnten, sind insgesamt vielversprechend: Die Klassifikationsgenauigkeiten sämtlicher getesteter Verfahren liegen deutlich über dem Zufallsniveau der jeweiligen Gruppierung. Als besonders zufriedenstellend können dabei die True Positive Rates der Kontrollgruppierung 0 eingeschätzt werden, bezüglich derer mithilfe eines Artificial Neural Networks (ANNs) eine Gesamtgenauigkeit von 63.5 % erreicht werden konnte, was dem Fünffachen des Zufallsniveaus entspricht. Im direkten Vergleich der Methoden bleiben die parametrischen hinter den übrigen zurück, wobei besonders auffällig ist, dass Gaussian-Mixture-Models (GMMs) nur geringfügig bessere Ergebnisse aufweisen als Naive-Bayes-Klassifizierer (NBs). Dies wird auf die Tatsache zurückgeführt, dass die Gewichte nicht mithilfe des EM-Algorithmus gelernt, sondern

durch eine Kreuzvalidierung ermittelt wurden.

Was die Leistungen der Klassifizierer bezüglich der verschiedenen Gruppierungen betrifft, konnten die Hypothesen bestätigt werden: Die Performanz der parametrischen Methoden ist bezüglich der Kontrollgruppierung 0 am geringsten, wohingegen KNNs und ANNs zwar in fast allen Gruppierungen die vorderen Plätze einnehmen, hier jedoch einen besonders deutlichen Vorsprung aufweisen.

Die Genauigkeit der Stimmmerkmalsmodelle ist wie erwartet höher als die der entsprechenden Sprechverhaltensmodelle, wobei in diesem Zusammenhang große Unterschiede zwischen den beiden betreffenden Gruppierungen bestehen. Während nämlich die Performanz der Sprechverhaltensmodells bezüglich Gruppierung 5 zumindest auf ähnlichem Niveau wie die des Stimmmerkmalsmodells liegt, sind bezüglich Gruppierung 3 gravierende Performanzeinbußen zu verzeichnen. Gemäß der oben genannten Hypothese geht unter Hinzunahme eines Sprechkontextes die Klassifikationsperformanz insgesamt zurück. Bei einem lauten Kontext verzeichnet das Sprechverhaltensmodell einen leichten Vorteil gegenüber dem Stimmmerkmalsmodell.

Abgesehen von diesem positiven Resultat zeichnet sich der AGENDER-Sprecherklassifikationsansatz durch die so genannte *Zweite Ebene* aus, die aus Dynamischen Bayes'schen Netzen (DBNs) basiert. Anhand von Beispielen wurde gezeigt, wie diese genutzt werden können, um erstens die klassifikationsinhärente Unsicherheit explizit zu modellieren, zweitens um Top-Down-Wissen in den Entscheidungsprozess einfließen zu lassen – wie z. B. die Tatsache, dass je nach Kontext die Resultate bestimmter Klassifizierer als zuverlässiger eingeschätzt werden sollten als die anderer – und drittens um eine Fusion multipler Klassifikationsergebnisse erreichen zu können. Mithilfe von DBNs kann die letztere Aufgabe sowohl in der statischen Variante gelöst werden, bei der sich die Ergebnisse auf dieselbe Äußerung beziehen, als auch in der dynamischen Variante, bei der die Ergebnisse bezüglich mehrerer Äußerungen in das Modell einfließen.
Im Rahmen eines konkreten, auf M3I aufbauenden Industrieprojektes, wird die Weiterentwicklung von AGENDER hauptsächlich in Richtung der in Kapitel 1.3 beschriebenen Telekommunikationsanwendungen betrieben, was unter anderem zur Folge hat, dass neben der Klassifikationsgenauigkeit das Hauptaugenmerk auf die Klassifikationszeit gerichtet sein wird. Dies betrifft zwar in erster Linie die AGENDER-Implementierung, die in Teil III beschrieben wird, bringt jedoch auch die Notwendigkeit mit sich, explizite Laufzeit-Studien (*Benchmarks*) der hier untersuchten Verfahren anzustellen. Derzeit wurden diesbezüglich lediglich allgemeine Aussagen auf Basis von deren theoretischer Komplexität getroffen. Die Benchmarks sollen außerdem Aufschluss darüber geben, ob die manuell konfigurierten Gruppierungen tatsächlich einen Laufzeit-Vorteil einbringen.

Darüber hinaus wird im Rahmen des genannten Projektes eine präzisere Einschätzung des Sprecheralters angestrebt. In Kapitel 6 wurde bereits aufgeführt, welche zusätzlichen Merkmale hierfür als Grundlage dienen sollen. Hinsichtlich der Mustererkennung müssen zusätzlich methodologische Veränderungen in Betracht gezogen werden. Das Performanzmaß der True Positive Rates wird z. B. ab einer gewissen Anzahl von Klassen nicht länger praktikabel sein und sollte durch ein Maß ersetzt werden, welches den Abstand zwischen der tatsächlichen Klasse und der geschätzten Klasse berücksichtigt. Möglicherweise rücken in diesem Zusammenhang auch ande-

re Klassifikationsverfahren in den Vordergrund, die von vornherein statt einer Zuordnung eines Merkmalsvektors zu einer diskreten Klasse eine lineare Regression durchführen, d. h. das Alter als numerischen Wert vorhersagen. Der Vorteil bestünde unter anderem darin, dass ein unmittelbarer Vergleich der Performanz des Modells mit der menschlichen Fähigkeit zur Einschätzung des Alters erfolgen könnte, wie sie beispielsweise von Braun und Cerrato (1999) untersucht worden ist. In dieser Studie wurde eine durchschnittliche Differenz von zehn bis zwölf Jahren zwischen tatsächlichem und geschätztem Alter festgestellt.

Teil III

Ein Client/Server-System zur Sprecherklassifikation für mobile Dialogsysteme mit angegliedertem Korpusanalyse-Werkzeug

11 Ein Client/Server-System zur Sprecherklassifikation für mobile Dialogsysteme

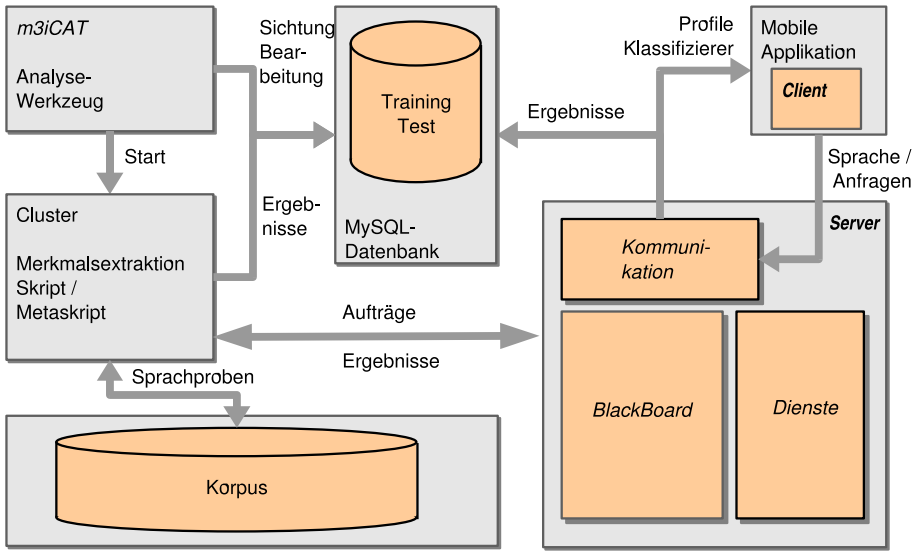


Abbildung 11.1: Gesamtarchitektur der AGENDER-Implementation.

Mobile Applikationen, die gesprochene, natürlichsprachliche Eingaben erlauben, bilden den Rahmen für die in der vorliegenden Arbeit vorgestellte Version von AGENDER. Bei den beiden Beispiel-Anwendungen handelt es sich um die im Rahmen des Projektes *m3i* entwickelten Systeme *m3i Mobile ShopAssist* und *m3i Personal Navigator* (vgl. Abschnitt 1.3). Das Anwendungsszenario sieht vor, dass das mobile Gerät – z. B. ein Pocket-PC oder ein Smartphone – über eine dauerhafte, breitbandige Netzwerkverbindung zu einem Server verfügt (UMTS, W-Lan), der Dienste bereitstellt, die auf dem Gerät selbst aufgrund seiner eingeschränkten Ressourcen entweder gar nicht oder nur eingeschränkt durchführbar wären. Die Sprecher- und Umgebungsklassifikation stellt ein Beispiel für einen solchen Dienst dar: Bei vorhandener Netzwerkverbindung werden die

Sprachproben der Benutzer an einen Server geschickt, der die Klassifikation durchführt und die daraus resultierenden Profile an das mobile Gerät zurücksendet. Falls die Netzwerkverbindung unterbrochen ist, sollte der Pocket-PC allerdings in der Lage sein, in eingeschränktem Maße eine eingebettete Klassifikation durchzuführen.

Gemäß diesen Vorgaben wurde – wie in Abbildung 11.1 dargestellt wird – die AGENDER-Implementierung als eine Client/Server-Lösung entworfen[1]. Die Hauptkomponente, der java-basierte *m3i Server*, wird in Kapitel 11.1 beschrieben. Der *m3i Client* stellt programmtechnisch eine Bibliothek in c++ dar, welche in die Anwendung integriert werden kann (vgl. Kapitel 11.2). Das *Corpus Analyzing Toolkit* m3iCAT ist ein im Rahmen des AGENDER-Projektes entwickeltes, PHP-basiertes System zur Durchführung von Korpusanalysen, Sichtung der Ergebnisse und Vorbereitung der Daten zum Trainieren von Modellen. Die Daten werden in einer zentralen MySQL-Datenbank gespeichert. Sowohl das Analysewerkzeug, das in Kapitel 12 beschrieben wird, als auch der m3i Server machen Gebrauch von einem *Cluster*, d. h. einem Verbund gleichartiger Rechner (*Knoten)*: Im ersten Fall wird die Korpusanalyse und im zweiten Fall die Merkmalsextraktion auf mehreren Knoten parallel durchgeführt.

[1]Die in dem akutellen und den nachfolgenden Kapiteln präsentierten Abbildung enthalten z.T. sowohl englischsprachliche als auch deutsche Etiketten. Ersteres ist immer dann der Fall, wenn es sich um die Nennung eines Moduls oder einer Klasse handelt, da diese im Quellcode ebenfalls englische Bezeichnungen haben.

11.1 Der m3i Server

Die Implementierung des *m3i Servers* wurde zeitgleich mit der Entwicklung des AGENDER-Ansatzes durchgeführt. Diese Vorgehensweise ermöglichte eine unmittelbare Überprüfung der Umsetzbarkeit der Konzepte, stellte jedoch besondere Anforderungen an die Architektur des Systems: Sie musste flexibel genug sein, um konzeptuelle Änderungen leicht umsetzen zu können, und transparent genug, damit das System trotz wachsender Komplexität überschaubar bleibt. Der Server wurde daher mit einer *Blackboard*-Architektur ausgestattet, welche diese Kriterien erfüllt, wie im Folgenden anhand von Beispielen deutlich gemacht wird. Blackboard-Architekturen wurden in den 70er Jahren zunächst für Sprachtechnologie-Systeme entwickelt, später jedoch auch in wissensbasierten Anwendungen eingesetzt (vgl. Corkill, Gallagher und Johnson, 1988). Besonders in sehr komplexen Systemen mit einer Vielzahl von Modulen zeigen sie Vorteile: In Verbmobil beispielsweise, das aus 69 hoch interaktiven Einheiten besteht, werden zur Handhabung des enormen Kommunikationsaufkommens eine Vielzahl von Blackboards eingesetzt – insgesamt 198 (vgl. Wahlster, 2001).

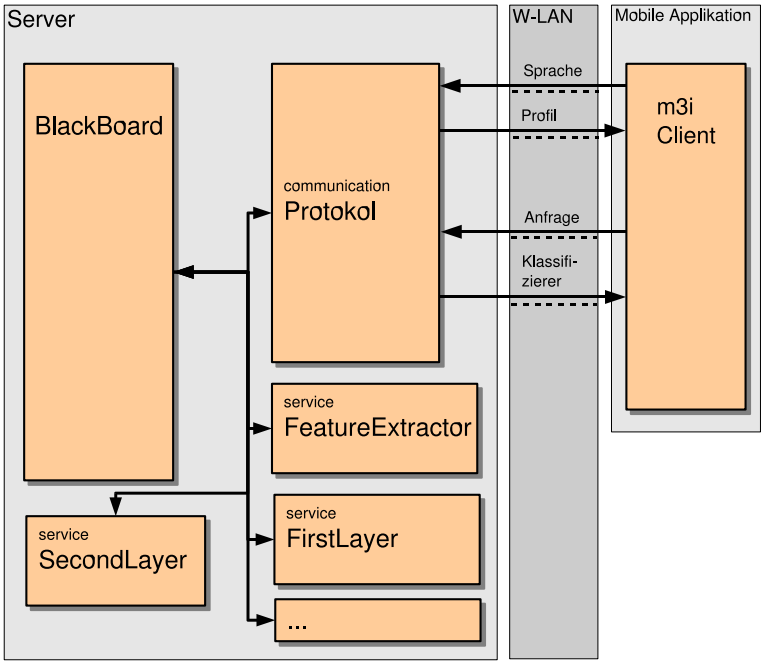


Abbildung 11.2: Blackboard-Architektur des m3i Servers.

Wie in Abbildung 11.2 dargestellt wird, stellt das „schwarze Brett“, das *BlackBoard*[2], auch in diesem System die zentrale Komponente der Architektur dar. Auf ihm werden typisierte Objekte abgelegt. Falls es Dienste (*Services*) gibt, die sich für diesen Objekt-Typ registriert haben, sendet das BlackBoard es ihnen zu und erhält die Ergebnisse wiederum als Objekt eines anderen Typs zurück. Der Vorteil dieser Architektur besteht darin, dass zur Einführung einer Komponente

[2]Mit Bezug auf die AGENDER-Implementierung wird die Schreibweise *BlackBoard* verwendet, was auch dem Java-Klassennamen entspricht.

lediglich ein neuer Dienst registriert werden muss, während das restliche System davon weitestgehend unbeeinflusst bleibt. Durch Ausnutzung von Vererbungsmechanismen kann zusätzlich erreicht werden, dass der Aufwand für die Formulierung eines neuen Dienstes minimiert wird, was beispielsweise den Test neuer Klassifizierer erleichtert (vgl. Abschnitt 11.1.1).

Der hauptsächliche Nachteil einer Blackboard-Architektur gegenüber Lösungen, die eine stärkere Integration der Komponenten beinhalten, ist eine geringere Performanz, welche durch die erforderliche Kommunikation zwischen den Diensten und dem Blackboard verursacht wird. In Kapitel 13 wird daher in Hinblick auf eine mögliche industrielle Anwendung von AGENDER eine auf Performanz hin optimierte Implementierung skizziert.

11.1.1 Domänenspezifische Aspekte

In dem folgenden Abschnitt werden zunächst die domänenspezifischen Aspekte des m3i Servers beschrieben, d. h. diejenigen Komponenten, die sich unmittelbar auf die Sprecher- und Umgebungsklassifikation beziehen. Dazu gehören die Dienste zur Merkmalsextraktion, Klassifikation (Erste Ebene) und Nachverarbeitung (Zweite Ebene). Darüber hinaus beinhaltet die Implementierung jedoch eine Reihe von generellen Aspekten, zu denen das eigentliche BlackBoard und die dazugehörigen Kontrollstrukturen gehören. Auf sie wird in Abschnitt 11.1.2 eingegangen.

Merkmalsextraktion

Die unmittelbar an der Merkmalsextraktion beteiligten *Services* sind der *AudioProcessor* und der *FeatureExtractor*. Beide Dienste greifen auf Dateien zu, die im lokalen Dateisystem des Servers abgespeichert wurden, wohin sie entweder über das *Protokoll* (siehe unten) gelangen – so fern es sich um eine Äußerung handelt, die vom *m3i Client* gesendet wurde – oder durch den internen Aufnahmemechanismus des Servers, welcher zu Testzwecken verwendet wird. Die Aufgabe des *AudioProcessors* besteht darin, die Aufnahme für die Merkmalsextraktion vorzubereiten, indem er die Qualität (Sample-Rate, Bit pro Sample, Kanal) an die der Korpusdateien (8000 KHz, 16 Bit, mono) anpasst. Durch diese Vorgehensweise wird sichergestellt, dass es nicht zu unvorhergesehenen Artefakten bei den zu betrachtenden Merkmalen durch unterschiedliche Kodierung der Dateien kommt. Die Vorverarbeitung erfolgt unter Verwendung des Programms *Sox*[3], das als externer Prozess in die Architektur integriert wurde (vgl. Abschnitt 11.1.2).

Wenn die Vorverarbeitung abgeschlossen ist, erhält der FeatureExtractor eine Nachricht vom BlackBoard. Er fordert eine Liste der verfügbaren Analyseskripte (*scripts*) an und verteilt die Datei an alle diejenigen von ihnen, welche als aktiv markiert worden sind (siehe unten). Die Zuordnung zu einem bestimmten Sprecher erfolgt über ein Objekt vom Typ *Speaker*, das vom FeatureExtractor angelegt wird. Jedes Zwischenergebnis enthält einen Verweis auf dieses Objekt.

Scripts / VoiceFeatureObjects

Skripte repräsentieren externe Prozesse mithilfe derer die Merkmale der Sprachprobe extrahiert werden. Ein Großteil dieser Prozesse basiert auf dem Programm PRAAT, wobei dieselben

[3] http://sox.sourceforge.net

PRAAT-Skripte verwendet werden, die bereits bei der Korpusanalyse zum Einsatz gekommen sind (vgl. 4). Darüber hinaus gibt es Skripte für ENRATE (Artikulationsgeschwindigkeit) und SRSAD (Sprechpausen). Das Ergebnis des Extraktionsprozesses wird in Form eines *VoiceFeatureObjects* auf das BlackBoard zurückgeschrieben.

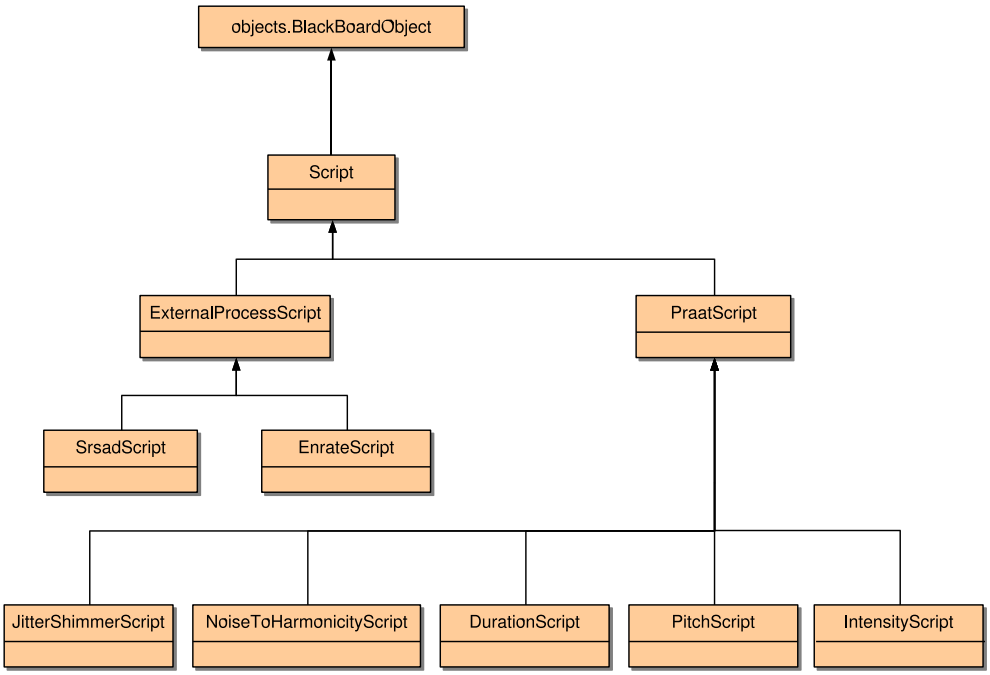


Abbildung 11.3: Vererbungshierarchie der Klasse Scripts.

Abbildung 11.3 stellt die Vererbungshierarchie der Skripte dar. Die Klasse *Script* selbst ist eine Unterklasse von *BlackBoardObject* und ist abstrakt, d. h. wird selbst nicht instanziiert. Alle PRAAT-Skripte erben von der Klasse *PraatScript*, während die beiden anderen der gemeinsamen Oberklasse *ExternalProcessScript* zugewiesen wurden. Die wesentlichen Methoden und Attribute der Klasse Script sind die folgenden:

multinode	Ein Attribut, das bestimmt, ob ein gegebenes Skript auf mehrere Cluster-Knoten (*nodes*) parallelisierbar ist oder nicht. *Multinode-Skripte* besitzen einen zusätzlichen Parameter, der dem jeweiligen Knoten mitteilt, welchen Einzelwert er berechnen soll. Da durch die Parallelisierung ein zusätzlicher Aufwand durch die Verwaltung der Einzelergebnisse entsteht, ist eine Realisierung als Multinode-Skript nur unter bestimmten Voraussetzungen sinnvoll (siehe unten).
filename	Verweist auf den Dateinamen des (externen) Skriptes, das letztlich ausgeführt wird.
createObject()	Stellt den eigentlichen Extraktionsprozess dar, d. h. die Erzeugung eines *VoiceFeatureObjects* aus einer vorverarbeiteten Klangdatei.
getFeatureNames()	Gibt eine Liste von Merkmalsbezeichnungen zurück, die das Skript extrahieren kann. Sie wird mit den Listen verglichen, in denen die von den Klassifizierern benötigten Merkmale verzeichnet sind. Bei (auch teilweiser) Übereinstimmung wird das Skript aktiviert.

PraatScripts

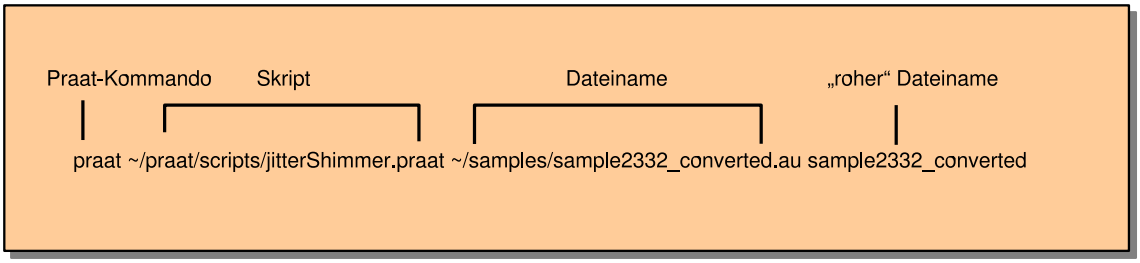


Abbildung 11.4: Syntax eines PRAAT-Aufrufs, wie er von Skripten vom Typ PraatScript verwendet wird.

Die Gemeinsamkeit der Unterklassen von *PraatScript* besteht darin, dass sie zur Extraktion der Merkmale einen externen Prozess mit der für PRAAT erforderlichen Syntax erzeugen (vgl. Abbildung 11.4). Sie unterscheiden sich im Wesentlichen dadurch, mit welchen *VoiceFeatureObjects* sie verbunden sind, und dementsprechend, welche Einträge die Liste *featureNames* umfasst:

JitterShimmerObject	*jitt_l*, *jitt_la*, *jitt_ppq*, *jitt_rap*, *jitt_ddp*, *shim_l*, *shim_ldb*, *shim_aq3*, *shim_apq11* und *shim_ddp*
NoiseToHarmonicityObject	*harm_mean*, *harm_min*, *harm_max* und *harm_stddev*
DurationObject	*dur*
IntensityObject	*intens_mean*, *intens_min*, *intens_max*, *intens_stddev*, *intens_ratio*
PitchObject	*pitch_min*, *pitch_max*, *pitch_quant*, *pitch_mean*, *pitch_stdev*, *pitch_mas*, *pitch_swoj*

Eine Besonderheit des *JitterShimmerScripts* besteht darin, dass es als *Multinode-Skript* realisiert wurde, also auf verschiedenen Rechnern parallel ausgeführt werden kann (vgl. Abschnitt 11.1.2).

ExternalProcessScripts

Die Syntax der Unterklassen von *ExternalProcessScript* ist einfacher, da lediglich das (externe) Skript selbst und der Dateiname der Sprachprobe benötigt werden. Der Aufruf von ENRATE lautet beispielsweise `enrate.sh samples/sample2332_converted.au`. Die dazugehörigen *VoiceFeatureObjects* sind das *SpeechRateObject* bzw. das *VoiceActivityObject* mit den folgenden Merkmalen:

SpeechRateObject	*ar_rate*
VoiceActivityObject	*pauses_onset*, *ar_time*, *pauses_dur*, *pauses_num*

VoiceFeatureObjects

Während die Besonderheiten der Unterobjekte sich hauptsächlich auf die unterschiedlichen Merkmalslisten beschränken, umfasst das übergeordnete *VoiceFeatureObject* einige funktionell relevante Attribute und Methoden. Mithilfe von *parseString()* beispielsweise wird der Rückgabewert eines externen Prozesses, der in Form einer Zeichenkette vorliegt, in einzelne reellwertige Zahlen zerlegt und den entsprechenden Merkmalen zugewiesen. Falls der Rückgabewert nicht zum Objekt passt, d. h. zu wenige, zu viele oder unpassende Werte enthält, wird eine entsprechende Fehlermeldung erzeugt. Darüber hinaus beinhaltet die Klasse die Methode *isComplete()*, welche überprüft, ob das Objekt bereits vollständig ist, d. h. ob alle Merkmale Werten zugeordnet werden konnten. Die Überprüfung ist besonders dann notwendig, wenn das Objekt wie im Fall des *JitterShimmerObjects* zu einem *Multinode-Skript* gehört, da in einem solchen Fall nicht vorhersehbar ist, in welcher Reihenfolge die Werte eintreffen.

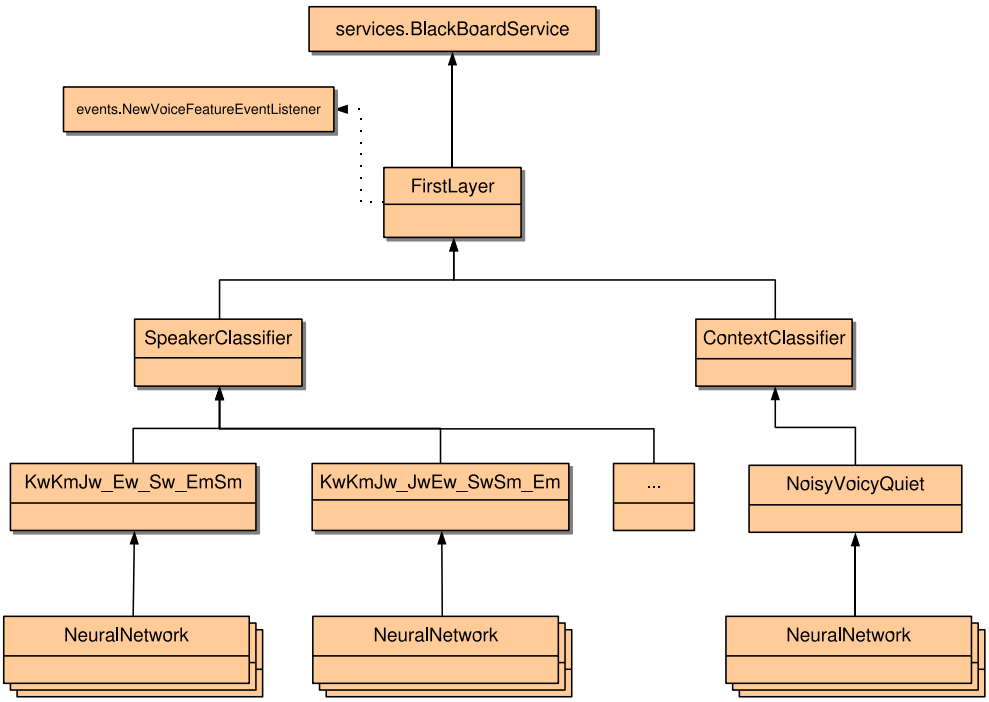


Abbildung 11.5: Klassendiagramm (Vererbungshierarchie) der *Ersten Ebene*.

Klassifikation (FirstLayer)

Die Mustererkennung, d. h. die *Erste Ebene*, wird durch die Klasse *FirstLayer* implementiert, welche eine Unterklasse von *BlackBoardService* ist. Als solche wird sie beim Start am BlackBoard registriert und empfängt Objekte vom Typ VoiceFeatureObject. Die Klasse FirstLayer an sich ist jedoch abstrakt, d. h. sie selbst wird nicht instanziiert, sondern ihre Unterklassen (vgl. Abbildung 11.5). Im Folgenden werden die wesentlichen Funktionen des FirstLayers beschrieben.

Laden und Speichern von Klassifizierern

Die Unterklassen von FirstLayer können Klassifizierer entweder neu trainieren oder – sofern das Training bereits erfolgt ist – aus einer Datei laden, wobei der Dateiname durch das Attribut *classifierFile* repräsentiert wird. Das Laden und Speichern erfolgt mithilfe der Methoden *loadClassifier()* bzw. *saveClassifier()*. Die Existenz der Datei wird vor dem Laden überprüft. Falls sie nicht vorhanden ist, wird der Klassifizierer automatisch neu trainiert.

Mithilfe der Methode *saveMetaData()* wird zusätzlich eine zweite Datei angelegt, die Angaben über den Klassifizierer enthält, wie Datum des Trainings, Anzahl der Trainingsinstanzen und Genauigkeit. Darüber hinaus beinhaltet sie eine Zuordnung von Klassenindex und Klassenetikett, was deshalb besonders wichtig ist, weil sich diese Zuordnung durch das Auslesen der Datenbank in zufälliger Reihenfolge bei jedem Training ändern kann. Die Metadaten werden beim Laden des

Klassifizierers eingelesen.

Training und Evaluation

Das Training eines Klassifizierers wird durchgeführt, wenn keine passende Datei vorhanden ist oder beim Start des Servers ein entsprechender Parameter gesetzt wurde (dies kann auch über die grafische Benutzeroberfläche erfolgen). Die Trainingsdaten werden mit *readInstances-FromDatabase()* direkt aus der Datenbank gelesen, wobei das SQL-Statement, wie in Abbildung 11.6 dargestellt wird, aus den Attributen *table* (Datenbanktabelle), *featureNames* (Liste der benötigten Merkmale) und *filter* (Beschreibung der Trainingssätze) zusammengesetzt wird. Die Einschränkung der Gesamtanzahl der Trainingsinstanzen (*limit*) wird verwendet, um ein verkürztes Training zu Testzwecken durchzuführen. Nach dem Auslesen der Datensätze wird das Training mithilfe der Methode *trainClassifier()* angestoßen, nach dessen erfolgreichem Abschluss die Klassifizierer-Datei zusammen mit den Metadaten abgespeichert wird.

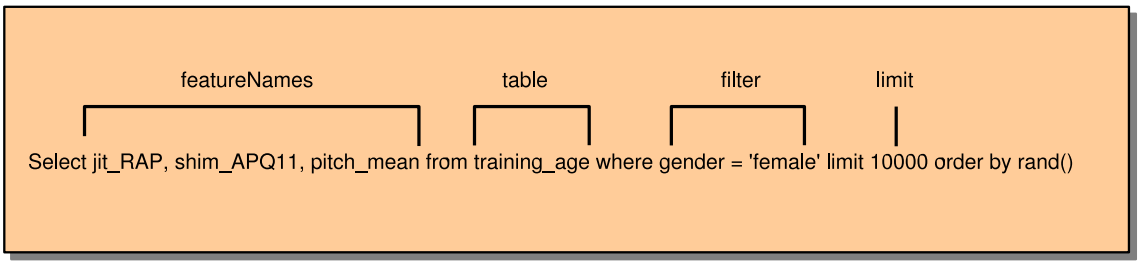


Abbildung 11.6: Beispiel eines SQL-Statements zur Auswahl der Trainingsinstanzen.

Die Methode *evaluateClassifier()* stößt eine zehnfache Kreuzvalidierung des Klassifizierers an. Diese erfolgt ausschließlich im Zusammenhang mit einem Training und kann über die Benutzeroberfläche hinzugeschaltet werden. Das Ergebnis der Evaluation kann über die grafische Repräsentation des fertigen Klassifizierers angesehen werden und wird darüber hinaus zusammen mit den anderen Metadaten abgespeichert. Auf Basis dieser Angaben werden u.a. die CPTs der Dynamischen Bayes'schen Netze auf der Zweiten Ebene konfiguriert.

Aktivierung von Analyseskripten

Neben der Auswahl der Datenbankfelder für das Training sind die Angaben in *featureNames* relevant für die Aktivierung von Analyseskripten, mithilfe derer die Merkmale extrahiert werden. Für jeden Eintrag in der Liste der Merkmale wird ein passendes Skript gesucht und dieses ggf. aktiviert.

Klassifikation

Alle Unterklassen von FirstLayer, die sich am BlackBoard registriert haben, erhalten Objekte vom Typ VoiceFeatureObject. Ein einzelnes VoiceFeatureObject enthält jedoch nicht alle Merkmale, die von einem bestimmten Klassifikator benötigt werden. Daher werden die *featureNames* des

Objektes mit denjenigen des Dienstes verglichen und bei Übereinstimmung die entsprechenden Werte gelesen. Sind alle Werte vorhanden, wird die Klassifikation angestoßen, ansonsten wartet der Dienst auf weitere VoiceFeatureObjects.

Das Ergebnis der Klassifikation wird in Form eines Objekts vom Typ *FirstLayerResultObject* auf das BlackBoard geschrieben. Dabei ist nicht nur eine Referenz auf den Sprecher notwendig (SpeakerObjekt), sondern auch auf die Äußerung (*UtteranceObject*). Auf der Zweiten Ebene werden die Ergebnisse bezüglich eines UtteranceObjects in einer Zeitscheibe zusammengefasst.

Besonderheiten der Unterklassen

Die wesentlichen Bestandteile eines Klassifizierers werden von der abstrakten Klasse FirstLayer geerbt – die Unterklassen überschreiben lediglich die für sie besonderen Attribute. Diese Methode hat den Vorteil, dass der Aufwand für die Implementierung eines neuen Klassifizierers sehr gering ist: Die jeweiligen Klassen umfassen in der Regel nur wenige Zeilen Code. Entsprechend gering ist der Aufwand zum Testen einer bestimmten Klassifikationsmethode.

Um dies zu verdeutlichen sei noch einmal auf die Vererbungshierarchie in Abbildung 11.5 verwiesen: Die direkten Unterklassen von FirstLayer, nämlich *SpeakerClassifier* und *ContextClassifier*, definieren eine Trainingstabelle, deren Unterklassen definieren die Klassenbezeichnungen und Merkmalsets und deren Unterklassen wiederum (die Blätter des Baumes) definieren einen Klassifizierer. Wie in Abbildung 11.5 angedeutet wird, ist die Anzahl der Klassifikationsmethoden beliebig. Dadurch ist ein unaufwendiges Hinzufügen, Testen und Entfernen von unterschiedlichen Klassifizierern möglich, was sich in der Praxis als sehr nützlich erwiesen hat.

Visualisierung

Die Klassifizierer erben von FirstLayer eine grafische Repräsentation aus mehreren Oberflächen, welche in Form von Registermappen angeordnet sind: (1) *Commentary* ist eine HTML-Oberfläche, die zur Erläuterung des Klassifizierers verwendet werden kann. Es handelt sich dabei um statische *URLs*, die fest mit dem Objekt, d. h. in dem Fall dem Klassifizierer, verbunden sind. (2) *Classifier* stellt die grafische Darstellung des Klassifizierers an sich dar. Welcher Art diese Darstellung ist, hängt vom Klassifizierer ab: Abbildung 11.7 repräsentiert beispielsweise die eines Neuronalen Netzes. (3) Die Mappe *Scatterplot* enthält eine n-dimensionale Darstellung der zugrunde liegenden Trainingsdaten, wobei jeweils zwei Merkmale in einem zweidimensionalen Diagramm dargestellt werden (vgl. Abbildung 11.8). (4) Die Mappe *Log* enthält eine Oberfläche, auf der die Log-Meldungen in HTML dargestellt werden. Auf diesen Mechanismus wird in Abschnitt 11.1.2 genauer eingegangen.

SecondLayer

Die Zweite Ebene wird hauptsächlich durch die Klasse *SecondLayer* realisiert, einem BlackBoardService, der sich beim Start für Objekte vom Typ FirstLayerResultObject registriert. Seine Funktion besteht darin, die Klassifikationsergebnisse zu sammeln und als Beobachtungen in das Bayes'sche Netz einzutragen, welches mithilfe des Systems HUGIN (Madsen, Jensen, Kjarulff und

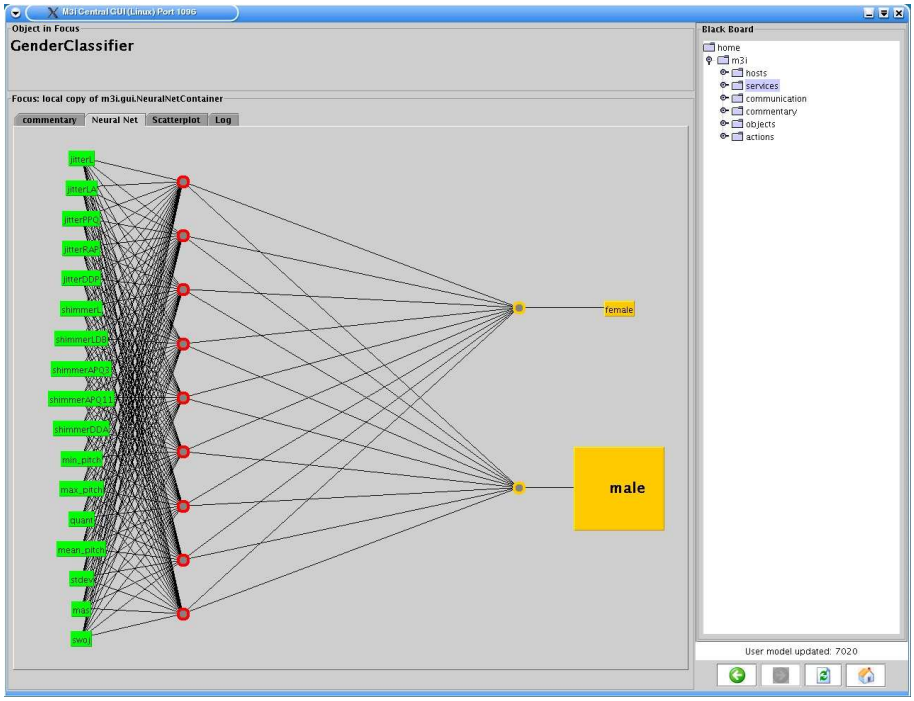


Abbildung 11.7: Classifier-Mappe einer FirstLayer-Visualisierung.

Lang, 2005) implementiert wurde. Wenn eine ausreichende Anzahl von Beobachtungen vorliegt, wird eine Rückwärtspropagierung angestoßen und die entsprechende Likelihood für die einzelnen Klassen im Benutzermodell (*UserModelObject*) gespeichert.

Neue Beobachtungen

Beim Eintreffen neuer Beobachtungen, d. h. Ergebnissen von Klassifizierern, wird zunächst überprüft, ob für den betreffenden Sprecher bereits ein Benutzermodell existiert. Dazu wird die Methode *getUserModel()* des SpeakerObjects genutzt, die – falls vorhanden – das UserModelObject zurückgibt oder andernfalls ein neues konstruiert. Darüber hinaus prüft der Dienst, zu welcher Äußerung die Beobachtung gehört, und legt ggf. eine neue Zeitscheibe an. Die Methode *addObservation()* identifiziert anschließend in der aktuellen Zeitscheibe denjenigen Knoten, der den Klassifizierer repräsentiert. Mithilfe von *setState()* wird derjenige Zustand instanziiert, der dem Klassifikationsergebnis entspricht.

In Abbildung 11.9 wird eine Zeitscheibe einer Sprecherklassifikation mit zwei-mal-zwei Klassen (JUNG und ALT, WEIBLICH und MÄNNLICH) dargestellt. Die unteren Knoten repräsentieren die Klassifikationsergebnisse aus der Ersten Ebene. Wenn der Geschlechtsklassifizierer beispielsweise das Ergebnis MALE zurückgibt, wird der entsprechende Zustand in dem Knoten unten links instanziiert. Anschließend wird der Klassifizierer für die aktuelle Zeitscheibe in die Liste *observations* eingetragen. Sobald die Liste alle obligatorischen Klassifizierer enthält, wird das Netz

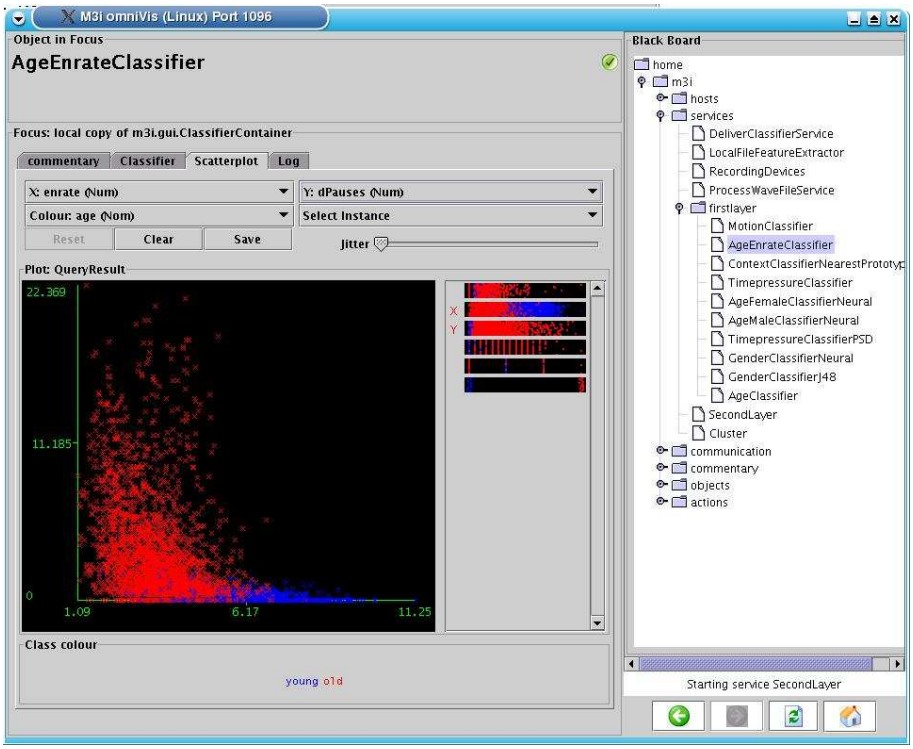


Abbildung 11.8: Scatterplot-Mappe einer FirstLayer-Visualisierung.

propagiert.

Neue Äußerungen

Wenn für einen vorhandenen Sprecher eine Beobachtung eintrifft, die zu einer neuen Äußerung gehört, wird die Methode *addUtterance()* aufgerufen, die eine neue Zeitscheibe konstruiert. Dabei werden mithilfe von *copyNodes()* sämtliche Knoten im Netz kopiert und die zwischen ihnen bestehenden Verbindungen mit *createEdges()* auf die neue Zeitscheibe übertragen. Die Knoten des Netzes werden hierfür aus der HUGIN-Repräsentation in ein *Array* geladen, wo sie der Reihe nach abgearbeitet werden. Dabei werden nur diejenigen Knoten betrachtet, die aus der unmittelbar vorangehenden Zeitscheibe stammen. Zusätzliche Kanten werden zwischen den Verbindungsknoten der Zeitscheiben hergestellt, die mithilfe von *isConnectionNode()* identifiziert werden. Neben den Knoten und den Verbindungen müssen auch die CPTs in die neue Zeitscheibe übertragen werden. Dazu wird die Methode *copyCPT()* verwendet, welche die Wahrscheinlichkeitstabelle eines Knotens aus der alten Zeitscheibe ausliest und sie auf den entsprechenden neuen Knoten überträgt.

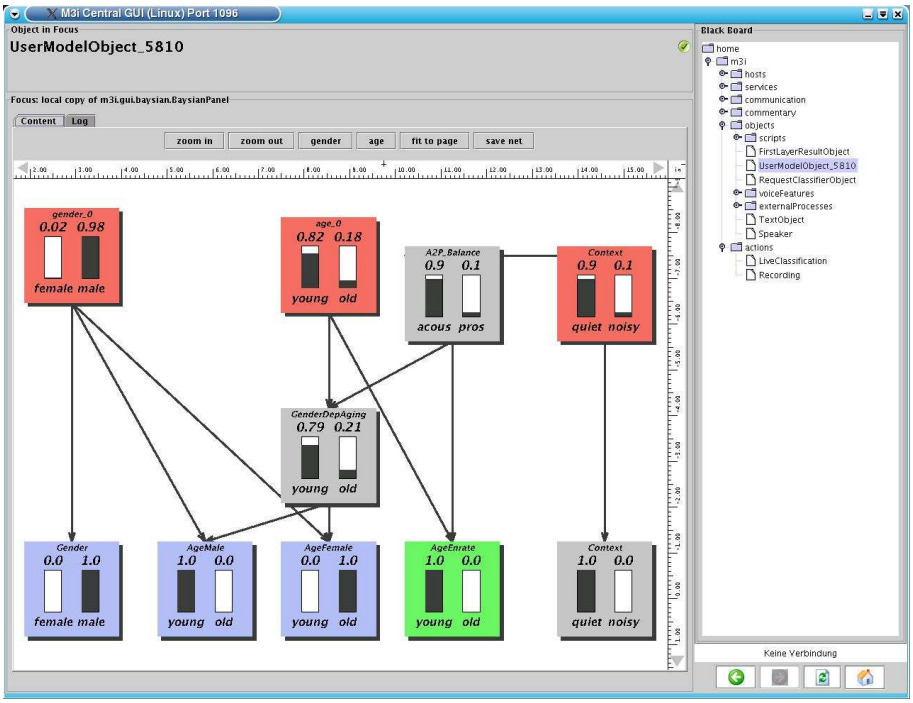


Abbildung 11.9: Grafische Darstellung der ersten Zeitscheibe des Bayes'schen Netzes (Zweite Ebene).

Repräsentation des Benutzerprofils

Das Ergebnis der Zweiten Ebene, also die Likelihood-Werte derjenigen Knoten, die in dem Beispiel in Abbildung 11.9 rot dargestellt werden, muss zur Auslieferung an die Applikation in eine geeignete Repräsentation überführt werden. Dazu wird die Methode *updateProfile()* verwendet, die in der aktuellen AGENDER-Version eine Merkmalsstruktur in Form einer Zeichenkette erzeugt, wie z. B. *user::gender::female*. Diese wird dann über das Protokoll an den m3i Client übertragen. Wird von der Applikation eine andere Form benötigt, muss die Methode updateProfile() entsprechend angepasst werden.

11.1.2 Generelle Aspekte

Generelle Aspekte der Architektur des m3i Servers stellen diejenigen Komponenten dar, die nicht spezifisch für die Sprecherklassifikation sind, d. h. auch für die Implementierung anderer serverseitiger Dienste verwendet werden können. Zu den wichtigsten allgemeinen Komponenten zählen die Kommunikation zwischen Server und Client sowie das eigentliche Blackboard mit dazugehörigen Kontrollstrukturen.

Kommunikation mit dem m3i Client

Die Kommunikation mit dem m3i Client basiert auf einer *tcp/ip*-Netzwerkverbindung. Der Austausch von Informationen wird mithilfe eines speziellen übergeordneten Kommunikationsprotokolls geregelt, welches im Folgenden genauer erläutert wird. Die Kommunikation stellt bezüglich der Quellcode-Organisation ein separates *Paket* dar und bezüglich der Laufzeit einen separaten *Thread*. Zu dem Paket gehören die Klassen *Connection*, *Protocol* und *ProtocolManager*.

Die Klasse Connection implementiert im Wesentlichen eine standardmäßige *Socket*-Verbindung. Wenn diese zur Laufzeit erfolgreich aufgebaut wurde, konstruiert Connection ein Objekt vom Typ *Client*, welches das verbundene Gerät auf dem Blackboard repräsentiert und dessen Eigenschaften (Gerätetyp, Betriebssystem etc.) kapselt.

Die Klasse Protocol dient dazu, die Kommunikation zwischen m3i Client und m3i Server zu regeln. Zu einem gegebenen Zeitpunkt befindet sich das Protokoll stets in einem eindeutigen Zustand, in dem Informationen in fest vorgeschriebener Form vom Klienten gelesen bzw. an den Klienten gesendet werden. Abbildung 11.10 stellt ein Diagramm der verschiedenen Zustände des Protokolls dar.

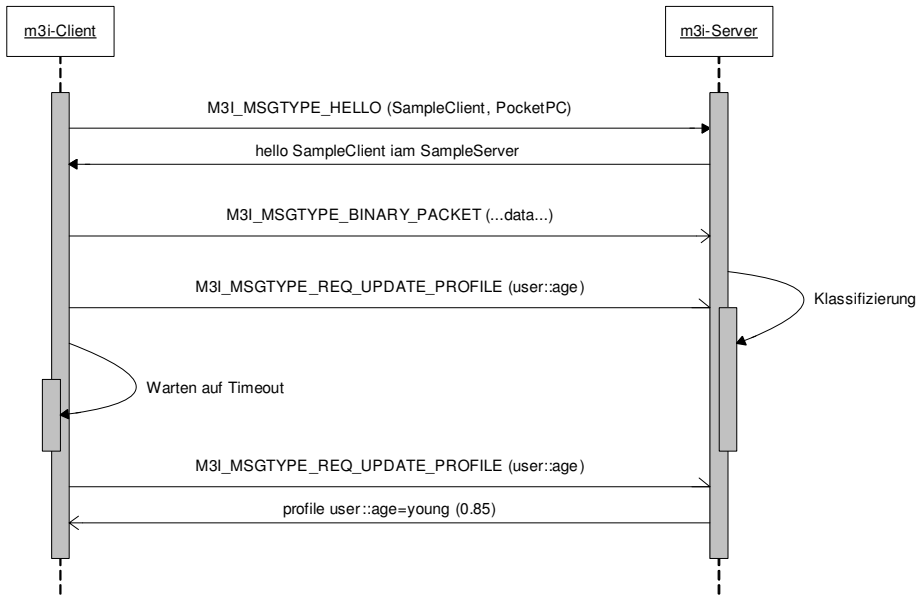


Abbildung 11.10: Sequenzdiagramm der Kommunikation zwischen m3i Client und m3i Server.

Der ProtocolManager, der als separater Thread läuft, verbindet Protocol und Connection miteinander. Dazu erzeugt er beim Programmstart eine Instanz von Connection und nach dem Aufbau der Verbindung eine Instanz von Protocol. Bricht die Verbindung ab, was kontrolliert durch eine Benutzeraktion oder unkontrolliert durch Abreißen der Netzwerkverbindung erfolgen kann, beendet der ProtocolManager die aktuelle Instanz von Connection und erzeugt eine neue, die dann wieder auf ein Signal vom m3i Client wartet.

BlackBoardObjects

Alle Objekte, die auf das BlackBoard geschrieben werden, müssen Unterklassen von *BlackBoard-Object* sein, von welchem sie die für ihre Verwaltung notwendigen Attribute und Methoden erben. Die Methode *log()* beispielsweise verwaltet für jedes Objekt ein HTML-Dokument, an das zeilenweise Log-Meldungen angehängt werden können und das zur Ansicht angeboten wird, wenn sich das entsprechende Objekt auf der grafischen Oberfläche im Fokus befindet. Der Vorteil dieser dezentralen Organisation der Log-Meldungen besteht darin, dass alle Meldungen, die ein bestimmtes Objekt betreffen, auf einen Blick sichtbar sind. Die Verwendung von HTML hat darüber hinaus den Vorteil, dass sie auf übersichtliche Art gestaltet werden können (vgl. Abbildung 11.11).

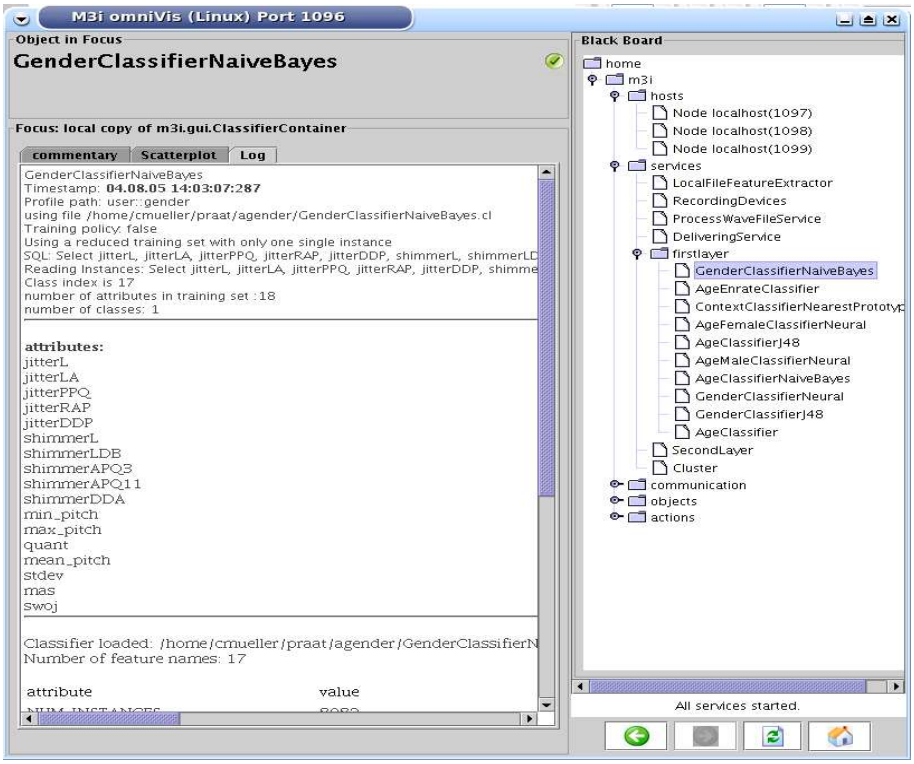


Abbildung 11.11: Log-Meldungen in HTML-Form.

Darüber hinaus werden alle BlackBoardObjects mit einem *Zeitstempel* versehen, der bei der Initialisierung mithilfe von *setTimestamp()* gesetzt wird. Der Zeitstempel dient zur Kontrolle des Alters eines BlackBoardObjects und wird automatisch als erster Eintrag in dessen Log-Dokument geschrieben (vgl. Abbildung 11.11).

Der *Schlüssel* eines BlackBoardObjects dient zur eindeutigen Kennzeichnung und wird darüber hinaus auf der grafischen Oberfläche für den Eintrag im Objekt-Baum sowie für die Überschrift genutzt. Er wird bei der Initialisierung automatisch mithilfe von *setKey()* erzeugt und leitet sich aus dem Klassennamen des entsprechenden Objektes ab. Bei denjenigen Objekten, bei denen mehrere Instantiierungen vorgenommen werden, wird ein eindeutiger Zusatz angehängt (vgl. Abbildung 11.12).

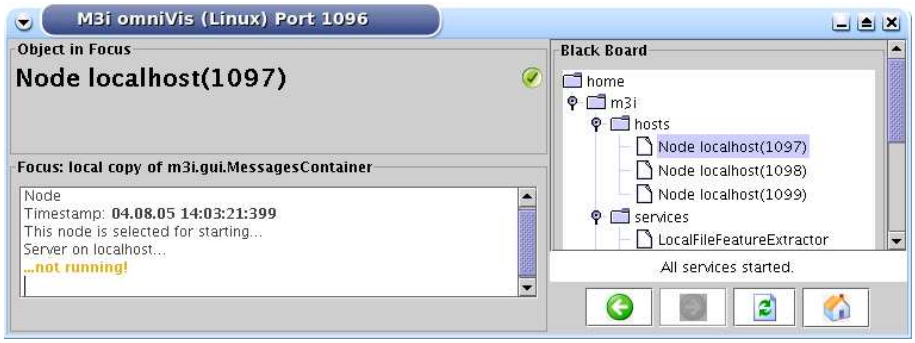


Abbildung 11.12: Oben und rechts: Im Titelbereich und im Baum werden die Schlüssel (keys) der Objekte verwendet. Links unten: Der Zeitstempel wird in lesbarer Form in die Log-Meldungen eingetragen.

Ein BlackBoardObject kann verschiedene *Zustände* annehmen: *aktiv / inaktiv* und *selektiert / nicht selektiert*. Bei der Initialisierung wird ein Objekt zunächst als inaktiv gekennzeichnet, was je nach Typ des Objektes eine andere Bedeutung hat: Ein Objekt vom Typ Script ist dann aktiv, wenn es zur Merkmalsextraktion verwendet wird – ein Service dagegen ist dann aktiv, wenn er erfolgreich gestartet wurde. Der Zustand wird auf der grafischen Oberfläche durch das Vorhandensein eines grünen Häkchen-Symbols für aktive Objekte repräsentiert. Der Zustand selektiert bedeutet, dass das entsprechende Objekt aktiviert werden soll. Im Fall der Skripte erfolgt eine Selektion dann, wenn es einen Klassifizierer gibt, der ein oder mehrere Merkmale dieses Scripts benötigt. Die Selektion bei Diensten ist von Benutzeraktionen abhängig: Über die Oberfläche zum Starten der *LiveClassification* (vgl. Abbildung 11.13) werden beispielsweise diejenigen Dienste aktiviert, die zur Sprecherklassifikation notwendig sind.

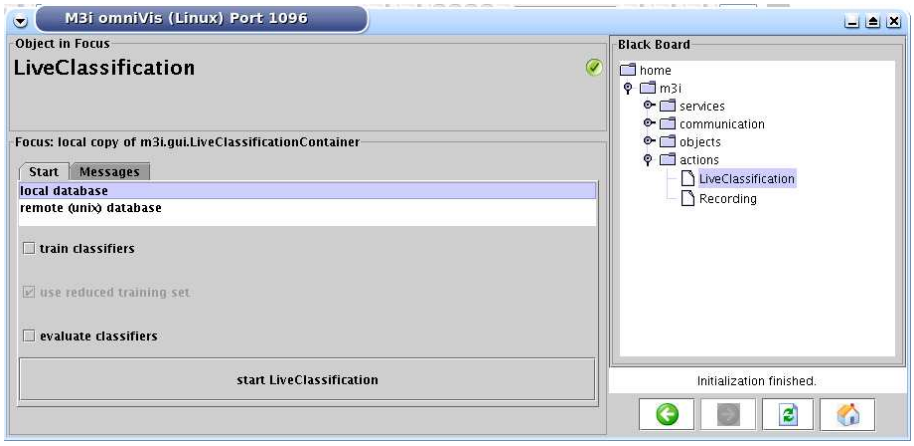


Abbildung 11.13: Oberfläche zum Starten der Dienste, die für die Sprecherklassifikation erforderlich sind.

Darüber hinaus verfügt das BlackBoardObject über Methoden, um Teile von sich in eine angeschlossene *Datenbank* zu schreiben. Zu diesen Methoden gehören *mySqlCreateTable()* zum Anlegen einer Datenbanktabelle, *toMySql()* zum Einfügen von Datensätzen und *mySqlExcecuteUpdate()* um ein beliebiges SQL-Statement auszuführen.

Schließlich stellt das BlackBoardObject eine einfache *Standard-Visualisierung* zur Verfügung. Dabei handelt es sich um eine einzelne Fläche, auf der die HTML Log-Meldungen angezeigt werden. Diese Darstellung wird verwendet, solange das Unterobjekt keine eigene definiert, was den Vorteil hat, dass bei der Einführung neuer Objekte nicht zwingend eine Visualisierung erzeugt werden muss.

BlackBoard

Zu den Aufgaben der Klasse BlackBoard gehört der Start von Diensten, die Registrierung und Deregistrierung der Dienste und die Zuteilung der BlackBoardObjects. Die *Initialisierung der Dienste* erfolgt beim Start des BlackBoards mithilfe der Methode *initializeServices()*. Welche Dienste gestartet werden sollen, wird in einer Konfigurationsdatei festgelegt. Auf diese Art und Weise kann zum Beispiel die aktuelle Auswahl der Klassifizierer angegeben werden.

Das *Hinzufügen von Objekten* erfolgt mithilfe einer separaten *add()*-Methode für jede Unterklasse von BlackBoardObject (z. B. *addUserModelObject()*). Je nach Objekt wird zunächst eine Konsistenzprüfung vorgenommen, wie z. B. im Fall des VoiceFeatureObjects, das auf seine Vollständigkeit hin überprüft wird. Die Objekte werden sowohl in einer Liste mit sequentieller Ordnung als auch in einer assoziativen Liste geführt. Auf Objekte, deren Schlüssel bekannt ist, kann über die assoziative Liste direkt zugegriffen werden – andernfalls wird die sequentielle Liste durchlaufen, bis das gewünschte Objekt gefunden wurde. In den meisten Fällen ist jedoch ein Zugriff auf gespeicherte Objekte nicht notwendig, da den Diensten neu eintreffende Objekte sofort zugesendet werden.

Bisher wurde vereinfachend davon gesprochen, dass sich die BlackBoardServices für BlackBoardObjects registrieren. Tatsächlich erfolgt die Kommunikation zwischen BlackBoard und den Diensten über *Ereignisse*, wobei von dem in *Java* vorhandenen Mechanismus der *Events* und *EventListeners* Gebrauch gemacht wird. Wenn beispielsweise ein VoiceFeatureObject auf das BlackBoard geschrieben wird, löst dies ein *NewVoiceFeatureObjectEvent* aus. Die BlackBoardServices implementieren EventListener für ein oder mehrere Ereignisse.

Für jedes Ereignis implementiert das BlackBoard eigene Methoden zur *Registrierung* und *Deregistrierung*, wie beispielsweise *registerForNewUserModelObjectEvent()* und *unregisterForNewUserModelObjectEvent()*. Als Argument wird ein entsprechender EventListener erwartet. Das BlackBoard implementiert des Weiteren für jedes Ereignis eine eigene *fire()*-Methode, die aufgerufen wird, um das Ereignis auszulösen. Für das NewUserModelObjectEvent gibt es dementsprechend die Methode *fireNewUserModelObjectEvent()*.

BlackBoardServices

Wie in Abbildung 11.14 dargestellt wird, erben – ähnlich wie die BlackBoardObjects – auch die Services die meisten Attribute und Methoden von einer übergeordneten Klasse. Dazu gehören die Methoden zur Registrierung und Deregistrierung am BlackBoard, zum Empfang und der Selektion von Ereignissen und zur Verarbeitung von BlackBoardObjects.

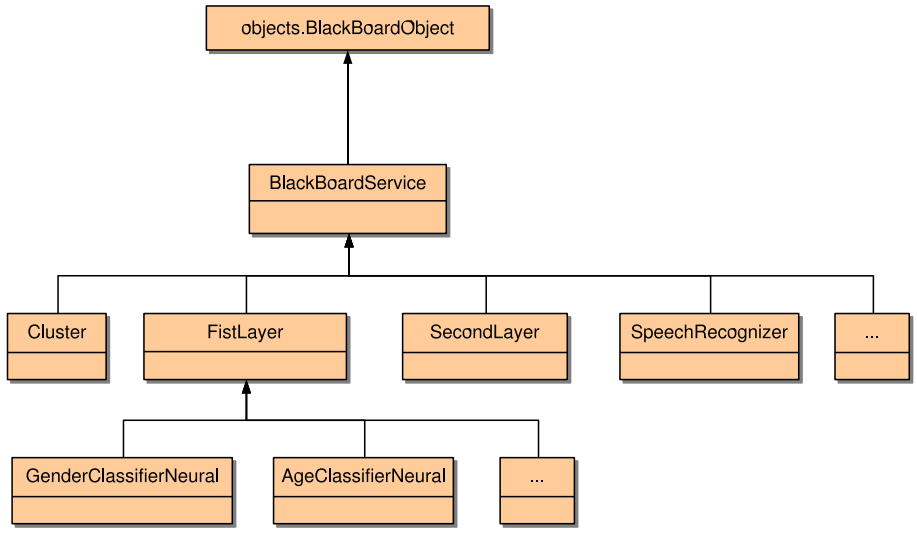


Abbildung 11.14: Vererbungshierarchie der Dienste (BlackBoardServices).

Bei der *Registrierung* eines Dienstes wird die entsprechende Methode des BlackBoards aufgerufen. Im Fall des Dienstes *FirstLayer*, der für die Klassifikation von Stimmmerkmalen zuständig ist, handelt es sich dabei beispielsweise um die Methode *registerForNewVoiceFeatureObjectEvent()*. Ein Dienst kann sich für mehrere Ereignisse registrieren, indem er mehrere Methoden aufruft. Bei der Registrierung werden in den meisten Fällen noch weitere Aktionen durchgeführt: Der Dienst wird beispielsweise aktiviert und erhält eine Repräsentation auf der grafischen Benutzeroberfläche. Die Deregistrierung erfolgt nach dem gleichen Prinzip: Der Dienst ruft die entsprechende Methode des BlackBoards auf, z. B. *unregisterForNewVoiceFeatureObjectEvent()*. Der Status des Dienstes wird dabei auf inaktiv gesetzt.

Für die Verarbeitung jedes *Ereignisses*, für das sich ein Dienst registriert hat, muss eine entsprechende Methode vorhanden sein. Im Fall von *FirstLayer* ist dies die Methode *newVoiceFeatureObjectEventOccured()*. Diese Methoden dienen dazu, zu entscheiden, ob das jeweilige Ereignis tatsächlich für den Dienst relevant ist. Dieser Mechanismus wurde eingeführt, um den Aufwand für die Formulierung von Events und EventListeners gering zu halten. Die Verarbeitung der BlackBoardObjects erfolgt mithilfe der Methode *run()*, deren Funktionsweise weitestgehend frei und von Dienst zu Dienst verschieden ist. Eine Gemeinsamkeit besteht darin, dass am Ende der Verarbeitung ein neues Objekt konstruiert und auf das BlackBoard geschrieben wird.

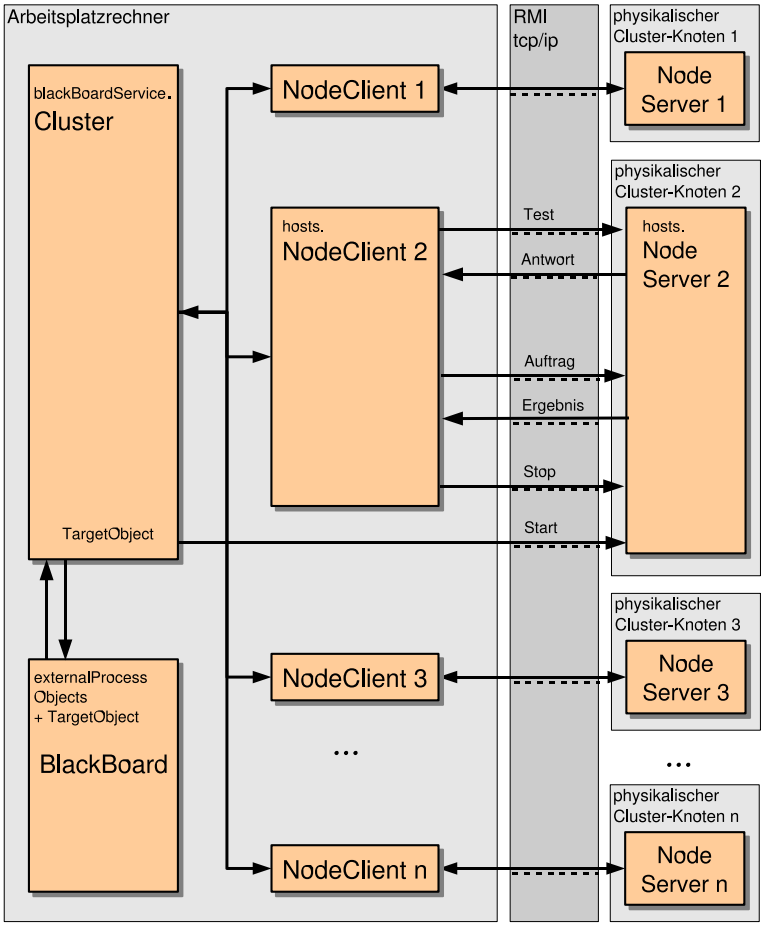


Abbildung 11.15: Ausschnitt aus der Architektur des Servers zur Verwaltung externer Prozesse auf dem Cluster.

Verwaltung externer Prozesse auf dem Cluster

Abbildung 11.15 stellt den Ausschnitt aus der Architektur des m3i Servers dar, der zur Verwaltung von externen Prozessen auf dem Cluster dient. Von Seiten des Servers erfolgt die Beauftragung externer Prozesse über den Dienst *Cluster*, der am BlackBoard für Objekte vom Typ *ExternalProcessObject* registriert wird. Diese Objekte werden zum Beispiel von VoiceFeatureObjects erzeugt (vgl. Abschnitt 11.1.1) und bestehen im Wesentlichen aus einer Liste von Kommandos.

Auf den (physikalischen) Cluster-Knoten laufen separate Prozesse, die *NodeServers*, welche auf dem Server durch die *NodeClients* repräsentiert werden. Die Kommunikation zwischen ihnen erfolgt über *Java RMI* (*Remote Method Invocation*), das auf einer *tcp/ip*-Verbindung aufsetzt (vgl. z. B. Grosso, 2001). Die RMI-Verbindung wird in Abbildung 11.15 durch die gestrichelten Linien angedeutet.

Bei der Initialisierung des *Clusters* werden sämtliche Knoten getestet. Falls die vorgeschrie-

bene Antwort erfolgt, wird der *NodeClient* aktiviert und erwartet Aufträge vom Cluster. Bleibt die Antwort jedoch aus, bleibt der Knoten zunächst inaktiv und eine entsprechende Warnmeldung wird geschrieben. Wie in Abbildung 11.15 bereits angedeutet wird, erfolgt der Start der *NodeServers* nicht über die *NodeClients*, da zu diesem Zeitpunkt noch keine RMI-Verbindung besteht. Diese wird aufgebaut, indem der Cluster über *SSH* die Prozesse auf den einzelnen Rechnern startet. Anschließend wird ein erneuter Test durchgeführt und der Knoten bei Erfolg aktiviert. Im Gegensatz dazu kann der Stopp-Vorgang über RMI veranlasst werden.

Nach erfolgreicher Aktivierung des Dienstes Cluster und mindestens eines Knotens (in der Regel mehrerer) steht der Dienst zur Ausführung externer Prozesse zur Verfügung. Dazu empfängt er vom BlackBoard Objekte vom Typ *ExternalProcessObject*, zu deren Attributen ein Kommando gehört, welches ausgeführt werden soll. Genauer gesagt handelt es sich um eine Liste von Kommandos, die mindestens ein Element enthält. Kommandolisten mit mehreren Einträgen werden zum Beispiel von Multinode-Skripten erzeugt (vgl. Abschnitt 11.1.1). In diesem Fall werden mehrere Knoten mit der Ausführung beauftragt, wobei jeder Knoten jeweils ein Kommando erhält.

Die Verteilung der Prozesse auf die Knoten erfolgt mithilfe der Methode *getNextNode()*. Dabei wird der Zeiger auf den Knoten k gesetzt, der zuletzt einen Auftrag erhalten hat. Wenn ein neues Objekt beim Cluster ankommt, wird das entsprechende Kommando an den Knoten $k+1$ gesendet und anschließend der Zeiger auf $k+1$ gesetzt. Bei einer Kommandoliste mit n Einträgen werden die Knoten $k+1, \ldots, k+n$ ausgewählt und der Zeiger auf $k+n$ gesetzt. Ist der Zeiger an dem letzten Element der Liste angekommen, wird der erste Knoten ausgewählt. Andere Verfahren, die die Last zwischen den Knoten besser verteilen, sind an der Stelle jedoch ebenfalls denkbar: Es könnte z. B. ein Modul in die Architektur integriert werden, das die jeweilige Last der einzelnen Knoten feststellt.

Die Ergebnisse externer Prozesse sind in der Regel Zeichenketten, die aus reellwertigen Zahlen und einem Trennzeichen bestehen. Damit sie von den nachfolgenden Diensten weiterverarbeitet werden können, müssen die Zeichenketten geparst werden.

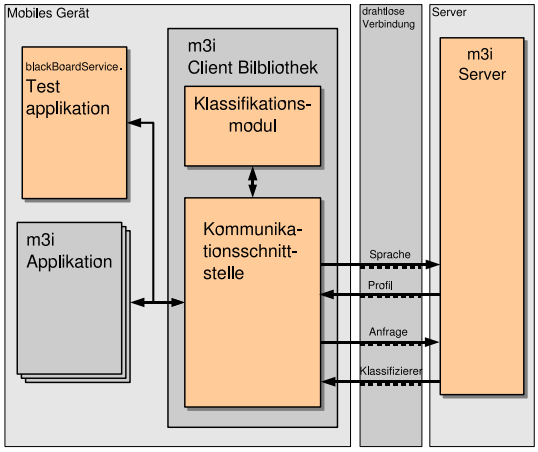


Abbildung 11.16: Die m3i Client-Architektur.

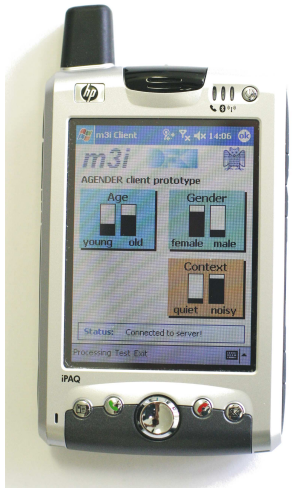


Abbildung 11.17: Die m3i Client-Testanwendung.

11.2 Der m3i Client

Die m3i-Gesamtarchitektur (vgl. Abbildung 11.16, Seite 225) sieht vor, dass der Benutzer mit einem mobilen Gerät, wie z. B. einem *Pocket-PC* oder einem *Smart-Phone* interagiert. Die Klassifikation der Stimme und des Kontextes soll auf dem Server erfolgen, falls eine breitbandige Netzwerkverbindung zur Verfügung steht. Falls nicht, soll der Client jedoch ebenfalls in der Lage sein, eine Klassifikation durchzuführen, wenn auch mit suboptimalem Ergebnis. Die Entwicklung des *m3i Clients* wurde im Rahmen einer studentischen Arbeit durchgeführt (vgl. Feld, 2006). Die wichtigsten implementatorischen Aspekte werden im Folgenden zusammenfassend beschrieben.

11.2.1 Architektur

Das Client-Modul besteht aus zwei Hauptteilen: einer Bibliothek, die in eine Anwendung integriert werden kann, und einer Testanwendung. Die Bibliothek umfasst die Funktionen zur Kommunikation mit dem Server sowie die eingebettete Merkmalsextraktion und Klassifikation. Die Testanwendung integriert die Bibliothek und stellt eine Oberfläche zu Testzwecken zur Verfügung (vgl. Abbildung 11.17).

Die Kommunikation mit dem Server stellt das Gegenstück zu der Verbindung dar, die in Abschnitt 11.1.2 aus Sicht des Servers beschrieben wurde: Auf einer zugrunde liegenden *Socket*-Verbindung, die in dem Fall mit *Embedded C++* statt mit *Java* programmiert wurde, setzt das *m3i*-Protokoll auf, das den Datenaustausch steuert. Nach dem Start überprüft der m3i Client zunächst, ob eine Verbindung zum Server möglich ist, und initialisiert diese gegebenenfalls. Nach der Aufzeichnung einer Äußerung über das interne Mikrofon wird diese dann mit der entsprechenden

Gerätebezeichnung	Prozessor	Taktfrequenz	Plattform
HP Jornada 568	ARM SA1110	206 MHz	Windows CE 3.0
HP Ipaq H6300	TI OMAP 1510	200 MHz	Windows CE 4.2
HP Ipaq H5450	Intel PXA250	400 MHz	Windows CE 4.2
HP Ipaq HX4700	Intel PXA270	624 MHz	Windows CE 4.21

Tabelle 11.1: Testgeräte für die Benchmarks.

Meta-Anweisung versehen und in Rohformat an den Server gesendet. Anschließend geht das Protokoll in einen Zustand über, in dem es auf eine Antwort, d. h. auf ein Sprecherprofil vom Server wartet. Darüber hinaus kann das Profil jedoch auch proaktiv vom m3i Client angefordert werden. Die Testanwendung stellt dafür eine entsprechende Schaltfläche bereit. Neben dem Senden von Sprachdaten und dem Empfang von Profilen sieht das Protokoll vor, Anfragen für bestimmte Klassifizierer an den Server zu senden. Diese Anfragen beschreiben den benötigten Klassifizierer gemäß Klassifikationszweck, Trainingsdatum (Neuheit) und Mindestgenauigkeit.

11.2.2 Merkmalsextraktion

Für die eingebettete Merkmalsextraktion wurde das Analyseprogramm PRAAT (vgl. Abschnitt 4.2.1) für die *Pocket-PC*-Plattform portiert. Als Entwicklungsgerät diente der *HP Jornada 568*, welcher mit einem *ARM SA1110*-Prozessor ausgestattet ist. Bezüglich des Betriebssystems wurde zwecks Kompatibilität mit den im Projekt *m3i* entwickelten mobilen Applikationen (vgl. Abschnitt 1.3) eine Kompatibilität mit Windows CE 3.0 und höher (Pocket-PC 2002 / Windows Mobile 2003) angestrebt. Da PRAAT nicht als Applikation auf dem *Pocket-PC* laufen, sondern ausschließlich als Merkmalsextraktor verwendet werden sollte, wurden zunächst sämtliche Komponenten entfernt, die die grafische Oberfläche und die Klangausgabe betreffen. Anschließend wurden alle hardwarenahen Funktionen im Quellcode so angepasst, dass sie den Anforderungen der Zielplattform entsprachen.

Der fertig portierte Quellcode wurde dahingehend modifiziert, dass er als Bibliothek in die Applikationen integriert werden konnte. Gegenüber einer Einbindung von PRAAT als externes Programm, wie es im Fall des Servers implementiert wurde, hat diese Variante vor allen Dingen Performanzvorteile, was in Anbetracht der weitaus geringeren Leistungsfähigkeit des mobilen Gerätes von zentraler Bedeutung ist.

Um die Performanz der ersten Version des fertigen Moduls zu testen, wurden Laufzeitanalysen auf verschiedenen Geräten durchgeführt (vgl. Tabelle 11.1). Zu diesem Zweck wurde eine Testanwendung entwickelt, welche aus einer Sprachprobe von 2.3 Sekunden Länge (8000 KHz, 16 Bit, mono) eine Reihe von Merkmalen extrahiert, die auch für AGENDER benötigt werden. Die Ergebnisse dieser Studie werden in Tabelle 11.2 zusammengefasst: Trotz der verhältnismäßig

Gerät	Benötigte Zeit
Jornada 568	82s
Ipaq H6300	106s
Ipaq H5450	59s
Ipaq HX4700	30s

Tabelle 11.2: Ergebnisse der initialen Laufzeittests mit einer Klangdatei von 2.3 Sekunden Länge (8000 Hz, 16 Bit, mono).

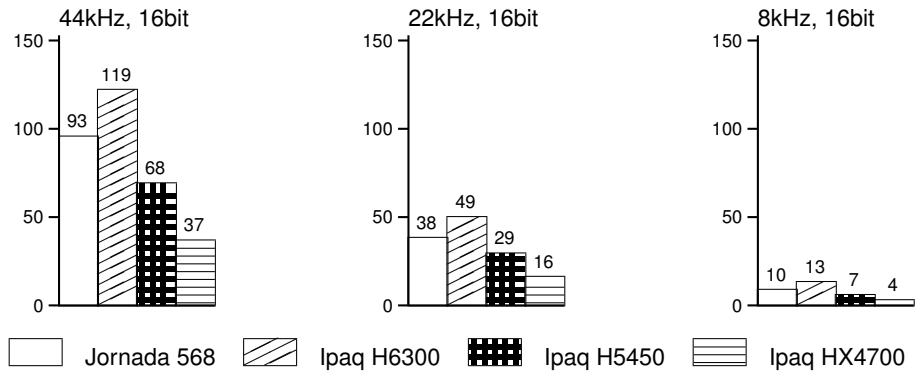


Abbildung 11.18: Laufzeitanalysen nach durchgeführter Optimierung mit einer Sprachprobe von 3.2 Sekunden Länge (mono).

geringen Dauer der Sprachprobe, benötigt die Merkmalsextraktion je nach Gerät zwischen einer halben Minute und fast zwei Minuten.

Eine genauere Analyse der Laufzeiten hat ergeben, dass 98 % der Rechenzeit für die Erstellung des *Pitch-Objektes* benötigt wird, welches in PRAAT die Basis für eine Vielzahl von Merkmalen darstellt. Daher erschien es sinnvoll, die Optimierungsbemühungen auf die fragliche Funktion zu konzentrieren, nämlich die Frequenzanalyse auf Basis der *Autokorrelationsmethode* (vgl. Abschnitt 2.2.2). Durch eine Änderung in dem besagten Algorithmus, bei der nur ein geringer Verlust der Genauigkeit in Kauf genommen werden musste, konnte die Gesamtlaufzeit erheblich verringert werden. Die Änderungen werden ausführlich in Feld (2005) beschrieben.

Abbildung 11.18 stellt die Ergebnisse eines Laufzeittests dar, der nach der Optimierung mit einer Sprachprobe von 3.2 Sekunden Länge durchgeführt wurde. Gegenüber der ursprünglichen Variante konnte eine Verbesserung um annähernd Faktor zehn erreicht werden. Dasselbe gilt auch für die Tests, die mit einer Sprachprobe von 8.4 Sekunden Länge durchgeführt wurden (vgl. Abbildung 11.19). Bei den neuerlichen Messungen wurden verschiedene Audioformate gegenübergestellt, wobei sich herausstellte, dass die Abtastfrequenz einen großen Einfluss auf die Verarbeitungszeit hat. Ausgehend von dem derzeitigen Stand der Optimierung ist Merkmalsextraktion auf

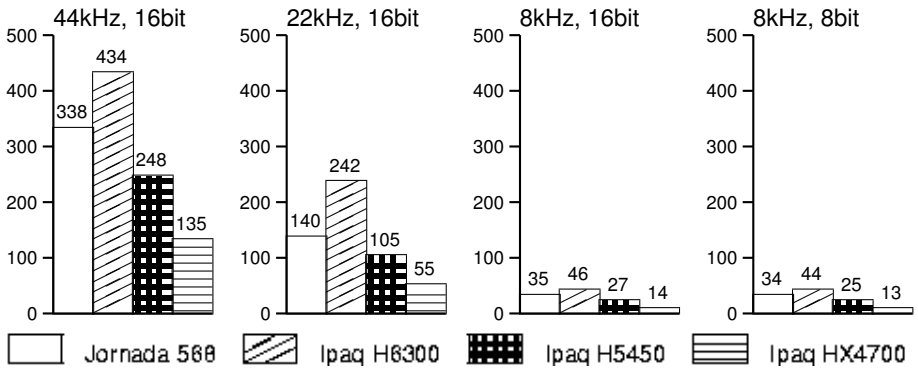


Abbildung 11.19: Laufzeitanalysen nach durchgeführter Optimierung mit einer Sprachprobe von 8.4 Sekunden Länge (mono).

dem Pocket-PC daher nur für Sprachproben mit maximal 22 kHz geeignet. Die Bitrate wirkt sich dagegen nur in geringem Maße auf die Verarbeitungszeit aus (vgl. Feld, 2005).

11.2.3 Klassifikation

Für eine Übertragung von Klassifikatoren auf die Pocket-PC-Plattform haben sich parametrische Klassifikationsmethoden (z. B. Gaussian-Mixture-Models) und Entscheidungsbäume als geeignet erwiesen. Der Vorteil parametrischer Klassifikationsmethoden besteht darin, dass lediglich der klassenspezifische Mittelwert μ, die klassenspezifische Standardabweichung Σ und ggf. ein Gewichtungsvektor auf den Pocket-PC übertragen werden müssen. Entscheidungsbäume sind ebenfalls gut *exportierbar*, da sie in eine Folge von *Wenn-Dann-Anweisungen* übersetzt und somit problemlos in eine beliebige Plattform integriert werden können. Aufgrund der höheren Klassifikationsgenauigkeit sind GMMs jedoch vorzuziehen (vgl. Feld, 2005).

12 Das Korpusanalyse-Werkzeug m3iCAT

Die Korpusanalyse stellte zunächst eine Komponente des m3i Servers dar, wurde jedoch im Laufe des AGENDER-Projektes ausgegliedert und als eigenständiges System auf Basis von PHP neu entwickelt. Dadurch konnte erstens die Komplexität des Servers in einem sinnvollen Rahmen gehalten werden und zweitens die Korpusanalyse an sich deutlich verbessert werden, so dass das System m3i CAT (Corpus Analyzing Toolkit) nun ein umfangreiches Werkzeug zur Durchführung und Überwachung von Korpusanalysen, zur Sichtung der Ergebnisse und zur Erzeugung von Trainingstabellen für die Klassifikation darstellt.

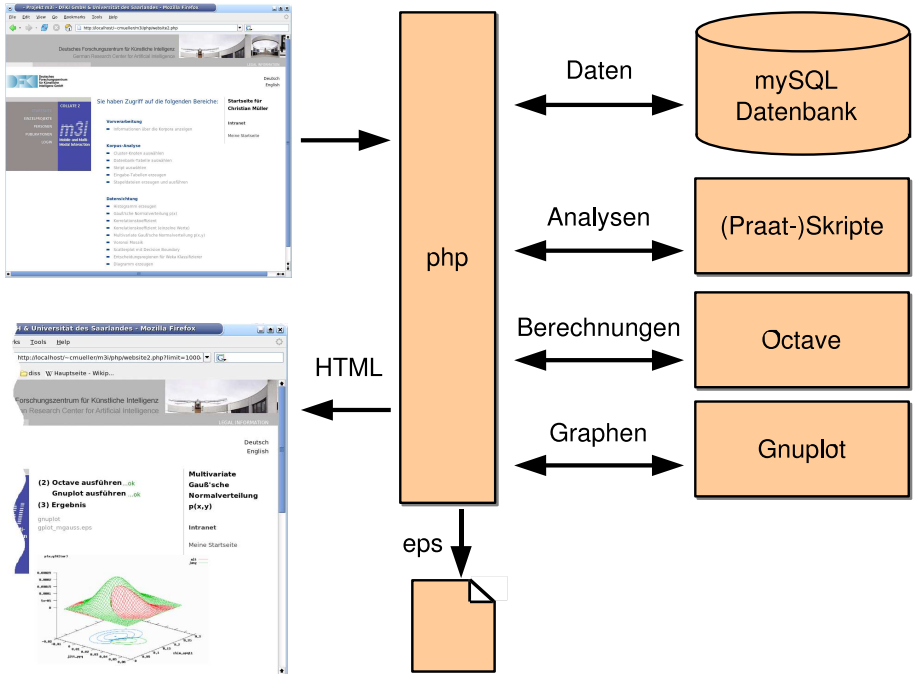


Abbildung 12.1: Aufbau des m3i CAT-Korpusanalyse-Systems.

In Abbildung 12.1 wird der Aufbau von m3i CAT dargestellt: Der Kern wird durch ein PHP-Programm gebildet, welches Anfragen vom Benutzer entgegennimmt und auf verschiedene ange-

gliederte Subsysteme verteilt. Die Analyse von Sprachproben erfolgt hauptsächlich mithilfe von PRAAT-Skripten, wobei es in m3i CAT keine prinzipielle Einschränkung auf eine bestimmte Art von Analyse-Skripten gibt. Die Ergebnisse der Analysen werden in einer MySQL-Datenbank[1] gespeichert. Zur Sichtung der Daten bietet m3i CAT eine Reihe von Funktionen, die unmittelbar auf SQL basieren. In den Fällen, in denen SQL nicht ausreicht, können Berechnungen mit OCTAVE[2] ausgeführt werden, einem frei verfügbaren Mathematik-Programm. Die Visualisierung der Daten erfolgt mithilfe von GNUPLOT[3], das ebenfalls frei verfügbar ist.

12.1 Domänenspezifische Aspekte

Wie bereits beim m3i Server, werden auch im Fall von m3i CAT zunächst die domänenspezifischen Aspekte des Systems beschrieben, d. h. diejenigen Komponenten, die sich unmittelbar auf die Sprecherklassifikation beziehen. Die Beschreibung der generellen Aspekte erfolgt im Anschluss daran in Abschnitt 12.2.

12.1.1 Vorbereitungen zur Korpusanalyse

Sprachkorpora sind – trotz bestehender Bemühungen zur Einführung von Konventionen (vgl. z. B. Schiel und Draxler, 2003) – sehr unterschiedlich organisiert, was die Verzeichnisstruktur, die Benennung der Dateien und die Verknüpfung zwischen Sprachproben und Annotationen betrifft. Um dennoch eine einheitliche Verarbeitung zu ermöglichen, werden in m3i CAT die zugrunde liegenden Korpora zunächst in einheitlichem Format in der Datenbank erfasst. Dazu wird die Verzeichnisstruktur des jeweiligen Korpus ausgelesen, die Annotationsdateien geparst und die Dateinamen inklusive Pfad werden zusammen mit den relevanten Sprechereigenschaften in die Tabelle CORPUS geschrieben. Für jeden der verwendeten Korpora (*BaS*, *Timit* und *Scansoft*) wird hierfür ein separater Parser benötigt. Bei diesem Vorgang werden auch die Angaben über die Sprecher in ein geeignetes Format übersetzt: Aus dem Geburtsdatum und dem Aufnahmedatum wird das Sprecheralter berechnet, aus dem sich wiederum die Altersklasse ergibt: Einträge mit $\text{ALTER} < 12$ erhalten die Angabe KINDER, Einträge mit $12 < \text{ALTER} < 20$ erhalten die Angabe JUGENDLICHE usw.

Nachdem alle Korpora in die Datenbank eingetragen wurden, können Analysen gestartet werden. Dazu wählt der Benutzer zunächst mithilfe eines von m3iCAT erzeugten Formulars ein Skript aus (vgl. Abbildung 12.2). Anschließend überprüft das System, welche Clusterknoten zur Verfügung stehen, indem es ein Test-Kommando absetzt, und präsentiert diese dem Benutzer zur Auswahl (vgl. Abbildung 12.3). Die Anzahl der Prozessoren pro Knoten kann bei der Konfiguration von m3i CAT angegeben werden.

Der nächste Schritt besteht darin, dass der Benutzer mithilfe eines SQL-Ausdrucks die Sprachproben auswählt, die analysiert werden sollen. Um eine parallele Bearbeitung der Datensätze zu

[1] http://www.mysql.com
[2] http://www.octave.org
[3] http://www.gnuplot.info

Abbildung 12.2: Auswahl eines Analyse-Skriptes.

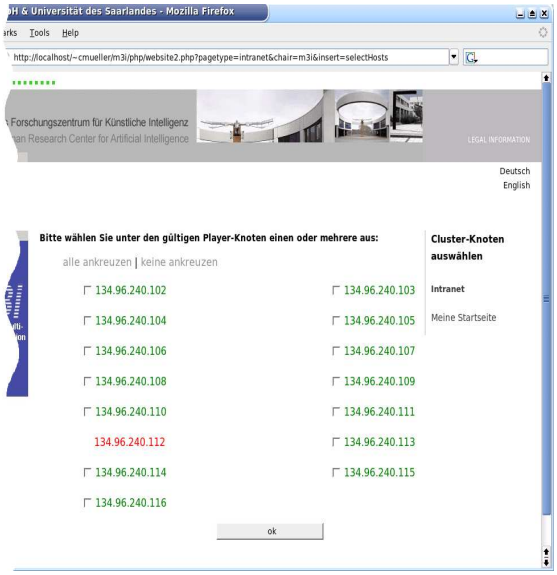


Abbildung 12.3: Auswahl der verfügbaren Cluster-Knoten.

ermöglichen, werden diese daraufhin vom System in mehrere Tabellen aufgeteilt, wobei für jeden ausgewählten Clusterknoten so viele Tabellen erzeugt werden, wie dieser Prozessoren besitzt. Dabei ist allerdings zu beachten, dass die Sprachproben in der Regel nicht gleich lang sind, so dass bei einer unausgeglichenen Verteilung einige Knoten längere Zeit benötigen als andere. Zur optimalen Parallelisierung der Analysen müssen die Daten so aufgeteilt werden, dass die Gesamtlängen der Äußerungen für jeden Knoten annähernd gleich sind. Da eine exakte Berechnung der Gesamtlänge jedoch mit erheblichem Aufwand verbunden wäre, bedient sich m3i CAT eines einfachen Prinzips: Die Datensätze werden nach Algorithmus 1 in einer zufälligen Reihenfolge aus der Tabelle CORPUS ausgelesen und auf die einzelnen Tabellen verteilt. Anschließend fasst m3i CAT das Ergebnis – wie in Abbildung 12.4 dargestellt – zusammen.

12.1.2 Durchführung der Analysen

Wie aus Abbildung 12.5 hervorgeht, besteht der Ablauf einer Korpusanalyse mit m3i CAT aus vier Schritten: Im ersten Schritt werden die nötigen Parameter des Skriptes aus unterschiedlichen Quellen zusammengestellt und die fertige Syntax anhand eines Beispiels überprüft. In Schritt zwei werden aus den Datenbanktabellen, die für jeden Knoten angelegt wurde (siehe oben), Stapelverarbeitungsdateien erzeugt, welche in Schritt drei ausgeführt werden. Die Ergebnisse der einzelnen Knoten werden in Schritt vier in SQL-Anweisungen übersetzt und in die zentrale Datenbank eingefügt.

Algorithmus 1 Aufteilung von Datensätzen auf die Clusterknoten.

```
k := Anzahl der ausgewählten Knoten
p := Anzahl der Prozessoren pro Knoten
s := Größe Eingabetabelle
s' := nächste kleinere Ganzzahl von (s / (m * k))
s'' := s - s'
Tabelle t := Kopie Eingabetabelle
for (i := 0 ; i < k ; i++) do
  b := Bezeichnung des i-ten Knotens;
  for (j := 0; j > p ; j++) do
  n := konkateniere (n, i);
    Tabelle n := Wähle zufällig s' Datensätze aus t aus
    Lösche aus t alle Datensätze in n
  od
od
Teile die restlichen s'' Datensätze auf die vorhandenen
Prozessoren der ausgewählten Knoten auf (dieser Schritt
ist trivial, da gilt: s'' < k * p).
```

Zusammenstellung der Parameter

Die Zusammenstellung der Parameter für ein bestimmtes Skript erfolgt auf der Basis des dazugehörigen Meta-Skriptes. Der Vorgang wird am Beispiel des Skriptes OVERLAY verdeutlicht, welches zur Erzeugung der Korpora für die Kontextklassifikation verwendet wird (vgl. Kapitel 4). In Abbildung 12.6 wird das Meta-Skript von Overlay dargestellt. Das Attribut `type` gibt an, ob es sich um ein Skript zur Korpusanalyse oder – wie in diesem Fall – zur Dateimanipulation handelt. Das Attribut `description` beinhaltet die Beschreibung des Skriptes, die unter anderem bei der Erzeugung des Auswahlformulars verwendet wird. Die Anzahl der Parameter wird über das Attribut `number_parameters` festgelegt. Overlay benötigt insgesamt sieben Angaben: 1. `prefix`, ein Präfix, mit dem das Verzeichnis des neuen Korpus markiert wird, um diesen von dem Originalkorpus zu unterscheiden; 2. `filename`, der Name der Datei, die im aktuellen Schritt bearbeitet wird; 3. `output_path`, der Ausgabepfad, der sich aus ursprünglichem Pfad, Präfix und Dateinamen zusammensetzt; 4. `raw_filename`, der Dateiname ohne Dateinamenserweiterung, der aufgrund einer Besonderheit von PRAAT für die meisten der Skripte benötigt wird; 5. `noise-file`, ein Verweis auf die Datei mit den Hintergrundgeräuschen; 6. `scaling_factor`, ein Faktor, der angibt, mit welcher Intensität der Hintergrund mit der Sprachprobe vermischt werden soll; 7. `sample_rate`, die Abtastfrequenz (Samplerate) der Ausgabe.

Das Attribut `Parameter_n_type` gibt an, ob sich der n-te Parameter von Datei zu Datei

Abbildung 12.4: Aufteilung der Eingabetabelle auf die einzelnen Cluster-Knoten.

verändert (`varying`), wie das bei `filename` der Fall ist, oder ob er für das gesamte zu verarbeitende Korpus gleich bleibt (`constant`). Bei variablen Parametern benötigt m3i CAT (im Gegensatz zu konstanten) eine Quellenangabe, die mithilfe des Attributs `Parameter_n_source` getätigt wird: Ein führendes '$'-Zeichen bedeutet, dass der Parameter direkt aus der Eingabetabelle gelesen werden kann, wobei das Vorhandensein der entsprechenden Spalte von m3i CAT überprüft wird. Ein führendes '&'-Zeichen bedeutet, dass der Parameter durch Anwendung der darauf folgenden Funktion ermittelt werden kann. Dabei kann es sich um eine Standard-PHP-Funktion handeln, wie im Fall von `Parameter_3_source`, oder um eine m3i CAT-eigene Funktion (`Parameter_4_source`). Die Existenz und Anwendbarkeit der Funktion werden ebenfalls vom System überprüft, indem es zufällig eine Zeile aus der Eingabetabelle ausliest und versucht, die Parameter schrittweise mit diesen Daten zu instantiieren (vgl. Abbildung 12.7).

Ein voranstehendes '%'-Zeichen bei der Angabe der Parameterquelle bedeutet, dass der Wert weder direkt aus der Datenbank gelesen noch durch eine Funktion ermittelt werden kann. In einem solchen Fall erzeugt m3i CAT ein Formular, mithilfe dessen der Benutzer den Wert angeben kann. Dabei werden verschiedene Formularfelder unterschieden: `text` erzeugt ein Textfeld, `file` ein Dateiauswahlfeld, `select` ein Listenfeld und `checkbox` ein Auswahlkästchen. Die Erzeugung des Formulars wird in Abschnitt 12.2.2 näher erläutert. Wenn ein Default-Wert angegeben wurde, wie z. B. im Fall von `Parameter_7_default`, werden die Formularfelder mit dem entsprechenden Wert vorbelegt. Das Attribut `syntactical` gibt an, ob es sich bei dem Parameter um einen „syntaktischen“ Parameter handelt, der beim Aufruf des Skriptes angeben wird, oder ob er wie im Fall des ersten Parameters lediglich zur Erzeugung anderer benötigt wird. Da m3i CAT standardmäßig annimmt, dass es sich um syntaktische Parameter handelt, fehlt diese Angabe ansonsten.

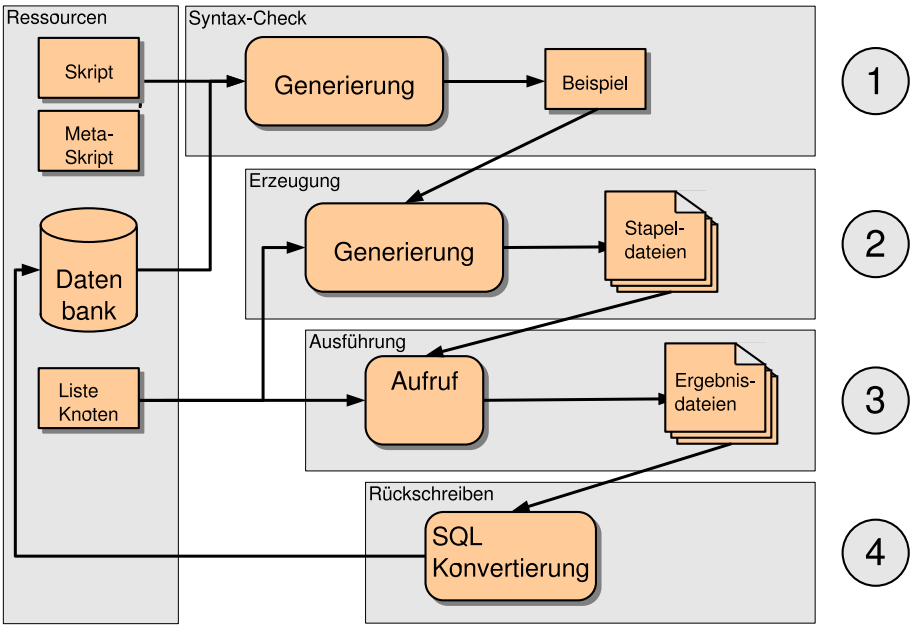


Abbildung 12.5: Ablauf einer Korpusanalyse mit Stapelverarbeitungsdateien.

Erzeugung und Ausführung der Stapelverarbeitungsdateien

In Schritt zwei wird auf Basis der Eingabetabellen und der Meta-Skripte für jeden Prozessor eine Stapeldatei erzeugt und anschließend von m3i CAT via *ssh* ausgeführt. Dabei ist zu beachten, dass es sich um Prozesse handelt, die über einen längeren Zeitraum – in der Regel mehrere Stunden – laufen. Da m3i CAT jedoch eine Web-basierte Anwendung ist, und die zulässigen Antwortzeiten daher auf wenige Minuten beschränkt sind, müssen die Prozesse im Hintergrund ausgeführt werden. Das System verwendet zu diesem Zweck das Unix-Kommando *at*, mithilfe dessen Prozesse zeitgesteuert ausgeführt werden können. Auf Basis der Verarbeitungsliste (*at-queue*) überwacht m3i CAT den Status der Hintergrundprozesse. Durch erneutes Laden der Seite kann der Benutzer überprüfen, ob die Analysen bereits abgeschlossen sind oder nicht. Dieses Prinzip, das auch bei anderen zeitintensiven Vorgängen angewendet wird, wird in Abschnitt 12.2.2 genauer beschrieben.

Einfügen der Ergebnisse in die Datenbank

Im vierten und letzten Schritt werden die Ergebnisse der Analysen in die Datenbank geschrieben, wobei m3i CAT erneut Gebrauch von den Angaben in den Meta-Skripten macht. In Abbildung 12.8 wird das Meta-Skript des Skriptes HARMONICITY dargestellt, welches zur Ermittlung der Harmonicity-to-Noise-Ratio verwendet wird. Die Angaben zu den Eingabeparametern wurden ausgespart, da hier lediglich die Ausgabewerte von Interesse sind. HARMONICITY erzeugt fünf Statistiken zur Harmonicity-to-Noise-Ratio: den Minimalwert, den Mittelwert, den Maximalwert und die Standardabweichung. Den Wert `class` erhält das Skript als Eingabe und gibt ihn

```
TYPE = file_manipulation

DESCRIPTION =
Überlagert eine gegebene Klangdatei mit
einem Hintergrundgeräusch. Art des
Hintergrundgeräusches und die Stärke der
Überlagerung können angegeben werden.

NUMBER_PARAMETERS = 7

PARAM_1_NAME = prefix
PARAM_1_TYPE = constant
PARAM_1_SYNTACTICAL = no
PARAM_1_SOURCE = %text
PARAM_1_DEFAULT = _overlay4113

PARAM_2_NAME = filename
PARAM_2_SOURCE = $filename
PARAM_2_TYPE = varying

PARAM_3_NAME = output_path
PARAM_3_TYPE = varying
PARAM_3_SOURCE = &getOutputPath
($filename, $prefix, 0)
```

```
PARAM_4_NAME = raw_filename
PARAM_4_SOURCE = &ereg_replace(".*/(.+)
\....?", "\\1", $filename)
PARAM_4_TYPE = varying

PARAM_5_NAME = noise_file
PARAM_5_SOURCE = %file
PARAM_5_TYPE = constant
PARAM_5_PREFIX = /
WW/home/cmueller/praat/noise/

PARAM_6_NAME = scaling_factor
PARAM_6_SOURCE = %text
PARAM_6_TYPE = constant
PARAM_6_DEFAULT = 0.5

PARAM_7_NAME = sample_rate
PARAM_7_SOURCE = %text
PARAM_7_TYPE = constant
PARAM_7_DEFAULT = 8000
```

Abbildung 12.6: Meta-Skript zum Skript *Overlay*, das zur Erzeugung der Korpora für die Kontextklassifikation verwendet wird.

unverändert aus. Das Attribut `Ouput_Value_5_Type` bestimmt den Datentyp: Im Gegensatz zu den Messwerten, die alle `double`-Werte sind, ist `class` vom Typ `varchar`. Da `double` als Standard-Datentyp angenommen wird, fehlt diese Angabe ansonsten.

Um die Ergebnisse in die Datenbank einzufügen, erzeugt m3i CAT eine Tabelle, deren Bezeichnung dem Namen des Skriptes entspricht. Anschließend parst das System die Ausgaben der Clusterknoten, konvertiert sie in INSERT-Kommandos und führt diese aus.

12.1.3 Datensichtung und Datenaufbereitung

Nach der Korpusanalyse werden die einzelnen Werte in der Tabelle RESULTS zusammengeführt, wobei die Dateinamen als Schlüssel dienen. Der Benutzer hat die Möglichkeit, die Ergebnisse mithilfe von SQL-Ausdrücken auszuwerten, indem beispielsweise die klassenspezifischen Mittelwerte und Standardabweichungen berechnet werden. Darüber hinaus bietet m3i CAT eine Reihe von Funktionen zur Datensichtung an, von denen die wichtigsten im Folgenden beschrieben werden: klassenspezifische Gauß'sche Normalverteilungen, Mittelwerttendenzen als Liniendiagramme, Streudiagramme und Entscheidungsgrenzen auf Basis verschiedener Klassifikationsverfahren.

Abbildung 12.9 stellt ein prototypisches Ablaufschema eines m3i CAT-Datensichtungsskriptes d dar, wobei die einzelnen Skripte durchaus an verschiedenen Stellen von diesem Schema abweichen können. In Schritt eins wird zunächst ein Formular erzeugt, mithilfe dessen der Benutzer die Parameter von d angeben kann. Die Werte werden in Form von *Cookies* gespeichert, was den Vorteil hat, dass sie über die Dauer der gesamten Sitzung erhalten bleiben. Wie in dem Beispiel aus

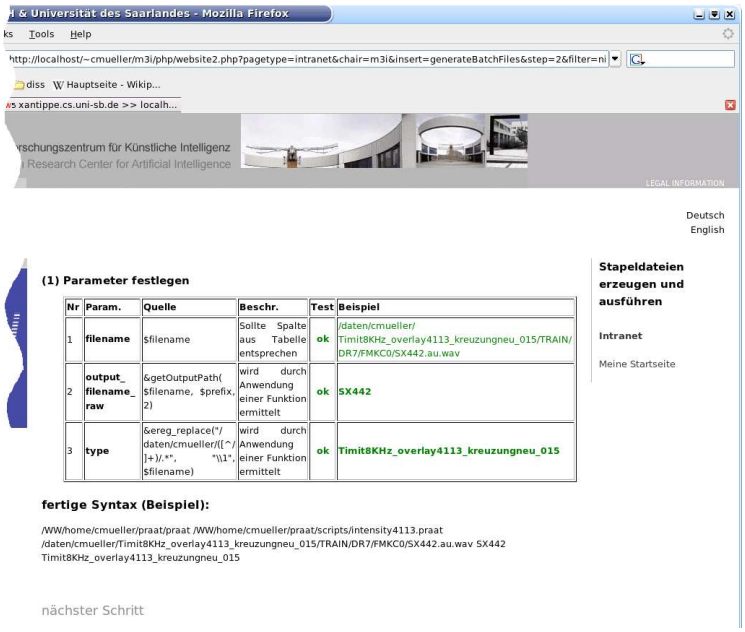


Abbildung 12.7: Test der Skriptsyntax anhand eines zufällig ausgewählten Beispieldatensatzes.

Abbildung 12.10 zu erkennen ist, können die einzelnen Werte auf Wunsch separat neu eingegeben werden. Die Erzeugung des Formulars erfolgt mithilfe einer Funktion, die in Abschnitt 12.2.2 genauer beschrieben wird, während d lediglich den Datentyp und die Default-Werte vorschreibt.

In Schritt zwei wird eine Auswahl von Datensätzen aus der Tabelle RESULTS gelesen, auf deren Basis die Sichtung erfolgen soll. Im Einzelfall kann die Auswahl auch die gesamte Tabelle umfassen. Schritt drei stellt für die meisten d eine Berechnung dar, wie z. B. von Gauß-Kurven oder Nullstellen von Entscheidungsgrenzen. Hierfür wird das frei verfügbare Mathematik-Programm Octave verwendet, das ähnlich wie das kommerzielle System *Mathlab* im Skript-Modus aufgerufen werden kann. Für jedes d wurden in m3i CAT eigene Octave-Skripte implementiert, allerdings zunächst in Form von Schablonen mit Platzhaltern. Zur Laufzeit lädt das System die passende Schablone, ersetzt die Platzhalter durch konkrete Werte, legt das fertig instantiierte Skript in einem temporären Verzeichnis ab und führt es anschließend aus. Das Ergebnis wird, z. B. in Form einer Matrix, ebenfalls in das temporäre Verzeichnis geschrieben. Die Visualisierung der Berechnungsergebnisse erfolgt mithilfe von Gnuplot. Das Verfahren ist ähnlich wie zuvor für Octave: Zunächst wird eine speziell für diesen Zweck erstellte Schablone geladen, und die Platzhalter werden mit konkreten Werten gefüllt. Zum Teil handelt es sich dabei um Werte, die aus der zuvor durchgeführten Berechnung stammen, zum Teil um solche, die als Parameter in Schritt eins eingegeben wurden (z. B. Größe des Diagramms oder Blickwinkel auf die Achsen). Andere Werte hängen wiederum direkt von den ausgewählten Datenbankfeldern ab (z. B. Achsenbeschriftungen oder Titel des Diagramms). Das fertige Gnuplot-Skript wird im temporären Verzeichnis gespeichert und ausgeführt. Die Skripte sind so ausgelegt, dass Gnuplot statt einer direkten Ausgabe

```
TYPE = corpus_analysis

[...]

NUMBER_OUTPUT_VALUES = 5
OUTPUT_VAL_1_NAME = mean
OUTPUT_VAL_2_NAME = minimum
OUTPUT_VAL_3_NAME = maximum
OUTPUT_VAL_4_NAME = standard_derivation
OUTPUT_VAL_5_NAME = class
OUTPUT_VAL_5_TYPE = varchar(50)
```

Abbildung 12.8: Meta-Skript zum Skript HARMONICITY, das zur Ermittlung der Harmonicity-to-Noise-Ratio verwendet wird.

auf den Bildschirm die Ergebnisse in eine Grafikdatei schreibt. Genauer gesagt wird die Grafik in den meisten Fällen in zwei verschiedenen Formaten erzeugt: im *png*-Format, welches im Browser dargestellt werden kann, und im *eps*-Format. Die *png*-Datei wird in Schritt sieben unmittelbar als <img>-tag in die HTML-Schablone eingesetzt. Die Postscript-Variante wird als Verweis angeboten.

Die Schritte eins bis sieben stellen den internen Ablauf dar und entsprechen nicht unmittelbar den Schritten, die der Benutzer durchführt. Nach der Bestimmung der Parameter gelangt dieser oftmals direkt zur Ergebnis-Seite. In anderen Fällen werden zunächst Zwischenergebnisse präsentiert, die der Benutzer mit einem „weiter“-Verweis quittiert.

Klassenspezifische Gauß'sche Normalverteilungen

In Abschnitt 7.2 wurde die Bayes'sche Entscheidungstheorie als eine statistische Grundlage von Verfahren des maschinellen Lernens eingeführt. In diesem Zusammenhang spielt die *Gauß'sche Normalverteilung* eine wichtige Rolle, da sie häufig als Grundlage für die Berechnung von klassenspezifischen Wahrscheinlichkeitsdichten dient. Die glockenförmige Normalverteilung stellt eine geeignete Ausgangsbasis dar, um eine Unterscheidbarkeit der Klassen mit den zur Verfügung stehenden Merkmalen zu überprüfen. Daher bietet m3i CAT diese Darstellung an und zwar sowohl in der univariaten als auch in der bivariaten Form.

Univariate Verteilungen

Anders als in dem in Abbildung 12.9 dargestellten Schema erfolgt die Erfassung der Parameter hier in mehreren Schritten: Zunächst bestimmt der Benutzer das zu betrachtende Merkmal, wobei er aus allen reellwertigen Datenbankfeldern der Tabelle RESULTS auswählen kann. Zur

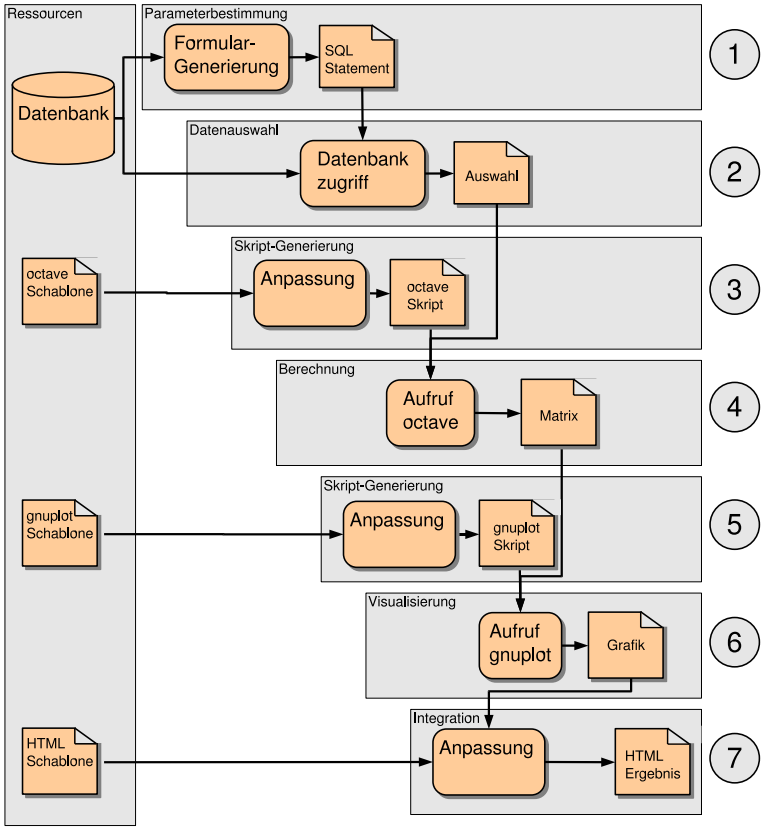


Abbildung 12.9: Typischer Ablauf einer Datenvisualisierung mit m3iCAT.

Bestimmung der Klassenbezeichnung, die als nächstes erfolgt, ermittelt das System alle nichtreellwertigen Spalten der Tabelle außer dem Dateinamen und bietet diese zur Auswahl an. In der aktuellen Version wird zusätzlich zu AGECLASS und GENDER für jede Gruppierung eine separate Klassenbezeichnung geführt (KW-KM-JW-JM-EW-EM-SW-SM, KWKMJW-EW-SW-EMSM usw.).

Anschließend ermittelt m3i CAT mithilfe von SQL-Anweisungen die klassenspezifischen Mittelwerte und Standardabweichungen des ausgesuchten Merkmals und trägt sie in ein Octave-Skript ein, das eine Implementierung der in Abschnitt 7.2 aufgeführten Formel für die univariate Gauß'sche Wahrscheinlichkeitsdichte darstellt. Dessen Ausgabe ist ein Vektor von Funktionswerten für einen informativen Wertebereich, dessen untere Grenze definiert wird als der niedrigste Mittelwert minus der dazugehörigen Standardabweichung multipliziert mit drei, und dessen obere Grenze definiert wird als der höchste Mittelwert plus die dazugehörige Standardabweichung multipliziert mit drei. Der Vektor wird mithilfe eines entsprechenden Gnuplot-Skriptes in eine grafische Darstellung überführt, die, wie oben beschrieben, als Pixelgrafik in die HTML-Seite integriert (vgl. Abbildung 12.11) und als Vektorgrafik zum Download angeboten wird.

Abbildung 12.10: Festlegung der Parameter zur Anzeige eines Streudiagramms mit Entscheidungsgrenzen.

Bivariate Verteilungen

Die Darstellung der bivariaten Gauß'schen Normalverteilung erfordert mehr Parameter als der univariate Fall: Erstens wird auf Basis der Tabelle RESULTS ein Formular zur Auswahl zweier Werte statt nur eines Wertes erzeugt, und zweitens werden neben der Klassenbezeichnung und der Anzahl der Datensätze noch eine Reihe weiterer Parameter benötigt, die die Form des Graphen betreffen: der Rand als ein Faktor der Standardabweichung, der Blickwinkel auf die Achsen und der Abstand des Diagramms von der z-Achse (vgl. Abbildung 12.12).

Die Berechnung der Verteilung basiert auf Gleichung 7.2 (Seite 119). Obwohl die Anzahl der Merkmale aufgrund der eingeschränkten Darstellbarkeit auf zwei begrenzt ist, gilt diese Einschränkung nicht grundsätzlich für die Berechnung der Werte: Das zugrunde liegende Octave-Skript kann auch für die Berechnung einer multivariaten Verteilung herangezogen werden.

Zunächst werden die klassenspezifischen Mittelwert-Vektoren mithilfe von *SQL-Statements* auf Basis der Datenauswahl ermittelt und in die Octave-Schablone eingetragen. Das Skript selbst ermittelt daraus jeweils die Kovarianzmatrix, die inverse Matrix und die Determinante. Der betrachtete Wertebereich wird bestimmt durch die Ränder, die als Parameter im Formular angegeben wurden. Für jeden der Datenpunkte (x,y) ermittelt Octave durch Anwendung von Gleichung 7.2 die Werte $z1$ (für Klasse 1) und $z2$ (für Klasse 2). Der größere der beiden Werte wird zusammen mit den xy-Koordinaten in die Matrix der entsprechenden Klasse eingetragen. In der jeweils anderen Matrix wird an der Stelle eine Null eingetragen, da unvollständige Matrizen von Gnuplot nicht korrekt verarbeitet werden können.

Nachdem der gesamte Wertebereich abgearbeitet wurde, werden die beiden Matrizen in ei-

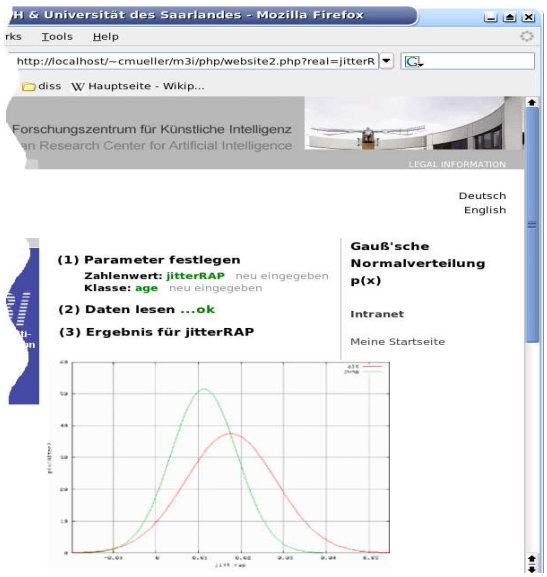


Abbildung 12.11: Univariate Gauß'sche Normalverteilung.

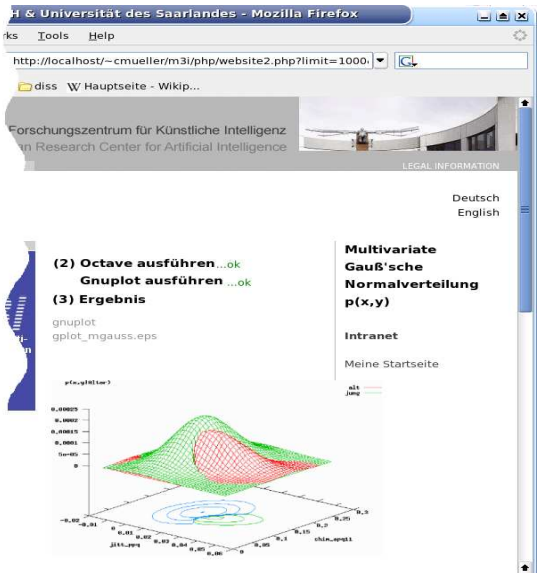


Abbildung 12.12: Bivariate Gauß'sche Normalverteilung.

ne Datei geschrieben, wobei sie als separate Datenblöcke gekennzeichnet werden. Die Gnuplot-Schablone wird mit den oben genannten Parametern zur Steuerung der Darstellung sowie den Achsenbeschriftungen und Klassenbezeichnungen gefüllt und als Skript abgespeichert. Gnuplot verarbeitet die beiden Blöcke der Eingabematrix separat, wodurch zwei übereinander angeordnete dreidimensionale Kurven entstehen, wie sie in Abbildung 12.12 dargestellt werden. Die Grafik kann ebenfalls im eps-Format abgespeichert werden.

Mittelwerttendenzen als Liniendiagramme

Während Gauß-Kurven zur Feststellung der Unterscheidbarkeit von Klassen auf Basis eines bestimmten Merkmals geeignet sind, können die Mittelwerttendenzen mithilfe von Liniendiagrammen übersichtlicher dargestellt werden. Die Erzeugung der beiden Diagramm-Typen unterscheidet sich in m3iCAT lediglich durch die Verwendung verschiedener Octave- und Gnuplot-Schablonen. Das hier verwendete Octave-Skript berechnet die statistische Signifikanz gemäß einem t-Test, wie er in Bosch (1987) beschrieben wird. Die Darstellung kann sowohl auf Basis der normalisierten als auch der nicht-normalisierten Werte erfolgen. Der Vorteil der Verwendung normalisierter Werte besteht darin, dass die Tendenzen unterschiedlicher Merkmale direkt miteinander verglichen werden können (vgl. Kapitel 5). In Abbildung 12.13 wird ein Beispiel für ein mit m3i CAT erzeugtes Liniendiagramm dargestellt. Die Tendenzen werden oberhalb der Linien angegeben, wobei in den Fällen, in denen sie nicht signifikant sind, ein entsprechender Hinweis in Klammern angehängt wird.

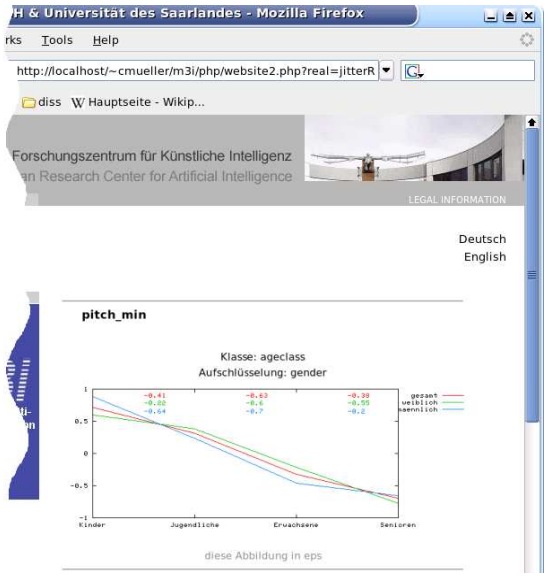
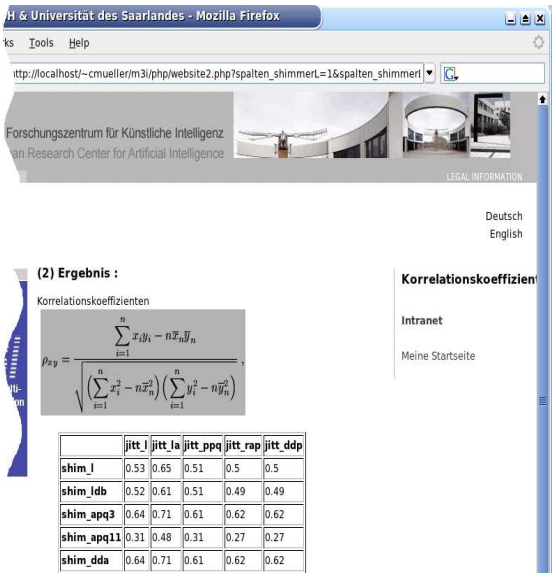


Abbildung 12.13: Liniendiagramm mit normalisierten Werten.



	jitt_l	jitt_la	jitt_ppq	jitt_rap	jitt_ddp
shim_l	0.53	0.65	0.51	0.5	0.5
shim_ldb	0.52	0.61	0.51	0.49	0.49
shim_apq3	0.64	0.71	0.61	0.62	0.62
shim_apq11	0.31	0.48	0.31	0.27	0.27
shim_dda	0.64	0.71	0.61	0.62	0.62

Abbildung 12.14: Bestimmung der Korrelationskoeffizienten einer Auswahl von Merkmalen.

Korrelationskoeffizienten

Darüber hinaus bietet m3i CAT ein Skript zur Ermittlung von Korrelationskoeffizienten an, deren Berechnung auf Basis von Gleichung 4.2 (Seite 68) erfolgt. Da in diesem Fall keine grafische Darstellung erforderlich ist, entfallen Schritt sieben und acht. Stattdessen werden die Korrelationskoeffizienten der ausgewählten Merkmale in tabellarischer Form präsentiert (vgl. Abbildung 12.14). Dazu wählt der Benutzer in Schritt eins jeweils eines oder mehrere Merkmale für die Zeilen bzw. Spalten aus.

Streudiagramme mit Entscheidungsgrenzen

Streudiagramme (*Scatterplots*) sind ein geeignetes Mittel zur Visualisierung von Trainingsinstanzen, wobei von einem n-dimensionalen Vektor jeweils zwei Dimensionen als Punkt in einem xy-Diagramm repräsentiert werden. Bei einer entsprechend großen Datenauswahl entstehen Punktwolken, die durch verschiedene Einfärbungen die Zugehörigkeit zu einer Klasse signalisieren. Diese Art der Darstellung wird auch von m3i CAT angeboten, wobei zusätzlich die Entscheidungsgrenzen eines Bayes'schen Klassifizierers auf Basis einer multivariaten Gauß'schen Wahrscheinlichkeitsdichte angezeigt werden.

Die Funktion läuft ebenfalls nach dem Schema ab, das in Abbildung 12.9 dargestellt wird. In Schritt eins werden zunächst zwei reellwertige Datenbankfelder, die Klassenbezeichnung und die Anzahl der Datensätze bestimmt, welche aus der Tabelle RESULTS ausgewählt werden sollen.

```
Vorzeichen (Wert x, Wert y) {
  d := f1(x,y) - f2(x,y)
     if (d == 0)
    return d
  else
    return d/|d|
  fi
}

Finde Nullstelle (Vektor X, Vektor Y) {
  n = Länge von x
  m = Länge von y
  for (i := 0 ; i < n ; i++) do
    for (j := 0; j < m; j++) do
      v = Vorzeichen(X[i],Y[j])
      if(v == 0)
        Nullstelle gefunden
        return (X[i],Y[j])
      fi
      if (v <> letzes Vorzeichen)
        //Vorzeichenwechsel hat stattgefunden. Dieses muss in
        //dem Intervall zwischen Y[j] und Y[j-1] liegen
        h:= Hälfte zwischen Y[j] und Y[j-1]
        v' = Vorzeichen(X[i], h)
        if (v' == 0)
          Nullstelle gefunden
          return (X[i],Y[j])
        fi
        X' = Vektor mit X[i] als einzigem Element
        if (v' == letztes Vorzeichen)
          //Nullstelle muss im Intervall h bis Y[j] liegen
          H = Vektor mit k Elementen h bis Y[j] wobei k > 2
        else
          //Nullstelle muss im Intervall von Y[j-1] bis h liegen
          H = Vektor mit k Elementen Y[j-1] bis H wobei k > 2
        fi
        Finde Nullstelle (X',H)
      fi
    od
  od
}
```

Abbildung 12.15: Algorithmus zum Auffinden von Nullstellen.

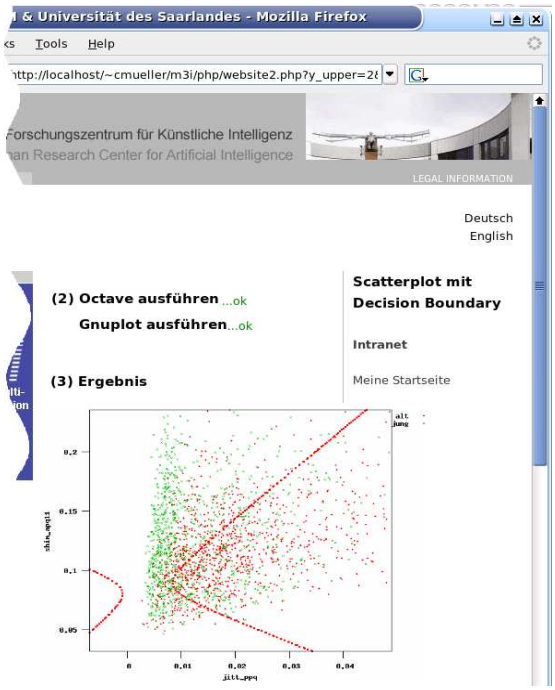


Abbildung 12.16: Anzeige eines Streudiagramms mit Entscheidungsgrenzen.

Wie bereits bei den zuvor besprochenen Datensichtungs-Skripten, werden die Wertebereiche des Diagramms auf Basis von Vielfachen der jeweiligen Standardabweichungen eingeschränkt, die ebenfalls an der Stelle als Parameter eingegeben werden können (die Standardwerte liegen hier bei einem Faktor von 2.5).

Die Matrix für das Streudiagramm wird direkt aus der Datenbank gelesen und für die Darstellung mit Gnuplot (Schritt sechs) in einer Datei abgespeichert. Für die Berechnung der Entscheidungsgrenzen mit Octave wird dasselbe Verfahren angewendet, wie es oben für die Darstellung der bivariaten Gauß'schen Wahrscheinlichkeitsdichte beschrieben wurde. Der Unterschied ist jedoch der, dass statt der Entscheidungsregionen hier die Entscheidungsgrenze(n) relevant sind, so dass ein einfacher Größenvergleich der Funktionswerte nicht ausreichend ist. Stattdessen müssen im angezeigten Wertebereich die Nullstellen der Gleichung 7.2 gefunden werden. Hierzu wurde in Octave der Algorithmus implementiert, der in Abbildung 12.15 dargestellt wird. Um die Laufzeit so gering wie möglich zu halten, wurde die Rekursionstiefe auf 100 begrenzt, wodurch nicht alle Nullstellen gefunden werden. Dies wirkt sich dahingehend aus, dass die Kurven der Entscheidungsgrenzen in der Ergebnisgrafik relativ grob aufgelöst sind (vgl. Abbildung 12.10). Es sei darauf hingewiesen, dass eine mathematisch korrekte Nullstellenanalyse in diesem Fall nicht erforderlich ist, da diese hier ausschließlich Visualisierungszwecken dient. Für die Klassifikation ist es nicht notwendig, die Nullstellen der Entscheidungsfunktion zu finden; hier reicht ein direkter

Größenvergleich der Ergebnisse auf Basis der alternativen Klassen aus.

In Abbildung 12.15 wird ein vereinfachter Algorithmus dargestellt: Da die Diskriminantenfunktion in dem bivariaten Fall zwei Eingabewerte nimmt, muss zum Auffinden aller Vorzeichenwechsel der Wertebereich einmal in x-y-Richtung und einmal in y-x-Richtung durchlaufen werden. Falls bis zum Erreichen des Limits keine Nullstelle gefunden wurde, wird das aktuelle h als Näherungswert zurückgegeben.

Entscheidungsregionen beliebiger Weka-Klassifizierer

Das Skript zur Darstellung der Entscheidungsregionen beliebiger Weka-Klassifizierer weicht insofern von dem in Abbildung 12.9 dargestellten allgemeinen Schema ab, als keine Berechnung mit Octave erfolgt. Stattdessen wird auf Basis der ausgewählten Merkmale ein Weka-Klassifizierer trainiert, der auf *Java* basiert, wie in Abschnitt 11.1.1 genauer beschrieben wird. Zunächst wird auf Basis der ausgewählten Datenbanktabelle ein Formular erzeugt, mithilfe dessen der Benutzer zwei reellwertige Felder auswählen kann. Weitere für das Training notwendige Parameter sind die Klassenbezeichnung und die Klassifikationsmethode. Für die Darstellung werden darüber hinaus die Auflösung und die unteren und oberen Ränder des Wertebereiches abgefragt (siehe oben). Für die Angabe der Klassifikationsmethode stehen derzeit die folgenden Alternativen zur Verfügung: k-Nearest-Neighbor, C 4.5 Entscheidungsbaum, Neuronales Netz, Support-Vector-Machine und Naive-Bayes (vgl. Kapitel 8).

Anstelle der Schritte drei und vier wird dementsprechend ein Java-Modul aufgerufen, welches die Methoden zum Trainieren eines Klassifizierers aus dem m3i Server beinhaltet. Man beachte, dass für die Darstellung der Entscheidungsregionen nicht auf die Klassifizierer zurückgegriffen wird, auf deren Basis später die Sprecherklassifikation erfolgen soll. Vielmehr werden jeweils für zwei Merkmale eigene Klassifizierer erzeugt. Neben den Methoden für das Training umfasst das besagte Java-Modul auch eine Methode, mithilfe derer der Wertebereich des Diagramms klassifiziert und das Ergebnis in einer externen Datei abgespeichert wird. Da das Training der Klassifizierer trotz der eingeschränkten Anzahl der Merkmale einige Zeit in Anspruch nehmen kann, laufen die Prozesse ähnlich wie bei der Korpusanalyse im Hintergrund ab (vgl. Abschnitt 12.2.2).

Die von dem Java-Modul generierte Datei stellt – wie ansonsten die Ausgabe von Octave – eine Matrix dar, die von Gnuplot zu einer Grafik weiterverarbeitet wird. Das Diagramm, das in Abbildung 12.17 dargestellt wird, wurde für einen C 4.5 Entscheidungsbaum mit den Werten $jitt\_ppq$ und $shim\_aqp11$ und der Klassenbezeichnung AGECLASS erzeugt.

Übersicht über die Eigenschaften der Korpora

Neben dem bisher vorgestellten Datensichtungs-Skript beinhaltet m3i CAT auch ein Skript zur Anzeige von Informationen über die verwendeten Korpora. Da hierbei keine Berechnungen durchgeführt werden, entfallen die Schritte drei und vier aus dem Schema in Abbildung 12.9. Das Skript stellt eine Sammlung von SQL-Abfragen dar, deren Ergebnisse in aufbereiteter Form teils tabellarisch, teils grafisch dargestellt werden: 1. Anzahl der Sprecher und Anzahl der Sprachproben pro Sprecher (vgl. Abbildung 12.18), 2. Anzahl der Sprachproben nach Geschlecht und 3. An-

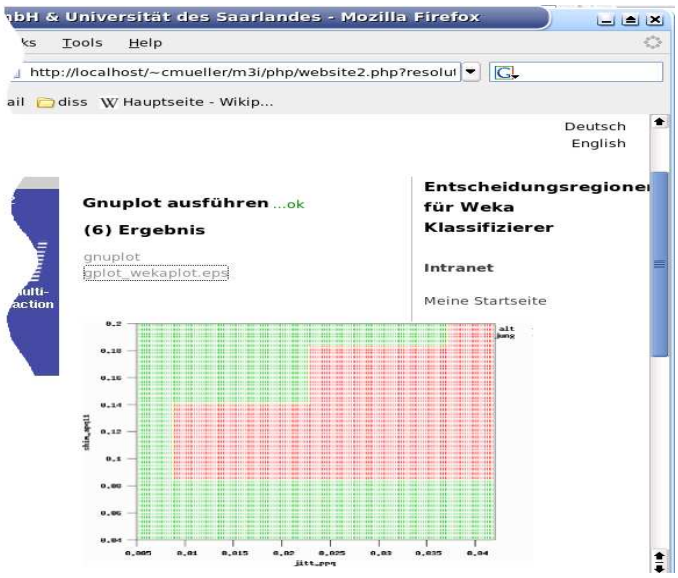


Abbildung 12.17: Anzeige von Entscheidungsregionen beliebiger Weka-Klassifizierer.

zahl der Sprachproben pro Lebensalter (vgl. Abbildung 12.19). Die Grafiken können wie üblich in eps-Format gespeichert werden.

12.2 Generelle Aspekte

12.2.1 Architektur

Während in Abbildung 12.1 die Architektur von m3i CAT bezüglich eines spezifischen Skriptes dargestellt wird, zeigt Abbildung 12.20 die übergeordnete Architektur. Es ist zu erkennen, dass eine weitestgehende Trennung von Konfiguration, Funktion und Layout vorgenommen wurde: Die Konfiguration erfolgt auf Basis des angeschlossenen Datenbank-Servers, wobei nicht die Datenbank verwendet wird, in der die Korpusanalysen abgespeichert werden, sondern eine separate. Die Konfigurationsdatenbank enthält Angaben über die einzelnen Skripte. Außerdem wird darin festgelegt, welche Benutzer auf welche Skripte Zugriff haben.

Das Layout wird mithilfe von HTML-Schablonen erzeugt, wobei zwischen äußeren und inneren Schablonen unterschieden wird. Die äußere Schablone bestimmt das globale Layout des Systems. Die aktuelle Version basiert beispielsweise auf der DFKI-Website (vgl. http://www.dfki.de). Da es sich um eine einzelne HTML-Datei handelt, kann das gesamte Erscheinungsbild von m3i CAT sehr leicht verändert werden. Die inneren Schablonen bestimmen das Layout der Ausgaben einzelner Skripte. Sie betreffen den variablen Bereich im Zentrum der Seite. Die Zuordnung von Layout-Schablonen und Skripten beinhaltet einen einfachen Vererbungsmechanismus: Falls für ein gegebenes Skript keine spezifische Schablone vorhanden

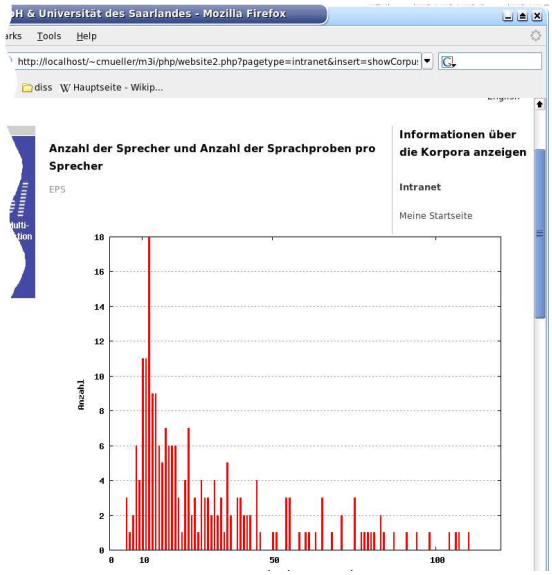


Abbildung 12.18: Darstellung der Anzahl der Äußerungen pro Sprecher als Histogramm.

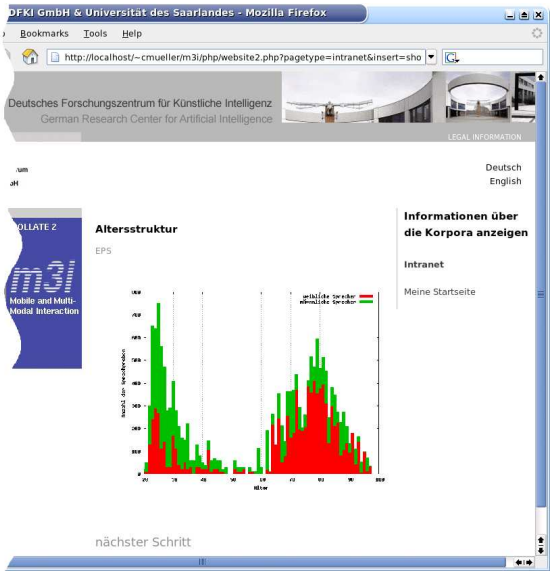


Abbildung 12.19: Anzeige der Altersstruktur eines ausgewählten Korpus.

ist, wird eine allgemeine Schablone verwendet. Diese Vorgehensweise hat sich bereits bei der Visualisierung des m3i Servers bewährt, da der Entwickler nur dann eine Schablone erzeugen muss, wenn ein spezielles Layout erforderlich ist.

m3i CAT ist als ein Intranet-System konzipiert worden, das heißt die Benutzer müssen sich mit Login und Passwort anmelden. Die Authentifizierung basiert dabei auf der MySQL-eigenen Benutzerverwaltung: Ist ein SELECT-Zugriff auf die Datenbank mit den angegebenen Benutzerdaten nicht möglich, scheitert der Login-Versuch. Ein nicht autorisierter Zugriff wird somit zuverlässig verhindert. Auf Basis der Angaben über die Zugriffsrechte der Benutzer stellt m3i CAT für jeden Benutzer eine „persönliche Startseite“ zusammen (vgl. Abbildung 12.21).

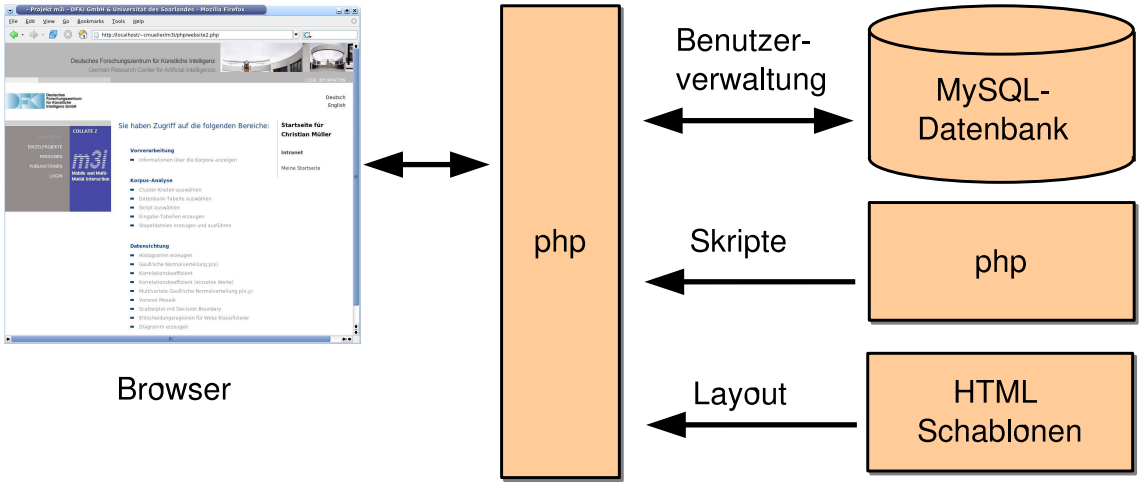


Abbildung 12.20: Übergeordnete Architektur von m3i CAT.

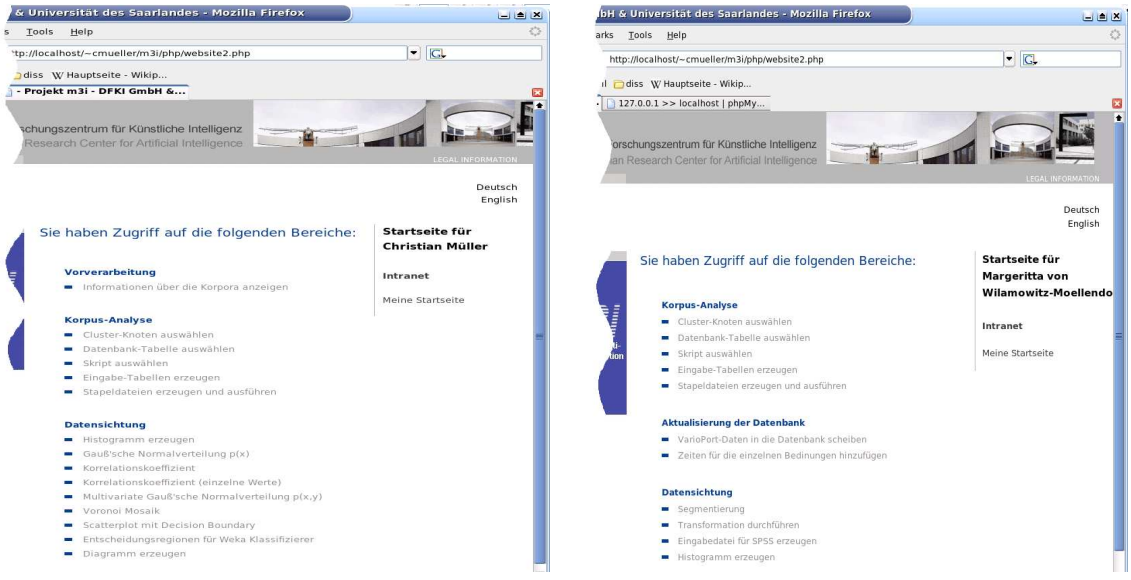


Abbildung 12.21: Persönliche Startseiten in m3i CAT.

```
$names['reals'] = "Zahlenwerte";
$types['reals'] = "choice";
$limits['reals'] = 2;

$names['class'] = "Klasse";
$types['class'] = "text";

$names['classmethod'] = "Klassifikationsmethode (knn,j48,neuralnet,svm,bayes)";
$types['classmethod'] = "text";
$defaults['classmethod'] = "j48";

$names['limit'] = "Anzahl der Datensätze";
$types['limit'] = "text";
$defaults['limit'] = "1000";

$names['x_lower'] = "Untere Grenze für x-Werte (Vielfaches stddev)";
$types['x_lower'] = "text";
$defaults['x_lower'] = "2.5";
```

Abbildung 12.22: Beispiel für die Spezifikation eines Parameter-Formulars.

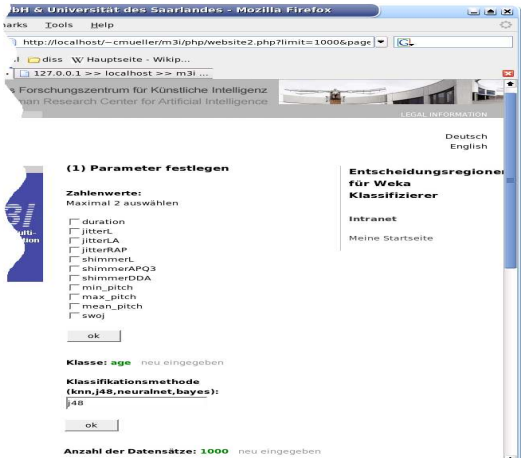


Abbildung 12.23: Ein vom Formulargenerator erzeugtes Formular mit verschiedenen Eingabetypen. Oben: eine auf Basis der ermittelten reellwertigen Datenbankfelder zusammengestellte Auswahl. Unten: teilweise vorbelegte Textfelder. Die grün dargestellten Werte wurden bereits eingegeben.

12.2.2 Verschiedene zentrale Funktionen

Generierung von Formularen

Bei den meisten der oben aufgeführten m3i CAT-Skripte wird im ersten Schritt ein Formular erzeugt, damit der Benutzer die Parameter für die nachfolgenden Schritte angeben kann. Hierfür wird eine Funktion zur Formularerzeugung verwendet, so dass innerhalb der einzelnen Skripte lediglich festgelegt werden muss, welche Parameter benötigt werden, in welcher Art sie eingegeben werden sollen und was die Standardwerte sind.

Abbildung 12.22 stellt ein Beispiel für eine solche Formular-Spezifikation dar: Der erste Parameter trägt den internen Variablennamen *reals*, seine Bezeichnung ist „Zahlenwerte“. Der Parameter-Typ ist *choice*, d. h. der Formulargenerator stellt fest, welche reellwertigen Spalten es in der betreffenden Datenbanktabelle gibt und listet sie zusammen mit einem Auswahlkästchen auf. Die Angabe *limit* legt fest, wie viele Einträge maximal selektiert werden dürfen. Die nachfolgenden Parameter sind alle vom Typ *text*, d. h. es wird für die Eingabe ein Textfeld in das Formular eingefügt. Einige Parameter besitzen darüber hinaus *Default-Werte*, was bedeutet, dass die Formularfelder mit den jeweiligen Werten vorbelegt werden. Neben Auswahlkästchen und Textfeldern kann der Formulargenerator *textareas* (mehrzeilige Textfelder) sowie Listenfelder verarbeiten.

Zusätzlich zur eigentlichen Formularerzeugung hat diese Funktion die Aufgabe, die eingegebenen Werte in geeigneter Form zu speichern. Sie werden zunächst über die *get*-Methode über-

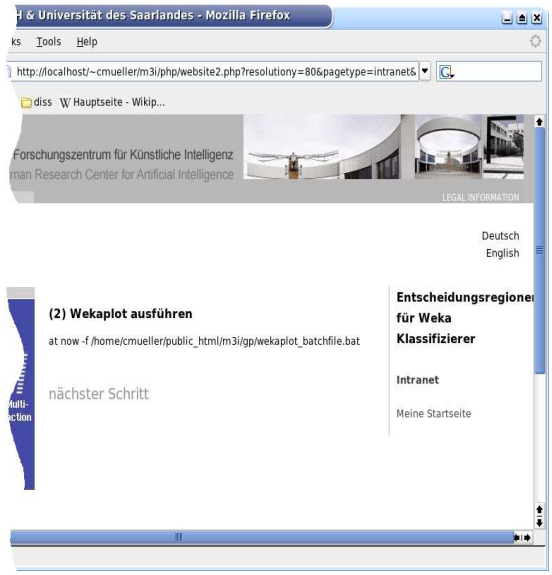


Abbildung 12.24: Aufruf externer Prozesse mit hoher Laufzeit mithilfe von 'at'.

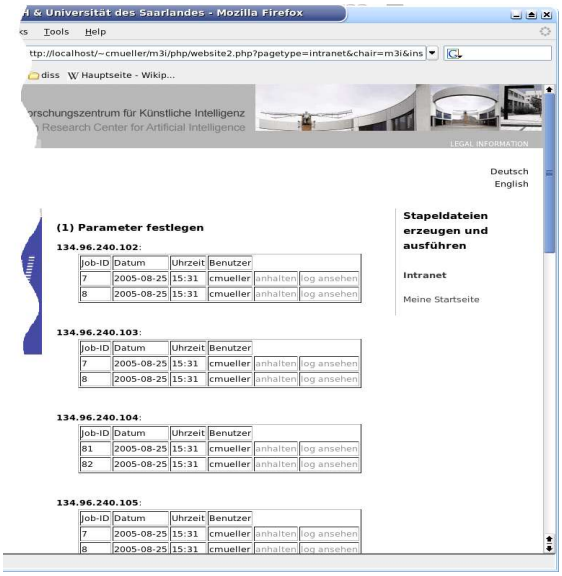


Abbildung 12.25: Die 'at queue' bei paralleler Ausführung des Prozesses auf dem Cluster.

geben und dann als *cookies* abgespeichert. Der Formulargenerator überprüft dabei, ob die Werte bereits gespeichert wurden und stellt sie ggf. grün dar. Falls nicht, wird nach wie vor das entsprechende Formularfeld angezeigt. Erst wenn alle Werte angegeben wurden, wird ein Verweis auf den nachfolgenden Schritt angeboten (vgl. Abbildung 12.23).

Externe Prozesse mit hoher Laufzeit

Vor allem bei der Korpusanalyse, aber auch bei einigen Datensichtungsfunktionen ist damit zu rechnen, dass die von m3i CAT gestarteten Prozesse eine hohe Laufzeit haben. Die Korpusanalyseprozesse, die im Rahmen der vorliegenden Arbeit mit m3i CAT durchgeführt wurden, hatten trotz der Parallelisierung auf dem Cluster eine Laufzeit von durchschnittlich einer Stunde. Die Berechnung der Entscheidungsregionen für k-Nearest-Neighbor-Klassifizierer (vgl. Abschnitt 12.1.3) nimmt auf einem Einzelplatzrechner etwa zehn Minuten in Anspruch. Da m3i CAT jedoch eine Web-basierte Anwendung ist, und die zulässigen Antwortzeiten daher auf wenige Minuten beschränkt sind, müssen die Prozesse im Hintergrund ausgeführt werden.

Aus diesem Grund wurde in m3i CAT eine Funktion implementiert, um externe Prozesse mit hoher Laufzeit im Hintergrund ablaufen zu lassen. Die Funktion macht Gebrauch von dem Unix-Kommando *at*, das gewöhnlich genutzt wird, um Prozesse zeitversetzt zu starten. Der Vorteil von *at* besteht darin, dass es über eine eigene Prozessverwaltung (*atq*) verfügt, über welche die derzeit laufenden Prozesse abgerufen und ggf. gestoppt werden können. Die betreffenden Prozesse werden nicht direkt durch einen Systemaufruf gestartet, sondern an *at* übergeben, so dass der Browser die Kontrolle sofort wieder erhält (vgl. Abbildung 12.24). Anschließend wird eine Seite präsen-

tiert, welche die *atq* auswertet (vgl. Abbildung 12.25). Durch erneutes Laden der Seite kann der Benutzer sich darüber informieren, ob die Hintergrundprozesse noch laufen und sie ggf. stoppen. Wenn sie abgeschlossen sind, d. h. die *atq* leer ist, wird statt der Tabelle ein Verweis auf die folgenden Verarbeitungsschritte angeboten. Werden die Prozesse auf mehreren Rechnern (Clusterknoten) parallel ausgeführt, startet m3i CAT diese einzeln und zeigt auf der Folgeseite die *atq* aller betreffenden Rechner an.

13 Zusammenfassung der Implementierung

Die Implementierung des in der vorliegenden Arbeit vorgestellten Ansatzes zur Sprecher- und Umgebungsklassifikation AGENDER umfasst insgesamt drei Komponenten: den java-basierten m3i Server, den m3i Client, der als c++-Bibliothek realisiert wurde, und das PHP-basierte Korpusanalysewerkzeug m3i CAT. Während das zuletzt genannte System weitestgehend unabhängig von der Anwendung des AGENDER-Ansatzes ist, wurden die beiden zuerst genannten Komponenten für mobile, natürlichsprachliche Dialogsysteme entwickelt, wie beispielsweise den Mobile ShopAssist oder den Personal Navigator.

Wie in Kapitel 11.1 gezeigt wurde, zeichnet sich der m3i Server besonders durch seine flexible Blackboard-Architektur aus, die eine unmittelbare Umsetzung und Evaluation von Weiterentwicklungen auf der konzeptuellen Ebene ermöglicht. Durch die weitgehende Ausnutzung von Vererbungsmechanismen für häufig benötigte Funktionen, die in Java vorhanden sind, wurde der Programmieraufwand bei der Erweiterung des Systems – z. B. durch das Hinzufügen neuer Klassifizierer – erheblich verringert. Die Wahl von Java hat sich auch deshalb als vorteilhaft erwiesen, weil dadurch auf eine umfangreiche Sammlung von Programm-Bibliotheken zurückgegriffen werden konnte, wie beispielsweise das WEKA-Paket. Um vorhandene Programme, die nicht auf Java basieren – wie PRAAT, ENRATE oder SRSAD – einsetzen zu können, wurde der m3i Server mit einem Mechanismus ausgestattet, der die Verwaltung externer Prozesse sowohl auf einem einzelnen Rechner als auch auf einem Cluster übernimmt. Der hohe Grad an Flexibilität hat allerdings seinen Preis: Die Laufzeit des m3i Servers ist vergleichsweise hoch. Läuft das System auf einem einzelnen Rechner, so dass eine parallele Durchführung der Merkmalsextraktion nicht möglich ist, nimmt die Klassifikation einer Äußerung durchaus bis zu zehn Sekunden in Anspruch – wobei die Visualisierung der einzelnen Mustererkennungsphasen nicht unwesentlich hierzu beiträgt.

Die Notwendigkeit des sparsamen Umgangs mit den beschränkten Ressourcen eines Pocket-PCs hat andererseits dazu geführt, dass bei der Entwicklung des in Kapitel 11.2 beschriebenen m3i Clients gerade auf die Laufzeit ein Hauptaugenmerk gelegt wurde. Im Zuge dessen wurden Optimierungen bei der Merkmalsextraktion durchgeführt, die Komponenten stärker integriert, komputationell weniger komplexe Klassifikationsmethoden ausgewählt und die Implementierung in dem schnelleren c++ durchgeführt. Durch diese Maßnahmen ist es gelungen, eine AGENDER-Version für den Pocket-PC zu entwickeln, die – bei fehlender Verbindung zum Server

– eine eingebettete Sprecherklassifikation durchführen kann. Dass dies nur auf Kosten einer wesentlich geringeren Flexibilität erreicht werden konnte, ist in dem Fall nicht problematisch, da konzeptuelle Entwicklungen nach wie vor auf Grundlage des Servers durchgeführt werden.

Als Grundvoraussetzungen für den Erfolg von AGENDER können die weitestgehende Automatisierung der Korpusanalysen, die Verwaltung der Ergebnisse in einer relationalen Datenbank sowie die intensive Arbeit mit den Daten unter Verwendung diverser Sichtungsmethoden angesehen werden. Dabei hat sich die separate Entwicklung des in Kapitel 12 beschriebenen Korpusanalysewerkzeuges m3i CAT als eine sinnvolle Entscheidung erwiesen. Das System basiert auf PHP und integriert eine Reihe von frei verfügbaren Komponenten: MySQL als Datenbank, Octave für Berechnungen sowie Gnuplot für die Erzeugung von Graphen. Dank seines Web-Interfaces ist das System intuitiv zu bedienen und kann aufgrund seiner modularen Architektur leicht erweitert werden, was beides bereits bei einer Anwendung im Rahmen eines anderen Projektes bestätigt werden konnte: Wilamowitz-Moellendorff et al. (2005) beschreiben Ergebnisse eines Experimentes über die Messung von Benutzerzuständen auf Basis von Biosensoren, bei dem mithilfe von m3i CAT die Rohdaten aufbereitet und die Ergebnisse ausgewertet wurden. Allerdings muss klar festgestellt werden, dass sich das System im Stadium eines Prototyps befindet. Bevor es für andere Wissenschaftler zugänglich gemacht werden kann, muss der Quellcode angepasst, Installationsskripte geschrieben und eine ausführliche Dokumentation erstellt werden.

Die neuerliche Konzentration auf Anwendungen im Telekommunikationsbereich wirkt sich auch auf die Weiterentwicklung der Implementierung aus. In einem AGENDER-System für Callcenter wird es die Komponenten m3i Server und m3i Client in dieser Form nicht geben, da vollkommen andere Anforderungen an das System gestellt werden. Was in dem Fall benötigt wird, ist ein Klassifikationsmodul, das in eine bestehende Callcenter-Plattform integriert werden kann, multiple Kanäle unterstützt – d. h. eine hohe Anzahl von gleichzeitig eingehenden Anrufen bearbeiten kann – und vor allen Dingen schnell ist. Um die Sprecherklassifikation sinnvoll in einen Callcenter-Dienst oder in ein Voice-Response-System einbinden zu können, werden von den Auftraggebern Antwortzeiten gefordert, die im Bereich von einhundert bis zweihundert Millisekunden liegen. Gleichzeitig müssen, wie in Kapitel 10.1 beschrieben wird, auch konzeptuelle Weiterentwicklungen vollzogen werden, was bisher ausschließlich auf Basis des Servers geschehen ist.

Um den neuen Anforderungen gerecht zu werden, wird die Gesamtarchitektur des AGENDER-Systems, wie in Abbildung 13.1 dargestellt modifiziert. Der jetzige m3i Server wird zu einer *Entwicklungsplattform* ausgebaut, auf der die Klassifizierer entworfen und getestet werden. Hierfür besteht nach wie vor eine Verbindung zur zentralen Datenbank, aus der die Trainings- und Testsätze ausgelesen werden. Die Entwicklungsplattform verfügt über einen *Build*-Mechanismus, der die so genannten *Eingebetteten Klassifikationsmodule* (EKM) erzeugt. Diese basieren auf dem jetzigen m3i Client und stellen kompakte, starre Einheiten dar. Der Build-Prozess kann dabei als eine Weiterentwicklung des bisherigen Mechanismus zum Download von Klassifizierern angesehen werden. Die EKM basieren auf c++ und werden für verschiedene Ziel-Plattformen erzeugt. Als Inferenzmechanismus für Dynamische Bayes'sche Netze wird statt HUGIN das schnellere Verfahren

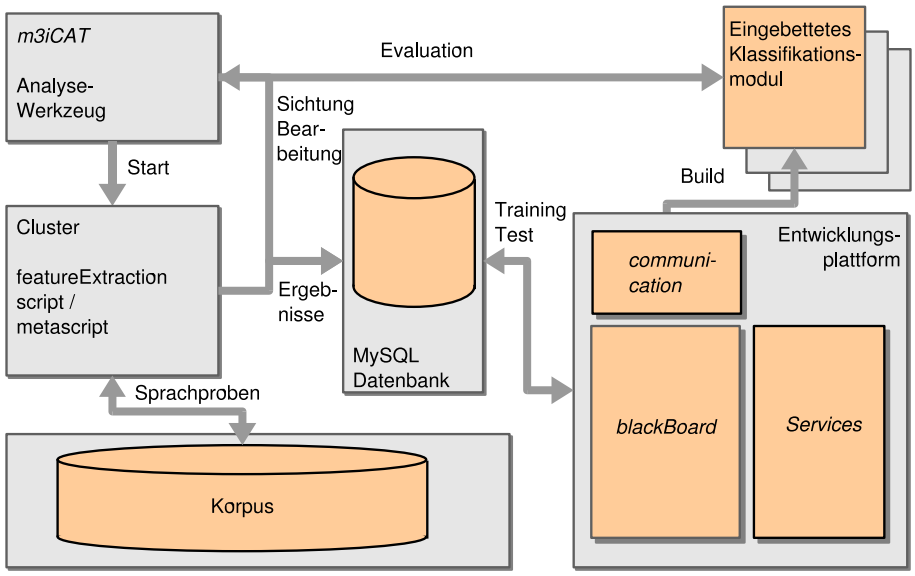


Abbildung 13.1: Modifizierte Gesamtarchitektur der AGENDER-Implementation für Telekommunikationsanwendungen

von Brandherm und Jameson (2004) verwendet.

Eine erste Version dieser neuen Architektur wurde im Rahmen des erwähnten Projektes für die telefonbasierte Anwendung von AGENDER bereits implementiert und befindet sich derzeit in der Testphase. Eine detaillierte Beschreibung erfolgt durch Feld (2006).

Wie aus Abbildung 13.1 ersichtlich ist, bleibt m3i CAT in der neuen Gesamtarchitektur als Korpusanalysewerkzeug bestehen. Tatsächlich wird dem System, dadurch dass es auch zur Evaluation genutzt wird, eine weitere Rolle zuteil. Der Grund dafür besteht darin, dass durch die Kreuzvalidierung auf der Entwicklungsplattform zwar die Genauigkeit der Klassifizierer, jedoch nicht die Performanz des EKM getestet werden kann. Die Äquivalenz der Modelle kann zwar theoretisch nachgewiesen werden, jedoch werden dadurch Fehler im Build-Prozess nicht ausgeschlossen. Zu diesem Zweck wird m3iCAT entsprechend erweitert, was bisher in Form zweier Evaluations-Skripte geschehen ist. Eines der Skripte stellt eine Übersicht von den klassenspezifischen univariaten Gauß'schen Wahrscheinlichkeitsdichten für alle Merkmale aller in dem betreffenden EKM enthaltenen Klassifizierer dar. Zusätzlich wird eine Sprachprobe aus der Datenbank ausgewählt und zur Klassifikation an das EKM gesendet. Neben dem Gesamtergebnis auf der Zweiten Ebene sendet das Modul die Einzelergebnisse der Ersten Ebene sowie die Werte der Merkmalsextraktoren zurück. Letztere werden als senkrechte Linien in die Gauß-Kurven eingetragen. Auf diese Art und Weise kann nicht nur die Plausibilität der Klassifikation überprüft, sondern darüber hinaus können etwaige Abweichungen in den Werten der Merkmale erkannt werden. Das zweite Skript führt eine Evaluation mit einer festgesetzten Anzahl von Testsätzen durch und ermittelt die Klassifikationsgenauigkeit anhand von Konfusionsmatrizen.

Teil IV

Gesamtzusammenfassung und Ausblick

Die vorliegende Arbeit leistet einen Beitrag zur Beantwortung der folgenden Forschungsfragen:

1. Wie manifestiert sich das Specheralter und -geschlecht in der Stimme und dem Sprechverhalten?

Die Beantwortung dieser humanwissenschaftlichen Fragestellung wurde auf Basis umfangreicher Korpusanalysen angestrebt, deren Ergebnisse wie folgt zusammengefasst werden können:

Bezüglich der Stimmen von Kindern im vorpubertären Alter konnte ein geschlechtsspezifischer Unterschied in der mittleren Höhe der Grundfrequenz festgestellt werden: Der gemessene Mittelwert bei den Mädchen beträgt 249.26 Hz, während er bei den Jungen um etwa 10 Hz darunter liegt (239.37 Hz). Die normierte Tendenz ist mit -0.28 zwar recht schwach, aber dennoch signifikant. Bei der Interpretation dieses Ergebnisses ist jedoch zu bedenken, dass keine Sprecher unter zehn Jahren betrachtet wurden.

Die Stimmen der Jugendlichen zeigen erhöhte Jitter- und Shimmerwerte, was möglicherweise auf den Kontrollverlust aufgrund der rapiden anatomischen Veränderungen während der Pubertät zurückzuführen ist. Für diese Interpretation spricht, dass die Werte sowohl bei den Kindern als auch bei den jüngeren Erwachsenen geringer sind. Kinder und Jugendliche unterscheiden sich von Erwachsenen und Senioren durch harmonischere Stimmen: Zusammengenommen beträgt die Harmonicity-to-Noise-Ratio für Kinder und Jugendliche durchschnittlich 14.0 dB, während sie bei Erwachsenen und Senioren im Mittel nur bei 9.4 dB liegt.

Die Grundfrequenz als das Hauptunterscheidungskriterium zwischen erwachsenen Frauen und Männern bestätigt sich auch in dieser Studie: Bei den Frauen ist der Durchschnittswert bei 197.68 Hz angesiedelt und bei den Männern bei 125.78 Hz. Was die Maße der Stimmqualität betrifft, zeigt sich hier jedoch ein komplementäres Bild im Vergleich zu der Studie von Pützer (2001): Bei den Sprecherinnen ist weniger Jitter, weniger Shimmer und eine höhere Harmonicity-to-Noise-Ratio zu beobachten. Die Unterschiede in der Standardabweichung der Grundfrequenz sind nicht signifikant.

Die Stimmen der Senioren enthalten erwartungsgemäß mehr Jitter und mehr Shimmer als alle anderen Altersgruppen, was sowohl für die Frauen als auch für die Männer gilt. Beide Maße können hauptsächlich als ein Symptom der Stimmalterung angesehen werden. Die leistungsbezogenen Maße entsprechen nur zum Teil den Vorhersagen: Die untere Grenze des Stimmumfangs bei Frauen geht erwartungsgemäß nach unten. Allerdings ist das – wenn auch in geringerem Maße – bei den Männern ebenso der Fall, was nicht mit den Hypothesen konform geht. An der oberen Grenze des Stimmumfangs sind bei beiden Geschlechtern nur sehr schwache Veränderungen messbar. Was die altersbedingten Veränderungen der mittleren Stimmtonhöhe betrifft, werden die Hypothesen bestätigt: Die Grundfrequenz sinkt bei den Frauen ab, wohingegen sie bei den Männern ansteigt. Letztere Tendenz ist zwar schwach, aber statistisch signifikant.

Die Artikulationsgeschwindigkeit der Senioren beider Geschlechter ist erwartungsgemäß deutlich geringer als die der jüngeren Erwachsenen. Dennoch kann sie nicht ausschließlich als ein Symptom der Veränderungen des Sprechverhaltens im Alter angesehen werden, da sie bei Kindern und Jugendlichen auf einem ähnlichen Niveau liegt. Was die Sprechpausen betrifft, wer-

den die Hypothesen ebenfalls bestätigt: Die meisten Pausen sind bei den Senioren zu verzeichnen, die wenigsten bei den jüngeren Erwachsenen. Das Niveau der Kinder und Jugendlichen liegt dazwischen. Die Dauer der Sprechpausen ist dagegen bei Kindern, Jugendlichen und Erwachsenen in etwa gleich, während sie bei den Senioren durchschnittlich länger ist.

Die erzielten Resultate sind geeignet, um als Beitrag zum Aufbau einer Referenzbasis für die Analyse der Stimme und des Sprechverhaltens zu dienen – im Rahmen der vorliegenden Arbeit bilden sie jedoch in erster Linie die Basis für ein System zur automatischen Sprecherklassifikation. Hierzu konnten insgesamt in genügendem Maße sowohl alters- als auch geschlechtsspezifische Unterschiede bei den betrachteten Merkmalen gemessen werden.

Wie kann das Specheralter und -geschlecht von einem System automatisch erkannt werden? und allgemeiner: **Welches ist ein geeigneter Ansatz zur Erkennung von Sprechereigenschaften auf Basis der Sprache?**

Zur Beantwortung dieser beiden Kernfragen wurde ein zweistufiger Ansatz zur Sprecherklassifikation mit der Bezeichnung AGENDER vorgestellt, mithilfe dessen das Geschlecht des Sprechers sowie die Altersklassen KINDER, JUGENDLICHE, (jüngere) ERWACHSENE und SENIOREN unterschieden werden können. Um die durch die wechselseitige Abhängigkeit von Sprecheralter und -geschlecht verursachte Komplexität zu verringern, wurde das aus vier mal zwei Klassen bestehende Problem (vier Altersklassen und zwei Geschlechter) in ein Acht-Klassen-Problem umformuliert. Die Klassen werden bezeichnet mit KW, KM, JW, JM, EW, EM, SW und SM, wobei der erste Buchstabe für die Altersklasse steht (Kinder, Jugendliche, Erwachsene und Senioren) und der zweite Buchstabe für das Geschlecht (weiblich und männlich).

Bezüglich der veränderten Problemstellung wurde eine erneute Auswertung der Korpusanalyse-Ergebnisse vorgenommen, wobei festgestellt wurde, dass die Bildung multipler Gruppierungen der Klassen möglich ist, die jeweils von unterschiedlichen Kombinationen von Merkmalen unterstützt werden. Gruppierung 1 wird beispielsweise definiert als KWKMJW-EW-SW-EMSM, d. h. die Kinder und (weiblichen) Jugendlichen bilden eine Gruppe, die jüngeren, erwachsenen Frauen und die Frauen im Seniorenalter werden separat betrachtet, und die jüngeren und älteren Männer bilden wiederum eine Gruppe. Gruppierung 1 wird hauptsächlich von den Maßen $jitt\_ppq$ und $pitch\_mean$ unterstützt, bedingt jedoch auch von $shim\_apq11$ und $pitch\_min$. Insgesamt wurden fünf solcher Gruppierungen gefunden, die zusammen mit der Kontrollgruppierung 0, bei der alle acht Klassen separat betrachtet werden, das Klassifikationsproblem der *Ersten Ebene* definieren. Um einen direkten Vergleich mit Systemen zur Einschätzung des Sprechergeschlechts zu ermöglichen, wurde darüber hinaus die Gruppierung EW-EM betrachtet.

Mit dem Begriff Erste Ebene werden in AGENDER diejenigen Phasen der Mustererkennung bezeichnet, welche die Merkmalsextraktion und die Klassifikation betreffen. Die Merkmalsextraktoren wurden auf Basis der Verfahren bestimmt, die bereits bei den oben erwähnten Korpusanalysen zum Einsatz gekommen sind. Bezüglich der Klassifikation wurden die folgenden bekannten Methoden des maschinellen Lernens untersucht: 1. Naive Bayes (NB), 2. Gaussian-Mixture-Models (GMMs), 3. k-Nearest-Neighbor (KNN), 4. C 4.5 Entscheidungsbäume

(C 4.5), 5. Support-Vector-Machines (SVMs) und 6. Künstliche Neuronale Netze (Artificial Neural Networks, ANNs). Die ersten beiden werden zu den parametrischen Methoden gezählt, da sie vollständig durch eine bestimmte Anzahl von Parametern definiert werden. Im Fall von NBs sind dies die klassenspezifischen Mittelwerte und Standardabweichungen – im Fall von GMMs kommen noch die Gewichte der einzelnen Merkmale hinzu. Die Methode KNN gehört dagegen zu den instanzbasierten Methoden, da in dem Fall die Zuordnung eines Merkmalsvektors zu einer Klasse nicht anhand eines zuvor generalisierten Modells, sondern anhand der einzelnen Trainingsinstanzen erfolgt. Entscheidungsbäume nehmen insofern eine Sonderstellung ein, als sie ursprünglich für die Klassifikation auf Grundlage von nominellen Merkmalen entwickelt wurden (z. B. ROT/GRÜN oder RUND/ECKIG). Sie gelten dennoch als ebenso geeignet für Klassifikationsprobleme mit reellwertigen Merkmalen, was auch bei dieser Untersuchung bestätigt werden konnte. SVMs stellen eine besondere Form von Klassifikationsverfahren auf Basis linearer Diskriminantenfunktionen dar, bei denen der ursprüngliche Merkmalsraum zunächst in einen höherdimensionalen Raum überführt wird, um auch nichtlineare Probleme lösen zu können. ANNs schließlich sind einem Mechanismus nachempfunden, bei dem man davon ausgeht, dass er bei der menschlichen Informationsverarbeitung angewendet wird. Die Entscheidungen werden durch künstliche Synapsen herbeigeführt, die über mehrere Ebenen miteinander vernetzt sind und deren Aktivierungsfunktionen auf Basis der Trainingsdaten angepasst werden.

Alle genannten Verfahren wurden auf das gegebene Klassifikationsproblem angewendet, um festzustellen, welches am besten geeignet ist. Im Vordergrund stand dabei die globale Klassifikationsgenauigkeit. Die Studie sollte jedoch ebenfalls Aufschluss darüber geben, inwiefern erstens Vorteile aus den Gruppierungen gezogen werden können, und zweitens sich die Leistungen der Stimmmerkmalsmodelle von denjenigen der Sprechverhaltensmodelle unterscheiden. Was den zuerst genannten Aspekt betrifft, ließen die jeweiligen theoretischen Eigenschaften der Verfahren unterschiedliche Hypothesen zu: Es wurde erwartet, dass besonders die parametrischen Methoden von einer Gruppierung der Klassen und einer vorherigen manuellen Selektion der Merkmale profitieren, wohingegen dies bei Verfahren mit hoch entwickelten Trainingsprozessen, wie z. B. ANNs, in geringerem Maße der Fall sein sollte. Bezüglich des Vergleichs der Stimmmerkmals- und Sprechverhaltensmodelle wurde angenommen, dass in einem ruhigen Kontext erstere eine bessere Performanz aufweisen, sich dieser Vorzug jedoch durch einen lauten Kontext relativiert und evtl. umkehrt.

Die Ergebnisse, die erzielt werden konnten, sind insgesamt vielversprechend: Die Klassifikationsgenauigkeiten sämtlicher getesteter Verfahren liegen deutlich über dem Zufallsniveau der jeweiligen Gruppierung. Als besonders zufriedenstellend können dabei die True Positive Rates der Kontrollgruppierung 0 eingeschätzt werden, bezüglich derer mithilfe eines Artificial Neural Networks (ANNs) eine Gesamtgenauigkeit von 63.5 % erreicht werden konnte, was dem Fünffachen des Zufallsniveaus entspricht. Im direkten Vergleich der Methoden bleiben die parametrischen hinter den übrigen zurück, wobei besonders auffällig ist, dass Gaussian-Mixture-Models (GMMs) nur geringfügig bessere Ergebnisse aufweisen als Naive-Bayes-Klassifizierer (NBs). Dies wird auf

die Tatsache zurückgeführt, dass die Gewichte nicht mithilfe des EM-Algorithmus gelernt, sondern durch eine Kreuzvalidierung ermittelt wurden.

Was die Leistungen der Klassifizierer bezüglich der verschiedenen Gruppierungen betrifft, konnten die Hypothesen bestätigt werden: Die Performanz der parametrischen Methoden ist bezüglich der Kontrollgruppierung 0 am geringsten, wohingegen KNNs und ANNs zwar in fast allen Gruppierungen die vorderen Plätze einnehmen, hier jedoch einen besonders deutlichen Vorsprung aufweisen. Die Genauigkeit der Stimmmerkmalsmodelle ist wie erwartet höher als die der entsprechenden Sprechverhaltensmodelle, wobei in diesem Zusammenhang große Unterschiede zwischen den beiden betreffenden Gruppierungen bestehen.

Abgesehen von diesem positiven Resultat zeichnet sich der AGENDER-Sprecherklassifikationsansatz durch die so genannte *Zweite Ebene* aus, die auf Dynamischen Bayes'schen Netzen (DBNs) basiert. Anhand von Beispielen wurde gezeigt, wie diese genutzt werden können, um erstens die klassifikationsinhärente Unsicherheit explizit zu modellieren, zweitens um Top-Down-Wissen in den Entscheidungsprozess einfließen zu lassen und drittens um eine Fusion multipler Klassifikationsergebnisse erreichen zu können. Mithilfe von DBNs kann die letztere Aufgabe sowohl in der statischen Variante gelöst werden, bei der sich die Ergebnisse auf dieselbe Äußerung beziehen, als auch in der dynamischen Variante, bei der die Ergebnisse bezüglich mehrerer Äußerungen in das Modell einfließen. Der AGENDER-Ansatz ist prinzipiell nicht auf die Variablen Alter und Geschlecht beschränkt. Das Verfahren kann auch zur Ermittlung weiterer Sprechercharakteristika eingesetzt werden, worauf im anschließenden Ausblick genauer eingegangen wird.

Welchen Einfluss auf die Erkennung von Sprechereigenschaften hat der auditive Kontext und auf welche Art und Weise kann dieser berücksichtigt werden?

Neben der Ermittlung der Sprechercharakteristika wurde in der vorliegenden Arbeit auch ein Kontextmodell vorgestellt. Hierzu wurden zunächst Vorstudien anhand von Aufzeichnungen verschiedener Umgebungen durchgeführt: einer Straßenkreuzung (KREUZUNG), eines Autobahnparkplatzes (AUTOBAHN), einer Buchhandlung (BUCHHANDLUNG), einer Bibliothek (BIBLIOTHEK), einer Baustelle, an der mit einem Presslufthammer gearbeitet wurde (KOMPRESSOR) und eines Raums, in dem sich mehrere Personen in einiger Entfernung vom Mikrofon unterhielten (STIMMEN). Die Vorstudien ergaben, dass auf Basis der Merkmale Harmonicity-to-Noise-Ratio und Intensity-Ratio Gruppen von Umgebungen gebildet werden können, die als QUIET (leise), NOISY (laut) und VOICY (stimmenähnlich) bezeichnet wurden. Zu den Kontexten, die als NOISY eingestuft wurden, gehören KOMPRESSOR, AUTOBAHN und KREUZUNG, während die VOICY-Kontexte STIMMEN, BIBLIOTHEK und BUCHHANDLUNG umfassen. Der Kontext QUIET repräsentiert die Bedingung, in der kein Hintergrund vorhanden ist.

Durch Überlagerung der ursprünglichen Sprachproben mit den verschiedenen Kontexten wurde ein künstliches Korpus erzeugt, auf Grundlage dessen C 4.5 Entscheidungsbäume trainiert wurden. Diese sind in der Lage, die Kontextklasse mit einer Gesamtgenauigkeit von 93.77 % zu identifizieren. Des Weiteren wurde untersucht, inwiefern sich der Kontext auf die Genauigkeit der Sprecherklassifikation auswirkt. Es wurde erwartungsgemäß festgestellt, dass die Performanz bei

den Kontexten NOISY und VOICY gegenüber QUIET zurückgeht. Des Weiteren konnte die Hypothese bestätigt werden, dass die Sprechverhaltensmodelle bei NOISY einen Vorteil gegenüber den Stimmmerkmalsmodellen aufweisen.

Anhand eines Beispiels wurde gezeigt, wie eine kontextsensitive Sprecherklassifikation auf der Zweiten Ebene des AGENDER-Ansatzes realisiert werden kann, ohne dass über das Dynamische Bayes'sche Netz hinaus ein weiterer Mechanismus eingeführt werden muss: Erstens werden die vom Kontext abhängigen Evaluationsergebnisse der Klassifizierer in den CPTs der entsprechenden Knoten berücksichtigt und zweitens wird im Fall von NOISY den Sprechverhaltensmodellen ein höheres Gewicht beigemessen, während ansonsten die Stimmmerkmalsmodelle in stärkerem Maße berücksichtigt werden.

Welche Anforderungen an ein Sprecherklassifikationssystem ergeben sich aus dem zugrunde liegenden Anwendungsszenario und wie kann das System diesen auf der Implementierungsebene gerecht werden?

Die Anforderungen an das konkrete Sprecherklassifikationssystem ergaben sich aus dem Anwendungsszenario der mobilen, sprachbasierten Dialogsysteme. Dieses sieht vor, dass das mobile Gerät – z. B. ein Pocket-PC oder ein Smartphone – über eine dauerhafte, breitbandige Netzwerkverbindung zu einem Server verfügt (UMTS, W-Lan), der Dienste bereitstellt, die auf dem Gerät selbst aufgrund seiner eingeschränkten Ressourcen entweder gar nicht oder nur eingeschränkt durchführbar wären. Die Sprecher- und Umgebungsklassifikation stellt ein Beispiel für einen solchen Dienst dar: Bei vorhandener Netzwerkverbindung sollen die Sprachproben der Benutzer an einen Server geschickt werden, der die Klassifikation durchführt und die daraus resultierenden Profile an das mobile Gerät zurücksendet. Falls die Netzwerkverbindung unterbrochen ist, sollte der Pocket-PC allerdings in der Lage sein, in eingeschränktem Maße eine eingebettete Klassifikation durchzuführen.

Das in Teil III vorgestellte Sprecherklassifikationssystem wurden gemäß dieser Vorgaben als Client/Server-Lösung entworfen. Es umfasst insgesamt drei Komponenten: den java-basierten m3i Server, den m3i Client, der als c++-Bibliothek realisiert wurde, und das PHP-basierte Korpusanalysewerkzeug m3i CAT.

Der m3i Server zeichnet sich besonders durch seine flexible Blackboard-Architektur aus, die eine unmittelbare Umsetzung und Evaluation von Weiterentwicklungen auf der konzeptuellen Ebene ermöglicht. Durch die weitgehende Ausnutzung von Vererbungsmechanismen für häufig benötigte Funktionen, die in Java vorhanden sind, wurde der Programmieraufwand bei der Erweiterung des Systems – z. B. durch das Hinzufügen neuer Klassifizierer – erheblich verringert. Die Wahl von Java hat sich auch deshalb als vorteilhaft erwiesen, weil dadurch auf eine umfangreiche Sammlung von Programm-Bibliotheken zurückgegriffen werden konnte, wie beispielsweise das WEKA-Paket. Um vorhandene Programme, die nicht auf Java basieren – wie PRAAT, ENRATE oder SRSAD – einsetzen zu können, wurde der m3i Server mit einem Mechanismus ausgestattet, der die Verwaltung externer Prozesse sowohl auf einem einzelnen Rechner als auch auf einem Cluster übernimmt. Der hohe Grad an Flexibilität hat allerdings seinen Preis: Die Laufzeit des m3i Servers ist vergleichsweise hoch. Läuft das System auf einem einzelnen Rechner, so dass eine

parallele Durchführung der Merkmalsextraktion nicht möglich ist, nimmt die Klassifikation einer Äußerung durchaus bis zu zehn Sekunden in Anspruch – wobei die Visualisierung der einzelnen Mustererkennungsphasen nicht unwesentlich hierzu beiträgt.

Die Notwendigkeit des sparsamen Umgangs mit den beschränkten Ressourcen eines Pocket-PCs hat andererseits dazu geführt, dass bei der Entwicklung des in Kapitel 11.2 beschriebenen m3i Clients gerade auf die Laufzeit ein Hauptaugenmerk gelegt wurde. Im Zuge dessen wurden Optimierungen bei der Merkmalsextraktion durchgeführt, die Komponenten stärker integriert, komputationell weniger komplexe Klassifikationsmethoden ausgewählt und die Implementierung in dem schnelleren c++ durchgeführt. Durch diese Maßnahmen ist es gelungen, eine AGENDER-Version für den Pocket-PC zu entwickeln, die – bei fehlender Verbindung zum Server – eine eingebettete Sprecherklassifikation durchführen kann. Dass dies nur auf Kosten einer wesentlich geringeren Flexibilität erreicht werden konnte, ist in dem Fall nicht problematisch, da konzeptuelle Entwicklungen nach wie vor auf Grundlage des Servers durchgeführt werden.

Als Grundvoraussetzungen für den Erfolg von AGENDER können die weitestgehende Automatisierung der Korpusanalysen, die Verwaltung der Ergebnisse in einer relationalen Datenbank sowie die intensive Arbeit mit den Daten unter Verwendung diverser Sichtungsmethoden angesehen werden. Dabei hat sich die separate Entwicklung des in Kapitel 12 beschriebenen Korpusanalysewerkzeuges m3i CAT als eine sinnvolle Entscheidung erwiesen. Das System basiert auf PHP und integriert eine Reihe von frei verfügbaren Komponenten: MySQL als Datenbank, Octave für Berechnungen sowie Gnuplot für die Erzeugung von Graphen. Dank seines Web-Interfaces ist das System intuitiv zu bedienen und kann aufgrund seiner modularen Architektur leicht erweitert werden, was beides bereits bei einer Anwendung im Rahmen eines anderen Projektes bestätigt werden konnte: Wilamowitz-Moellendorff et al. (2005) beschreiben Ergebnisse eines Experimentes über die Messung von Benutzerzuständen auf Basis von Biosensoren, bei dem mithilfe von m3i CAT die Rohdaten aufbereitet und die Ergebnisse ausgewertet wurden. Allerdings muss klar festgestellt werden, dass sich das System im Stadium eines Prototyps befindet. Bevor es für andere Wissenschaftler zugänglich gemacht werden kann, muss der Quellcode angepasst, Installationsskripte geschrieben und eine ausführliche Dokumentation erstellt werden.

Das primäre Ziel zukünftiger Korpusanalysen wird eine genauere Differenzierung der Altersklassen sein. Um dies zu erreichen, bieten sich eine Reihe von Möglichkeiten zur Verbesserung des Verfahrens an. Einer der Ansatzpunkte betrifft beispielsweise die Verwendung multilingualer Spachproben (Englisch und Deutsch). Trotz der Annahme einer weitestgehenden Sprachunabhängigkeit der untersuchten Merkmale, muss davon ausgegangen werden, dass die sprecherabhängigen Effekte zumindest teilweise durch sprachabhängige Effekte überlagert werden. Im Zuge der Weiterentwicklung des AGENDER-Ansatzes sind Analysen geplant, denen ausschließlich unilinguale Sprachproben zugrunde liegen werden.

Des Weiteren ist das Potenzial zur Erweiterung der Menge der Merkmale bisher nicht erschöpfend genutzt worden. Es bestehen Pläne, in Zusammenarbeit mit der Universität von Lund (Schweden) die Möglichkeiten einer automatischen Extraktion weiterer akustischer Manifestationen des Sprecheralters zu eruieren. Dazu gehören *cepstrale* Merkmale und vor allen Dingen die Formantenfrequenzen F1 und F2. Die Betrachtung subtilerer Eigenschaften der Sprache stellt jedoch neue Anforderungen an das zugrunde liegende Verfahren: Statt eine Erweiterung der Datenbasis vorzunehmen, wird diesbezüglich zunächst eine kleine, besser zu kontrollierende Menge von Daten betrachtet. Auf diese Art und Weise können zusätzliche Probleme bei der automatischen Extraktion besser gelöst werden, wie z. B. im Fall der Formantenfrequenzen die Segmentierung des Eingabesignals, die nötig ist, um die informativen Bereiche (Vokale und Frikative) zu identifizieren. Die initiale Studie wird sich daher auf die Aussprache einzelner, für alle Sprecher gleicher Wörter beschränken, wobei als Quelle das entsprechend aufbereitete *Swedia*-Korpus vorgesehen ist.

Im Rahmen eines konkreten, auf M3I aufbauenden Industrieprojektes, wird die Weiterentwicklung von AGENDER hauptsächlich in Richtung Telekommunikationsanwendungen betrieben, was unter anderem zur Folge hat, dass neben der Klassifikationsgenauigkeit das Hauptaugenmerk auf die Klassifikationszeit gerichtet sein wird. Dies betrifft zwar in erster Linie die AGENDER-Implementierung, bringt jedoch auch die Notwendigkeit mit sich, explizite Laufzeit-Studien (*Benchmarks*) der hier untersuchten Verfahren anzustellen. Derzeit wurden diesbezüglich lediglich allgemeine Aussagen auf Basis von deren theoretischer Komplexität getroffen. Die Benchmarks sollen außerdem Aufschluss darüber geben, ob die manuell konfigurierten Gruppierungen tatsächlich einen Laufzeit-Vorteil einbringen.

Darüber hinaus wird im Rahmen des genannten Projektes eine präzisere Einschätzung des Sprecheralters angestrebt, wofür die oben aufgeführten Studien als Grundlage dienen sollen. Hinsichtlich der Mustererkennung müssen zusätzlich methodologische Veränderungen in Betracht gezogen werden. Das Performanzmaß der True Positive Rates wird z. B. ab einer gewissen Anzahl von Klassen nicht länger praktikabel sein und sollte durch ein Maß ersetzt werden, welches den Abstand zwischen der tatsächlichen Klasse und der geschätzten Klasse berücksichtigt. Möglicherweise rücken in diesem Zusammenhang auch andere Klassifikationsverfahren in den Vordergrund, die von vornherein statt einer Zuordnung eines Merkmalsvektors zu einer diskreten Klasse eine lineare Regression durchführen, d. h. das Alter als numerischen Wert vorhersagen. Der Vorteil bestünde unter anderem darin, dass ein unmittelbarer Vergleich der Performanz des Modells mit der menschlichen Fähigkeit zur Einschätzung des Alters erfolgen könnte.

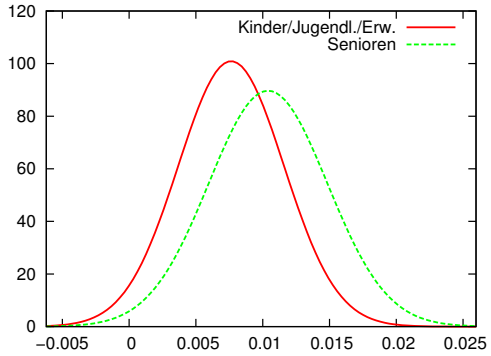
Die neuerliche Konzentration auf Anwendungen im Telekommunikationsbereich wirkt sich auch auf die Weiterentwicklung der Implementierung aus. In einem AGENDER-System für Callcenter wird es die Komponenten m3i Server und m3i Client in dieser Form nicht geben, da vollkommen andere Anforderungen an das System gestellt werden. Was in dem Fall benötigt wird, ist ein Klassifikationsmodul, das in eine bestehende Callcenter-Plattform integriert werden kann, multiple Kanäle unterstützt – d. h. eine hohe Anzahl von gleichzeitig eingehenden Anrufen bearbeiten kann – und vor allen Dingen schnell ist. Um die Sprecherklassifikation sinnvoll in einen Callcenter-Dienst oder in ein Voice-Response-System einbinden zu können, werden von den Auftraggebern Antwortzeiten gefordert, die im Bereich von einhundert bis zweihundert Millisekunden liegen. Gleichzeitig müssen auch konzeptuelle Weiterentwicklungen vollzogen werden, was bisher ausschließlich auf Basis des Servers geschehen ist.

Teil V

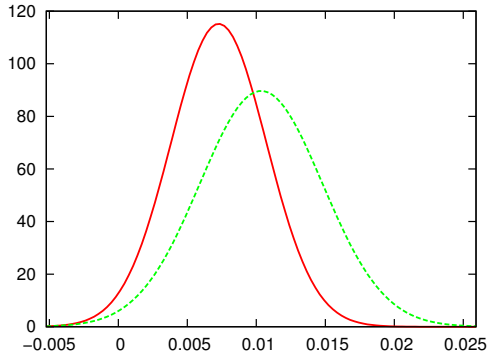
Anhang

A Weitere Ergebnisse der Korpusanalysen

Jitter $jitt\_rap$

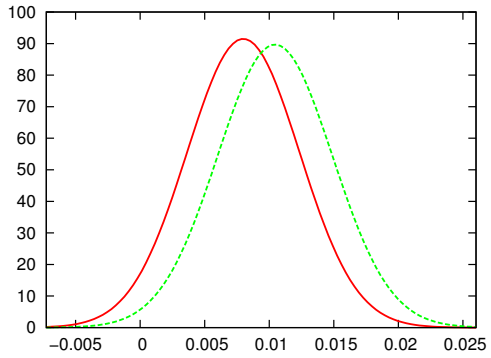


	μ	Σ	$\bar{\mu}$
gesamt	0.0083	0.0043	0
Kinder/Jugendl./Erw.	0.0076	0.004	-0.1625
Senioren	0.0104	0.0044	0.4875



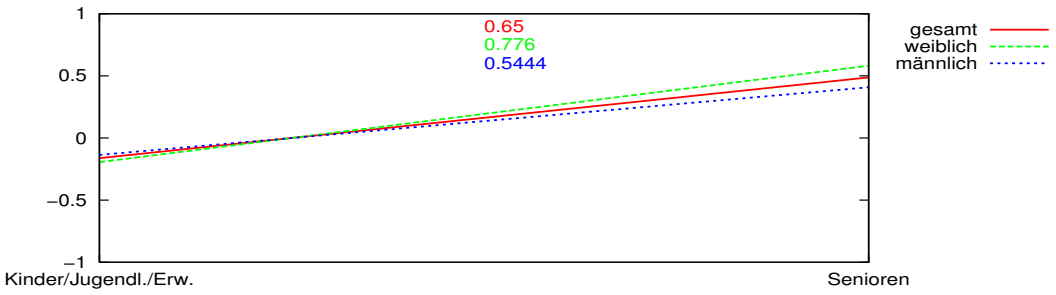
weibliche Sprecher

gesamt	0.008	0.004	0
Kinder/Jugendl./Erw.	0.0073	0.0035	-0.194
Senioren	0.0104	0.0044	0.582

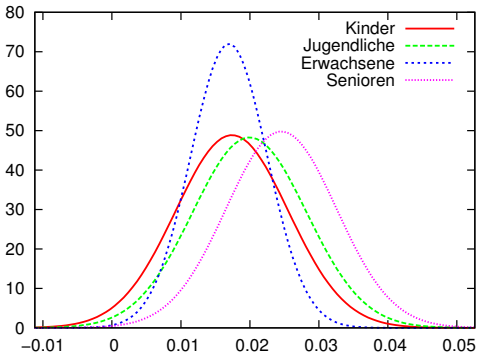


männliche Sprecher

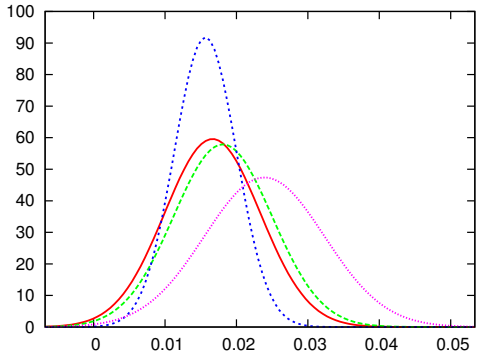
gesamt	0.0086	0.0045	0
Kinder/Jugendl./Erw.	0.008	0.0044	-0.1361
Senioren	0.0104	0.0044	0.4083



Jitter $jitt\_l$

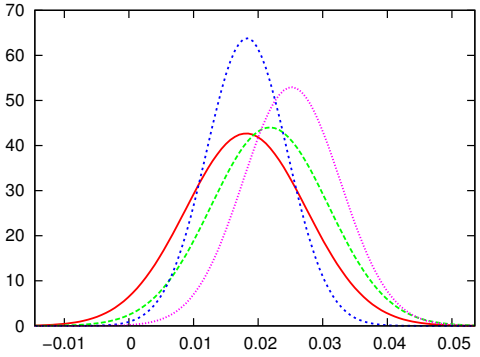


	μ	Σ	$\bar{\mu}$
gesamt	0.0197	0.0082	0
Kinder	0.0174	0.0082	-0.2842
Jugendliche	0.02	0.0083	0.0293
Erwachsene	0.017	0.0055	-0.3378
Senioren	0.0246	0.008	0.5927



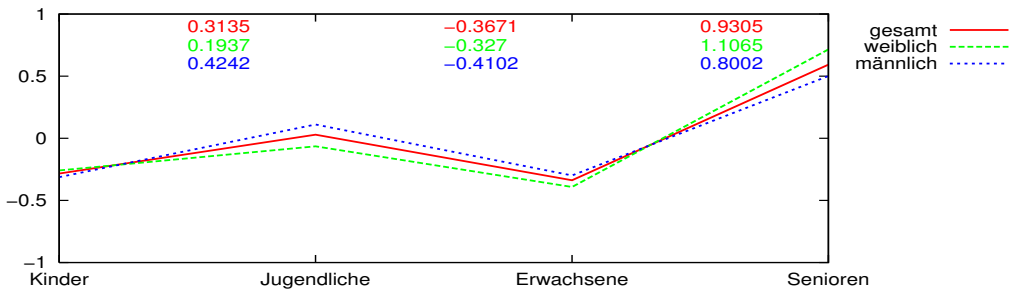
weibliche Sprecher

gesamt	0.0186	0.0075	0
Kinder	0.0167	0.0067	-0.2584
Jugendliche	0.0181	0.0069	-0.0647
Erwachsene	0.0157	0.0043	-0.3917
Senioren	0.0239	0.0084	0.7148

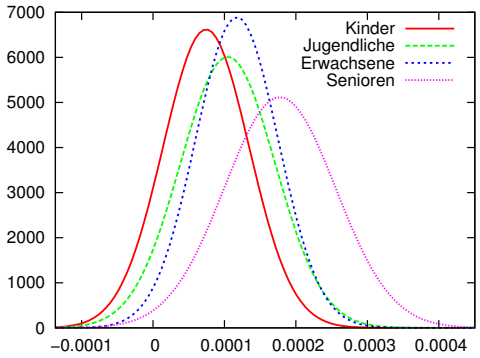


männliche Sprecher

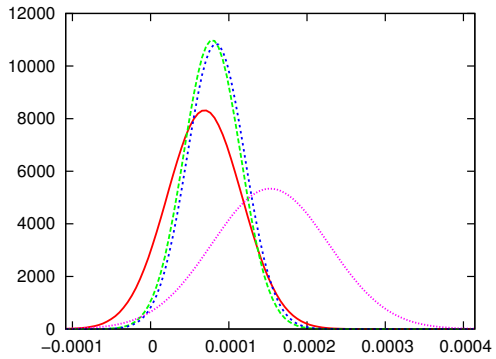
gesamt	0.0209	0.0086	0
Kinder	0.0181	0.0093	-0.3131
Jugendliche	0.0218	0.0091	0.1111
Erwachsene	0.0183	0.0062	-0.2991
Senioren	0.0252	0.0075	0.5011



Jitter *jitt_la*

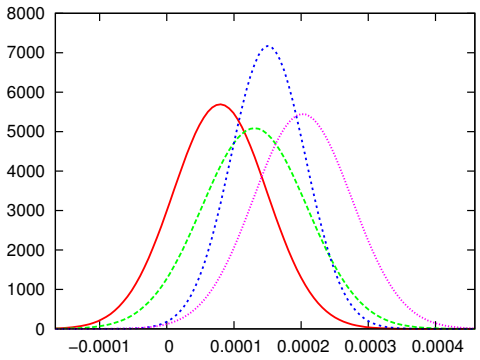


	μ	Σ	$\bar{\mu}$
gesamt	0.0001	0.0001	0
Kinder	0.0001	0.0001	-0.5816
Jugendliche	0.0001	0.0001	-0.1818
Erwachsene	0.0001	0.0001	-0.0149
Senioren	0.0002	0.0001	0.7782



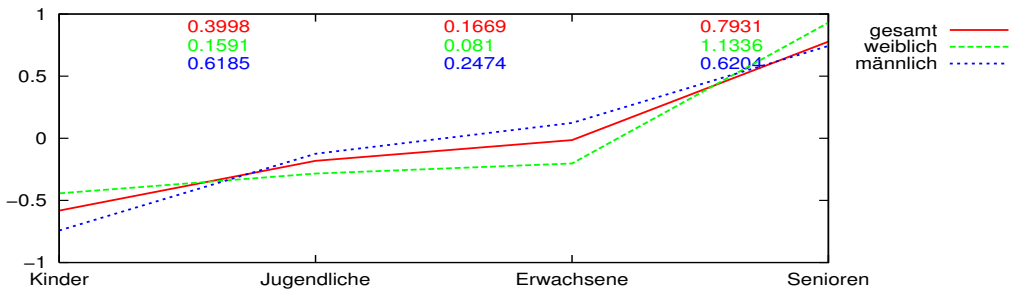
weibliche Sprecher

gesamt	0.0001	0.0001	0
Kinder	0.0001	0	-0.4432
Jugendliche	0.0001	0	-0.2841
Erwachsene	0.0001	0	-0.2031
Senioren	0.0002	0.0001	0.9305

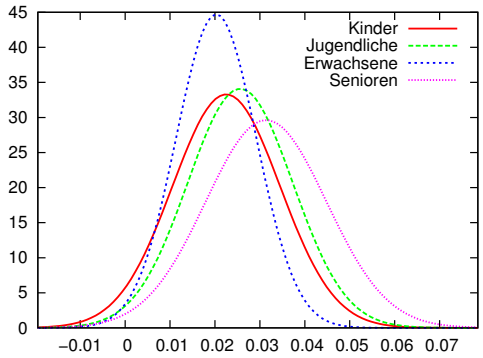


männliche Sprecher

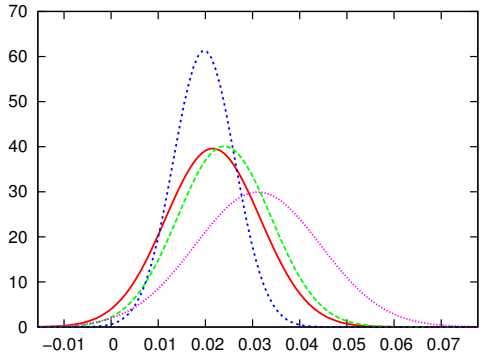
gesamt	0.0001	0.0001	0
Kinder	0.0001	0.0001	-0.7427
Jugendliche	0.0001	0.0001	-0.1242
Erwachsene	0.0002	0.0001	0.1232
Senioren	0.0002	0.0001	0.7436



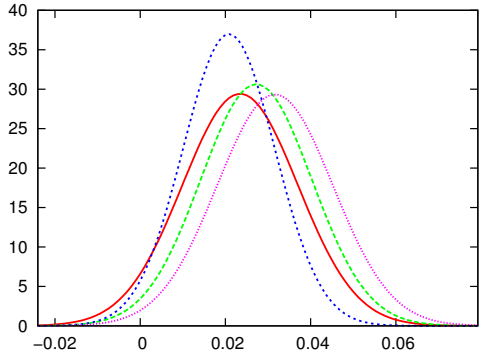
Jitter *jitt_ddp*



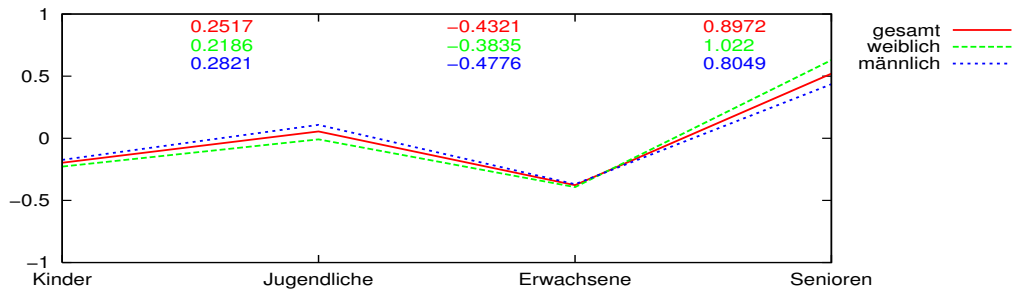
	μ	Σ	$\bar{\mu}$
gesamt	0.0249	0.0124	0
Kinder	0.0225	0.012	-0.197
Jugendliche	0.0256	0.0117	0.0547
Erwachsene	0.0203	0.0089	-0.3774
Senioren	0.0314	0.0135	0.5198



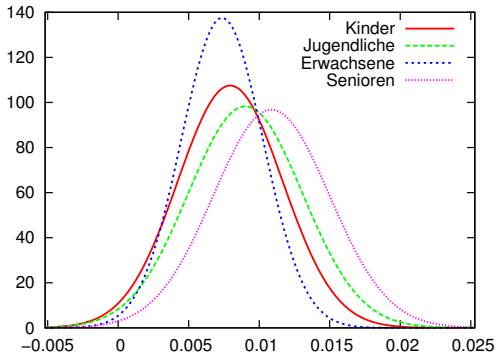
weibliche Sprecher			
gesamt	0.0241	0.0111	0
Kinder	0.0216	0.0101	-0.2277
Jugendliche	0.024	0.0099	-0.0091
Erwachsene	0.0197	0.0065	-0.3926
Senioren	0.0311	0.0133	0.6294



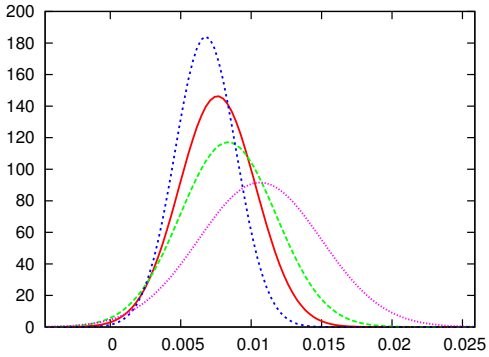
männliche Sprecher			
gesamt	0.0258	0.0134	0
Kinder	0.0235	0.0136	-0.174
Jugendliche	0.0272	0.013	0.1081
Erwachsene	0.0208	0.0108	-0.3695
Senioren	0.0316	0.0136	0.4354



Jitter *jitt_ppq*

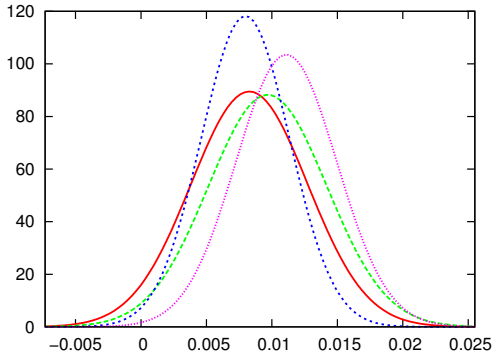


	μ	Σ	$\bar{\mu}$
gesamt	0.0088	0.004	0
Kinder	0.0079	0.0037	-0.2163
Jugendliche	0.009	0.0041	0.0553
Erwachsene	0.0074	0.0029	-0.3607
Senioren	0.0109	0.0041	0.5216



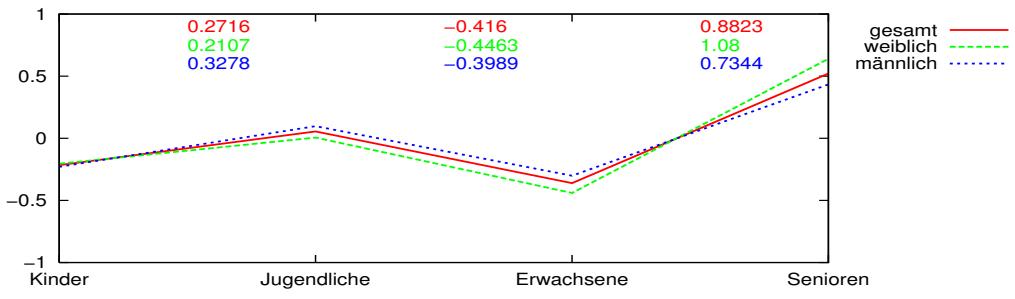
weibliche Sprecher

	μ	Σ	$\bar{\mu}$
gesamt	0.0083	0.0036	0
Kinder	0.0076	0.0027	-0.2049
Jugendliche	0.0084	0.0034	0.0058
Erwachsene	0.0068	0.0022	-0.4405
Senioren	0.0106	0.0044	0.6395

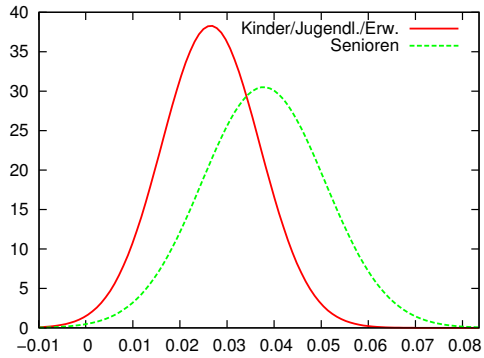


männliche Sprecher

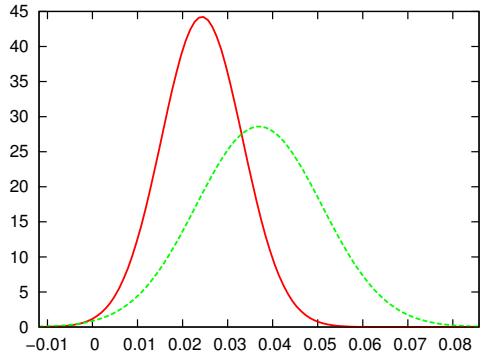
	μ	Σ	$\bar{\mu}$
gesamt	0.0093	0.0043	0
Kinder	0.0083	0.0045	-0.23
Jugendliche	0.0097	0.0045	0.0978
Erwachsene	0.008	0.0034	-0.3011
Senioren	0.0111	0.0039	0.4333



Shimmer $shim\_apq3$

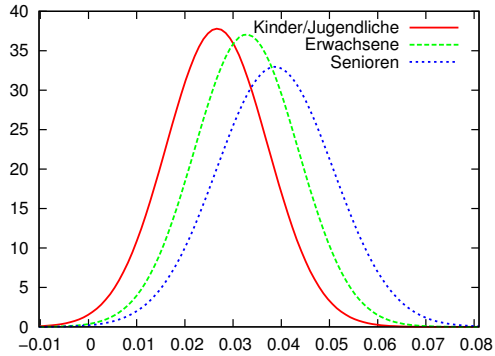


	μ	Σ	$\bar{\mu}$
gesamt	0.0293	0.0122	0
Kinder/Jugendl./Erw.	0.0265	0.0104	-0.2307
Senioren	0.0378	0.0131	0.6921



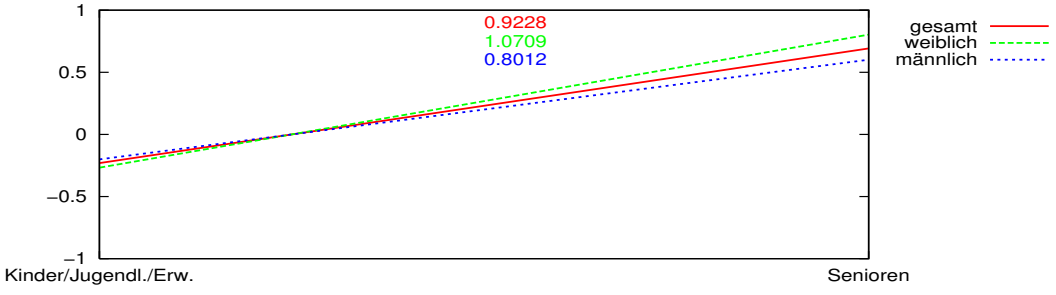
weibliche Sprecher

gesamt	0.0275	0.0118	0
Kinder/Jugendl./Erw.	0.0243	0.009	-0.2677
Senioren	0.037	0.014	0.8032

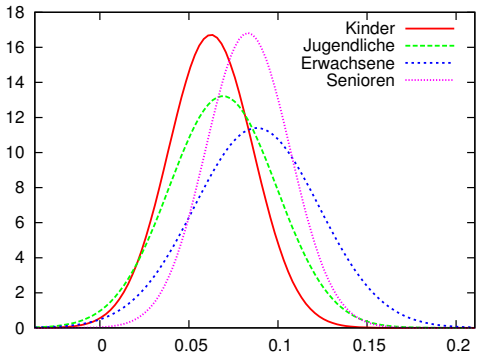


männliche Sprecher

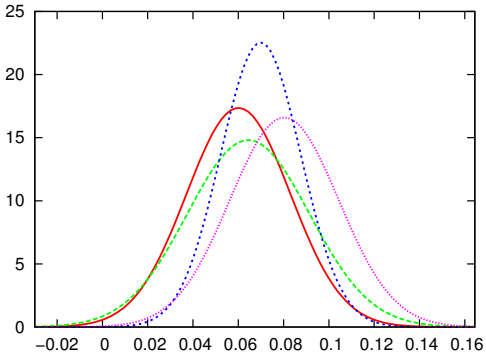
gesamt	0.0312	0.0121	0
Kinder/Jugendliche	0.0267	0.0106	-0.3741
Erwachsene	0.0328	0.0108	0.1295
Senioren	0.0387	0.0121	0.6187



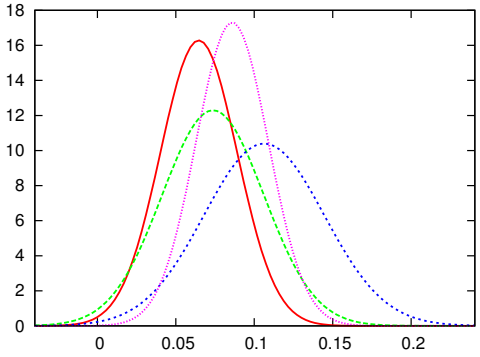
Shimmer — $shim\_apq11$



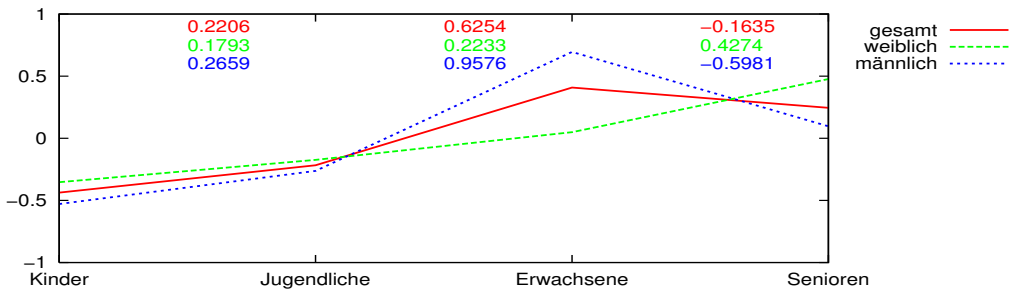
	μ	Σ	$\bar{\mu}$
gesamt	0.0757	0.0304	0
Kinder	0.0624	0.0239	-0.4373
Jugendliche	0.0691	0.0302	-0.2167
Erwachsene	0.0881	0.035	0.4087
Senioren	0.0831	0.0237	0.2452



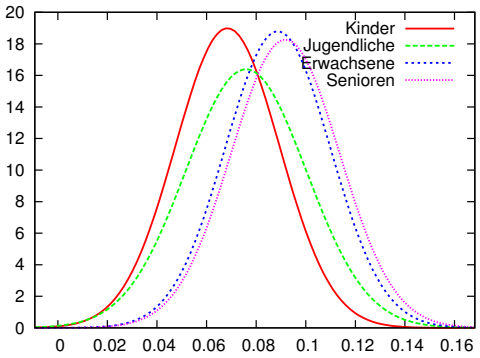
weibliche Sprecher			
gesamt	0.0687	0.0244	0
Kinder	0.0601	0.023	-0.353
Jugendliche	0.0645	0.0269	-0.1737
Erwachsene	0.0699	0.0177	0.0496
Senioren	0.0803	0.024	0.477



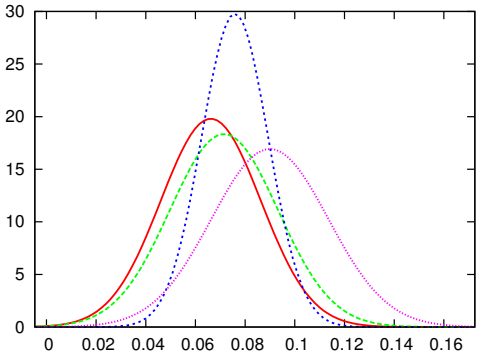
männliche Sprecher			
gesamt	0.0826	0.034	0
Kinder	0.0647	0.0245	-0.5287
Jugendliche	0.0737	0.0324	-0.2628
Erwachsene	0.1063	0.0384	0.6948
Senioren	0.0859	0.0231	0.0967



Shimmer $shim\_l$

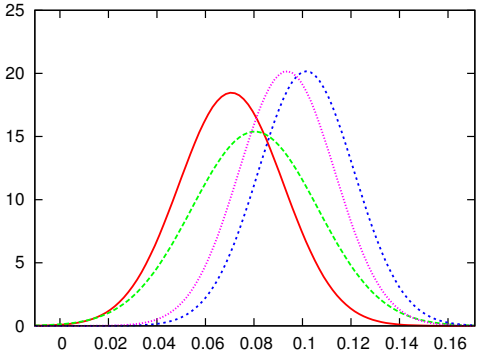


	μ	Σ	$\bar{\mu}$
gesamt	0.0812	0.0241	0
Kinder	0.0684	0.021	-0.531
Jugendliche	0.0759	0.0243	-0.2197
Erwachsene	0.0887	0.0212	0.312
Senioren	0.0918	0.0218	0.4387



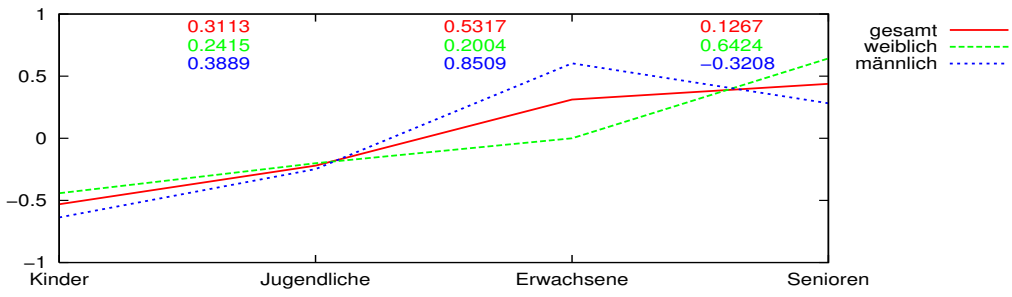
weibliche Sprecher

gesamt	0.0759	0.0219	0
Kinder	0.0662	0.0201	-0.4419
Jugendliche	0.0715	0.0218	-0.2004
Erwachsene	0.0759	0.0134	-0
Senioren	0.09	0.0236	0.6424

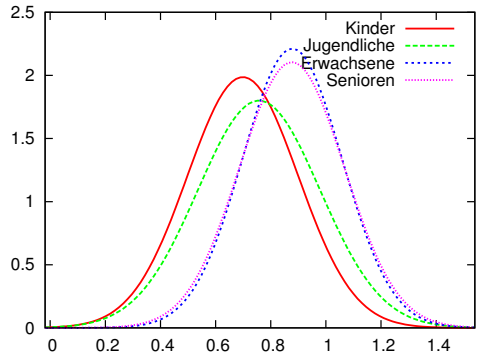


männliche Sprecher

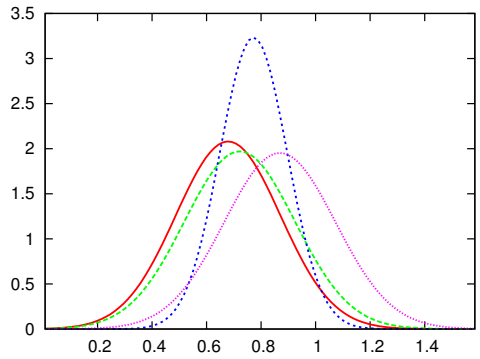
gesamt	0.0865	0.0249	0
Kinder	0.0707	0.0216	-0.6369
Jugendliche	0.0803	0.0259	-0.248
Erwachsene	0.1016	0.0198	0.6029
Senioren	0.0936	0.0198	0.2821



Shimmer — *shim_ldb*

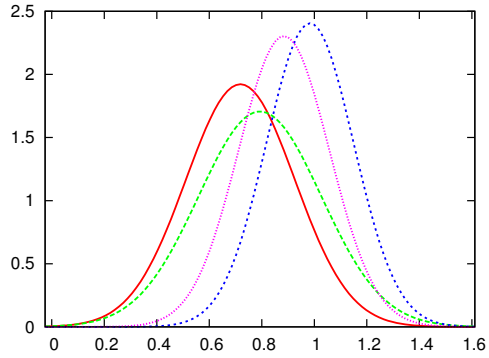


	μ	Σ	$\bar{\mu}$
gesamt	0.8026	0.2133	0
Kinder	0.6987	0.2008	-0.4874
Jugendliche	0.7573	0.2216	-0.2127
Erwachsene	0.8779	0.1804	0.3531
Senioren	0.8766	0.1896	0.347



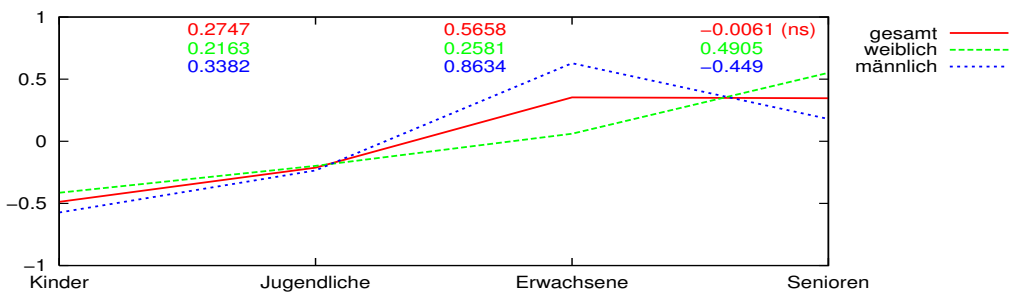
weibliche Sprecher

gesamt	0.7604	0.1967	0
Kinder	0.679	0.1918	-0.4139
Jugendliche	0.7215	0.2025	-0.1976
Erwachsene	0.7723	0.1234	0.0605
Senioren	0.8688	0.2043	0.551

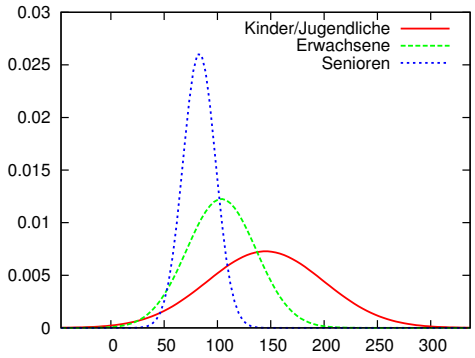


männliche Sprecher

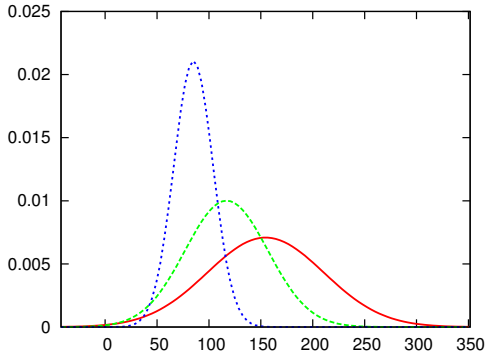
gesamt	0.8448	0.2207	0
Kinder	0.7184	0.2076	-0.5731
Jugendliche	0.793	0.2338	-0.2349
Erwachsene	0.9835	0.1659	0.6285
Senioren	0.8845	0.1732	0.1795



Minimale Stimmtonhöhe *pitch_min*

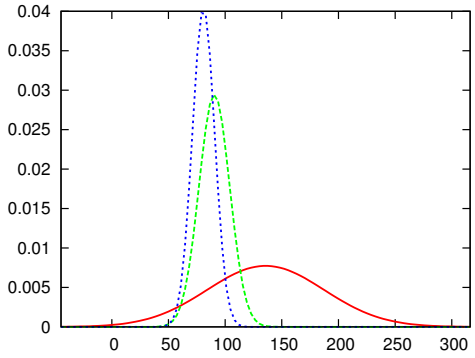


	μ	Σ	$\bar{\mu}$
gesamt	119.24	50.49	0
Kinder/Jugendliche	145.13	54.81	0.51
Erwachsene	103.69	32.56	-0.31
Senioren	83.01	15.23	-0.72



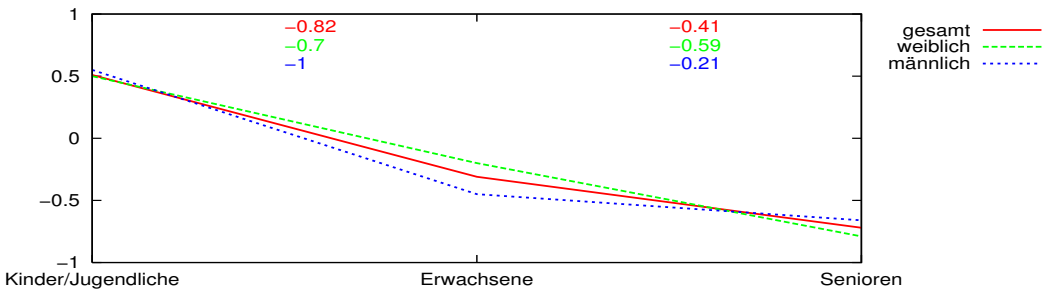
weibliche Sprecher

gesamt	127.77	54	0
Kinder/Jugendliche	154.56	56.31	0.5
Erwachsene	116.89	39.85	-0.2
Senioren	85.08	18.91	-0.79

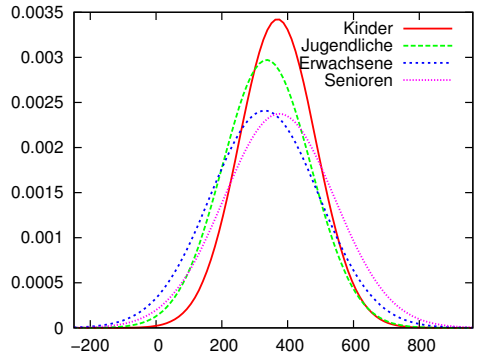


männliche Sprecher

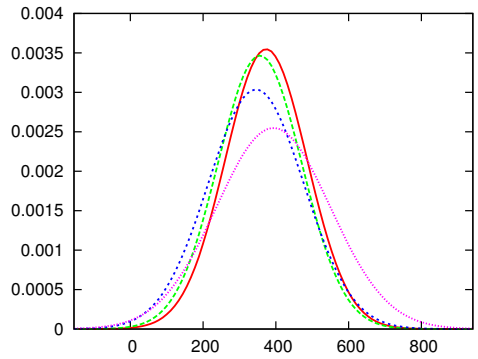
gesamt	110.71	45.12	0
Kinder/Jugendliche	135.7	51.57	0.55
Erwachsene	90.5	13.55	-0.45
Senioren	80.93	9.88	-0.66



Maximale Stimmtonhöhe — *pitch_max*

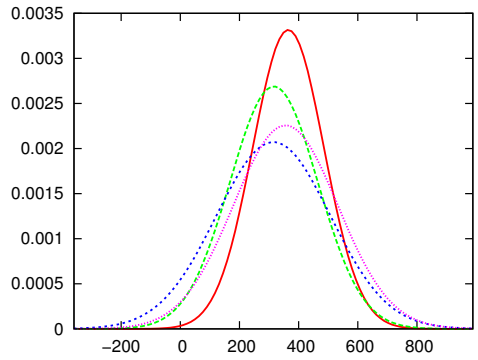


	μ	Σ	$\bar{\mu}$
gesamt	352.58	148.99	0
Kinder	369.31	116.54	0.11
Jugendliche	336.25	134.24	-0.11
Erwachsene	330.12	165.57	-0.15
Senioren	374.64	168.02	0.15



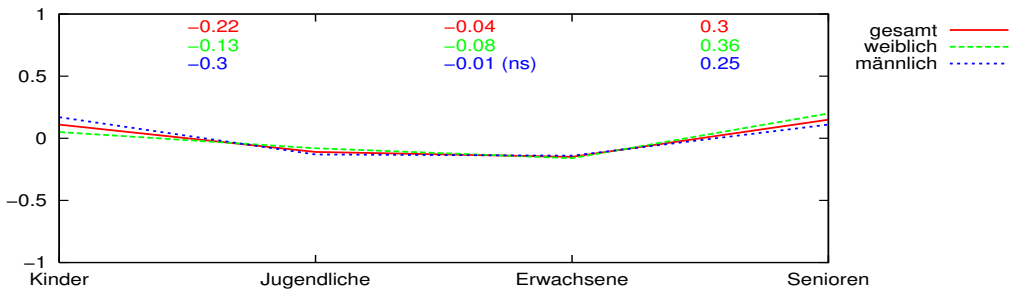
weibliche Sprecher

gesamt	367.01	131.34	0
Kinder	373.33	112.48	0.05
Jugendliche	356.05	115.1	-0.08
Erwachsene	345.42	131.41	-0.16
Senioren	393.24	156.58	0.2

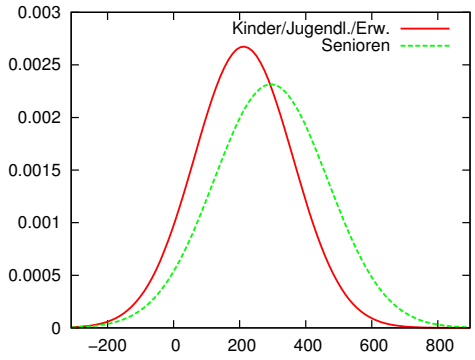


männliche Sprecher

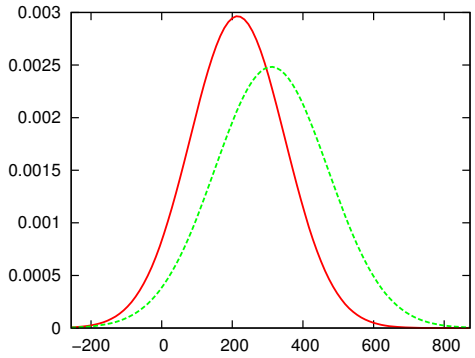
gesamt	338.15	163.48	0
Kinder	365.3	120.32	0.17
Jugendliche	316.45	148.36	-0.13
Erwachsene	314.83	192.59	-0.14
Senioren	356.04	176.79	0.11



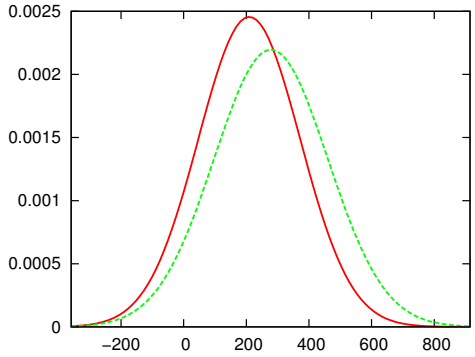
Stimmumfang *pitch_diff*



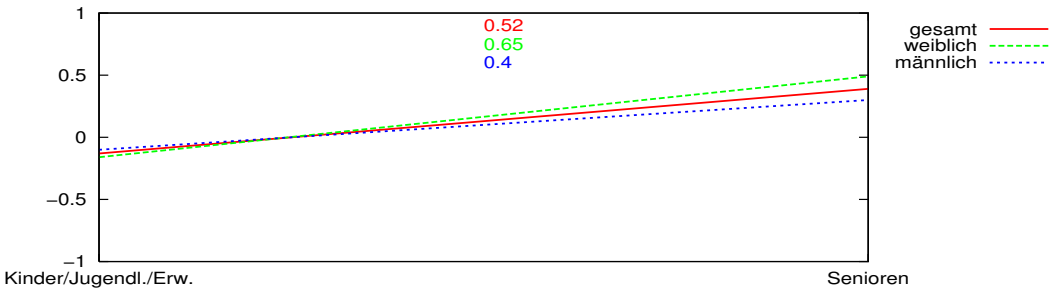
	μ	Σ	$\bar{\mu}$
gesamt	232.75	159.32	0
Kinder/Jugendl./Erw.	212.14	149.18	-0.13
Senioren	294.61	172.23	0.39



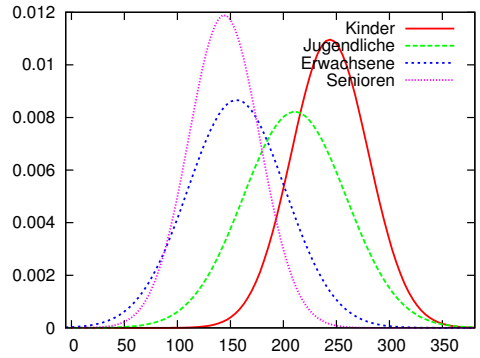
weibliche Sprecher			
gesamt	238.8	147.5	0
Kinder/Jugendl./Erw.	214.84	134.56	-0.16
Senioren	310.69	160.67	0.49



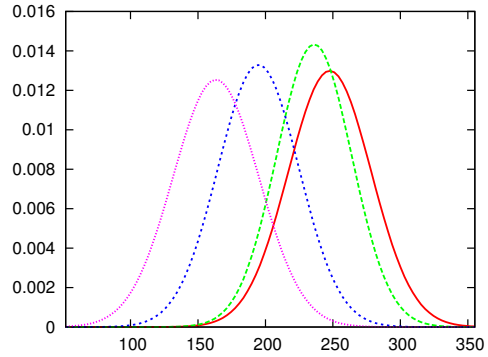
männliche Sprecher			
gesamt	226.71	170.11	0
Kinder/Jugendl./Erw.	209.43	162.45	-0.1
Senioren	278.53	181.65	0.3



Mittlere Stimmtonhöhe — *pitch_quant*

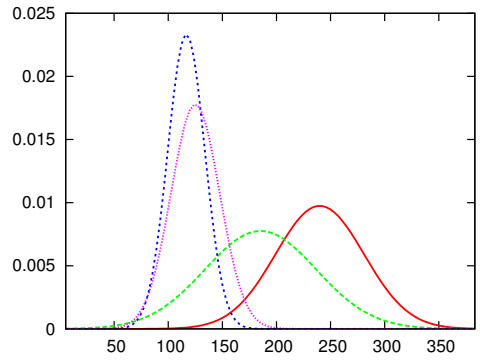


	μ	Σ	$\bar{\mu}$
gesamt	188.63	58.13	0
Kinder	243.86	36.39	0.95
Jugendliche	210.7	48.54	0.38
Erwachsene	155.78	46.05	-0.57
Senioren	144.2	33.53	-0.76



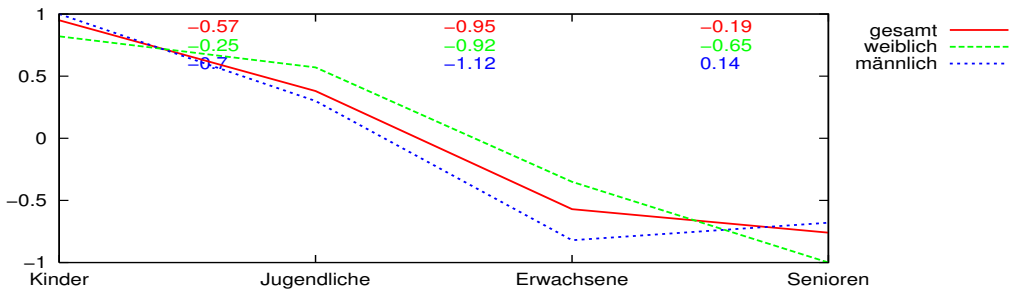
weibliche Sprecher

gesamt	210.49	45.11	0
Kinder	247.7	30.72	0.82
Jugendliche	236.11	27.84	0.57
Erwachsene	194.82	30.01	-0.35
Senioren	163.34	31.82	-1

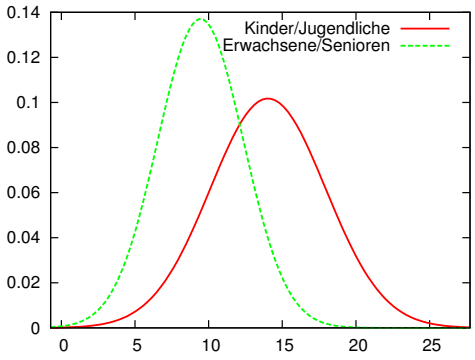


männliche Sprecher

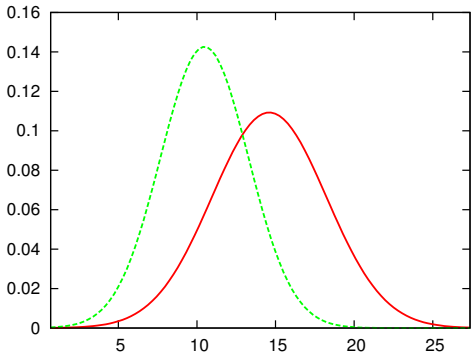
gesamt	166.78	61.38	0
Kinder	240.02	40.94	1
Jugendliche	185.3	51.44	0.3
Erwachsene	116.74	17.07	-0.82
Senioren	125.06	22.45	-0.68



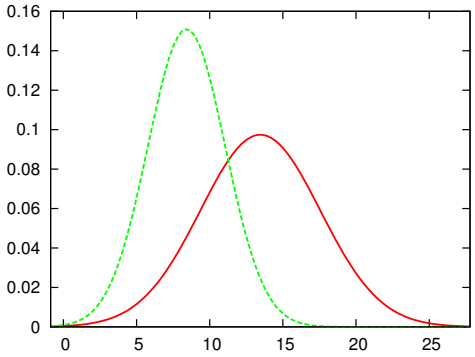
Harmonizität *harm_mean*



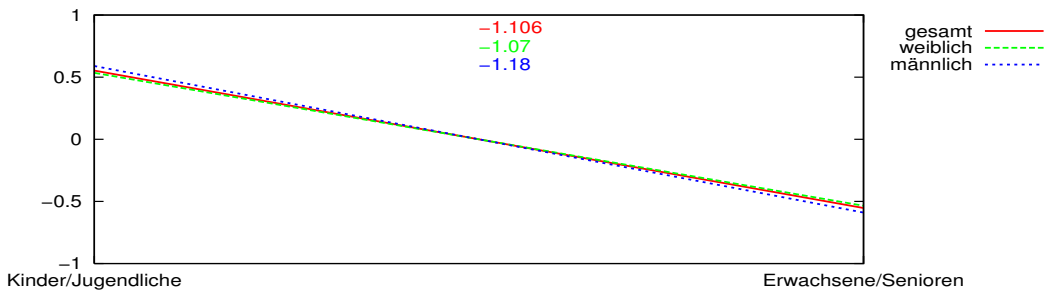
	μ	Σ	$\bar{\mu}$
gesamt	11.737	4.142	0
Kinder/Jugendliche	14.027	3.921	0.553
Erwachsene/Senioren	9.448	2.909	-0.553



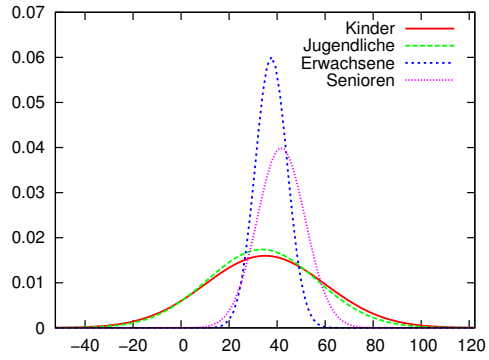
weibliche Sprecher			
gesamt	12.536	3.849	0
Kinder/Jugendliche	14.595	3.651	0.535
Erwachsene/Senioren	10.477	2.797	-0.535



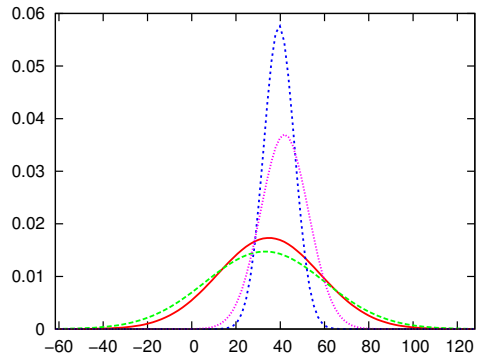
männliche Sprecher			
gesamt	10.939	4.269	0
Kinder/Jugendliche	13.459	4.095	0.59
Erwachsene/Senioren	8.418	2.643	-0.59



Harmonizität — $harm\_max$

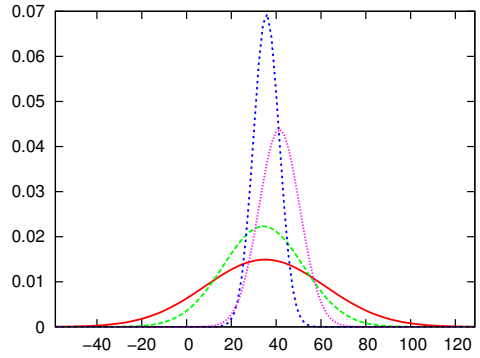


	μ	Σ	$\bar{\mu}$
gesamt	36.97	18.24	0
Kinder	34.98	24.95	-0.11
Jugendliche	33.69	22.95	-0.18
Erwachsene	37.55	6.64	0.03
Senioren	41.67	9.99	0.26



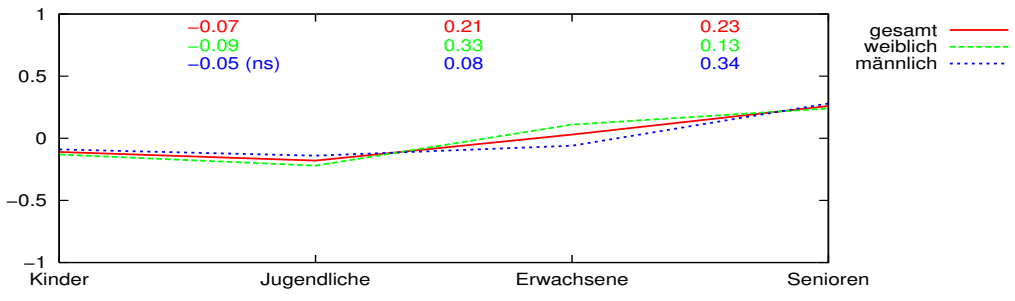
weibliche Sprecher

gesamt	37.36	19.21	0
Kinder	34.9	23.02	-0.13
Jugendliche	33.15	27.08	-0.22
Erwachsene	39.48	6.91	0.11
Senioren	41.92	10.78	0.24

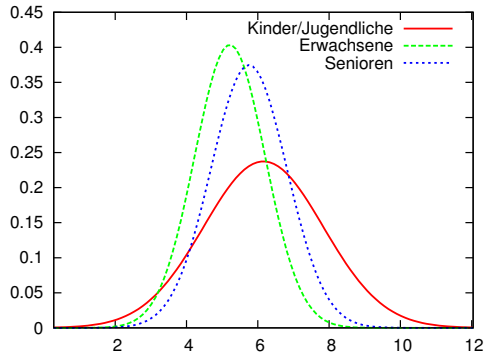


männliche Sprecher

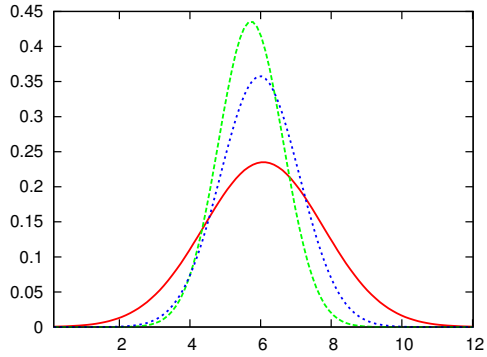
gesamt	36.59	17.2	0
Kinder	35.07	26.75	-0.09
Jugendliche	34.23	17.87	-0.14
Erwachsene	35.63	5.74	-0.06
Senioren	41.42	9.11	0.28



Harmonizität $harm\_stddev$

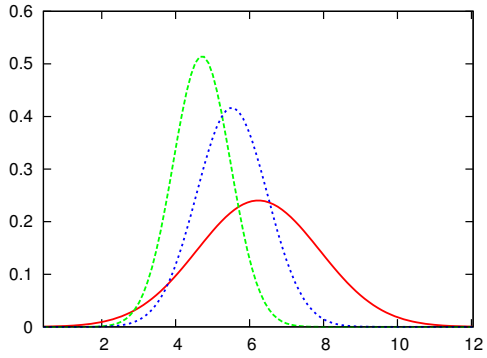


	μ	Σ	$\bar{\mu}$
gesamt	5.821	1.445	0
Kinder/Jugendliche	6.158	1.681	0.233
Erwachsene	5.217	0.988	-0.418
Senioren	5.751	1.065	-0.048



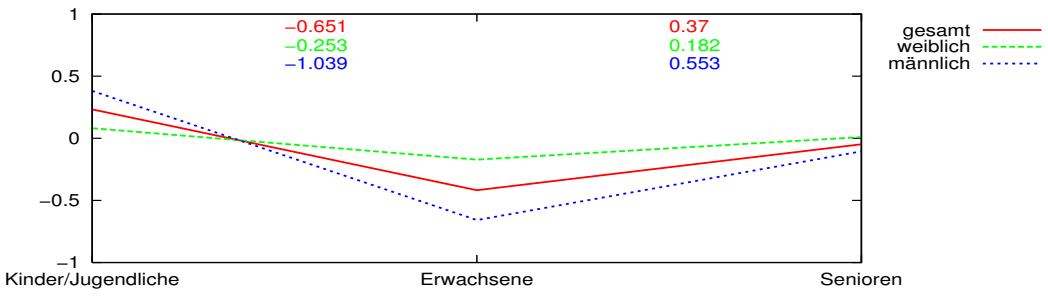
weibliche Sprecher

gesamt	5.967	1.407	0
Kinder/Jugendliche	6.081	1.697	0.081
Erwachsene	5.725	0.915	-0.172
Senioren	5.981	1.115	0.01

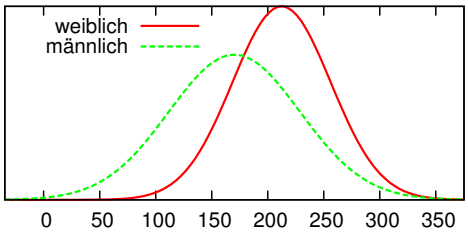


männliche Sprecher

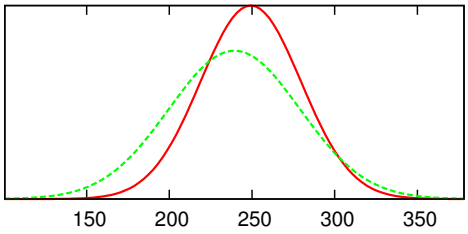
gesamt	5.676	1.468	0
Kinder/Jugendliche	6.235	1.661	0.381
Erwachsene	4.709	0.775	-0.658
Senioren	5.522	0.958	-0.105



Mittlere Stimmtonhöhe — *pitch_mean*

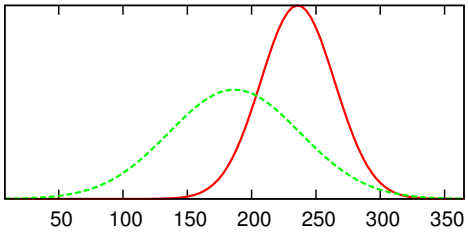


	μ	Σ	$\bar{\mu}$
gesamt	191.31	55.87	0
weiblich	212.38	43.9	0.38
männlich	170.24	58.55	-0.38



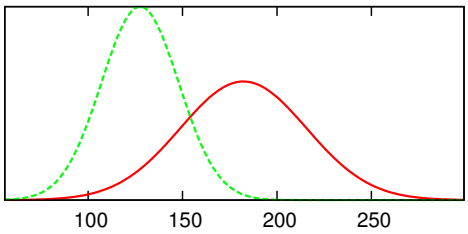
Kinder

gesamt	244.31	35.68	0
weiblich	249.26	30.39	0.14
männlich	239.37	39.67	-0.14



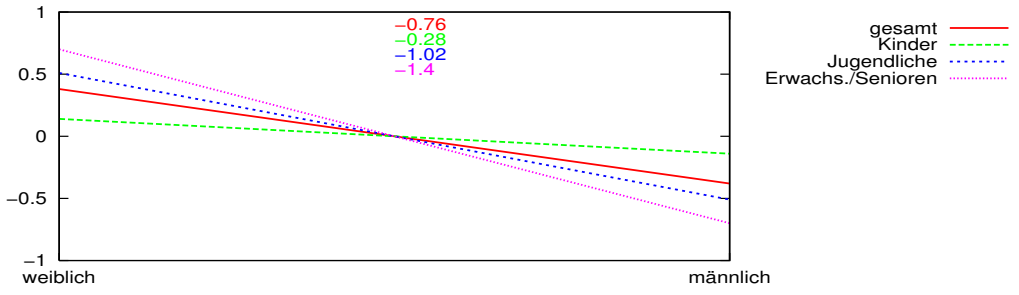
Jugendliche

gesamt	211.21	48.17	0
weiblich	235.81	28.79	0.51
männlich	186.61	51.01	-0.51

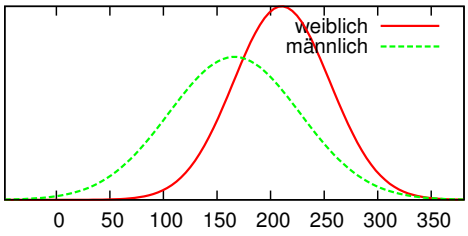


Erwachs./Senioren

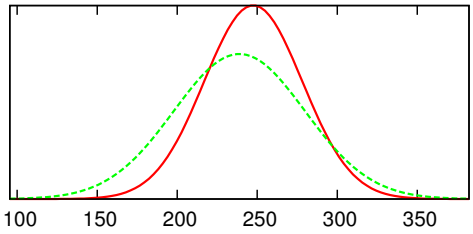
gesamt	154.86	38.92	0
weiblich	182.22	33.36	0.7
männlich	127.49	20.46	-0.7



Mittlere Stimmtonhöhe *pitch_quant*

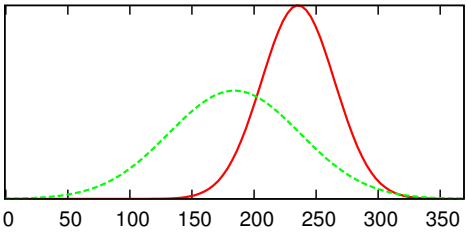


	μ	Σ	$\bar{\mu}$
gesamt	188.31	58.19	0
weiblich	210.4	45.23	0.38
männlich	166.22	61.24	-0.38



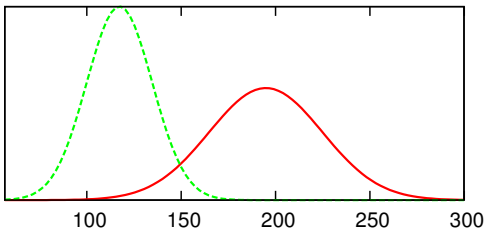
Kinder

gesamt	243.18	36.47	0
weiblich	247.51	30.7	0.12
männlich	238.85	40.99	-0.12



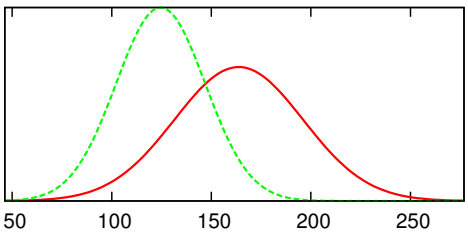
Jugendliche

gesamt	209.83	49.91	0
weiblich	235.49	29.54	0.51
männlich	184.18	52.85	-0.51



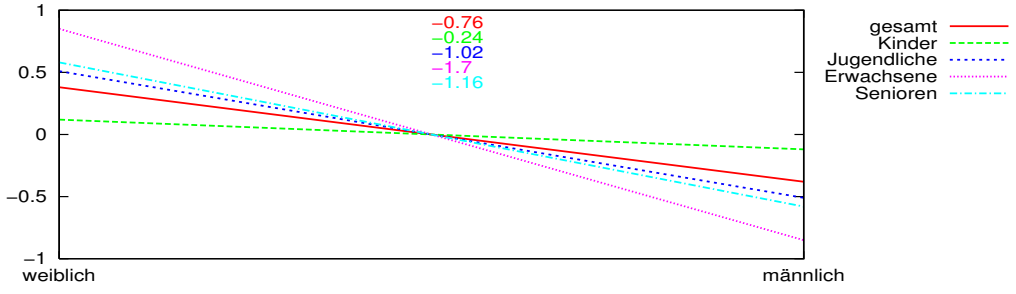
Erwachsene

gesamt	156.06	45.86	0
weiblich	194.82	30.01	0.85
männlich	117.29	17.37	-0.85

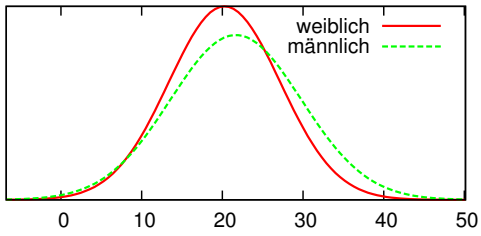


Senioren

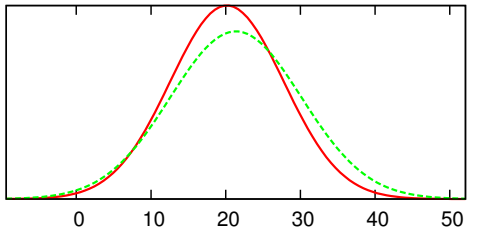
gesamt	144.16	34.01	0
weiblich	163.77	32.32	0.58
männlich	124.56	22.36	-0.58



Stimmtonhöhen-Veränderungen *pitch_swoj*

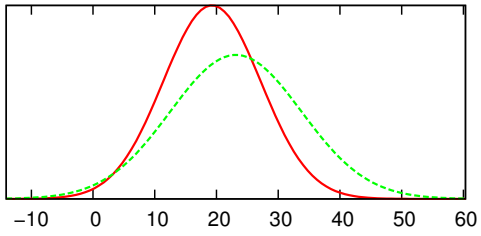


	μ	Σ	$\bar{\mu}$
gesamt	20.976	7.585	0
weiblich	20.271	6.925	-0.093
männlich	21.681	8.131	0.093



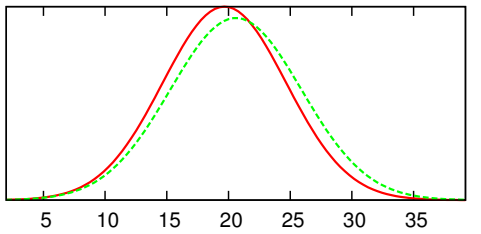
Kinder

gesamt	20.766	8.236	0
weiblich	20.17	7.606	-0.072
männlich	21.362	8.782	0.072



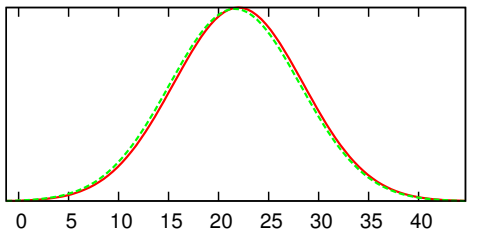
Jugendliche

gesamt	21.2	9.566	0
weiblich	19.262	7.907	-0.203
männlich	23.138	10.629	0.203



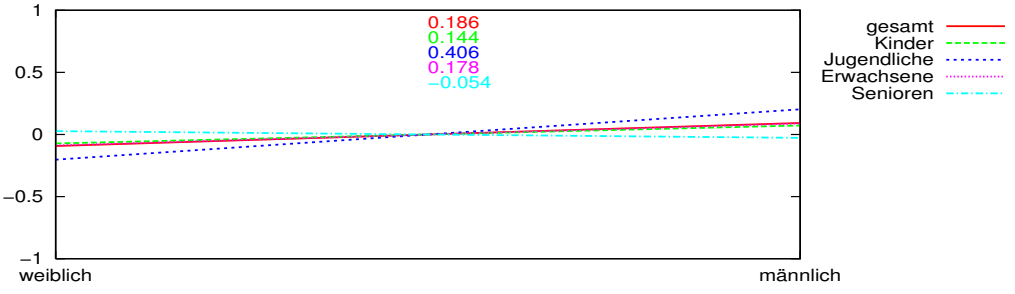
Erwachsene

gesamt	20.133	5.183	0
weiblich	19.67	5.004	-0.089
männlich	20.597	5.317	0.089

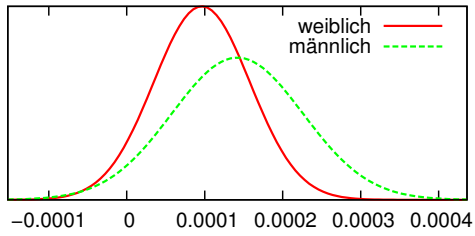


Senioren

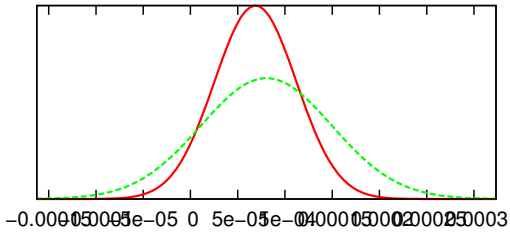
gesamt	21.804	6.513	0
weiblich	21.983	6.488	0.027
männlich	21.626	6.532	-0.027



Jitter *jitt_la*

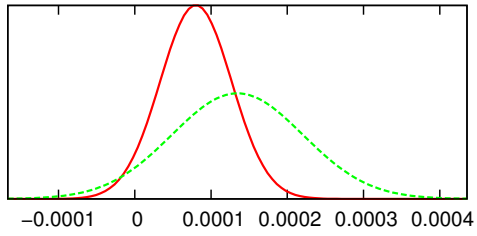


	μ	Σ	$\bar{\mu}$
gesamt	0.0001	0.0001	0
weiblich	0.0001	0.0001	-0.2955
männlich	0.0001	0.0001	0.2955



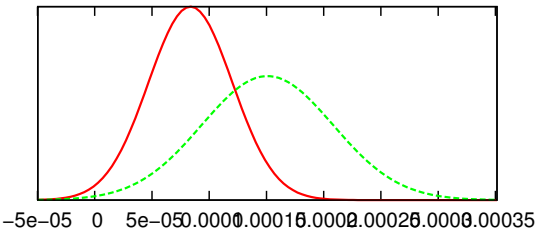
Kinder

gesamt	0.0001	0.0001	0
weiblich	0.0001	0	-0.1
männlich	0.0001	0.0001	0.1



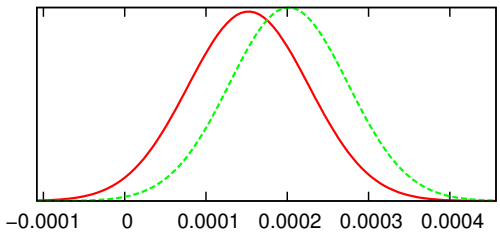
Jugendliche

gesamt	0.0001	0.0001	0
weiblich	0.0001	0	-0.3657
männlich	0.0001	0.0001	0.3657



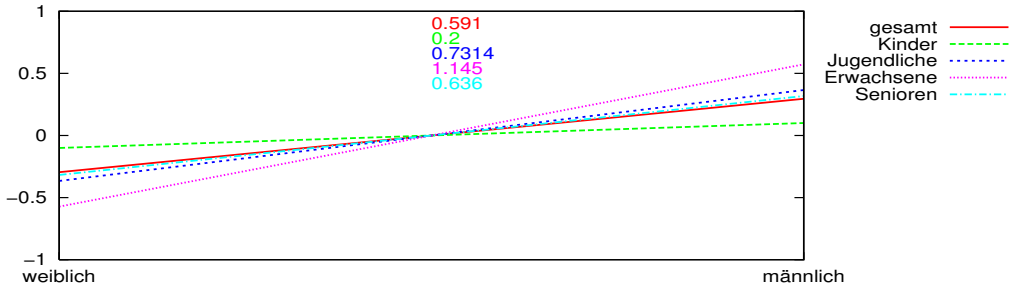
Erwachsene

gesamt	0.0001	0.0001	0
weiblich	0.0001	0	-0.5725
männlich	0.0002	0.0001	0.5725

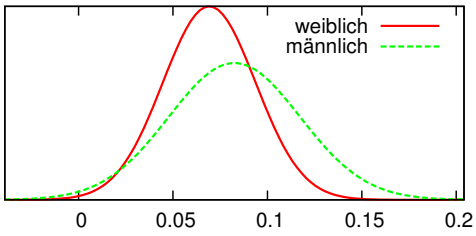


Senioren

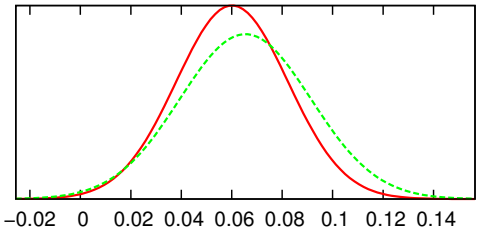
gesamt	0.0002	0.0001	0
weiblich	0.0002	0.0001	-0.318
männlich	0.0002	0.0001	0.318



Shimmer — $shim\_apq11$

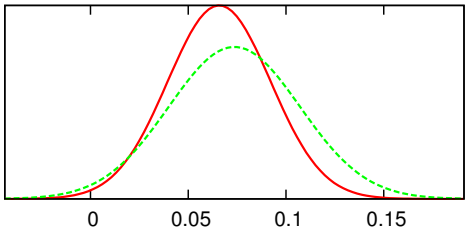


	μ	Σ	$\bar{\mu}$
gesamt	0.0758	0.0308	0
weiblich	0.0692	0.0246	-0.2163
männlich	0.0825	0.0347	0.2163



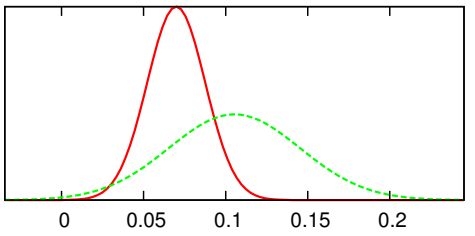
Kinder

	μ	Σ	$\bar{\mu}$
gesamt	0.0628	0.0243	0
weiblich	0.0601	0.0222	-0.1086
männlich	0.0654	0.026	0.1086



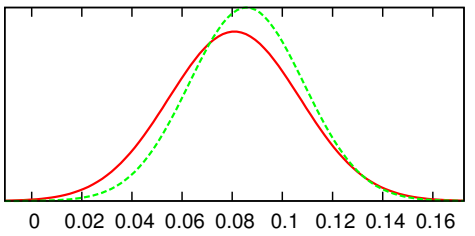
Jugendliche

	μ	Σ	$\bar{\mu}$
gesamt	0.0697	0.0304	0
weiblich	0.0658	0.0264	-0.1281
männlich	0.0736	0.0336	0.1281



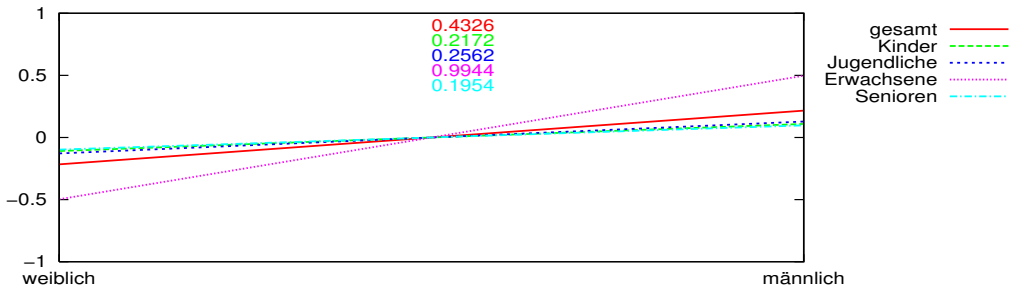
Erwachsene

	μ	Σ	$\bar{\mu}$
gesamt	0.0876	0.0356	0
weiblich	0.0699	0.0177	-0.4972
männlich	0.1053	0.04	0.4972

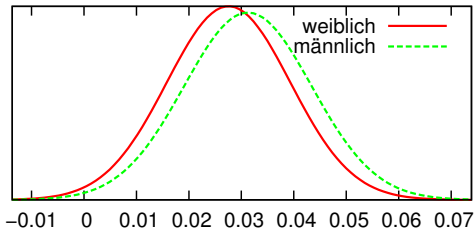


Senioren

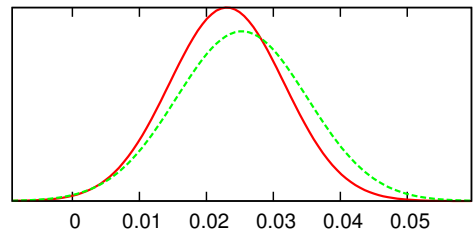
	μ	Σ	$\bar{\mu}$
gesamt	0.0833	0.0247	0
weiblich	0.0809	0.0262	-0.0977
männlich	0.0857	0.0229	0.0977



Shimmer $shim\_apq3$

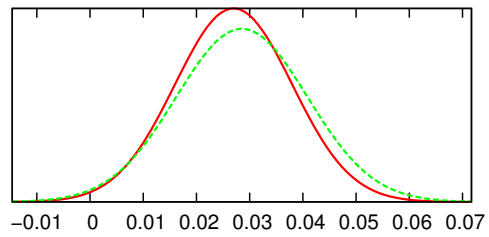


	μ	Σ	$\bar{\mu}$
gesamt	0.0294	0.0121	0
weiblich	0.0276	0.0118	-0.152
männlich	0.0312	0.0122	0.152



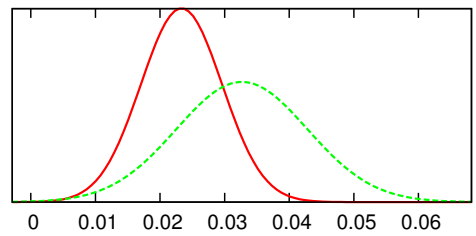
Kinder

gesamt	0.0242	0.0093	0
weiblich	0.0231	0.0086	-0.1207
männlich	0.0253	0.0098	0.1207



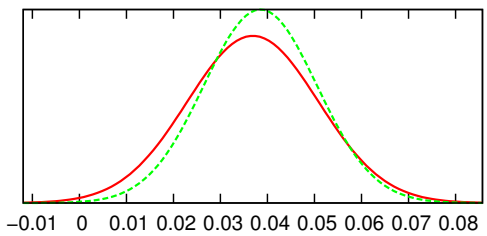
Jugendliche

gesamt	0.0278	0.0117	0
weiblich	0.027	0.011	-0.0665
männlich	0.0285	0.0123	0.0665



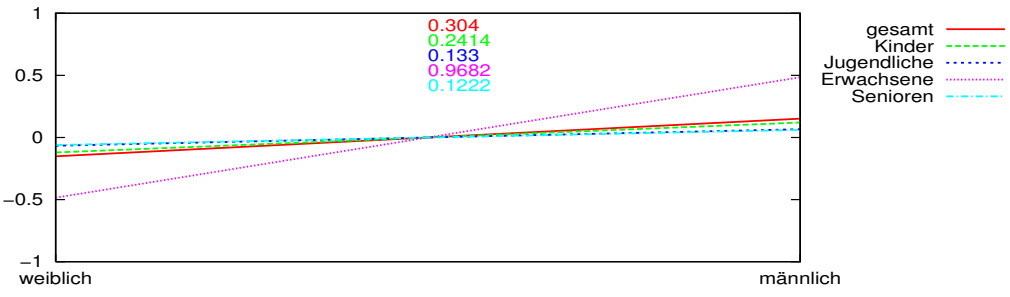
Erwachsene

gesamt	0.028	0.0097	0
weiblich	0.0233	0.0063	-0.4841
männlich	0.0327	0.0102	0.4841

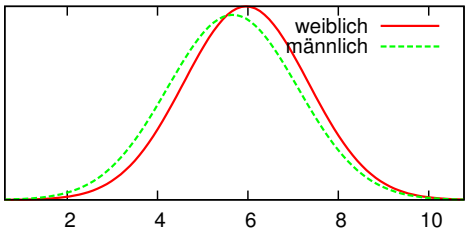


Senioren

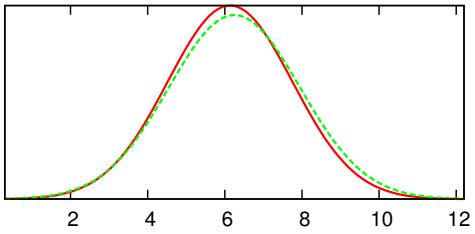
gesamt	0.0377	0.0131	0
weiblich	0.0369	0.014	-0.0611
männlich	0.0385	0.0121	0.0611



Harmonizität — $harm\_stddev$

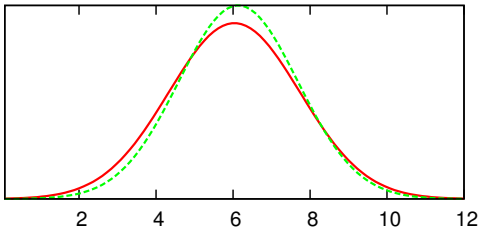


	μ	Σ	$\bar{\mu}$
gesamt	5.81	1.417	0
weiblich	5.957	1.378	0.104
männlich	5.662	1.441	-0.104



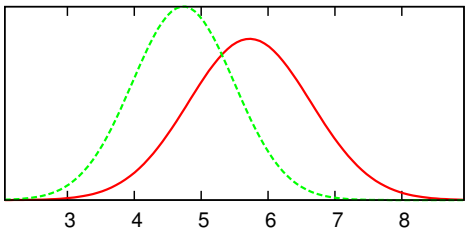
Kinder

gesamt	6.192	1.66	0
weiblich	6.137	1.617	-0.034
männlich	6.248	1.701	0.034



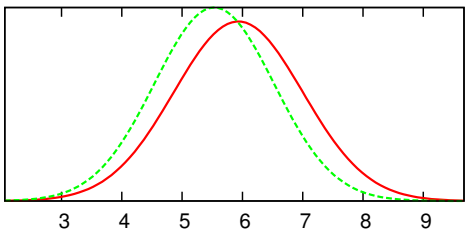
Jugendliche

gesamt	6.085	1.629	0
weiblich	6.035	1.704	-0.031
männlich	6.136	1.548	0.031



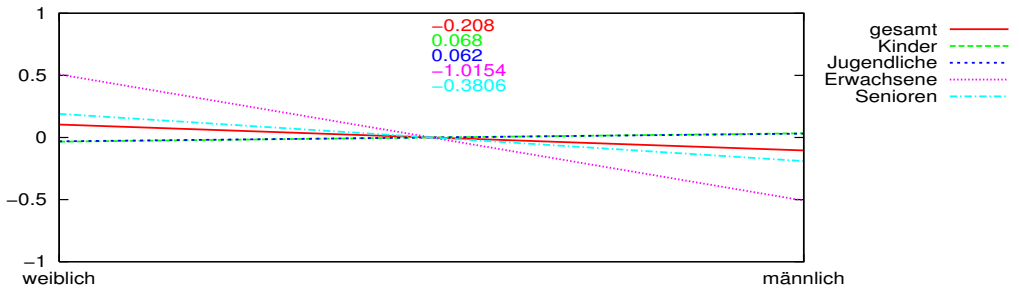
Erwachsene

gesamt	5.229	0.9778	0
weiblich	5.7254	0.9155	0.5077
männlich	4.7326	0.7623	-0.5077



Senioren

gesamt	5.7319	1.0506	0
weiblich	5.9319	1.0692	0.1903
männlich	5.532	0.9921	-0.1903



B Korrelationskoeffizienten

Korrelationskoeffizienten der Jitter-Maße für WEIBLICH/JUNG

	jitt_l	*jitt_la*	*jitt_ppq*	*jitt_rap*	*jitt_ddp*
jitt_l	1	0.87	0.96	0.98	0.98
jitt_la		1	0.78	0.77	0.77
jitt_ppq			1	0.98	0.98
jitt_rap				1	1
jitt_ddp					1

Korrelationskoeffizienten der Jitter-Maße für WEIBLICH/ALT

	jitt_l	*jitt_la*	*jitt_ppq*	*jitt_rap*	*jitt_ddp*
jitt_l	1	0.94	0.98	0.99	0.99
jitt_la		1	0.9	0.93	0.93
jitt_ppq			1	0.99	0.99
jitt_rap				1	1
jitt_ddp					1

Korrelationskoeffizienten der Jitter-Maße für MÄNNLICH/JUNG

	jitt_l	*jitt_la*	*jitt_ppq*	*jitt_rap*	*jitt_ddp*
jitt_l	1	0.89	0.97	0.98	0.98
jitt_la		1	0.88	0.86	0.86
jitt_ppq			1	0.98	0.98
jitt_rap				1	1
jitt_ddp					1

Korrelationskoeffizienten der Jitter-Maße für MÄNNLICH/ALT

	jitt_l	*jitt_la*	*jitt_ppq*	*jitt_rap*	*jitt_ddp*
jitt_l	1	0.91	0.95	0.98	0.98
jitt_la		1	0.81	0.85	0.85
jitt_ppq			1	0.97	0.97
jitt_rap				1	1
jitt_ddp					1

Korrelationskoeffizienten der Shimmer-Maße für WEIBLICH/JUNG

	shim_l	*shim_ldb*	*shim_apq*3	*shim_apq*11	*shim_ddp*
shim_l	1	0.92	0.94	0.95	0.94
shim_ldb		1	0.84	0.88	0.84
*shim_apq*3			1	0.88	1
*shim_apq*11				1	0.88
shim_ddp					1

Korrelationskoeffizienten der Shimmer-Maße für WEIBLICH/ALT

	shim_l	*shim_ldb*	*shim_apq*3	*shim_apq*11	*shim_ddp*
shim_l	1	0.91	0.91	0.82	0.91
shim_ldb		1	0.76	0.83	0.76
*shim_apq*3			1	0.62	1
*shim_apq*11				1	0.62
shim_ddp					1

Korrelationskoeffizienten der Shimmer-Maße für MÄNNLICH/JUNG

	shim_l	*shim_ldb*	*shim_apq*3	*shim_apq*11	*shim_ddp*
shim_l	1	0.91	0.86	0.85	0.86
shim_ldb		1	0.78	0.78	0.78
*shim_apq*3			1	0.63	1
*shim_apq*11				1	0.63
shim_ddp					1

Korrelationskoeffizienten der Shimmer-Maße für MÄNNLICH/ALT

	shim_l	*shim_ldb*	*shim_apq*3	*shim_apq*11	*shim_ddp*
shim_l	1	0.91	0.91	0.82	0.91
shim_ldb		1	0.76	0.83	0.76
*shim_apq*3			1	0.62	1
*shim_apq*11				1	0.62
shim_ddp					1

Index

Literaturverzeichnis

Abdulla, W. und Kasabov, N. (2001). Improving speech recognition performance through gender separation. In *Proceedings of the Fifth Biannual Conference on Artificial Neural Networks and Expert Systems (ANNES 2001)* (S. 218-222). Dunedin, New Zealand.

Aha, D. W., Kibler, D. und Albert, M. K. (1991). Instance-Based Learning Algorithms. *Machine Learning*, *6*(1), 37–66.

Baken, R. und Orlikoff, R. (2000). *Clinical Measurement of Speech and Voice* (2. Aufl.). San Diego, Ca, USA: Singular Publishing Group.

Batliner, A., Huber, R., Niemann, H., Nöth, E., Spilker, J. und Fischer, K. (2000). The Recognition of Emotion. In W. Wahlster (Hrsg.), *Verbmobil: Foundations of Speech-to-Speech Translations* (S. 122–130). New York - Berlin: Springer.

Benjamin, B. J. (1988). Changes in Speech and Linguistic Behaviour with Aging. In B. B. Shadden (Hrsg.), *Communication Behavior and Aging: A Sourcebook for Clinicians.* Baltimore, USA: Williams and Wilkins.

Bennett, S. und Weinberg, B. (1979). Acoustic correlates of perceived sexual identity in preadolescent children's voices. *Journal of the Acoustical Society of America*, *66*(4), 989–1000.

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *Proceedings of the Dutch Institute of Phonetic Sciences (IFA)* (S. 97–110). Amsterdam, Netherlands.

Boersma, P. (2001). PRAAT, a system for doing phonetics by computer. *Glot International*, *9*(5), 341–345.

Bosch, K. (1987). *Elementare Einführung in die Angewandte Statistik.* Braunschweig, Germany: Vieweg.

Bou-Ghazale, S. und Hansen, J. (1996). Synthesis of Stressed Speech from Neutral Speech using HMM-based Models. In *Proceedings of the International Conference on Spoken Language Processing* (S. 1860–1863). Philadelphia, PA, USA.

Brandherm, B. und Jameson, A. (2004). An Extension of the Differential Approach for Bayesian Network Inference to Dynamic Bayesian Networks. *International Journal of Intelligent Systems*, *19*(8), 727–748.

Braun, A. und Cerrato, L. (1999). Estimating speaker age across languages. In *Proceedings of the 14th Conference on Phonetic Sciences ICPhS '99* (S. 1369–1372). San Francisco, Ca, USA.

Byrd, D. (1992). Preliminary results on speaker-dependant variation on the TIMIT database. *Journal of the Acoustical Society of America*, *92*(1), 593–596.

Campbell, W. M., Campbell, J. P., Reynolds, D. A., Jones, D. A. und Leek, T. R. (2003). Phonetic Speaker Recognition with Support Vector Machines. In *Proceedings of the Neural Information Processing Systems Conference* (S. 1377–1384). Vancouver, British Columbia, Canada.

Carstensen, K.-U., Ebert, C., Endriss, C., Jekat, S., Klabunde, R. und Langer, H. (Hrsg.). (2004). *Computerlinguistik und Sprachtechnologie* (2. Aufl.). München - New York: Elsevier.

Cohen, A. und Lapidus, V. (1995). Unsupervised Text Independent Speaker Classification. In *Proceeding of the 18th Convention of Electrical and Electronics Engineering* (S. 1745–1749). Tel-Aviv, Israel.

Corkill, D., Gallagher, K. und Johnson, P. (1988). Achieving Flexibility, Efficiency, and Generality in Blackboard Architectures. In A. Bond und L. Gasser (Hrsg.), *Readings in Distributed Artificial Intelligence.* Morgan-Kaufman.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L. und Spiegelhalter, D. J. (1999). *Probabilistic Networks and Expert Systems.* Berlin - Heidelberg - New York: Springer.

Duda, R. O., Hart, P. E. und Stork, D. G. (2000). *Pattern Classification* (2. Aufl.). New York, USA: Wiley-Interscience.

Feld, M. (2005). *Portierung von Merkmalsextraktion auf die PocketPC-Plattform* (Tech. Rep. Nr. TM-05-01). Saarbrücken, Germany: DFKI, Deutsches Forschungszentrum für Künstliche Intelligenz GmbH.

Feld, M. (2006). *Erzeugung von Sprecherklassifikationsmodulen für multiple Plattformen.* Unveröffentlichte Diplomarbeit, Fachbereich 6.2 Informatik, Universität des Saarlandes, Deutschland. (in Druck)

Ferrand, C. T. (2002). Harmonics-to-Noise Ratio: An Index of Vocal Aging. *Journal of Voice*, *16*(4), 480–487.

Fitch, W. und Giedd, J. (1999). Morphology and development of the human vocal tract: A study using magnetic response imaging. *Journal of the Acoustical Society of America*, *103*(3), 1511–1522.

Garg, A., Pavlovic, V. und Rehg, J. M. (2000). Audio-Visual Speaker Detection using Dynamic Bayesian Networks. In *Proceedings of the 4th International Conference of Automatic Face and Gesture Recognition* (S. 374–471). Grenoble, France.

Garofolo, J. e. A. (1998). *DARPA TIMIT CD-ROM: An Acoustic Phonetic Continous Speech Database.* Gaithersburg, MD, USA: National Institute of Standards and Technology.

Germesin, S. (2006). *Spracherkennung mit alters- und geschlechtsspezifischen Akustikmodellen.* Bachelorarbeit, Fachbereich 6.2 Informatik, Universität des Saarlandes, Deutschland. (in Druck)

Green, M. C. L. (1982). Aging of the Voice: a Review. In M. Edwards (Hrsg.), *Communication Changes in Elderly People.* London, UK: College of Speech Therapists.

Greff, G. und Fojut, S. (2003). *Das ABC des Call Center Management. Die wichtigsten Fachbegriffe rund um Call Center und Kundenservice.* Wiesbaden, Germany: Gabler.

Grosso, W. (2001). *Java RMI.* Cambridge, USA: O'Reilly and Associates.

Hahn, W. v., Henskes, D., Hoeppner, W. und Wahlster, W. (1975). HAM-RPM: Ein Redepartnermodell als Simulationsprogramm. In *Grammatik* (S. 337–357). Tübingen, Germany: Niemeyer.

Hartmann, D. E. und Danhauer, J. L. (1976). Perceptual features of speech for males in four perceived age decades. *Journal of the Acoustic Society of America*, *59*(3), 713–715.

Hasek, C. S. und Singh, S. (1980). Acoustic attributes of preadolescent voices. *Journal of the Acoustical Society of America*, *68*(5), 1262–1265.

Haselager, G., Slis, I. und Rietveld, A. (1991). An alternative method of studying the development of speech rate. *Clinical Linguistics and Phonetics*, *5*(1), 53–63.

Hickey, M. (2003). *When talking is not enough* (Tech. Rep.). Bristol, UK: HP Mobile and Media Systems Laboratory.

Holmes, J. und Holmes, W. (2001). *Speech Synthesis and Recognition* (2. Aufl.). London, New York: Taylor and Francis.

Huang, D., Minifie, F., Kasuya, H. und Lin, S. (1995). Measures of vocal function during changes in vocal effort level. *Journal of Voice*, *9*, 429–438.

Huang, H.-J. und Hsu, C.-N. (2002). Recognizing 100 Speakers using Homologous Naive Bayes. In *Proceedings of the Seventh Pacific Rim International Conference on Artificial Intelligent (PRICAI 2002)* (S. 395 – 403). Tokyo, Japan.

Iurgel, U., Werner, S. und Gerhard, R. (2003). Vergleich von automatischer und manueller Segmentierung von Fernsehnachrichten und deren Einfluss auf die Sprach- und Themenerkennung. In *Tagungsband der 14. Konferenz für Elektronische Sprachsignalverarbeitung (ESSV)* (S. 67 – 74). Karlsruhe, Germany.

Jameson, A., Großmann-Hutter, B., Müller, C., Wittig, F., Kiefer, J. und Rummer, R. (2005). Recognition of Psychologically Relevant Aspects of Context on the Basis of Features of Speech. In *Proceedings of the Second International Workshop on Modeling and Retrieval of Context in conjunction with IJCAI'05.* Edinburgh, UK.

Jande, P.-A. (2004). Pronunciation variation modelling using decision tree induction from multiple linguistic parameters. In *Proceedings of FONETIK 2004* (S. 12 – 15). Stockholm, Sweden.

Jensen, F. V. (1997). *An Introduction to Bayesian Networks.* New York, USA: Spinger.

Kakumanu, P., Gutierrez-Osuna, R., Esposito, A., Bryll, R., Goshtasby, A. und Garcia, O. (2001). Speech Driven Facial Animation. In *Proceedings of the Workshop on Perceptive User Interfaces* (S. 1–5). Orlando, Florida, USA.

Karlsson, I. (1987). Sex differentiation cues in the voices of young children of different language background. *Journal of the Acoustical Society of America*, *81*, 68–69.

Kent, R. (1976). Anantomical and neuromuscular maturation of the speech mechanism: Evidence from acoustic studies. *Journal of Speech and Hearing Research*, *19*, 421–447.

Klein, C. (2004). *Acoustic and perceptual gender characteristics in the voices of pre-adolescent children.* Magisterarbeit, Phonetik, Universität des Saarlandes, Deutschland.

Kobsa, A. und Wahlster, W. (Hrsg.). (1989). *User Models in Dialog Systems* (Bd. -). Berlin - Heidelberg - New York: Springer.

Konig, Y., Morgan, N. und Chandra, C. (1991). *GDNN : a gender-dependent neural network for continuous speech recognition* (Bd. TR-91-071; Report). Berkeley, CA, USA: International Computer Science Institute (ICSI).

Levelt, W. (1989). *Speaking: From Intention to Articulation.* Cambridge, USA: MIT Press.

Linville, S. E. (2001). *Vocal Aging.* San Diego, Ca, USA: Singular Publishing Group.

Linville, S. E. und Fisher, H. B. (1985). Acoustic characteristics of perceived versus actual vocal age in controlled phonation by adult females. *Journal of the Acoustic Society of America*, *70*(1), 40–48.

Lisetti, C., Nasoz, F., LeRouge, C., Oyzer, O. und Alvarez, K. (2003). Developing multimodal intelligent affective interfaces for tele-home health care. *International Journal of Human-Computer Studies*, *59*, 245–255.

Madsen, A. L., Jensen, F., Kjarulff, U. B. und Lang, M. (2005). HUGIN - The Tool for Bayesian Networks and Influence Diagrams. *International Journal of Artificial Intelligence Tools*, *14*(3), 507–543.

Markowitz, J. A. (2000). Voice Biometrics. *Communications of the ACM*, *43*(3).

Minematsu, N., Yamauchi, K. und Hirose, K. (2003). Automatic Estimation of Perceptual Age Using Speaker Modeling Techniques. In *Proceedings of the Eighth European Conference on Speech Communication and Technology (Eurospeech 2003)* (S. 3005 – 3008). Geneva, Switzerland.

Mirghafori, N., Fosler, E. und Morgan, N. (1996). Towards Robustness To Fast Speech In Asr. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'96)* (S. 335–338). Atlanta, Georgia, USA.

Morgan, N. (1990). *Considerations for the electronic implementation of artificial neural networks* (Bd. TR-90-003; Report). Berkeley, CA, USA: International Computer Science Institute (ICSI).

Morgan, N., Fosler, E. und Mirghafori, N. (1997). Speech Recognition using On-line Estimation of Speaking Rate. In *Proceedings of the 5th European Conference on Speech Communication and Technology, EUROSPEECH* (S. 2079–2082). Rhodes, Greece.

Müller, C. (2002). Multimodal Dialog in a Pedestrian Navigation System. In *Proceedings of ISCA Tutorial and Research Workshop on Multi-Modal Dialogue in Mobile Environments* (S. 42 – 44). Kloster Irsee, Germany.

Müller, C. (2005). *Estimating the Acoustic Context to Improve Speaker Classification.* (Presented as poster at the Fifth International and Interdisciplinary Conference on Modeling and Using Context (CONTEXT-05). Paris, France.)

Müller, C., Großmann-Hutter, B., Jameson, A., Rummer, R. und Wittig, F. (2001). Recognizing Time Pressure and Cognitive Load on the Basis of Speech: An Experimental Study. In M. Bauer, P. Gmytrasiewicz und J. Vassileva (Hrsg.), *UM2001, User Modeling: Proceedings of the Eighth International Conference* (S. 24 – 33). New York - Berlin: Springer.

Müller, C. und Wittig, F. (2003). Speech as a Source for Ubiquitous User Modeling. In *Proceedings of the Workshop on User Modeling in Ubiquitous Computing in conjunction with the Ninth International Conference on User Modeling (UM 2003)* (S. 46 – 50). Pittsburgh, USA.

Müller, C., Wittig, F. und Baus, J. (2003). Exploiting Speech for Recognizing Elderly Users to Respond to their Special Needs. In *Proceedings of the Eighth European Conference on Speech Communication and Technology (Eurospeech 2003)* (S. 1305 – 1308). Geneva, Switzerland.

Nairn, M. (1997). *Acoustic and perceptual gender differences in the speech of 4.5 to 5.5 year old children.* Dissertation, Queen Margareth University College, Edinburgh, UK.

Neppert, J. und Petursson, M. (1986). *Elemente einer akustischen Phonetik* (2. Aufl.). Hamburg, Germany: Helmut Buske.

Newell, A. (2003). Spoken language and e-inclusion. In *Proceedings of the Eighth European Conference on Speech Communication and Technology (Invited Keynote)* (S. 1309–1312). Geneva, Switzerland.

Nix, D., Fairweather, P. und Adams, B. (1998). Speech Recognition, Children, and Reading. In *Proceeding of the CHI 98 Conference on Human Factors in Computing Systems* (S. 245–246). Los Angeles, Ca, USA: ACM.

Orio, N. und SistiSette, M. (2003). An HMM-based pitch tracker for audio queries. In *Proceeding of the Fourth International Conference on Music Information Retrieval* (S. 249 – 250). Baltimore, Maryland, USA.

Parris, E. und Carey, M. J. (1996). Language Independant Gender Identification. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP'96)* (S. 685 – 688). Atlanta, Georgia, USA.

Peskin, B., Navratil, J., Abramson, J., Jones, D., Klusacek, D., Reynolds, D. A. und Xiang, B. (2003). Using Prosodic and Conversational Features for High Performance Speaker Recognition: Report from JHU WS'02. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing* (S. 792–795). Hong Kong, China.

Petursson, M. und Neppert, J. (1991). *Elementarbuch der Phonetik.* Hamburg, Germany: Helmut Buske.

Pompino-Marschall, B. (2003). *Einführung in die Phonetik.* Berlin - New York: Walter de Gruyter.

Pützer, M. (2001). Multiparametrische Stimmqualitätserfassung männlicher und weiblicher Normalstimmen. *Folia Phoniatrica et Logopaedica*, *53*, 73–84.

Pützer, M. und Marasek, K. (2000). Differenzierung gesunder Stimmqualitäten und Stimmqualitäten bei Rekurrensparese mit Hilfe elektroglottographischer Messungen und RBH-System. *Sprache Stimme Gehör*, *24*, 154–163.

Quinlan, R. (1993). *C4.5: Programs for Machine Learning.* San Mateo, CA, USA: Morgan Kaufmann Publishers.

Reynolds, D. A., Campbell, J. P., Campbell, W. M., Dunn, R. B., Gleason, T. P., Jones, D. A., Quatieri, T. F., Quillen, C. B., Sturim, D. E. und Torres-Carrasquillo, P. A. (2003). Beyond Cepstra: Exploiting High-Level Information in Speaker Recognition. In *Proceedings of the Workshop on Multimodal User Authentication* (S. 223–229). Santa Barbara, Ca, USA.

Reynolds, D. A., Quatieri, T. F. und Dunn, R. B. (2000). Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, *10*(1-3), 19–41.

Rienks, R., Poppe, R. und Poel, M. (2005). Speaker prediction based on head orientations. In *Proceedings of the Fourteenth Dutch-Belgian Conference on Machine Learning BeneLearn* (S. 73 – 79). Enschede, Netherlands.

Robb, M. P. und Simmons, J. O. (1990). Gender comparison of children's vocal fold contact behaviour. *Journal of the Acoustical Society of America*, *88*(3), 1318–1322.

Rojas, P. (1996). *Neural Networks – A Systematic Introduction.* Berlin - New York: Springer.

Russel, S. J. und Norvig, P. (1995). *Artificial intelligence: a modern approach.* Englewood Cliffs, New Jersery, USA: Prentice-Hall.

Russel, S. J. und Norvig, P. (2003). *Artificial intelligence: a modern approach* (2. Aufl.). Upper Saddle River, New Jersery, USA: Prentice-Hall.

Sachs, J., Liebermann, P. und Erickson, D. (1973). Anatomical and cultural determinants of male and female speech. In R. Shuy und R. Fasold (Hrsg.), *Language attidudes: Current trends and prospects* (S. 74–84). Georgetown, USA: Georgetown University Press.

Schiel, F. (1998). Speech and Speech-Related Resources at BAS. In *Proceedings of the First International Conference on Language Resources and Evaluation* (S. 343–349). Granada, Spain.

Schiel, F. und Draxler, C. (2003). *Production and Validation of Speech Corpora.* Berlin, Germany: Books on Demand GmbH.

Schötz, S. (2003). Towards synthesis of speaker age: A perceptual study with natural, synthesized and resynthesized stimuli. In *Proceedings of Fonetik 2003* (S. 153 – 156). Umea, Sweden.

Schötz, S. (2004a). Prosodic Cues in Human and Machine Estimation of Female and Male Speaker Age. In *Proceedings Nordic Prosody IX* (S. 215–223). Lund, Sweden.

Schötz, S. (2004b). Some Acoustic Cues to Human and Machine Estimation of Speaker Age. In *Proceedings of Fonetik 2004* (S. 40–43). Stockholm, Sweden.

Schötz, S. (2004c). The Role of F0 and Duration in Perception of Female and Male Speaker Age. In *Proceedings of Speech Prosody 2004* (S. 379–382). Nara, Japan.

Smith, D., Townsend, J., Nelson, D. und Richman, D. (1999). A Multivariate Speech tivity Detector Based on the Syllable Rate. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'99)* (S. 73–76). Phoenix, USA.

Tanner, J. (1962). *Wachstum und Reifung des Menschen.* Stuttgart, Germany: Georg Thieme Verlag.

Trouvain, J. (2004). *Tempo Variation in Speech Production. Implications for Speech Synthesis.* Dissertation, Department of Phonetics, Saarland University, Germany.

Vergin, R., Farhat, A. und Shaughnessy, D. O. (1996). Robust gender-dependant acoustic-phonetic modelling in continuous speech recogntion based on a new automatic male/female classification. In *Fourth International Conference on Spoken Language Processing* (Bd. 2, S. 1081–1084). Philadelphia, PA, USA.

Wahlster, W. (2000a). Künstliche Intelligenz: Werden Computer zu intelligenten Assistenten für jedermann? In Brockhaus-Redaktion (Hrsg.), *Visionen 2000* (S. 172 – 175). Mannheim, Germany: Brockhaus.

Wahlster, W. (2000b). *Verbmobil: Foundations of Speech-To-Speech Translation.* Berlin - Heidelberg - New York: Springer.

Wahlster, W. (2001). Robust Translation of Spontaneous Speech: A Multi-Engine Approach. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence* (Bd. 2, S. 1484 – 1493). Seattle, Washington, USA: San Francisco: Morgan Kaufmann.

Wasinger, R., Krüger, A. und Jacobs, O. (2005). Integrating Intra and Extra Gestures into a Mobile and Multimodal Shopping Assistant. In *Proceedings of the 3rd International Conference on Pervasive Computing (Pervasive)* (S. 297–314). Munich, Germany.

Wasinger, R., Stahl, C. und Krüger, A. (2003). M3I in a Pedestrian Navigation & Exploration System. In *Proceedings of the Fourth International Symposium on Human Computer Interaction with Mobile Devices* (S. 481–485). Pisa, Italy.

Wilamowitz-Moellendorff, M., Müller, C., Jameson, A., Brandherm, B. und Schwartz, T. (2005). Recognition of Time Pressure via Physiological Sensors: Is the User's Motion a Help or a Hindrance? In *Proceedings of the Workshop on Adapting the Interaction Style to Affective Factors in conjunction with the User Modeling (UM05) conference* (S. 43–48). Edinburgh, UK.

Wilcox, K. A. und Horii, Y. (1980). Age and Changes in Vocal Jitter. *Journal of Gerontology*, *35*(2), 194–198.

Witten, I. H. und Frank, E. (1999). *Data Mining: Practical Machine Learning Tools and Techniques With Java Implementations.* San Mateo, CA, USA: Morgan Kaufmann Publishers.

Yacoub, S., Simske, S., Lin, X. und Burns, J. (2003). Recognition of Emotions in Interactive Voice Response Systems. In *Proceedings of the Eighth European Conference on Speech Communication and Technology (Eurospeech 2003)* (S. 729–732). Geneva, Switzerland.

Yamazawa, H. und Hollien, H. (1992). Speaking Fundamental Frequency Pattern of Japanese Women. *Phonetica*, *49*, 128–140.

Yu, F., Chang, E., Xu, Y. und Heung-Shum, Y. (2001). Emotion Detection from Speech to Enrich Multimedia Content. In *Proceedings of the Second IEEE Pacific Rim Conference on Multimedia: Advances in Multimedia Information Processing* (S. 550 – 557). Beijing, China.

Zervas, P., Xydas, G., Fakotakis1, N., Kokkinakis1, G. und G., K. (2004). Evaluation of Corpus Based Tone Prediction in Mismatched Environments for Greek TtS Synthesis. In *Proceedings of the 8th International Conference on Spoken Language Processing (INTERSPEECH - ICSLP)* (S. 761–764). Jeju, Korea.

Zwicker, E. und Fastl, H. (1999). *Psychoacoustics - Facts and Models* (2. Aufl.). New York - Berlin: Springer.