Lehrstuhl für Mensch-Maschine-Kommunikation
Technische Universität München

Automatische Emotionserkennung aus sprachlicher und manueller Interaktion

Björn Schuller

Vollständiger Abdruck der von der Fakultät für Elektrotechnik und Informationstechnik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor-Ingenieurs (Dr.-Ing.)

genehmigten Dissertation.

Vorsitzende:		Univ.-Prof. Dr. rer. nat. Doris Schmitt-Landsiedel
Prüfer der Dissertation:	1.	Univ.-Prof. Dr. rer. nat. Manfred K. Lang, i. R.
	2.	Univ.-Prof. Dr.-Ing. Joachim Hagenauer

Die Dissertation wurde am 24.11.2005 bei der Technischen Universität München eingereicht und durch die Fakultät für Elektrotechnik und Informationstechnik am 25.04.2006 angenommen.

Vorwort

Die vorliegende Arbeit ist das Ergebnis meiner Forschungstätigkeit als wissenschaftlicher Assistent am Lehrstuhl für Mensch-Maschine-Kommunikation an der Technischen Universität München.

Mein ganz besonderer Dank gilt meinem Doktorvater UNIV.-PROF. DR. RER. NAT. MANFRED LANG, für die Inspiration und Möglichkeit zur Durchführung dieser Arbeit an seinem Lehrstuhl. Das von ihm geförderte fruchtbare interdisziplinäre Arbeitsumfeld und seine herausragende, stete wissenschaftliche Betreuung mit genügend Freiraum zur Verwirklichung meiner eigenen Ideen, aber auch dem Blick für das Wesentliche, haben den Erfolg dieser Arbeit ermöglicht.

Ebenso besonders bedanken möchte ich mich bei seinem Nachfolger als Lehrstuhlinhaber, Herrn UNIV.-PROF. DR.-ING. HABIL. GERHARD RIGOLL für die hervorragende fachliche Beratung und Anregungen.

Für die aktive Unterstützung bei meiner Teilnahme an wissenschaftlichen Fachtagungen im In- und Ausland und zahlreichen Publikationen möchte ich beiden Lehrstuhlinhabern herzlich danken.

Außerordentlich gerne erinnere ich mich an die sehr gute Zusammenarbeit mit meinen Kollegen PROF. (FH) DR. RER. NAT. JÖRG HUNSINGER, DR.-ING. MICHAEL GEIGER, DR.-ING. MARC HOFMANN, DR.-ING. ROBERT NEUSS, RALF NIESCHULZ, MARTIN ZOBL, DR. RER. NAT. FRANK ALTHOFF und GREGOR MCGLAUN. Ein lieber Dank geht auch an meine Kollegen FRANK WALLHOFF, JAN STADERMANN, DEJAN ARSIC, DR. RER. NAT. FRED NENTWICH und STJEPAN SKRTIC für viele anregende fachübergreifende Diskussionen. Für die Hilfe in allen technischen Bedürfnissen danke ich Herrn DR.-ING. CLAUS VON RÜCKER, PETER BRAND und HEINER HUNDHAMMER.

Ferner danke ich Herrn UNIV.-PROF. DR.-ING. JOACHIM HAGENAUER für die Erstellung des Zweitgutachtens, dem FERMUS Firmenkonsortium, Herrn PROF. DR.-ING. HABIL. GÜNTHER RUSKE und allen Kollegen für ihre Diskussionsbereitschaft und tatkräftige Unterstützung sowie allen beteiligten Versuchspersonen.

Herzlich danken für ihre Beiträge möchte ich auch den Diplomanden RAQUEL JIMENEZ-VILLAR, JUAN WANG, FLORIAN HÖRGER, RONALD MÜLLER, STEPHAN REITER, YUPING SHI, GUNTHER SPAHN, und TING-YAP TONG, den Studienarbeitern MARC KIRSCH, ELMAR SOMMER und MARTIN WIESER, den Interdisziplinären Projektarbeitern MELTEM AKYAZI, TAYFUR COSKUN, STEFAN KUGELE, ANDREAS SCHOLZ, STEPHAN THOMA, den Werkstudenten SEBASTIAN GRAF und ANDREAS MAVREDAKIS sowie den Praktikanten BERNARDO JOSÉ BRÜNING SCHMITT und MARKUS FRONIUS.

Für ihre Inspiration, Motivation und Unterstützung möchte ich ganz besonders meiner Frau DAGMAR und meiner Familie danken.

München, im November 2005

Björn Wolfgang Schuller

Kurzfassung

Integration emotionaler Aspekte ist Basis natürlicher und zukunftsweisender Mensch-Maschine-Kommunikation. Vor diesem Hintergrund werden in dieser Arbeit innovative Verfahren zur robusten maschinellen Erkennung menschlicher Gefühle aus Sprache und Daten der manuellen Interaktion vorgestellt.

Die Besonderheit der Schätzung aus dem akustischen Sprachsignal liegt in der Applikation, automatischen Generierung und Selektion neuartiger Merkmale mit Techniken der evolutionären Programmierung und gleitenden sequentiellen Suchverfahren. Ein wesentlicher Bestandteil ist ferner die quantitative Gegenüberstellung einer Modellierung in Zeitreihen und nach Verfahren der deskriptiven Statistik gebildeten Funktionalen sowie ein extensiver Vergleich diverser Lernverfahren. Hierzu gehören instanzbasierte Erkennung, statistische Modellierung, Kernelmaschinen, Neuronale Netze und Entscheidungsbäume. Der Einsatz von Techniken der Ensembleklassifikation wie MultiBoosting oder Stacking zur Optimierung und Vereinigung unterschiedlicher Stärken dieser Ansätze des maschinellen Lernens rundet die Betrachtung ab.

Eine ergänzende inhaltliche Interpretation hinsichtlich ausgedrückter Emotion des gesprochenen Texts unterstützt die Erkennung und ermöglicht diese auch aus geschriebener Eingabe. Hierzu werden in diesem Gebiet innovative Ansätze wie eine grafische Modellierung oder Vektorraumrepräsentation mit klassischer N-Gramm Darstellung verglichen. Der Problematik der realen Texterfassung wird in einem Exkurs in die Spracherkennung, Handschrifteingabe und Soft-String-Matching Rechnung getragen.

Als Novum soll auch konventionelle Bedienung über eine Computermaus ohne ergänzende Hardware oder die Verwendung eines Touchscreens hinsichtlich einer Eignung zur Erkennung des Benutzerzustands als ergänzende Modalität betrachtet werden. Analog zur akustischen Modellierung werden geeignete Größen und Modelle zur Schätzung vorgestellt und in Bezug auf eine potenzielle Erkennung kritisch analysiert.

Der Diskussion dieser singulären Ansatzpunkte folgt die Vorstellung von Methoden der synergetischen Integration. Verfahren der Multistream- und multimodalen Fusion auf Merkmals- sowie semantischer Ebene werden im Einsatz untersucht. Im Hinblick auf ein reales Umfeld wird des Weiteren eine Adaption an den aktuellen Benutzer zur Steigerung der Robustheit vorgestellt.

Szenarien zur Demonstration eines praxistauglichen Einsatzes und Transfers der erarbeiteten Methodik bilden den Abschluss. Diese sind robuste automatische Spracherkennung, Musiksuche und der Einsatz im Fahrzeug.

Als Ergebnis dieser Forschungstätigkeit kann Emotion unter idealen Bedingungen aus dem akustischen Sprachsignal maschinell vergleichbar einem menschlichen Entscheider klassifiziert werden. Die Integration einer Inhaltsanalyse bringt eine signifikante Verbesserung der Robustheit. Schließlich erscheint eine Erkennung auch aus der manuellen Interaktion prinzipiell möglich.

Abstract

Affective Computing establishes the basis of natural future Human-Computer Communication. Within this context this thesis focuses on a variety of innovative approaches towards a robust *Automatic Emotion Recognition out of Spoken and Manual Interaction*.

The specialty of propagated recognition methods within acoustic processing is defined by the application, automatic generation and selection of novel features with advanced techniques of Evolutionary Programming and Floating Search Methods. Another important aspect is the comparison of modelling the information stream by a multivariate time series or by functionals derived out of the series by means of descriptive statistics. Additionally, an exhaustive search for the optimal classifier is provided. Among such instance based learning, stochastic modelling, Kernel Machines, Neural Nets, and Decision Trees can be found. The application of ensemble construction techniques such as MultiBoosting or Stacking in order to enhance and combine the individual strengths of base classifiers further supports the improvement in connection with general performance.

The following analysis of the spoken content in association with emotional information enhances recognition accuracy, and enables accessory processing of written text. Thereby novel approaches within this field such as Graphical or Vector Space Modelling are compared to classical N-Gram representation. Capturing of the text itself is discussed within a brief digression on Soft-String-Matching, Automatic Handwriting and Speech Recognition.

Innovatively, also conventional interaction by a computer-mouse without additional hardware, and the usage of a touch-screen will be considered as further modalities to estimate an underlying user affect. Analogical to acoustic processing relevant attributes and models for the recognition will be introduced and critically evaluated.

The processing of single information streams is refined by the introduction of methods for their synergistic integration. Thereby multi-stream and multimodal fusion on a feature and semantic level are dealt with. Considering a real-life application, adaptation to the actual user in order to improve overall accuracy is furthermore performed.

Finally, three use-cases are presented, which are Robust Automatic Speech Recognition, multimodal Music Information Retrieval, and affective interaction in an automotive environment.

Concluding, emotion can be recognized close to human performance based on acoustic information given ideal conditions, integration of spoken content analysis significantly boosts performance, and estimation out of manual interaction seems generally feasible.

Inhaltsverzeichnis

1

Einleitung

„Nur wer sein Ziel kennt, findet den Weg.“

LAO-TSE (6. JH. V. CHR.)

1.1 Motivation und Zielsetzung

Der Mensch ist das wichtigste Element eines jeden Informationssystems [LAN94]. Der Umgang mit technischen Systemen soll von daher vorrangig für den Anwender natürlich, intuitiv und effizient gestaltet werden [LAN01]. Weiterhin kann mit hinreichender Sicherheit davon ausgegangen werden, dass in der menschlichen Interaktion mit technischen Systemen, insbesondere mit Computern, Nutzer ihre interpersonellen sozialen Verhaltensmuster extrapolieren [REE96]. So gaben etwa im Rahmen dieser Arbeit durchgeführter Studien von insgesamt 39 täglich Computer benutzenden Personen 35 an, ihr Gerät regelmäßig lautsprachlich zu ermahnen[1] sollte es zu Fehlern kommen. Dieser menschlichen Komponente muss seitens der Entwicklung von Mensch-Maschine-Schnittstellen der nächsten Generation durch soziale Kompetenz begegnet werden. Dabei wird unter dem Begriff der *emotionalen Intelligenz* die Bedeutung emotionaler Einflüsse in rationalen Entscheidungsprozessen, sozialer Interaktion, Wahrnehmung, Memorisierung, Lernvorgängen und kreativen Abläufen zusammengefasst und verstanden [PIC98]. Die Forschungsarbeiten in diesem Rahmen erschließen das junge Gebiet des *Affective Computings*, welches sich mit der Erkennung von Benutzergefühlen und anderen Stimuli sowie adäquaten Reaktionen auf diese beschäftigt. Speziell die robuste automatische Einschätzung aktueller Benutzeremotionen bildet den Kernaspekt dieser Arbeit.

Im Vordergrund modernen Ingenieurdenkens steht die Frage nach potentieller Anwendung angestrebter Forschungsaktivitäten. Speziell im Rahmen der Mensch-Maschine-Kommunikation fällt dabei ein besonderes Gewicht auf die Gebrauchstauglichkeit von Zielapplikationen. Unter

[1] Die befragten Personen waren im Alter von 21 a bis 39 a mit einem Mittelwert von 26,7 a, sechs Personen waren weiblich. Es handelte sich dabei ausschließlich um Studenten der Ingenieurswissenschaften.

diesem Aspekt sollen im Folgenden zunächst konkrete Szenarien für den Einsatz der anvisierten automatischen Emotionserkennung vorgestellt werden. Im Anschluss wird die Wahl der Modalitäten[2] unter den Gesichtspunkten der Eignung, aber auch der Benutzerfreundlichkeit, motiviert werden. Abschließend wird die Zielsetzung der vorliegenden Arbeit, die in der finalen Diskussion kritisch hinsichtlich ihrer Erfüllung hinterfragt werden wird, im Detail herausgearbeitet.

1.1.1 Anwendungsgebiete der automatischen Emotionserkennung

In bestehenden Arbeiten zur Emotionserkennung wird eine Vielzahl potentieller Einsatzszenarien angeführt. Die folgende Aufstellung gibt einen möglichst umfassenden Einblick hierzu. Eine vollständige Aufzählung ist aber gleichermaßen an dieser Stelle nicht möglich, weswegen der Schwerpunkt bei der Auswahl interessanter Beispiele auf den Bereich der Mensch-Maschine-Kommunikation gelegt ist. Des Weiteren werden in dieser Arbeit im Fokus des Interesses stehende, und teilweise neuartig konzipierte Einsatzgebiete vorgestellt.

- **Interpretation der Benutzerintention:** In [COW95] wird zwischen dem expliziten und impliziten Kanal einer zwischenmenschlichen Kommunikation unterschieden. Neben dem in der Forschung bisher hauptsächlich betrachteten expliziten Kanal in der Interaktion zwischen Anwender und System beinhaltet der implizite Information über die Person selbst. Diese wird als essentiell für die Interpretation der realen Intention bewertet. In diesem Zusammenhang ist der Benutzerzustand etwa für eine richtige Beurteilung ironischer oder nicht wahrheitsgemäßer Äußerungen unabkömmlich. Speziell hinsichtlich später vorgestellter Verfahren der Mustererkennung kann im Vorfeld auch eine Einschränkung der Interpretationsmöglichkeiten auf Grund der Emotion geschehen. Dies wird in dieser Arbeit im Bereich der Suche von Musiktiteln eingesetzt, um je nach Gemütslage des Anwenders bestimmte Titel im Vorfeld auszuschließen (siehe Kap. 7.2.3).

- **Lernen aus der emotionalen Reaktion:** In modernen Bedienschnittstellen wird meist eine Vielzahl maschineller Lernverfahren eingesetzt, um etwa Sprache oder Handgesten zu erkennen. Diese erzielen in der Regel höhere Erkennungsraten, wenn eine Adaption eines generischen Lernmodells an den aktuellen Benutzer möglich ist. Da eine Trainingsphase vor der Anwendung für die konkrete Person aber aus Komfortgründen unerwünscht ist, erfolgt oft eine automatische Anpassung während der Bedienung. Hier soll die erkannte Emotion als rückgekoppelte Überwachungsgröße der Adaption dienen (siehe Kap. 7.3.2). Um der Gefahr einer sich aufschaukelnden Fehladaption vorzubeugen, ist eine zuverlässige Schätzung über beispielsweise Zufriedenheit oder Unzufriedenheit der Zielperson Voraussetzung.

- **Akkommodation in der Kommunikation:** Nach der Theorie der *sprachlichen Akkommodation*[3] [GIL87] wird es zwischen Gesprächspartnern teils als unsozial betrachtet, wenn die Sprechweise hinsichtlich akustischer Parameter wie der Lautstärke nicht dem

[2] Unter dem Begriff der *Modalität* wird im Kontext dieser Arbeit ein menschlicher Kommunikationskanal auf verschiedenen Abstraktionsebenen verstanden. Diese Definition steht in keinem Zusammenhang mit dem Pendant aus dem Bereich der Sprachwissenschaft.

Gegenüber angepasst wird. Zur Integration sozialer Kompetenz in der Mensch-Maschine-Kommunikation bildet daher die Erkennung der Benutzeremotion die Basis für die potentielle Anpassung einer Ausgabe.

- **Emotionale Observation:** Der emotionale Zustand des Anwenders oder einer Zielperson soll automatisch überwacht werden. In der Mensch-Maschine-Interaktion ist dies etwa für die Erkennung von Hilfebedarf im Falle eines irritierten Benutzers von Interesse. Im Falle eines im Anschluss an eine Systemreaktion verärgerten Benutzers können ferner Maßnahmen zur Fehlerauflösung eingeleitet werden (siehe Kap. 7.3.2). In [COW01] werden in diesem Zusammenhang auch die Erkennung von Langweile oder Stress für die Anpassung der Lehrgeschwindigkeit in Edutainment-Programmen genannt. Allgemein besteht auch in sicherheitskritischen Bereichen hohes Interesse an einer Überwachung. Oft wird die Erkennung von Trunkenheit oder Müdigkeit von Fahrzeugführern, Piloten oder Kernkraftwerk-Operateuren genannt. Seit den terroristischen Anschlägen vom 11. September 2002 auf amerikanische Einrichtungen wird dies auch zur maschinellen Erkennung von potentiellen Attentätern anhand von auffälligem emotionalem Verhalten, wie zum Beispiel in Flugzeugen gewünscht [ARS05A], [ARS05C]. Dabei soll nach Möglichkeit die Dichte und Güte der Überwachung ohne zusätzlichen Personalbedarf gesteigert werden. Erste reale Einsätze gibt es auch im Bereich automatisierter Callcenter, in denen verärgerte Anrufer an menschliche Operateure weitergeleitet werden sollen, um diese nicht als Kunden zu verlieren [PET99], [LEE02A], [DEV03]. Darüber hinaus wird die Anwendung bei medizinischen Problemstellungen angedacht [COW01]. Hierbei könnten beispielsweise Patienten mit Messgeräten ausgestattet werden, die ein Psychologe nach einer gegebenen Beobachtungsdauer auswertet. Tendenziell cholerische Patienten mit mangelnder Selbstkontrolle könnten so ferner von einer portablen Einheit zur emotionalen Selbstüberwachung frühzeitig gewarnt werden. Abschließend sei die Beobachtung der Emotion bei der Kundenbetreuung beispielsweise in Online-Portalen angeführt. Hier soll verfolgt werden, welche Produkte den Kunden mehr oder weniger interessieren, oder wo Anlass zur Verärgerung gegeben scheint [MOR97], [MUE01], [SCH04C], da effizientes Kundenbindungsmanagement hier von zunehmender ökonomischer Bedeutung ist [HES05]. Unabhängig von der Anwendung lässt sich aber sagen, dass die emotionale Observation vor einem gesellschaftspolitischen Hintergrund nicht grundsätzlich unkritisch erscheint.

- **Objektive Messung:** Zur Erweiterung menschlichen Urteilsvermögens sind hier unter Anderem Detektion von Lügen für juristische Zwecke [HIR05], Beurteilung von Zuneigung (siehe Kap. 2.6) oder Stützgrößen für medizinische Diagnosen zu nennen. Dieser Bereich stellt jedoch die höchsten Ansprüche an Zuverlässigkeit und Leistung, da er menschliche Fähigkeiten im Allgemeinen überschreiten soll.

- **Transmission von Emotionen:** In bestimmten Kommunikations-Situationen ist eine Übertragung der Emotion zwischen Menschen nur bedingt möglich. Als erstes Beispiel sei hier die Übertragung bei spontaner textueller Konversation wie beim sogenannten *Internet-*

[3] In der entsprechenden Publikation *Speech Accommodation Theory*, oder kurz: *Accommodation Theory* genannt.

Relate-Chat[4] [ZHE02], [HOL03], oder in SMS und E-Mails genannt. Zwar haben sich hier eigene Formen wie sogenannte *Emoticons*[5], aus Sonderzeichen zusammengesetzte emotionale Gesichter, etabliert - um Gefühle aber noch lebhafter darzustellen, besteht die Aufgabe der automatischen Emotionserkennung hier darin, die Emotion des Gesprächspartners zunächst zu erkennen und dann ein Bild desselben mit passender Emotion gemeinsam mit dem Text zu versenden. In vergleichbarer Weise wird dies auch bei Computerspielen über Netzwerk versucht [RAO03]. Als weiteres Beispiel sei ein Fahrzeug (siehe. Kap. 2.6) genannt, das die Fahreremotion nach außen unter Anderem durch Änderung der Fahrzeugfarbe übertragen soll. Auch hier wird die Emotion erst von einer Maschine aus Daten des Fahrverhaltens erkannt, um sie anschließend anderen Verkehrsteilnehmern zu kommunizieren. Durch manuelle Auswahl einer Emotion durch den Benutzer soll eine Übertragung von dieser auch in Mobiltelefonen der nächsten Produktgeneration der Firma Samsung möglich werden: Die Emotion soll dem Gesprächspartner haptisch durch einen Vibrationsmotor anhand von Parametern wie Intensität und Frequenz vermittelt werden. Hier würde sich eine automatische Erkennung des Affekts, etwa durch lautsprachliche Parameter, zur Steigerung des Nutzerkomforts anbieten.

- **Multimediale Suche:** Im Zuge der rasant wachsenden digitalen Archive multimedialen Text-, Bild-, Video- und Audio-Materials werden effiziente automatische Suchverfahren unabkömmlich. Dabei erscheint es unter Anderem von Interesse, nach ausgeprägten oder bestimmten emotionalen Passagen [TOI02], etwa in Meeting-Aufzeichnungen [SCU01], Sport-Events [PEK02], oder Zeitungen und Büchern, direkt suchen zu können. Als weitere Möglichkeit bietet sich hier eine Suggestion an, die zusätzlich die Emotion des Anwenders erkennt und übereinstimmend eine Vorauswahl trifft (siehe auch Kap. 7.2.3).

- **Unterhaltungselektronik:** In diesem Bereich sind die ersten Endverbraucherprodukte zu beobachten (siehe Kap. 2.6) und weiterhin zu erwarten. Dies liegt in erster Linie an der höheren Toleranz gegenüber Fehlern im Bereich von Trendprodukten. Beispiele möglicher Applikationen sind Videospiele, in denen der Protagonist im Spiel auf die Emotion des Spielers reagiert, Roboter-Spielzeuge die auf Emotion eingehen [BRA00] oder interaktives Fernsehen. Darüber hinaus wird auch im Bereich des Edutainment versucht Lernsoftware durch Einsatz emotionaler Agenten interessanter zu gestalten [PAD96].

Abschließend sei erwähnt, dass neben der Erkennung von Emotionen auch der Synthese dieser seitens der Ausgabe eine tragende Rolle zukommt [PIW03]. Dieser Aspekt ist jedoch nicht Gegenstand der Arbeit.

1.1.2 Wahl der Modalitäten

Es existiert eine Reihe von Modalitäten, die grundsätzlich zur Erkennung der menschlichen

[4] Geläufig unter der Abkürzung *IRC*.

[5] Zusammengesetztes Kunstwort aus *Emotion* und *Icon*. Bildliche Beispiele sind etwa ☺,☹,😐.

Emotion geeignet sind. Traditionell einzuordnen ist dabei vor allem das generell invasive Messen physiologischer Daten wie Hautleitfähigkeit, Herzfrequenz, Temperatur, oder Blutdruck [PIC00], [NAS03]. Obwohl dieser Ansatz geeignet ist sinnvolle Ergebnisse zu liefern, erscheint er unnatürlich, da er einem menschlichen Gegenüber im Allgemeinen nicht zugänglich ist. Darüber hinaus ist durch die in der Regel erforderliche Verkabelung oder den Kontakt mit einem Sensor [PIC97] ein geringerer Komfort für den Anwender gewährleistet [PAN03]. In der modernen Mensch-Maschine-Kommunikation werden daher berührungslose Verfahren zur Analyse des auditiven und des visuellen Kanals in Anlehnung an die menschliche Kommunikation bevorzugt.

Im Rahmen dieser Arbeit wird vorrangig der Sprachkanal betrachtet, der sich in besonderem Maße zur Erkennung eignet [YAN01]. Dabei soll, im Gegensatz zur allgemeinen Fokussierung auf akustische Größen[6], auch linguistische Information in die Betrachtung mit einbezogen werden. Sprache erlaubt dem Benutzer ein besonders hohes Maß an Kontrolle über die gezeigte Emotion, und vermeidet somit das Gefühl einer zu starken Beobachtung. Mikrophone sind in der Fertigung kostengünstig erhältlich und als Standardhardware in Geräten wie Mobiltelefonen vorhanden. Durch die zu erwartende künftige Verbreitung von Spracherkennungstechnologie in Fahrzeugsystemen oder an Rechnerarbeitsplatzen ist hier mutmaßlich noch weiteres Potential gegeben. Durch die angesprochene Betrachtung von Information auf Ebene der Linguistik ergibt sich zusätzlich die Möglichkeit geschriebenen Text hinsichtlich der zu Grunde liegenden Emotion hin zu interpretieren.

Als weitere Modalität wird in dieser Arbeit die manuelle Interaktion durch Zeigen und Auswählen innerhalb einer grafischen Benutzeroberfläche über eine konventionelle Computermaus[7] oder einen Touchscreen als Eingabegerät betrachtet. In den Arbeiten von Ark et al. [ARK99] werden neben der Erfassung der Herzfrequenz durch einen Brustgürtel, der Temperatur und des Hautwiderstands[8] durch ein Thermoelement und ein digitales Multimeter in der linken Hand Emotionen auch über die somatische Bewegung einer Maus ausgewertet. In dieser Arbeit soll gezeigt werden inwiefern es möglich ist, die Emotion allein aus der Interaktion mit einer Maus ohne ergänzende Hardware zu erfassen, wodurch Emotionserkennung auf rein softwaretechnischer Ebene an nahezu jedem Computerarbeitsplatz ermöglicht würde. Als besonderer Vorteil erweist sich dabei die exakte Erfassung des Eingangssignals im Vergleich etwa zur Aufzeichnung von Sprach- oder Videosignalen. Um dem steigenden Einsatz der stift- oder berührungsorientierten Eingabe in Taschencomputern wie PDAs oder Tablet PCs[9] sowie in öffentlichen Informationsterminals Rechnung zu tragen, erscheint es sinnvoll, Emotionserkennung auch aus berührsensitiven manuellen Bewegungsdaten zu erschließen.

---

[6] In [MEH68] wird unter Anderem eine Kommunikationsrate von 38% über die Stimmlage angegeben.

[7] Im Folgenden wird das Eingabegerät Computermaus verkürzt als *Maus* bezeichnet.

[8] In der anglistischen Literatur als *GSR* für *Galvanic Skin Response* oder *EDR* für *Electrodermal Response* bezeichnete aktive oder passive Messung des Hautwiderstands an den Fingerspitzen oder dem Handballen, die die Aktivität der Schweißdrüsen widerspiegelt. Der Hautwiderstand wurde unter Anderem 1967 von Fenz und Epstein zur Messung von Furcht und Stress sowie 1973 von Raskin zur Detektion von Lügen angewandt [FUL77].

[9] Aktuelle Laptop Variante, die auf handschriftliche Eingabe ausgerichtet ist.

Neben den bereits vorgestellten Modalitäten existiert eine Reihe weiterer Ansatzpunkte wie bildbasierte Verfahren zur Erkennung der Mimik und Gestik [PAN03], oder Kontextanalyse [EAG04], wie die Interpretation von SMS und Telefonierverhalten.

Um jedoch ein Maximum an Robustheit zu erreichen, sollte eine möglichst optimale Vereinigung der aus verschiedenen Kanälen stammenden Information realisiert werden.

1.1.3 Defizite im Bereich praktischer Anwendbarkeit

Die Erkennung von Emotion im Rahmen der maschinellen Interaktion eröffnet wie beschrieben eine Vielzahl von Einsatzgebieten, welche auch in ökonomischer Hinsicht erhebliche Bedeutung erlangen könnten. Der Nutzen des Erkennens von Emotion kann also unstreitig von großer Bedeutung sein - jedoch stellt sich die Frage, welche Verfahren und Modalitäten im jeweilig konkreten Bereich unter Berücksichtigung der Robustheit, Effizienz und unter ökonomischen Aspekten tatsächlich praktisch anwendbar sind.

Eine maschinelle Erkennung kann generell nicht fehlerfrei garantiert werden. Als Ausgangsziel, um eine ausreichende Akzeptanz bei Anwendern im Alltagsgebrauch zu gewährleisten, scheint zunächst eine Schätzleistung im Bereich menschlicher Sicherheit anstrebenswert. Einige der geschilderten Szenarien verlangen jedoch noch höhere Genauigkeiten. Dabei sind Einschränkungen wie Personenabhängigkeit oder optimale Erfassungsbedingungen der Ausgangssignale in der Regel eher vertretbar als mangelhafte Präzision.

Aktuell werden als Leistungen automatischer Systeme auf öffentlichen Datenbanken Genauigkeiten im Bereich von 51,6% im Vergleich zu 67,3% menschlicher Sicherheit bei fünf Emotionen [VER04A] bis zu 77,4% im entsprechenden Vergleich zu 84,3% bei sieben Emotionen [VOG05] aus akustischer Analyse berichtet (vgl. Kap. 3.12). Die eingesetzten Systeme haben dabei als Vorteil jeweils vorab Sprachdaten zum Lernen typischer Eigenschaften des Sprechers zur Verfügung. Zur linguistischen Analyse existieren noch keine allgemein üblichen Datenbanken zum Vergleich, und Perzeptionstests sind rar. Auf individuellen Daten werden unter Anderem die Leistungen 75,3% bei zwei Emotionen [LEE02A], 46,7% entgegen 55,0% des Menschen bei drei Emotionen [POL00], und 64,5% im Gegensatz zu 100% entsprechend der Übereinstimmung mehrerer Annotatoren bei sieben Emotionen [ANG02] angegeben. Für die Erkennung aus der manuellen Interaktion können nur Angaben aus dem Bereich der Interessenserkennung herangezogen werden. Ähnliche Arbeiten liegen sonst nicht vor. Hier werden 66,6% als Maximum korrekter Zuordnung genannt [GOC00] (vgl. Kap. 2.7.3).

Die Akkuratheit heutiger Systeme liegt somit zwar vereinzelt unter Einschränkungen nur noch ca. 10% unter der menschlichen Perzeptionsleistung – Fehlerraten von durchgängig mehr als 20% erscheinen dennoch für kaum eine der genannten Applikationen einsetzbar. Hieraus begründet sich der Forschungsbedarf bezüglich einer gesteigerten Sicherheit und Fehlerrobustheit in der automatischen Erkennung von Emotion.

Daher wird diese Arbeit zuerst kritisch den Stand der gegenwärtigen Forschung analysieren, um danach im Rahmen eigener Forschungsleistung Wege zur gesteigerten Robustheit aufzuzeigen und

Ergebnisse hinsichtlich der Effizienz und Nutzbarkeit unterschiedlicher Methoden darzustellen, um das angesprochene Defizit zwischen Anwendungsmöglichkeiten und aktueller Praktikabilität zu reduzieren.

1.1.4 Ziele der Arbeit

Das primäre Ziel der Arbeit ist es, Maschinen im Sinne informationsverarbeitender Systeme eine robuste Einschätzung über die aktuell zu Grunde liegende Emotion ihrer Anwender im Hinblick auf eine soziale Kompetenz zu ermöglichen. Im Rahmen der Themenstellung soll dies aus der Analyse von Daten sprachlicher und manueller Interaktion durch die Konzeption, Implementierung und qualitative Beurteilung neuartiger rechnergestützter Ansätze vollzogen werden.

In Bezug auf die akustische Emotionserkennung, zu deren Realisierung ein breiteres Feld an Forschungsaktivitäten besteht, sollen aktuelle Ansätze durch innovative Merkmale und deren Generierung und quantitative Selektion vorangetrieben werden. Ein extensiver Vergleich diverser Klassifikationsstrategien sowohl dynamischer als auch statischer Natur soll die günstigste Modellierungsform aufzeigen. Hierzu sollen wesentliche aktuelle Verfahren rivalisierend gegenübergestellt werden.

Eine sprachliche Inhaltsanalyse soll als Ansatzpunkt ebenfalls in die Betrachtung der Emotion einfließen. Die auf diesem Gebiet noch wenig zahlreich zu findenden Arbeiten sollen durch innovative Ansätze ergänzt werden. Im Zentrum des Interesses ist dabei ferner eine robuste Handhabung einerseits unsicherer Eingangsgrößen aus der automatischen Verarbeitung lautsprachlicher oder textueller Eingabe, und andererseits auch die Möglichkeit einer synergetischen Zusammenführung mit ergänzenden Informationskanälen.

Bei der Betrachtung der manuellen Interaktion soll eine Erkennung als Novum ohne zusätzlichen Hardwareaufwand über eine Standardcomputermaus oder einen Standardtouchscreen hinaus erfolgen. Dabei soll analysiert werden, inwiefern eine Schätzung aus diesen Daten grundsätzlich möglich erscheint.

Als sekundäres Ziel soll gezeigt werden, wie die erzielten Ergebnisse angewandt und in verwandte Domänen wie Sprachverstehen und Musikverarbeitung transferiert werden können.

1.2 Lösungsansatz

Zur Lösung der Aufgabenstellung werden primär geeignete Verfahren der Signalverarbeitung und des maschinellen Lernens beziehungsweise der Mustererkennung eingesetzt. Auf Grund des stark interdisziplinär geprägten Charakters des Themas sind des Weiteren Ansätze, Betrachtungen und Methodiken der Disziplinen Psychologie, Phonetik und Linguistik eine obligatorische Ergänzung. Im abschließenden Transfer in den Bereich der Musikverarbeitung wird analog Fachwissen in den Modellierungsprozess integriert werden.

Um im Sinne des vorrangigen Ziels der hier durchgeführten Forschungstätigkeit die bisher maximal erzielbare Leistung bei der automatischen Erkennung von Emotion zu steigern, wird, bis auf die hardwareseitig orientierte Erfassung der Signale, im Rahmen dieser Arbeit jeder weitere Teilaspekt

der automatischen Erkennung beleuchtet werden. Hierzu gehört zunächst die Festlegung auf ein Modell zur Beschreibung von Emotion unter der Prämisse einer sinnvollen Einbettung in den späteren Systemkontext. Weiterhin folgt die Erstellung qualitativ hochwertiger Datenbanken zum Trainieren und Evaluieren maschineller Verfahren. Von entscheidender Rolle ist ferner eine der Charakteristik des Signaltyps und der Umgebung angepasste Vorverarbeitung und Merkmalsextraktion. Hier werden entsprechend der jeweiligen Modalität geeignete Größen erarbeitet und gegebenenfalls ergänzt. Des Weiteren werden im Sinne einer methodischen Vorgehensweise Verfahren der automatischen Generierung, Selektion und Reduktion von informationstragenden Attributen eingesetzt. Diese sind im Einzelnen Genetische Algorithmen, Sequentielle Gleitende Suchverfahren, Entropieberechnungen und auf Hauptachsentransformation basierende Reduktion. Um die Klassifikation aufsetzend auf gebildeten Funktionalen optimal zu gestalten, werden instanzbasierte Lerner, Support-Vektor-Maschinen, künstliche Neuronale Netze, Entscheidungsbäume, und Bayessche Netze im Einsatz verglichen. Zur weiteren Steigerung und Vereinigung der Leistung werden die Verfahren Bagging, Boosting und Stacking aus dem Bereich der Ensembleklassifikation eingesetzt. Im unmittelbaren Vergleich hierzu erfolgt eine Schätzung aus dem direkten Verlauf von Merkmalen mit Hilfe von Hidden-Markov-Modellen.

Schließlich wird eine Steigerung der Sicherheit einer Hypothese durch die Fusion von Information verschiedener Kanäle durch frühe Vereinigung auf Merkmalsebene oder späte auf semantischer Ebene vorgestellt.

1.3 Aufbau der Arbeit

In **Kapitel 2** werden die wesentlichen theoretischen Grundlagen im Zusammenhang mit dem Thema der Arbeit behandelt. Neben psychologischen Ansätzen zum Verständnis der menschlichen Emotion und deren Modellierung wird auf die Kollektion emotionaler Beispiele zum Training und Test der automatischen Erkennung eingegangen. Eng im Zusammenhang mit dieser Datensammlung sind die im Anschluss behandelten entwickelten und angewandten Verfahren des Usability Engineering zu sehen. Der für die automatische Erkennung essentielle Bereich der Mustererkennung wird ebenfalls kurz vorgestellt. Den Abschluss des Kapitels bildet ein Überblick über den Stand der Forschung in der automatischen Emotionserkennung. Er wird gemeinsam für die verschiedenen verfolgten Ansatzpunkte vorgestellt und greift teilweise der später erfolgenden ausführlichen Beschreibung der Methodiken voraus.

In den folgenden drei Kapiteln wird auf die verschiedenen betrachteten Größen als Ansatzpunkt für eine maschinelle Bewertung über Benutzeremotion im Detail eingegangen. Neben eingesetzten Merkmalen und ihrer Extraktion werden auch angewandte Verfahren zur automatischen Klassifikation vorgestellt. **Kapitel 3** beschäftigt sich dabei mit der Erkennung aus dem akustischen Sprachsignal, **Kapitel 4** analysiert den linguistischen Inhalt gesprochener oder textueller Eingaben und **Kapitel 5** betrachtet die manuelle Interaktion mit der Maus oder über einen Touchscreen. Die Kapitel schließen jeweils mit durchgeführten Experimenten und erzielten Ergebnissen.

Im Sinne einer Vereinigung aller zur Verfügung stehender Wissensquellen zeigt **Kapitel 6** wie die bisher gewonnene Information unter maximaler Erhaltung von Teilergebnissen integriert werden kann. Um eine möglichst robuste Schätzung zu ermöglichen, werden auch Wege zur Adaption an

den aktuellen Anwender aufgezeigt.

Kapitel 7 beschreibt, wie die vorgeführten Ansätze in der Praxis eingesetzt werden können. Neben der Anwendung in der automatischen Spracherkennung werden zwei implementierte Demonstratoren vorgestellt: Eine Bedienschnittstelle zur Suche in Musikarchiven sowie eine weitere zur Steuerung von Diensten des Infotainments im Fahrzeug. Darüber hinaus wird eine Reihe von erfolgreich durchgeführten Transfers der vorgestellten Methodiken in den Bereich der automatischen Musikverarbeitung vorgestellt.

In **Kapitel 8** folgt eine ausführliche Diskussion der vorgestellten Verfahren und erreichten Ziele. Ergänzend wird Potential für weitere Forschungsarbeiten aufgelistet.

2

Grundlagen

„Nur das Gefühl versteht das Gefühl.“

HEINRICH HEINE (1797-1856)

In diesem Kapitel sollen einige grundlegende Begriffe definiert, Verfahren im Ansatz erklärt, und der Stand der Forschung berichtet werden. Diese Ausführungen bilden die Basis der in den nachfolgenden Kapiteln erfolgenden Beschreibungen durchgeführter Studien.

2.1 Mensch-Maschine-Kommunikation

Die Mensch-Maschine-Kommunikation versteht sich als Bindeglied an der Schnittstelle zwischen Mensch und Maschine [JSN93]. Der Brückenschlag selbst erfolgt dabei durch ein sogenanntes Mensch-Maschine-Interface, kurz *MMI* [CHA94]. Mit einer stetig wachsenden Nutzerzahl technisch zunehmend komplexer Geräte und einem immer breiteren Spektrum von Anwendergruppen steht die Bedienbarkeit heute oft im Vordergrund technischer Systeme [LAN94]. Ziel der jungen Disziplin der Mensch-Maschine-Kommunikation ist es dabei die Nutzung für das Individuum so intuitiv, effizient und natürlich wie möglich zu gestalten. Auf eine Einlernphase sollte möglichst vollständig verzichtet werden können, und ein System sollte jederzeit maximale Transparenz aufweisen. Dies gilt in erster Linie für die vielschichtige Nutzerschaft unterschiedlichen fachlichen Hintergrunds, deren vermehrtem Gebrauch informationstechnischer Geräte im alltäglichen Einsatz Rechnung zu tragen ist. Nicht zuletzt ist dies auch als wirtschaftlicher Faktor zu sehen, da mit steigender Funktionsvielfalt auf dem Markt konkurrierender Produkte eine intuitive Bedienbarkeit im Interesse des Kunden signifikant gestiegen ist [NIL93]. Darüber hinaus gelten die genannten Maxime jedoch auch für hoch technisierte Systeme, die durch Experten ihres Anwendungsgebiets gehandhabt werden sollen.

In jüngerer Zeit sind wesentliche Erfolge in der intuitiven Gestaltung von Bedienprozessen durch die Adaption an menschliche Kommunikationsformen zu verzeichnen. Der technische Fortschritt im weiteren Gebiet der Rechnertechnik erlaubt den Einsatz von Sprach- und Bildverarbeitung auf

hohem Niveau [LAN02]. Darüber hinaus ist es möglich, den Einsatz multipler Informationskanäle des Menschen parallel auszuschöpfen. Diesem Prinzip der sogenannten *Multimodalität* folgend, kann mit einem System etwa durch Interaktion mittels natürlicher Spracheingabe, fließender Handschrift, Gestenbedienung und herkömmlicher manueller Bedienung parallel kommuniziert werden [OVI00A], [TIM00].

Um den eingangs genannten Zielen einer optimalen Bedienbarkeit gerecht zu werden, stellt aus heutiger Sicht die Integration emotionaler Faktoren und sozialer Kompetenz als zusätzliche Größe in die Interaktion eine unabdingbare Komponente künftiger Mensch-Maschine-Schnittstellen dar. Erst durch sie wird eine reale Natürlichkeit in Anlehnung an die zwischenmenschliche Kommunikation möglich.

2.2 Definition der Emotion

Eine einheitliche Definition des Begriffs Emotion[10] ist in der Psychologie sehr umstritten. Grundsätzlich beschreiben Emotionen subjektive Empfindungen kürzerer Zeiträume, die sich auf bestimmte Ereignisse, Personen oder Objekte beziehen. Im Gegensatz hierzu stehen Stimmungen, die sich über längere Abschnitte erstrecken und tendenziell diffusere Bezugspunkte aufweisen. Vier unterschiedliche theoretische Ansätze zur Entstehung und Natur von menschlichen Emotionen haben sich in erster Linie heraus kristallisiert [COR00], [COW01], [PAN03]:

- **Evolutionstheoretischer Ansatz:** Nach der Darwinschen Sichtweise [DAR72] sind Emotionen ein Ergebnis der allgemeinen menschlichen Evolution. Ihnen wird eine essentielle Bedeutung zum Überleben einer Spezies zugeschrieben. In der Konsequenz sind bestimmte Verhaltensmuster direkt mit einer zugeordneten emotionalen Empfindung verknüpft.

- **Stimulativer Ansatz:** Dieser Ansatz begründet sich in den Arbeiten von James [Jam84]. James ist der Ansicht, dass Gefühle die Wahrnehmung der menschlichen Reaktion auf Ereignisse darstellen. Somit entsteht eine Emotion durch die Stimulation von Sinnesorganen durch ein Objekt. Die Perzeption selbst vollzieht sich durch afferente, zum Gehirn hinführende, Impulse, sobald diese den Kortex erreichen. In der Konsequenz werden innere Organe und Muskeln durch efferente Impulse angeregt. Durch die Rückleitung in Form erneuter afferenter Impulse von den Organen und Muskeln zur Großhirnrinde schließlich kommt es zu der beschriebenen Wahrnehmung der körperlichen Veränderung in Form einer Emotion. Eine emotionale Empfindung ist somit nur in Kombination mit einer vorausgehenden körperlichen Reaktion möglich.

- **Kognitiver Ansatz:** Ähnlich dem stimulativen Ansatz wird auch hier von Vertretern dieser Theorie wie Schlachter oder Arnold davon ausgegangen, dass Emotionen ursächlich auf Körperreaktionen entsprechend bestimmter Umstände rückführbar sind. Im Unterschied

[10] Alternativ als *Gefühl* oder lateinisch *Affekt* bezeichnet. Man spricht daher bei der technischen Integration von emotionalen Faktoren vom sogenannten *Affective Computing*.

dazu spielt jedoch die Bewertung dieser Umstände eine entscheidende Rolle. Vor der Reaktion erfolgt somit eine Einschätzung des Erlebten. Entsprechend dieser gestaltet sich die affektive Empfindung. Folglich sind Gefühle mit Bewertungen verbunden und somit bedingt eine Änderung der situativen Einschätzung eine unmittelbare Adaption der Emotion.

- **Sozial konstruktiver Ansatz:** Repräsentanten wie Averill [AVE80] und Harré [HAR86] vertreten die Ansicht, dass Gefühle das Resultat erlernter sozialer Verhaltensregeln widerspiegeln. Entscheidender Faktor dabei ist die zu Grunde liegende Kultur, da sie in bedeutendem Ausmaß die Bewertung von Umständen die zu einer Emotion führen beeinflusst. Entsprechend divergieren die Auslöser etwa für Ärger interkulturell und sogar interpersonell. Diesem Modell folgend spielt der kulturelle Kontext eine bedeutende Rolle für die Beurteilung von Emotionen. Der sozial konstruktive Ansatz gehört zu den jüngsten und mit am kontroversesten diskutierten psychologischen Theorien zu menschlichen Gefühlen. Für ihn spricht, dass manche Syndrome in einigen Kulturen eindeutig als Affekt erkannt werden, während dies in anderen Kreisen nur bedingt zutreffen kann. Im Widerspruch steht dieser Ansatz aber besonders zur Auffassung der Emotionen als Produkt der Evolution.

Abschließend lässt sich feststellen, dass sich diese Ansätze in Teilen zwar überlagern, jedoch wird keiner als allgemein richtig betrachtet. Darüber hinaus existiert eine Reihe von Bestrebungen die Theorien zu vereinen.

2.3 Kategorisierung von Emotion

Entsprechend den teils widersprüchlichen Versuchen zur Beschreibung der menschlichen Gefühle in der Psychologie gibt es mehrere unterschiedliche Ansätze zur Kategorisierung der Emotion. Aus der Perspektive eines hier betrachteten technischen Blickwinkels muss dabei eine pragmatische Entscheidung über das zugrunde liegende Modell getroffen werden. In Bezug auf die Komplexität und Beschreibungsform sollte die angestrebte Anwendung im Rahmen der Zielapplikation im Vordergrund stehen. Im Folgenden werden die zwei aus dieser Sicht wichtigsten Ansätze emotionaler Modellierung vorgestellt.

2.3.1 Dimensionales Modell

Ausgangspunkt dieses auch als Emotionsraum bezeichneten Modells ist ein orthogonales Basissystem mit diverser Zahl emotionaler Dimensionen [BUR00]. Weitestgehende Übereinstimmung herrscht dabei über die beiden Achsen Aktivität a und Valenz v [LNG95][11]. Letztere beschreibt wie angenehm oder positiv, beziehungsweise unangenehm oder negativ ein Gefühl sich gestaltet. Hinzu genommen wird oft eine Achse für Dominanz oder Kontrolle. Einzelne Emotionen, wie im Folgenden näher dargestellt, lassen sich dabei, wie in folgender Abbildung gezeigt, als Punkte $\underline{e}$ auffassen. So liegt etwa im Ursprung der neutrale emotionale Zustand und der skizzierte Punkt gesteigerter Aktivität und Valenz entspricht leichter Freude. Es sind des Weiteren

[11] Respektive *Arousal* und *Valence*.

die Emotionen des Emotionsrads nach Plutchik eingezeichnet [PLU94].

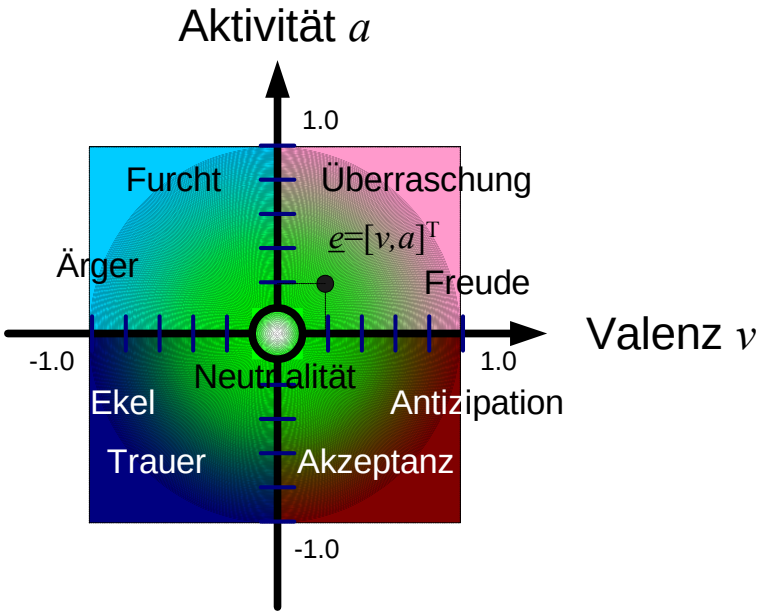


Abb. 2.1: Emotionale Ebene mit Achsen Aktivität und Valenz mit Emotionen des Emotionsrades nach Plutchik

Diese Form der Modellierung weist zunächst den Vorteil einer kontinuierlichen Darstellung auf. Ein System zur automatischen Erkennung, das in der Lage ist einen Punkt zu lokalisieren, erlaubt eine viel komplexere Erkennung als die Zuordnung zu einem, wie im Folgenden beschriebenen, kleinen Set von Emotionen. Zusätzlich sind einige physiologische oder auch akustische Messgrößen direkt zu diesen Dimensionen korreliert[12]. Dies hat den dimensionalen Ansatz zu einer beliebten Darstellungsform gemacht [BAL99]. Auf der anderen Seite übersteigt die Komplexität oft den applikationsspezifischen Bedarf, und erfordert so unnötiges Trainingsmaterial. Das größte Problem ist jedoch in der abstrakten und wenig intuitiven Beschreibung von Emotionen zu sehen. Dies ist vorrangig für einen Vergleich zu menschlicher Erkennungsleistung oder die manuelle Annotation problematisch. Der folgende, im Rahmen dieser Arbeit durchgeführte Versuch, soll dies manifestieren: 15 Studenten im mittleren Alter von 24,8 a (Minimum 22 a, Maximum 28 a), davon vier weiblich, wurden gebeten, 16 äquidistant über die gezeigte Ebene verteilte Punkte und den zusätzlichen Koordinatenursprung mit je einer Emotion zu benennen. Als einzig einheitlich bezeichneter Punkt ergab sich der neutrale Zustand im Ursprung. In einem zweiten Versuchsteil sollte jeder Proband die in Tab. 2.2 aufgelisteten Adjektive als Punkt $\underline{e} = [v,a]^T$ in der beschriebenen Ebene mit $v \in [-1;1]$ und $a \in [-1;1]$ zuordnen. Die Tabelle zeigt auch das Resultat dieser Studie. Angeführt sind der Mittelwert μ und die Standardabweichung σ der Aktivität und der Valenz. Während die Mittelwerte über die Probanden im Vergleich zu Plutchiks Rad für ein sinnvolles Modell sprechen, deutet die hohe mittlere Standardabweichung in Aktivität und Valenz

[12] Die Sprechrate oder mittlere Grundfrequenz stehen beispielsweise in starkem direktem Zusammenhang mit der Aktivität.

auf mangelnde Intuitivität für eine Annotation hin.

Adjektiv	$\mu(v)$	$\sigma(v)$	$\mu(a)$	$\sigma(a)$	Adjektiv	$\mu(v)$	$\sigma(v)$	$\mu(a)$	$\sigma(a)$
aggressiv	**-0,87**	0,19	**0,71**	0,56	**gefällig**	**0,39**	0,25	**0,18**	0,40
ängstlich	**-0,68**	0,21	**-0,15**	0,40	**gelangweilt**	**-0,27**	0,35	**-0,52**	0,42
aufgeregt	**0,07**	0,32	**0,65**	0,31	**interessiert**	**0,41**	0,19	**0,39**	0,56
aufmerksam	**0,38**	0,30	**0,33**	0,54	**missmutig**	**-0,58**	0,19	**-0,13**	0,44
besorgt	**-0,37**	0,25	**0,19**	0,52	**nachdenklich**	**0,03**	0,18	**0,05**	0,49
deprimiert	**-0,84**	0,21	**-0,67**	0,38	**nervös**	**-0,32**	0,31	**0,40**	0,45
dominant	**0,01**	0,29	**0,81**	0,18	**neugierig**	**0,24**	0,23	**0,40**	0,50
enthusiastisch	**0,62**	0,39	**0,72**	0,29	**resigniert**	**-0,70**	0,34	**-0,85**	0,12
entspannt	**0,42**	0,31	**-0,37**	0,35	**scherzend**	**0,42**	0,44	**0,53**	0,25
enttäuscht	**-0,55**	0,28	**-0,32**	0,41	**selbstsicher**	**0,62**	0,23	**0,29**	0,35
ernst	**-0,05**	0,39	**0,20**	0,44	**skeptisch**	**-0,29**	0,23	**-0,10**	0,42
erschöpft	**-0,14**	0,36	**-0,68**	0,22	**überrascht**	**0,27**	0,37	**0,46**	0,42
freundlich	**0,69**	0,19	**0,31**	0,34	**ungeduldig**	**-0,20**	0,38	**0,39**	0,36
fröhlich	**0,66**	0,21	**0,54**	0,33	**verwirrt**	**-0,37**	0,31	**0,02**	0,35
frustriert	**-0,72**	0,16	**-0,12**	0,57	**zufrieden**	**0,60**	0,30	**-0,03**	0,32

Tab. 2.2: Adjektive zur Zuordnung in der Aktivitäts- / Valenzebene mit Angabe des Mittelwerts μ und der Standardabweichung σ. Die mittlere Standardabweichung beträgt 0,28 für die Valenz v und 0,39 für die Aktivität a

2.3.2 Kategoriale Beschreibung

Anstelle eines Kontinuums werden alternativ diskrete emotionale Zustände definiert. Zurückgeführt wird dieser Ansatz eines Basissets von Emotionen auf die Ansichten von Descartes. Vor allem die Vertreter der Darwinschen Sichtweise sprechen von einer Zahl fundamentaler und ursprünglicher Emotionen als Grundlage erlebter Empfindung. Tab. 2.3 gibt einen Überblick über bedeutende Sets [ORT90], [PLU94], [COR96]. Es zeigt sich deutlich eine hohe Variabilität in Zahl und Art der als essentiell definierten Emotionen. Dabei sind teilweise Emotionen ähnlicher Art in mehreren Stufen der Ausgeprägtheit zu finden. Nach Shaver [SHV92] bestehen übergeordnete Basissets auch in einem gewissen Ausmaß über Kulturkreise hinweg. In seinen Arbeiten werden China, Italien und die USA als Kulturkreise mit sechs in hohem Maße übereinstimmenden Emotionen angeführt. Diese sind *Ärger*, *Freude*, *Überraschung*, *Furcht*, *Trauer* und *Liebe*. Sie finden sich in ähnlicher Weise auch in den genannten Sets unterschiedlicher Größe wieder. Dieser Ansatz bietet sich für einen technischen Einsatz als pragmatische Lösung an, wird jedoch unter Psychologen der kognitiven Theorieschule durchaus widersprüchlich bewertet [ORT88]. Im MPEG-4 Standard findet sich das von Ekman [EKM72] vertretene Set diskreter Basisemotionen aus einer Mimikstudie mit japanischen und amerikanischen Probanden wieder, das den Arbeiten von Shaver bis auf die Liebe, die durch *Ekel* ersetzt ist, entspricht [COR96]. Es wird die Grundlage für die Untersuchungen im Rahmen dieser Arbeit bilden. Grund hierfür ist, dass sich dieses Set zurzeit im Hinblick auf Vergleichbarkeit der Ergebnisse in der automatischen Emotionserkennung durchzusetzen scheint [VER03]. Als weitere Emotion wird *Langweile* betrachtet, die in Datenbanken als nächst häufigste

Emotion nach Ärger, Trauer, Freude, Furcht, Ekel und Überraschung zu finden ist [VER03]. Speziell in der Verarbeitung manueller Eingabe werden schließlich die beiden Zustände Irritation und Reflexion ergänzend hinzugefügt, die den gegebenen Möglichkeiten angepasst sind.

Vertreter	**Basisset der Emotion**
Arnold	Ärger, Aversion, Begierde, Courage, Dejektion, Furcht, Hass, Hoffnung, Liebe, Trauer, Verzweiflung
Ekman, Friesen, Ellsworth	Ärger, Ekel, Freude, Furcht, Trauer, Überraschung
Frijda	Begierde, Freude, Interesse, Sorge, Überraschung, Verwunderung
Gray	Freude, Furcht, Rage, Schrecken
Izard	Ärger, Ekel, Freude, Furcht, Interesse, Leid, Scham, Schuldgefühl, Überraschung, Verachtung
James	Furcht, Harm, Liebe, Rage
McDougall	Ärger, Ekel, Elation, Furcht, Unterwerfung, Verwunderung, Zärtlichkeit
Mowrer	Schmerz, Wohlgefallen
Oatley, Johnson-Laird	Ärger, Ängstlichkeit, Ekel, Freude, Trauer
Panksepp	Erwartung, Furcht, Panik, Rage
Plutchik	Akzeptanz, Antizipation, Ärger, Ekel, Freude, Furcht, Trauer, Überraschung
Shaver	Ärger, Freude, Furcht, Liebe, Trauer, Überraschung
Tomkins	Ärger, Ekel, Freude, Furcht, Interesse, Leid, Verachtung, Scham, Überraschung
Watson	Furcht, Liebe, Rage
Weiner, Graham	Freude, Trauer

Tab. 2.3: Auswahl an bedeutenden Sets diskreter emotionaler Zustände

Im Folgenden findet sich eine kurze Beschreibung der im Rahmen der Arbeit angewandten emotionalen Zustände, beginnend mit den Basisemotionen nach Ekman:

- **Ärger** wird oft in *Cold Anger* und *Hot Anger* unterteilt, da die Ausprägung in sehr unterschiedlicher Stärke von Missmut bis hin zu Wut auftritt. Ärger gilt als spontane und innere Reaktion auf eine Situation oder Erinnerung.

- **Ekel** wird als die Empfindung ausgeprägten Widerwillens oder extremer Abneigung gegen Gerüche, Nahrung, aber auch Personen oder Verhaltensweisen verstanden. Ekel zeigt sich im Gegensatz zu einer rationalen Ablehnung durch starke physiologische Reaktionen wie Übelkeit oder Ohnmacht. Ekel kann ferner, ähnlich der im Folgenden diskutierten Furcht, eine protektive Funktion zugeschrieben werden, etwa zum Schutz vor Vergiftungen. Freud ordnet Ekel als erlerntes Verhalten ein, doch beobachtbare interkulturelle Differenzen sprechen zumindest teilweise für eine soziokulturelle Komponente.

- **Freude** ist die einzige hier gewählte positive Emotion. Sie wird als spontane und innere Reaktion auf angenehme Ereignisse oder Erinnerungen betrachtet. In der Konsequenz sind temporär seelische Bedürfnisse erfüllt. Darüber hinaus gilt sie maschinell als eher schwer erkennbar und leicht mit Ärger verwechselbar.

- **Furcht**[13] wird als elementares Gefühl eingeordnet und in der Regel durch die Erwartung oder das Eintreten von Schmerz, Verlust oder Strafe ausgelöst. Dabei muss nicht zwangsläufig ein objektivierbarer Anlass gegeben sein. Auf Grund ihres in die Zukunft gerichteten Warncharakters kann auch ihr Schutzfunktionalität durch oft verbundenes Vermeidungsverhalten zugeschrieben werden. Typische Körperreaktionen in Verbindung mit Furcht sind erhöhte Transpiration, Gliederzittern, Muskelanspannungen, Atmungsstörungen, Kribbelgefühl und in ausgeprägten Fällen Durchfall und Blasenfunktionsstörungen. Als pathologische Form kann sie sich zur Phobie entwickeln.

- **Trauer** tritt als Reaktion auf Verlust auf. Sie führt zu einer Lähmung der normalen Aktivitäten des Menschen.

- **Überraschung** bezeichnet unvorhergesehene Situationen und Ereignisse. Als Reaktion auf diese tritt sie in der Regel spontan und kurz auf. Physiologische Anzeichen sind Aufschauen, Blickwechsel, Erröten, Lachen, und Bewegungen der Extremitäten oder des Kopfes sowie Zucken. Dem Kontext entsprechend kann sie sowohl angenehm wie auch unangenehm ausfallen.

Zu diesen angeführten Emotionen, die im MPEG-4 Standard festgehalten sind[14], wird oft abgrenzend ein neutraler Zustand festgelegt:

- **Neutralität**[15] wird als Anker- oder Basiszustand ohne ausgeprägte Emotion betrachtet.

Weiterhin werden hier verwendet:

- **Langweile**. Sie wird nicht zu den klassischen Basisemotionen gezählt und in vergleichbaren Arbeiten seltener betrachtet. Dieser Zustand weist tendenziell geringe Aktivität und leicht negative Valenz auf. Er kann unter Anderem Anteile von Ärger oder Traurigkeit beinhalten, was erwartungsgemäß zu entsprechenden Konfusionen führen kann. In der Literatur wird zwischen gegen- und zuständlicher Langweile unterschieden. Diese beziehen sich jeweils auf ein Objekt, eine Person oder einen Umstand, beziehungsweise auf rein selbstbezogene Langweile.

- **Irritation** ist eine schwache Ausprägungsform von Überraschung und Hilflosigkeit.

---

[13] Oder auch Angst, abstammend von lateinisch *angustia* für Enge. Der Begriff *Furcht* wurde hier wegen seiner phonetischen Nähe zum englischen *Fear* bevorzugt. Es sei jedoch darauf hingewiesen, dass teilweise zwischen Angst und Furcht differenziert wird.

[14] Dort als *Anger, Fear, Disgust, Joy, Surprise* und *Sadness* bezeichnet.

[15] Dem lateinischen *ne-utrum* für *„keines von beiden“* abstammend.

- **Reflexion** steht hier für Nachdenklichkeit.

2.4 Akquisition emotionaler Daten

Die Ausgangsbasis maschinellen Lernens sind Lernbeispiele. Erst mit ihnen ist es möglich, eine automatische Erkennung zu leisten. Darüber hinaus beeinflusst die Qualität dieser Lerndaten im entscheidenden Maße die Qualität der Leistung eines Systems. Schließlich werden Daten zum Testen benötigt, um die Güte eines gewählten Ansatzes beurteilen zu können. Im Folgenden ist zunächst eine Reihe von Anforderungen und Gütemerkmalen im Sinne optimaler Datenbanken beschrieben, die sich über das Gebiet der Emotionserkennung hinaus teilweise auf verwandte Problemstellungen übertragen lassen:

- **Adäquate Daten**. Bei der konkreten Akquisition emotionaler Muster kann grundsätzlich zwischen spontanen, elizitierten und gespielten Beispielen unterschieden werden. Je nach Anwendungsgebiet können zwar gespielte Daten präferiert werden, etwa bei der Suche in multimedialen Archiven, der Disambiguierung von Ironie oder zur Detektion von Lügen. Generell lässt sich jedoch sagen, dass spontane Reaktionen bezüglich einem potentiellen Umfeld und Verlauf idealer erscheinen.

- **Ideale Erfassungsbedingungen**, im Sinne eines Ausschlusses weiterer Einflussfaktoren. Für die Erkennung aus dem Sprachsignal bedeutet dies etwa eine getrennte Betrachtung unter Umgebungslärm.

- **Gezielte Störeinflüsse**, die in getrennten Sets zur separaten Analyse dieser ebenfalls erfasst sein sollten.

- **Hohe Gesamtzahl** an Lern- und Testbeispielen. In der Regel ist mit steigender Zahl an Lernbeispielen auch eine höhere Gesamtleistung eines Erkennungssystems auf Grund exakterer Modellierung der Problemstellung zu erwarten.

- **Sinnvolle Kategorisierung**. Die Annotation von Daten, beziehungsweise ihre Zuordnung zu Klassen ist ein essentieller Bestandteil einer Datenbank. Für viele Gebiete wie Spracherkennung erscheint dies unproblematisch. Die Diffizilität bei der Emotionserkennung wurde in Kap. 2.3 beschrieben.

- **Eindeutige Annotation** durch die Testperson selbst, Experten oder eine möglichst hohe Zahl von individuellen Annotatoren. Auch dies erweist sich als nicht trivial, da eine eindeutige Zuordnung beispielsweise einer lautsprachlichen Äußerung zu einer Emotion teilweise nur aus dem Kontext heraus, oder anhand der Empfindung des Sprechers selbst getroffen werden kann. Teilweise kommen hier auch begleitende physiologische Messungen zum Einsatz.

- **Gleichmäßige Verteilung** unter den gewählten Kategorien, um ausreichend Daten für jede Klasse zu gewährleisten. Darüber hinaus besteht die Gefahr der Bevorzugung einer überrepräsentierten Klasse.

- **Perzeptionstests** mit einer repräsentativen Zahl an Testpersonen, um eine Vergleichsbasis mit der menschlichen Leistung zu ermöglichen.

- **Hohe Diversität** bezüglich Personen, Altersklassen, Geschlecht, Bildungsgrad, kulturellen und sozialen Hintergründen, Kontext und Sprache für die sprachliche Analyse, um ein möglichst umfassendes Abbild der Gesamtheit zu leisten.

- **Beschreibung der Randbedingungen**. Ohne diese ist eine Datenbank nur bedingt gebrauchstauglich.

- **Freigabe der Daten** zur Gewährleistung internationaler Vergleichbarkeit und Reproduzierbarkeit erzielter Ergebnisse.

Möchte man diesen Kriterien bei der Kollektion von Datenmaterial möglichst umfassend gerecht werden, sind in erster Linie gespielte Daten von professionellen Schauspielern oder Laien nahe liegend. Diese bieten die Vorteile eindeutiger Zuordenbarkeit, Aufnahmen in gewünschter Qualität, eine hohe Zahl erzielbarer Beispiele, und eine ideale Verteilung. Auch bezüglich einer maximalen Vielfalt ist, im Rahmen gegebener Möglichkeiten, vergleichsweise hohe Flexibilität gewährleistet. Somit leistet diese Variante eine sinnvolle und allgemein gerne verwendete Grundlage, um zu einem großen Schatz an Beispielen zu gelangen, die einen Test auf grundsätzliche Eignung gewählter Verfahren erlauben. Als Nachteil ist die Gefahr unrealistisch übertriebener oder verzerrter Darstellung der Äußerungen sowie in Einzelfällen auch die mangelnde Ausprägung je nach schauspielerischer Begabung der Akteure zu nennen. Eine Aufnahme über einen größeren Zeitraum kann einem zu starken Lerneffekt entgegenwirken. Durch diesen fallen Beispiele unter Umständen zu ähnlich aus, um von diesen ausgehend später eine Verallgemeinerung leisten zu können. Im Hinblick auf die Evaluierung der Leistungsfähigkeit automatischer Ansätze zur Emotionserkennung ist dennoch umstritten, inwiefern diese gespielten Muster auf Grund des teilweise zu ausgeprägten Charakters nicht eine zu starke Vereinfachung des Problems darstellen. Nach [BAT00B] sind sie etwa leichter zu erkennen als beispielsweise elizitierte, da in Versuchen unter Laborbedingungen Emotion oft nicht richtig gezeigt wird.

Diese letztgenannte Provokation von Emotionen kann als Behelfslösung für eine größere Natürlichkeit gesehen werden. Stressversuche [FER00], eine emotionalen Einstimmung [MOZ98], Spiele, oder Geschmackstests sind hier beliebte Möglichkeiten. Oft wird jedoch auch eine falsche Emotion elizitiert. Dieser Umstand erfordert eine nachträgliche Kontrolle und Annotation des Materials. Des Weiteren ist es in der Regel sehr aufwändig Stichproben in ausreichender Zahl zu erhalten.

Um schließlich spontane Daten zu erhalten, können Beobachtungen vollzogen werden, in denen eine Versuchsperson kontinuierlich im Alltag beobachtet wird [CAM02]. Ziel ist es hier, das in Laborversuchen für die Testperson oft omnipräsente Gefühl der Observation zu reduzieren. Hierdurch entsteht jedoch ein besonders hoher Überschuss an Daten, was einen hohen Aufwand zur Segmentierung und Verschriftung mit sich bringt. Die Zuordenbarkeit ist hierbei erschwert, da der Proband erst nach einer längeren Zeit zu seiner Emotion befragt werden kann, um die Versuchsatmosphäre nicht zu präsent zu gestalten. Zusätzlich ergibt sich in der Regel eine verzerrte

Verteilung unter den Emotionen, welche sich aber über einen längeren Zeitraum gesehen der tatsächlichen Auftrittswahrscheinlichkeit annähert. Eine optimale Vermeidung von störenden Beobachtungseinflüssen lässt sich dabei ausschließlich durch eine verdeckte Observation erreichen. Auf Grund der rechtlichen Situation und zur Wahrung der Privatsphäre ist dies aber durchaus kritisch zu bewerten. Generell wird die größte Spontaneität nur auf Kosten geringerer Zuordenbarkeit, eines ebenfalls hohen Überschusses an unbrauchbaren Daten und in der Regel mit hohem Störeinfluss bei der Aufzeichnung erzielt. Ähnlich verhält sich dies bei Ausschnitten aus Medienbeiträgen, die sich speziell für Bild- und Tonmaterial anbieten. Auch hier ist ein hoher Aufwand bei der Segmentierung zu betreiben und es sind in der Regel überlagerte Störungen präsent. Die Emotionen sind je nach Beitragsart spontan wie in Liveübertragungen oder gespielt wie in Spielfilmen. Eine Beurteilung der Emotion kann aber bei echten Gefühlen nur aus der Situation heraus von Dritten eingeteilt werden. Bei gespieltem Affekt ist hier dafür von höherer Professionalität als bei Experimenten mit Laien auszugehen.

Speziell für die Erkennung aus der gesprochenen oder geschriebenen Sprache stellt sich des Weiteren die Frage nach geeigneten Testsätzen. Hierbei lassen sich vier Kategorien unterscheiden [BUR00]:

- **Sinnleere Texte**, die keinen semantischen Gehalt aufweisen. Dies können etwa fiktive Silben oder Alphabete sein.
- **Emotional neutrale Texte**, denen keine emotionale Bedeutung inhärent ist.
- **Emotional unbestimmte Texte,** denen auf semantischer Ebene keine Emotion eindeutig zuordenbar ist, wie etwa der Ausruf *„Ganz toll!“*, der ohne Betrachtung der Prosodie oder des Kontexts sowohl erfreut als auch ironisch verärgert bewertet werden kann.
- **Emotionale Texte,** deren Inhalt eindeutig auf eine Emotion Rückschluss erlaubt. Exemplarisch sei der Ausspruch *„Ich hasse ihn!“* angeführt.

Vorzugsweise werden in dieser Arbeit, soweit vorhanden, öffentlich zugängliche Datenbanken aus Gründen der Vergleichbarkeit bevorzugt. Für konkrete Bedarfssituationen werden jedoch eigene Sammlungen unter Berücksichtigung beschriebener Aspekte erstellt.

2.5 Usability Engineering

Unter *Usability Engineering* sind im weiteren Sinne Methoden der Akzeptanzuntersuchung und der Bedientauglichkeit von Eingabegeräten zu verstehen. Dabei definiert sich Usability[16], im Deutschen auch *Gebrauchstauglichkeit*, *Bedienfreundlichkeit* oder *Benutzbarkeit*, nach der Norm ISO 9241-11 als das Ausmaß, in dem ein Produkt durch Benutzer angewendet werden kann, um bestimmte Ziele effektiv, effizient und mit Zufriedenheit zu erreichen [ISO98]. Effektivität steht hier für Genauigkeit und Vollständigkeit, mit der ein Arbeitsergebnis erreicht werden kann. Effizienz hingegen bezieht

[16] Im IEEE-Standard 610.12-1990 wird Usability folgendermaßen festgelegt: *„The ease with which a user can learn to operate, prepare inputs for, and interpret outputs of a system or component“*.

sich auf einen dabei möglichst geringen Aufwand. Es kann nach zeitlicher, wirtschaftlicher und menschlicher Effizienz differenziert werden. Die Zufriedenheit als drittes Hauptziel der Bedienfreundlichkeit bezieht sich speziell auf die subjektive Bewertung durch den Anwender.

In dieser Arbeit kommen Methoden des Usability Engineerings in zweierlei Hinsicht zum Einsatz: Es sollen erstens vorgestellte neue Verfahren zur Erkennung der Benutzeremotion im realen Einsatz bezüglich Ihrer Akzeptanz bewertet werden. Zweitens sind aber Untersuchungen im besonderen Maße zur beschriebenen Erhebung von Daten als Referenzmaterial für das Training und die Evaluation der eingesetzten lernenden Verfahren von Nöten. Das Verständnis ausgewählter Methoden des Usability Engineerings ist daher primär Vorraussetzung für die Betrachtung der Sammlung emotionaler Daten aus dem gewünschten Kontext der späteren Anwendung.

Als eine spezielle Form der Versuchsdurchführung sei hier die *Wizard-Of-Oz (WOO)*[17] Studie vorgestellt. Es handelt sich dabei um die Simulation eines funktionsfähigen Systems durch einen menschlichen Operator, *Wizard* genannt. Letzterer kann dabei einzelne Teile eines Systems oder sogar die gesamte Funktionalität durch eine Fernsteuerung simulieren [NIL93]. Für die Testperson bleibt er dabei in der Regel verborgen. Die WOO Methode kann etwa zur Vortäuschung eines auf Emotion reagierenden Systems angewandt werden, um Reaktionen der Nutzer zu beobachten, bevor ein solches real existiert.

2.5.1 Laborräume

Die Versuche werden in zwei Labors durchgeführt: Der *reflexionsarme Raum* eignet sich besonders zur Aufnahme akustischer Schallbeispiele. Er unterstützt die freie akustische Ausbreitung unter nahezu gänzlicher Vermeidung von Reflexionen in starker akustischer Isolation. Zu diesem Zweck sind die Wände, Decke und der Boden, wie in Abb. 2.4 zu sehen, mit verschachtelten keilförmigen Mineralfibern gekleidet. Diese besitzen über einen weiten spektralen Bereich einen hohen Absorptionskoeffizient. Um die Kammer zusätzlich vor der Induktion externer Schwingungen zu sichern, ist sie in einem umgebenden Raum eingelagert[18]. Außerhalb dieser beiden Kammern befindet sich ein Kontrollbereich für einen Versuchsleiter.

---

[17] Benannt nach der Hauptfigur eines im amerikanischen Raum populären Kinderbuchs aus dem Jahr 1900 von Lyman Frank Baum mit dem Titel *„The Wizard of Oz“*. Die Namensgebung beruht auf der Dialogstelle *„Pay no attention to that man behind the curtain.“,* in direkter Anspielung auf den Versuchsleiter, der meist hinter einem Sichtschutz agiert. Die Figuren und Teile der Handlung werden auch in anderen Bereichen gerne als Allegorie verwendet. Eines der ersten und bekanntesten WOO Experimente ist der *„Listening Typewriter“* [GOU83].

[18] Das Nutzvolumen beträgt in Länge x Breite x Höhe 7,5m x 4,2m x 2,8m = 88,2 m³.

Abb. 2.4: Probandin im reflexionsarmen Raum bei der Aufnahme emotionaler Sprachsamples

Ein weiterer Teil der Versuche wurde in einer fahrzeugähnlichen Versuchsumgebung, die angepasste akustische Bedingungen garantieren und einer realen Fahrzeugsituation weitest gehend gerecht werden soll, durchgeführt. Hierfür wurde ein zweigeteilter Raum verwendet, der sich in eine Kontrollzentrale zur Beobachtung und Steuerung durch den Versuchsleiter, und den durch eine Glaswand und einen zusätzlichen Vorhang abgeschirmten Fahrzeugbereich gliedert (siehe Abb. 2.5). In letzterem befindet sich ein Fahrzeug des Typs BMW 750i, das mit einem adaptierten Lenkradcontroller der Firma Microsoft mit Kraftrückführung, Fabrikat SideWinder, einer Schaltung vom Typ Tiptronic, Gas- und Bremspedalen, einem Touchscreen und einer Bedieneinheit sowie Kameras und Mikrophonen ausgestattet ist. Vor dem Fahrzeug befindet sich eine große Projektionsfläche[19], die an den Rändern mit dem Sichtbereich des geradeaus schauenden Fahrers schließt. Auf diese kann die in Kap. A.5 vorgeführte Fahraufgabe projiziert werden, um Benutzerverhalten in Fahrsituationen zu evaluieren.

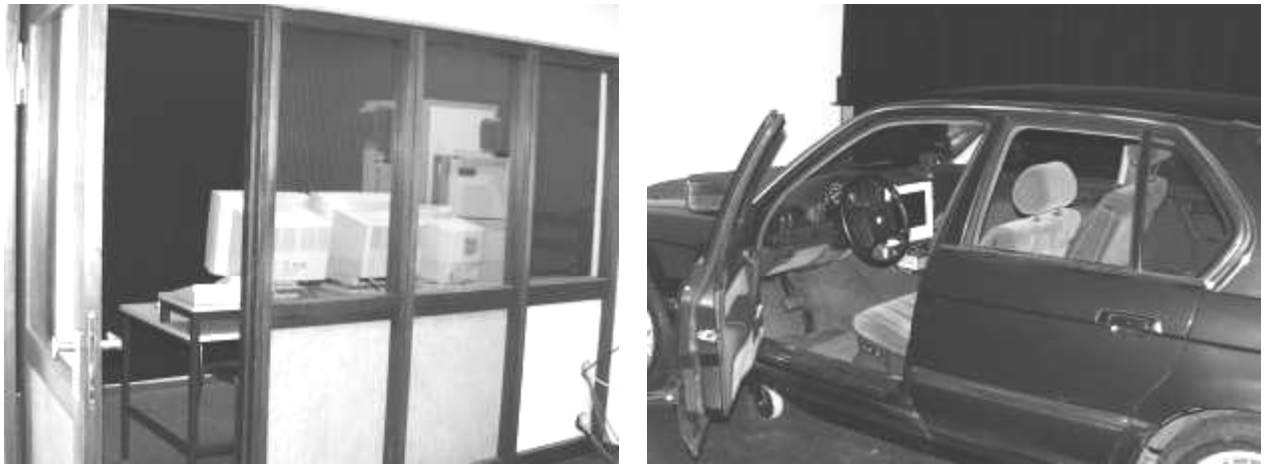


Abb. 2.5: Kontrollstand des Versuchsleiters und Fahrzeug für die Versuchsdurchführung

[19] Breite x Höhe 4m x 3m.

2.5.2 Semiautomatische Versuchsablaufsteuerung

Um der Anforderung nach reproduzierbaren und konstanten Testbedingungen [NIL93] gerecht zu werden, wird ein Konzept zum teilautomatisierten Versuchsablauf vorgestellt [NIE02], [SCH02E]. Besonders bei WOO Experimenten kann ein Versuchsleiter so während der Durchführung und bei der Nachbearbeitung einer Studie entlastet werden [BAL96]. Ziel ist es, die Steuerung des Ablaufs, die Erfassung, Segmentierung und Auswertung von Daten zu integrieren und zu automatisieren. Zu Grunde liegender Gedanke ist die Vorhersehbarkeit typischer Versuchabläufe. Soll beispielsweise bei vorgetäuschter Systemfunktionalität eine emotionale Benutzerreaktion in einer bestimmten Situation getestet werden, wird dem Probanden eine konkrete Bedienaufgabe hierzu gestellt. Zur Erfüllung dieser bietet sich nur eine im Vergleich zur Gesamtfunktionalität geringe Zahl von Möglichkeiten an. Da diese im entsprechenden Moment bekannt sind, kann so die Gesamtzahl zu einem bestimmten Zeitpunkt zu simulierender Systemreaktionen stark eingeschränkt werden. Dies kann die Reaktionszeit eines Versuchsleiters in WOO-Experimenten stark verkürzen und die Fehlerrate durch falsche Simulation reduzieren. Darüber hinaus können so komplexere Funktionalitäten oft bereits durch eine einzige Person simuliert werden. Abb. 2.6 veranschaulicht das Prinzip. In ihr findet sich auch der Fall einer Abweichung gegenüber dem vorab angenommenen Ablauffluss. Hieraus bedingt sich, dass jederzeit Zugriff auf die volle Systemfunktionalität bestehen muss.

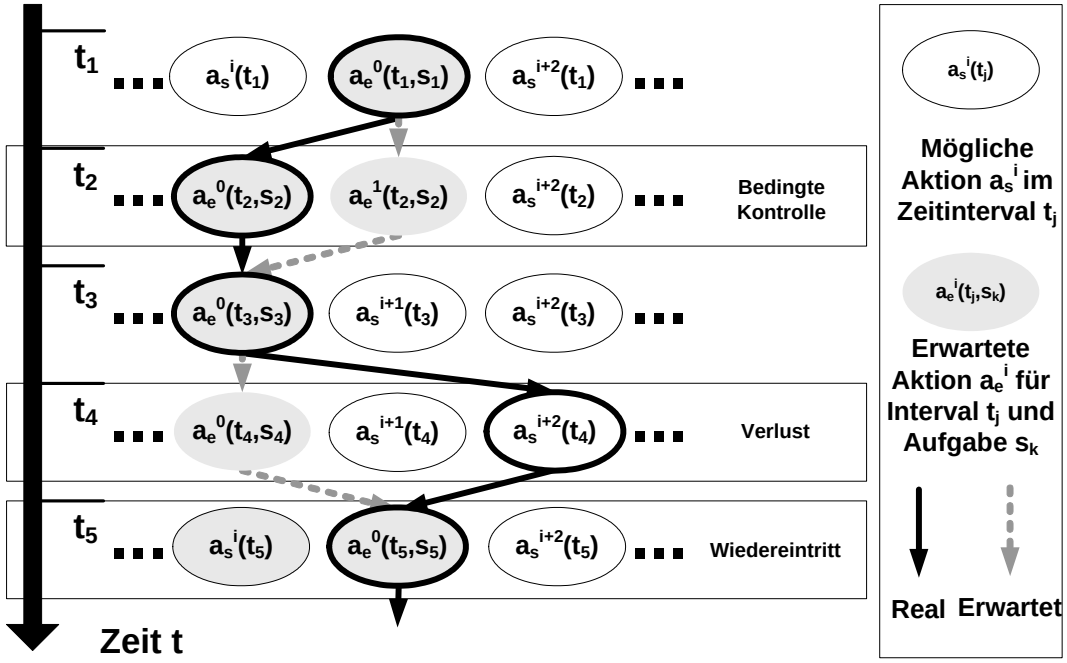


Abb. 2.6: Zeitlicher Ablauf erwarteter und potentieller Benutzeraktionen während einer Wizard-Of-Oz Studie

Die Einführung einer interpretierten Kontroll-Sprache, welche über bedingte Verzweigungs- und Schleifenstrukturen verfügt, erlaubt nun die Programmierung einer zentralen Steuereinheit für den Ablauffluss. Diese steuert neben dem Test-MMI weitere Einheiten, welche versuchsspezifische

Aufgaben wie die Ansage von Versuchsanweisungen oder die Kontrolle von Nebenaufgaben zur Bedienung, beispielsweise eine Fahrsimulation wie in Kap. A.5 beschrieben, übernehmen. In einer zentralen Logdatei werden anfallende Daten des MMI, der aufgabenspezifischen Einheiten und der Prozesskommunikation zusammengefasst. Eine Spezifikation in BNF dieser interpretierten Kontroll-Sprache findet sich in Kap. A.1. Von besonderer Bedeutung ist die Segmentierung der Versuche in Einzeletappen sowie die Synchronisation zu eventuell parallel erfolgenden Bild- und Tonaufzeichnungen für die vereinfachte Protokollanalyse. Abb. 2.7 gibt einen Überblick über die Architektur des Konzepts. Das Herzstück bildet die zentrale Steuereinheit, welche das Ablaufskript sukzessive in Echtzeit abarbeitet. An Verzweigungsstellen im Versuchsablauf wird auf Eingaben des Wizards gewartet. Letzterer segmentiert die Studie, indem er dem System mitteilt, wie weit der Proband vorangeschritten ist, für welche aktuell gültige Teilaktion er sich entschieden hat, und ob er das Versuchsziel erreicht oder verfehlt hat. Für Ausnahmefälle besitzt der Versuchsleiter die Möglichkeit, direkt in das Geschehen einzugreifen. Er kann ferner Teilziele wiederholen lassen oder überspringen. Die Kommunikation zwischen Einheiten erfolgt über TCP/IP und einen einfachen Protokollstandard[20], um eine Verteilung auf multiple Rechner zuzulassen.

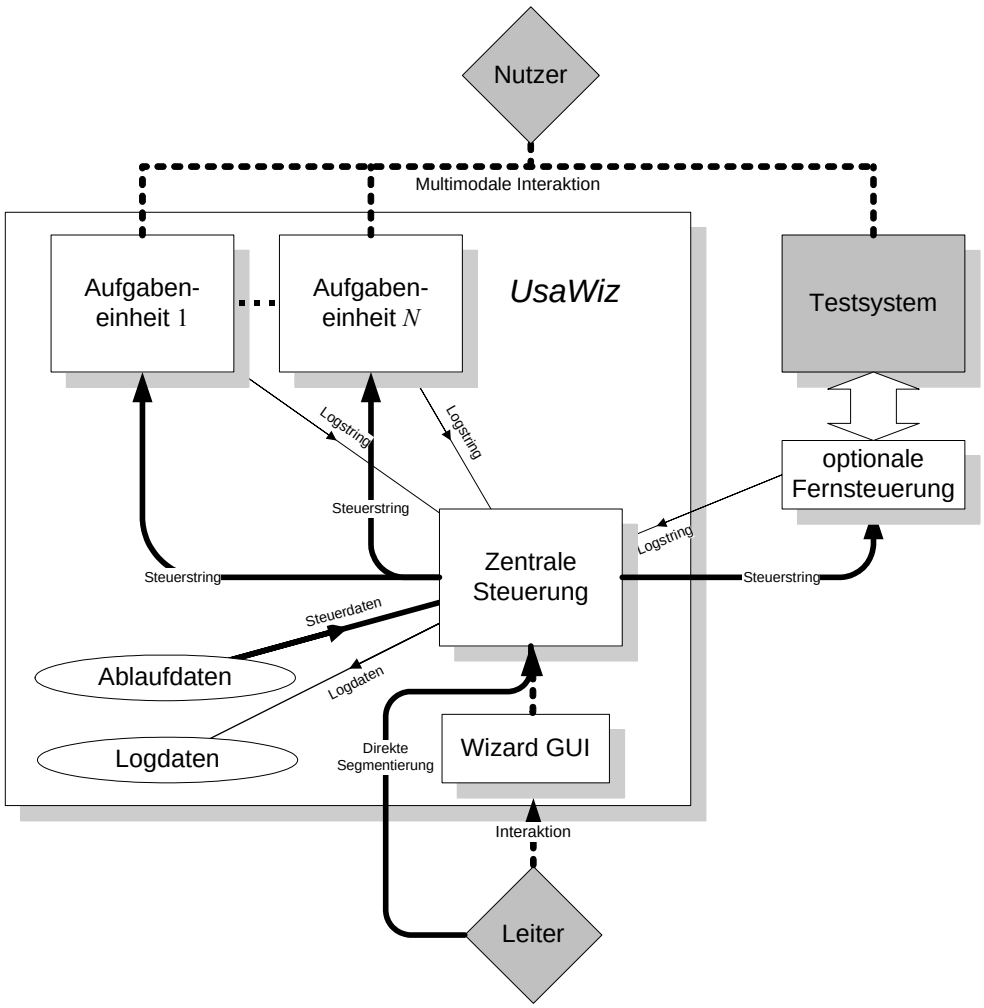


Abb. 2.7: Architekturübersicht zur semiautomatischen Versuchsablaufsteuerung

[20] Dieser besteht im Wesentlichen aus Quellen- und Senkenangabe sowie einem Eintrag für Anweisungen.

Tab. 2.8 zeigt zwei beispielhafte Szenarien, die die erzielte Effektivität demonstrieren sollen.

	Basis-Aktionen (Einzelaktionen)	**Reduktion Aktionen auf**	**Beschleunigungs-Faktor**
MMI im Fahrzeug	20 (>>10^4)	7%	13
3D-Navigation	116 (>10^3)	2%	21

Tab. 2.8: Evaluierung der semiautomatischen Versuchsablaufsteuerung

Vorneweg ist die Zahl der potenziell zur Verfügung stehenden Basisaktionen, und in Klammern die sich aus der Zahl der Parametervariationen ergebende Anzahl von Aktionen angegeben. Des Weiteren wird gezeigt, auf wie viel Prozent die Gesamtzahl im Mittel über den Ablauf konkreter Versuche[21] reduziert werden kann. Der sich daraus ergebende Beschleunigungsfaktor wurde anhand multipler Versuche mit und ohne teilautomatisierte Steuerung gemessen (siehe [SCH02E]).

2.6 Mustererkennung

Die stetige und weiterhin erwartete Steigerung der Leistungsfähigkeit von Digitalrechnern[22] ermöglicht es heute und in Zukunft komplexe Abläufe der Informationsverarbeitung zu analysieren und zu modellieren. Besonders von Interesse ist hier die Imitation perzeptiver Fähigkeiten des Menschen und anderer Lebewesen, wie die Erkennung von gesprochener Sprache, Gestik [ZOB03], oder die in dieser Arbeit betrachtete Einschätzung menschlicher Emotion. Im Vordergrund einer Simulation solcher perzeptiver Leistungen steht dabei weniger eine Nachbildung der Methodik der Natur. Ziel ist es vielmehr, eine mindestens adäquate Leistung in der Zuordnung zu erreichen.

Unter dem Begriff der *Mustererkennung* versteht man zusammengefasst die mathematisch-technische Analyse von Aspekten der automatischen Auswertung und Verarbeitung von Mustern [NIE03]. Ein *Muster* definiert sich dabei als Elemente $f(x)$ der Menge Ω von einem bestimmten, limitierten Anwendungsgebiet zugehöriger Funktionen, die mit geeigneten Sensoren erfasst werden können. Eine weitere Unterteilung der Menge $\Omega \subset U$ als Einschränkung der Umwelt U[23] führt zur Definition der Musterklassen, kurz *Klassen* Ω_κ. Diese Teilung erfolgt dabei in k Untermengen unter Einhaltung der folgenden Bedingungen:

[21] Eine nähere Beschreibung der Versuche zum Szenario Navigation in 3D-Welten findet sich in [ALT01] und [SCH01B]. Das MMI im Fahrzeug wird in Kap. 7.3.1 näher beschrieben.

[22] Diese verhält sich zurzeit noch nach dem *Mooreschen Gesetz*, einer Annahme des Intel® Gründers Gordon Moore im Jahr 1975, dass sich die Zahl der Transistoren auf einem Chip alle ein bis zwei Jahre verdoppeln wird. Dies hat sich bis jetzt erfüllt. Seit aufgetretenen Problemen bei den Chips *Prescott* und *Tejas* wird die weitere Entwicklung auf Grund physikalischer Grenzen jedoch skeptischer gesehen.

[23] Diese Einschränkung der Gesamtheit physikalisch messbarer Funktionen begründet sich in der Praktikabilität.

$$\begin{aligned} &\Omega_\kappa \neq \varnothing && mit && \kappa = 1,...,k \\ &\Omega_\kappa \cap \Omega_\lambda = \varnothing && für && \kappa \neq \lambda \\ &\bigcup_{\kappa=1}^{k} \Omega_\kappa = \Omega \end{aligned} \tag{2.1}$$

Die Klassen sind somit als disjunkt vorausgesetzt. Unter der Vielzahl möglicher Zerlegungen, die dem Genüge leisten, sei hier von einer sinnvollen Variante ausgegangen, bei der Muster einer zugehörigen Klasse sich untereinander ähnlich sind, und sich von denen anderer Klassen unterscheiden. Um Muster zu modellieren, die keiner Klasse zugeordnet werden können, kann eine zusätzliche Rückweisungsklasse Ω_0 eingeführt werden.

Der Einsatz von Verfahren der automatischen Mustererkennung im Rahmen dieser Arbeit begründet sich somit durch ihre Fähigkeit unbekannte Muster einer Klasse nach einer Lernphase zuzuordnen. Dabei entsprechen den Mustern unter Anderem Beobachtungen über die sprachliche und manuelle Interaktion des Anwenders und den Klassen die jeweiligen Emotionen. Es wird in der Regel zwischen den drei Hauptaufgaben *Vorverarbeitung*, *Merkmalsextraktion* und *Klassifikation* differenziert, wobei die beiden erstgenannten je nach Art der Aufgabenstellung zusammengefasst sein können. Sie sollen im Folgenden kurz erläutert werden. Eine detaillierte und auf die praktische Anwendung bezogene Darstellung findet sich in den jeweiligen Kapiteln zur automatischen Erkennung der Emotion aus Sprache und manueller Interaktion.

2.6.1 Vorverarbeitung und Merkmalsextraktion

Es existiert nahezu keine Aufgabe im Bereich der Mustererkennung, in der das originale beobachtete Signal direkt einem Klassifikator angeboten werden kann [RIG04]. Im Rahmen der Vorverarbeitung soll daher eine kanonische Repräsentation der Information erreicht werden. Dies gilt unabhängig von der Art des untersuchten Musters und beinhaltet in der Regel eine gemeinsame Normalisierung eines Datenvektors. Weiteres Ziel ist oft die Eliminierung oder Verminderung von Störungen [RUS88], wie etwa die Optimierung der SNR, oder das Selektieren relevanter Abschnitte.

Während der Merkmalsextraktion werden die vorverarbeiteten Daten in einen musterspezifischen Merkmalsvektor $\underline{x} \in \mathbb{R}^n$ der Dimension *n* unter Hervorhebung der Charakteristik und Trennbarkeit überführt. Oft resultiert dies in einer Reduktion der ursprünglichen Daten [RUS88]. Je nach Beschaffenheit des Quellsignals ergibt sich hier eine Vielzahl von Möglichkeiten.

2.6.2 Klassifikation

Die automatische Klassifikation hat die tatsächliche Zuordnung des Musters zur jeweiligen Klasse κ von *k* Klassen in der nun parametrisch günstigeren Darstellung als Aufgabe. Dies lässt sich formal wie folgt als Abbildung beschreiben:

$$\underline{x} \rightarrow \Omega_\kappa \tag{2.2}$$

Die hier allgemein als Ω_κ angeführte Klasse sei im Weiteren speziell durch die *erkannte* Klasse κ_e referenziert. Es werden somit Verfahren zur Zuordnung von Objekten unbekannter Typen oder

Klassen zu einer oder mehreren möglichen Klassen zusammengefasst. Hierzu kommen geeignete Methoden des maschinellen Lernens[24] zum Einsatz. Diese lassen sich in schablonen-basierte, statistische, syntaktische oder regelbasierte einteilen [JAI00].

Für die Bildung von Modellen für die Klassifikation ist eine Menge von Lernbeispielen $\mathcal{L}$ erforderlich (vgl. Kap. 2.4). Dies begründet sich durch den Bedarf an Kenntnis der von einem System zu behandelnden Objekte. Die Menge darf ausschließlich Muster aus dem Problemkreis des Interesses beinhalten, es sei denn, sie sind als Ausnahmen gekennzeichnet. Weiterhin ist zu fordern, dass die Menge repräsentativ für den Problemkreis ist, um eine Generalisierung von Beobachtungen zu erlauben. $\mathcal{L}$ setzt sich aus Tupeln $(\underline{x}_l, y_l)$ einerseits eines Merkmalsvektors $\underline{x}_l$ und andererseits seines zugehörigen kategorialen Attributs y_l zusammen. Letzteres kann dabei der Klassenname oder der Klassenindex κ der Klasse Ω_κ sein. In Sonderfällen ist lediglich eine binäre Entscheidung möglich. In diesem Fall gilt $y_l \in \{0,1\}$ [25], wobei im Falle multipler Klassen $y_l = 1$ die Zugehörigkeit zu einer Klasse und entsprechend $y_l = 0$ nicht vorhandene Zugehörigkeit anzeigt [26]. Somit sei die Lernmenge gegeben zu:

$$\mathcal{L} = \left\{ (\underline{x}_l, y_l) \mid l = 1,...,L \right\} \tag{2.3}$$

Die getroffene Notation mit den Designatoren $\underline{x}$ und y ist dabei auch als Ein-, beziehungsweise Ausgabe eines Klassifikators vorstellbar.

Die folgende Darstellung vermittelt ein Bild des wesentlichen Ablaufs bei der Klassifikation.

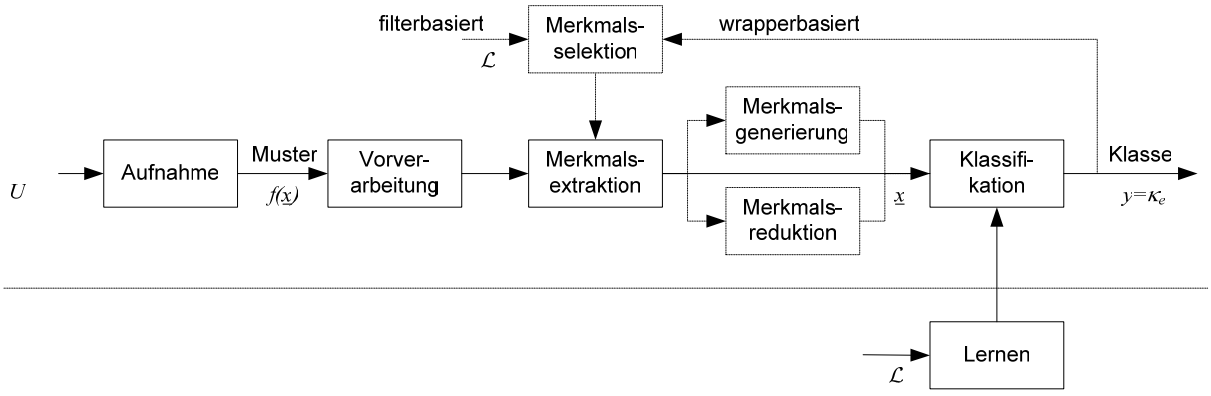


Abb. 2.9: Wesentlicher Ablauf bei der Klassifikation von Mustern

---

[24] Dieses Lernen als Teilbereich des interdisziplinären Felds künstlicher Intelligenz (*KI*) ist im Wesentlichen auf die Arbeiten von Alan Turing im Jahr 1950 zurückzuführen. Die künstliche Intelligenz selbst hat es nach John McCarthy zum Ziel, Maschinen zu bilden, die sich intelligent verhalten. Hier steht die sogenannte *schwache KI* im Vordergrund, die auf den philosophisch fragwürdigen Anspruch der Schaffung eines Bewusstseins verzichtet. Sie fokussiert sich auf konkrete Anwendungsparadigmen, die intelligente Leistung erfordern.

[25] Oder auch $y_l = \{-1,1\}$, wie speziell in Kap. 3.7.2.

[26] Der Sonderfall $y_m \in \mathbb{R}$ führt zum Regressionsproblem, welches in Kap. 3.8.2 behandelt wird.

Die eingezeichneten Module Merkmalsselektion, -Generierung und –Reduktion sind optional, und hier der Vollständigkeit halber eingezeichnet. Sie werden in Kap. 3.9 und Kap. 3.10 behandelt. Darüber hinaus können weitere Module auftreten, wie ein Vektorquantisierer nach der Merkmalsextraktion.

Eine Reihe unterschiedlicher Aspekte motiviert den im weiteren Verlauf der Arbeit vollzogenen Einsatz und Vergleich verschiedener Verfahren zur Klassifikation der betrachteten Merkmale. Tab. 2.10 soll dies exemplarisch veranschaulichen. In späteren Kapiteln wird unter Anderem auch versucht, die Stärken mehrerer Lernverfahren durch geeignete Ansätze zu vereinen.

Anforderung	**Beispiele**
Adäquate Modellierung	Statische oder dynamische Klassifikation
	Daten- oder wissensgetriebene Modellierung
	Umgang mit fehlenden Werten
	Umgang mit Unsicherheiten
	Trainingsstabilität
	Modell- oder instanzbasiert
	Transparenz
Optimale Erkennungsleistung	Lösbarkeit nichtlinearer Probleme
	Maximierte Diskriminativität bezüglich der Klassen
	Eigenständige Priorisierung von Merkmalen
	Toleranz gegenüber Dimensionserhöhung
	Nachtrainierbarkeit
	Hohe Verallgemeinerungsfähigkeit
	Adaption an die Komplexität
	Einsetzbarkeit diverser Sets von Merkmalen
Effizienz	Echtzeitfähigkeit in der Erkennung
	Kurze Trainingszeit
	Geringer Bedarf an Lernbeispielen
Ökonomische Faktoren	Geringer Bedarf an Rechenleistung
	Geringer Bedarf an Speicher
	Günstige Umsetzbarkeit in Hardware
	Merkmalsreduktion ohne Trainingsbedarf
Optimale Integration	Angabe über die Sicherheit der Klassifikation
	Klassenbezogene Bewertung über die Sicherheit
	Berücksichtigung von Eingangsunsicherheiten

Tab. 2.10: Beispielhafte Kriterien zur Wahl eines geeigneten Klassifikators

2.7 Stand der Forschung

Zurzeit erscheinen am Markt erste Produkte und Prototypen zu Produktideen bezüglich der kommerziellen Anwendung automatischer Emotionserkennung. Als erstes Gerät ist der sogenannte *Handytruster*, nach den Arbeiten von Amir Libermann, zu benennen. Hierbei handelt es sich um einen Lügendetektor, der auf Grund akustischer Stimmschwankungen Lügen und Stresszustände

eines telefonischen Gesprächspartners erkennen soll[27] [NEM04]. Aus derselben Quelle stammt als Nachfolgeprodukt der sogenannte *LoveDetector* auf reiner Softwarebasis. Mit ihm soll es möglich sein, die Zuneigung eines Gesprächspartners zu ermitteln. Auch über die Stimme werden beim Produkt *WhyCry*, nach der Forschung von Pedro Monagas entwickelt, Babyschreien Emotionen und Bedürfnisse zugeordnet. Dieses sind *Hunger, Langweile, Müdigkeit, Stress* und *Unzufriedenheit*. Als reiner Prototyp wurde von der Firma Toyota® im Jahr 2004 auf der Motorshow in Tokio ein Konzept-Fahrzeug namens *POD* vorgestellt, das die Emotion des Fahrers nach außen hin sichtbar machen soll. Abb. 2.11 zeigt die genannten Produktbeispiele.

Abb. 2.11: Erste Produkte zur Emotionserkennung, von links:
HandyTruster, LoveDetector, WhyCry, POD

Im Folgenden wird detaillierter auf den internationalen Stand der Technik bezüglich der Emotionserkennung aus den hier gewählten Nutzerinformationen eingegangen. Vorweg sei erwähnt, dass sich weder bezüglich der erkannten Gefühle, der gewählten Verfahren, oder der verwendeten Datenbanken ein eindeutiger Trend beobachten lässt.

2.7.1 Akustische Analyse

Am intensivsten neben bildbasierten Ansätzen wird in der internationalen Forschung bis heute vor Allem die Erkennung von Emotion aus dem akustischen Sprachsignal betrachtet. Hier sind auch die robustesten Ergebnisse vorzuweisen. Eine Hauptrichtung zur Vorgehensweise wie etwa bei der Spracherkennung, in der sich Merkmale wie MFCC oder Klassifikationsverfahren wie HMM (siehe Kap. 7.1) besonders durchgesetzt haben, ist noch nicht zu beobachten [KÜS04]. Lediglich bei der Wahl der Merkmale haben sich folgende feste Ansatzpunkte herauskristallisiert: Verlauf der Sprachgrundfrequenz, Verlauf der Lautheit, zeitliche Verhältnisse, spektrale Eigenschaften und Formanten höherer Ordnung. Was die zeitliche Modellierung dieser Information anbelangt, verfolgt die deutliche Mehrheit der Arbeiten die Betrachtung abgeleiteter Merkmale gegenüber einer direkten Klassifikation aus Konturverläufen.

Bezüglich des gewählten Klassifikationsansatzes der akustischen Merkmale kommt eine Vielzahl von Verfahren zum Einsatz. Trotz mehrerer Vergleiche lässt sich auch hier keine generelle Aussage über ein besonders geeignetes treffen. In Tab. 2.12 findet sich eine Auswahl an bisher eingesetzten Verfahren mit Angabe von Quellen, in denen diese, wie auch in dieser Arbeit, analysiert werden.

[27] Die Produktion wurde nach drei Produktzyklen mittlerweile eingestellt.

Die Verfahren selbst werden in Kap. 3.7 und 3.11 näher beschrieben.

Des Weiteren werden auch Ensembles von Klassifikatoren (siehe Kap. 3.8) in der Emotionserkennung eingesetzt. So wird Boosting in [PET99] und [LIT03B] neben Studien im Rahmen dieser Arbeit (vgl. [SCH05A], [SCH05C], [SCH05E] und [SCH05F]) erfolgreich eingesetzt. In [KAP04] wird ebenfalls Boosting angewandt, wobei dort nur visuell basierte Erkennung verfolgt wird.

Verfahren zur Klassifikation	**Beispiele der Anwendung**
Bayessche Netze	[BAL99]
Dynamic-Time-Warping	[SCH02D]
Entscheidungsbäume	[LIT03B], [YAC03]
Gaußsche-Mixtur-Modelle	[BRA00], [SCH03B]
Hidden-Markov-Modelle	[NOG01], [SCH02D], [KWO03], [SCH03A]
K-Nächste-Nachbarn-Verfahren	[DEL96], [PET99], [YUF01], [LEE02B], [LIT03B], [SCH03E], [YAC03]
Künstliche Neuronale Netze	[PET99], [MCI00], [YUF01], [TAT02], [ADE03], [YAC03]
Lineare Diskriminanz Klassifikatoren	[BAT00A], [MCI00], [LEE02B], [TOI02], [KWO03]
Naive-Bayes	[DEL96], [SCH05F]
Support-Vektor-Maschinen	[MCI00], [YUF01], [LEE02B], [KWO03], [YAC03], [SCH04A]

Tab. 2.12: Auswahl in der Literatur eingesetzter Klassifikationsverfahren zur akustischen Emotionserkennung

Ebenso werden heute diverse Verfahren zur Reduktion von Merkmalen (siehe Kap. 3.9) eingesetzt. So wird etwa eine sequentielle Vorwärtssuche mit k-Nächste-Nachbarn als Klassifikator in [DEL96] verwendet. Besonders beliebt ist die Reduktion mittels Hauptachsentransformation [LEE02B], [CHU04], [VER04A].

Hinsichtlich der Klassifikationsdauer wird auf einem konventionellen Rechner der heutigen Generation[28] bereits Echtzeitfähigkeit berichtet [LIT03B].

2.7.2 Linguistische Analyse

Die Betrachtung des gesprochenen Inhalts einer Äußerung ist seit Beginn der automatischen Sprachemotionserkennung mitverfolgt worden [ELL92], [MOR97], [BAT00A] und erfreut sich zurzeit steigender Beliebtheit [SCH02B], [SCH02C], [HOL03], [CHU04]. Die wohl einfachste Methode aus Text Emotion zu erkennen ist die Schlüsselwortsuche. Hierbei wird nach bestimmten Ausdrücken in einem Text gesucht, ohne Zusammenhänge zwischen Wörtern zu erfassen. Auf Grund seiner Einfachheit erfreut sich dieser Ansatz großer Verbreitung. Es ist jedoch nicht möglich, beispielsweise Verneinungen, wie in *„Ich bin überhaupt nicht glücklich.“,* zu handhaben. Darüber

[28] Hier ein Rechner vom Typ Intel Pentium® mit 4, 3 GHz Prozessortakt und 512 MB Hauptspeicher.

hinaus wird Emotion oft nicht direkt durch affektive Wörter ausgedrückt, sondern geht aus dem Verständnis des Inhalts hervor. Hierzu seien die Beispielsätze *„Ich kann nicht mal eine Sekunde Pause machen.“* und *„Ich hab' Pause!“* gegeben. Sehr beliebt sind daher auch Ansätze der statistischen Modellierung. Hier erfolgt eine Betrachtung lexikalischer Affinität einzelner Wörter zu bestimmten Emotionen. So wird dem Wort *Pause* etwa im angeführten Beispiel mutmaßlich eine Wahrscheinlichkeit für die Zustände Stress und Freude zugewiesen. Weiterhin wird versucht, das genante Problem eines mangelnden Kontextbezugs durch Sicht über *N* benachbarte Wörter zu lösen. Daneben wird, wie in Tab. 2.13 ersichtlich, vereinzelt eine Reihe von weiteren Verfahren, auf die hier nicht im Einzelnen eingegangen werden kann, angewandt. Dabei eignen sich einige, wie die latente semantische Analyse, ausschließlich auf längeren Textpassagen, was hier nicht Gegenstand der Betrachtung ist.

Verfahren	Beispiele der Anwendung
Unigramme	[LEE02A], [DEV03]
Bigramme	[POL00]
Trigramme	[ANG02]
Bayessche Netze	[BRS98], [SCH03E], [SCH04A]
Latente Semantische Analyse	[GOE00]
Regelbasierte Modellierung	[LIT03A]
Schlüsselwortsuche	[ELL92], [COW99]
Semantische Bäume	[ZHE02]
Transformationsbasiertes Lernen	[WUT02]
Vektorraumrepräsentation	[SCH05F]
Weltwissensmodell	[LIU03]

Tab. 2.13: Auswahl in der Literatur eingesetzter Verfahren zur linguistischen Emotionserkennung

Eine allgemeine Vorgehensweise hat sich also auch hier noch nicht durchgesetzt. Gemein ist jedoch das Problem der Erfassung des Inhalts, besonders bei gesprochener Sprache. Aber auch bei geschriebenem Text muss in der Regel erst eine Vorverarbeitung stattfinden, weil beispielsweise Text von einem Dokument oder einer handschriftlichen Eingabe ausgehend digitalisiert werden muss, oder um Schreibvarianten, Rechtschreibfehler und Interpunktionsprobleme in den Griff zu bekommen.

Die Fusion der Information aus Akustik und Linguistik geschieht bisher oft relativ rudimentär durch einen gewichteten Mittelwert [LEE04] sowie aufwändiger durch ein Neuronales [MÜL04], [SCH04A], [RIG05] oder Bayessches Netz [BRS98]. Generell lassen sich diese Ansätze um die Integration weiterer Information wie Bilddaten erweitern.

2.7.3 Verarbeitung manueller Interaktionsdaten

Besonders im Bereich des Internets wird Mausaktivität als Grundlage zur Erkennung von Interesse betrachtet. Hier soll für kommerzielle Dienste der Fokus des Anwenders erkannt werden. In [MUE01] wird von einer Studie mit 18 Nutzern berichtet, deren Internetsurfverhalten aufgezeichnet wurde. Die Analyse erfolgt dort manuell und nicht automatisiert. Über einer Webseite wird der

Verlauf der Mausbewegungen betrachtet, um bestimmte Interesse anzeigende Muster zu erkennen. Als Merkmale werden Hesitationen auf Hyperlinks, oder gezielte Bewegungen ohne Zögern genannt. Darüber hinaus wird festgestellt, dass Anwender gerne den Mauszeiger in leeren Räumen ruhen lassen, um nicht versehentlich auf einen falschen Link zu klicken. 70% Erkennungsrate werden im Mittel angegeben, um Objekte von Interesse zu identifizieren. In [GOC00] wird ebenfalls Mausaktivität und Scrollradnutzung als Grundlage zur Detektion von Interesse angewandt. Es wird eine Datenbank von insgesamt 200 Beispielen genannt, die durch Aufzeichnung von Surfverhalten generiert wurde. Anhand dieser wird nachvollzogen, wie sich Interesse im Vorfeld besuchter Seiten an dem Bewegungsmuster in der Nähe ihrer zugehörigen Hyperlinks ankündigt. Mit Hilfe eines Neuronalen Netzes erfolgt eine automatische Klassifikation auf Basis der drei Merkmale *gewählte Hyperlinks*, *Scrollaktivität* und *generelle Mausaktivität*.

Über eine reine Detektion von Interesse hinaus ist eine Erkennung von Emotion aus Daten der manuellen Interaktion mit einer grafischen Benutzeroberfläche bislang nicht ausschließlich aus der somatischen Information erfolgt. Es existieren jedoch Arbeiten, in denen die Computermaus als natürlicher Platz angesehen wird, um affektive Größen aufzufassen. Als Grund hierfür wird unter Anderem angegeben, dass Rund ein Drittel der Zeit, die vor einem Computerarbeitsplatz verbracht wird, auf den Gebrauch des Eingabemediums entfällt [ARK99]. In [ARK99] und [PIC00] wird zusätzlich Spezialhardware verwendet, um physiologische Daten zu messen. Die Aufzeichnung der Merkmale *Herzfrequenz*, *Temperatur*, *GSR* und *generelle somatische Aktivität (GSA)* erfolgt dabei, wenn der Anwender auf ein bestimmtes Icon zeigt. Die sechs Emotionen des MPEG-4 Standards werden mittels Diskriminanzfunktion erkannt. Es werden Daten von sechs Probanden, drei davon weiblich, verwendet. Diese mussten jeweils für fünf Minuten mittels Gesichtsausdruck eine Emotion spielen, während sie mit Sensoren und Maus vor dem Computer saßen. Als Erkennungsrate werden 66% angegeben.

In Studien im Rahmen der hier vorgestellten Forschungsaktivität (vgl. [SCH02C], [SCH04C]) wird hingegen ausschließlich eine standardmäßige Computermaus oder ein Touchscreen eingesetzt. Aus den Aufzeichnungen der Bewegungen werden Modelle zur Erkennung von vier Emotionen gebildet.

3

Akustische Emotionserkennung

„Das Verständliche an der Sprache ist nicht das Wort selber, sondern Ton, Stärke, Modulation, Tempo, mit denen eine Reihe von Worten gesprochen wird – kurz die Musik hinter den Worten, die Leidenschaft hinter dieser Musik, die Person hinter dieser Leidenschaft: alles das also, was nicht geschrieben werden kann.“

FRIEDRICH NIETZSCHE (1844-1900)

Ziel dieses Kapitels ist es, erarbeitete Verfahren zur automatischen Schätzung der Emotion des aktuellen Benutzers aus akustischen Merkmalen gesprochener Eingaben zu beschreiben und zu evaluieren. Bei der Betrachtung dieser Attribute ist eine Abgrenzung zwischen in physikalischen Maßeinheiten beschreibbaren Größen und solchen der perzeptiven Wahrnehmung zu treffen. Im Rahmen des Kapitels konzentrieren sich alle Beschreibungsebenen auf das akustische Signal unabhängig vom gesprochenen Text. Durch linguistische Ebenen bestimmte Parameter wie die Betonung sind daher differenziert zu betrachten. Grundlage des Verständnisses gewählter Charakteristika und Modellierungsformen ist zunächst eine Beschreibung der menschlichen Sprachproduktion und Perzeption. Ihr folgen Ausführungen zur Vorverarbeitung des Signals und Berechnung von relevanten Verläufen aus der Elongation über der Zeit. Im Anschluss folgt eine Extraktion von statischen Merkmalen auf höherer Ebene unter statistischen Betrachtungen berechneter Kurvenverläufe. Die so gewonnenen akustischen Merkmale lassen sich in die Kategorien *prosodischer*, *stimmqualitativer* und *artikulatorischer* Charakteristika untergliedern. Den Anschluss an die Beschreibung informationstragender Größen bildet die Vorstellung ausgewählter Verfahren zur automatischen Klassifikation. Diese sind dabei so ausgewählt, dass ein möglichst breites Spektrum unterschiedlicher Ansätze vertreten ist. Es werden darüber hinaus Varianten zur Optimierung und Integration diverser Stärken aufgezeigt. Die automatische Generierung, Selektion und Reduktion von Merkmalen ist ein weiterer wichtiger Bestandteil des Kapitels mit dem Ziel optimaler Erkennungsgüte bei gleichzeitig höchster Effizienz. Über die Zuordnung von statischen Mustern hinaus wird eine Variante zur direkten Schätzung aus dynamischen Verläufen aufgezeigt. Schließlich werden Experimente zur Beurteilung der Güte beschrieben und zugehörige Datenbanken emotionalen Sprachmaterials vorgestellt.

3.1 Menschliche Sprachproduktion und Perzeption

Für die Diskussion betrachteter akustischer Größen soll als Grundlage zunächst die humane Kommunikationskette vom Sender über die Transmission des Sprachsignals zum Empfänger vorgestellt werden. Als erstes erfolgt hierzu eine kurze Einführung in die natürliche Erzeugung von Sprache. Abb. 3.1 bietet hierzu einen Überblick über den menschlichen Vokaltrakt. Der supraglottal, das heißt oberhalb der Stimmbänder liegende Verbund aus Rachen-, Nasen- und Mundhöhle wird als Ansatzrohr bezeichnet. Zu ihm gehören auch die Sprechwerkzeuge Lippen, Zunge, Zähne, Gaumen mit Gaumensegel und Zäpfchen. Weiterhin entscheidend sind der Kehlkopf, oder *Larynx*, sowie die Organe der Atmung, beziehungsweise *Respiration*.

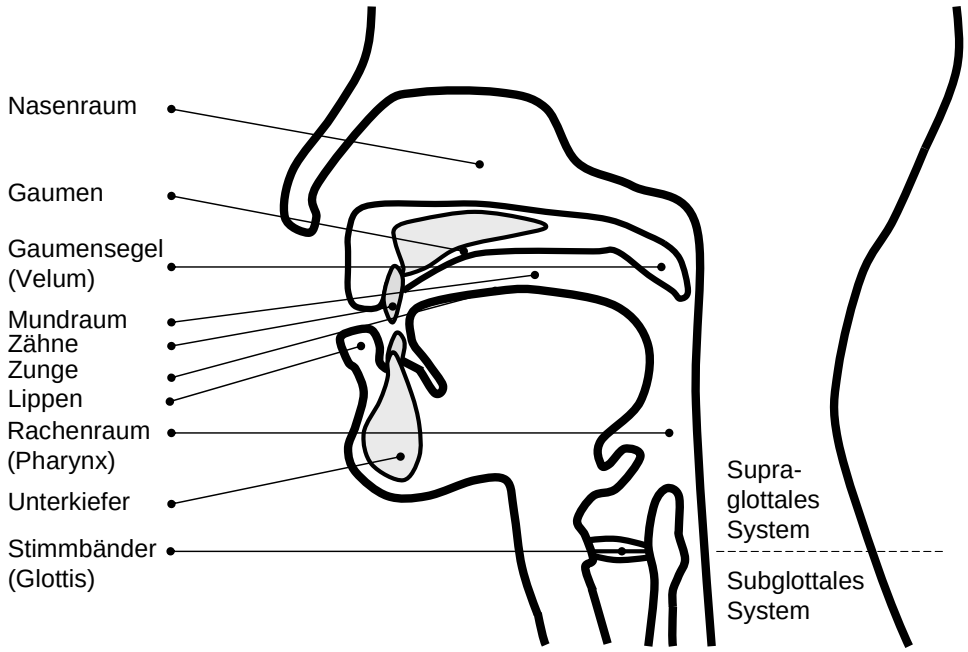


Abb. 3.1: Vokaltrakt mit Sprechwerkzeugen als Schema im Schnitt

Ausgangspunkt der Lauterzeugung ist die Generierung eines als *Phonationsstrom* bezeichneten Luftstroms. Dieser kann *ingressiv*, das heißt in den Körper hinein, oder *egressiv*, also entsprechend aus dem Körper heraus gerichtet sein. Die Kontrolle erfolgt dabei über den Kehlkopf mit dem schützenden Kehldeckel. Hier entscheidet sich, ob ein Expirationsstrom zum Phonationsstrom wird. Erzeugt wird dieser entweder durch die Organe der Atmung mittels Bewegungen von Zwerchfell- und Brustkorb, durch bei geschlossener Glottis erfolgendes abruptes Heben und Senken des Kehlkopfes, oder durch den Verschluss und einer Druckgenerierung folgende plötzliche Öffnung im Bereich des Ansatzrohrs. Man spricht entsprechend von *pulmonalem*, *glottalem* oder *oralem* Erzeugungsmechanismus, wobei am häufigsten ein egressiv pulmonaler zur Anwendung kommt.

Im Weiteren kann zwischen der Generierung stimmhafter, stimmloser und transienter Laute unterschieden werden. Entscheidend hierfür ist die Periodizität des Anregungssignals. Periodische Anregung wird dabei durch Modulation des subglottalen Luftstroms mit Hilfe von Vibration der

Stimmlippen erreicht. Im Gegensatz hierzu wird eine aperiodische Anregung durch Friktion an einer Engstelle im Ansatzrohr erzielt. Transiente Signale, etwa zur Erzeugung von Plosivlauten benötigt, ergeben sich aus der plötzlichen Öffnung eines geschlossenen Bereichs. Für die Emotionserkennung von besonderem Interesse ist im Speziellen die Form der Stimmlippenanregung, *Phonationsart* genannt [KLA00A]. Sie besitzt in der Deutschen Sprache keine linguistische Funktion. Sechs Basisarten lassen sich nach Laver [LAV80] unterscheiden, die teilweise kombiniert werden können:

- **Modale Anregung**, bei der durchschnittliche Tensionsverhältnisse beteiligter Muskeln gegeben sind. Dies entspricht einer normalen Sprechweise.

- **Falsett**, welches auf einer dünnen, steifen und leicht geöffneten Stellung der Stimmlippen beruht. Der Druck im subglottalen Bereich ist hierdurch im Vergleich zur modalen Anregung geringer. Charakteristisch sind eine hohe Grundfrequenz (vgl. Kap 3.4.3) mit einer hieraus resultierend geringen Anzahl von Obertönen, starke Anteile von Rauschen sowie ein dünner Klang.

- **Flüstern**, das durch starke Rauschanteile im Signal geprägt ist. Dieser Effekt wird durch eine Kombination mit modaler Anregung oder Falsettstimme verstärkt.

- **Hauchen**, durch eine entspannte Muskulatur des Kehlkopfs erzielt.

- **Knarren**, mittels geringem, aber stark variierendem subglottalen Druck im Vergleich zur modalen Phonation und unregelmäßig niederfrequent schwingenden Stimmlippen erreicht. Hieraus ergibt sich unter Anderem eine starke Streuung der Grundfrequenz.

- **Rauhe Anregung**, welche nur ein Modifikator anderer Phonationsarten ist. Der Stimmapparat weist dabei einen hohen Spannungszustand auf. Akustisch ähnelt sie einer Knarrstimme bei gleichzeitig höherer Grundfrequenz.

Hinzu kommt die sogenannte *Nullphonation*, bei der die Luft ungehindert durch die Stimmlippen fließt. Des Weiteren kann zwischen ge- und entspannter Stimme unterschieden werden, was sich auch direkt auf den Muskelzustand bezieht [BUR00]. Die gespannte Stimme ist dabei eher laut, die Grundfrequenz hoch, und die spektrale Energie zwischen 1 und 4 kHz höher. Durch die Spannung ergibt sich eine Tendenz zu einer rauen Stimme. Es findet eine Artikulation leicht über dem Zielbereich statt (vgl. Kap. 3.5.1), und die Sprachsegmente weisen eine durchschnittlich höhere Dauer auf. Bei der entspannten Stimme hingegen tritt durch die lockeren Muskelzustände am ehesten behauchte modale Phonation auf. Im supraglottalen Bereich führt die Entspannung unter Umständen zu einer nasalen Sprechweise. Die weiteren Parameter verhalten sich umgekehrt der angespannten Sprechweise. Die Grundfrequenz und Energie sind somit niedriger und die Formantenlagen (vgl. Kap. 3.5.1) zentralisierter.

Im Anschluss an die Anregung erfolgt eine frequenzabhängige Filterung im Artikulationstrakt zur eigentlichen Lautbildung. Maßgebend sind dabei die sogenannten *Artikulatoren*. Dies sind die

relativ beweglichen Teile des Ansatzrohrs: Unterlippe sowie Zungenkranz, -rücken, und –wurzel, beziehungsweise *Korona*, *Dorsum* und *Radix*[29]. Aus ihrer Stellung ergeben sich unterschiedliche Resonanzeigenschaften des Artikulationstrakts. Zur Beschreibung des Verhaltens des Vokaltrakts existiert eine Reihe von akustischen und phonetischen Modellen, unter denen mit das populärste auf Grund seiner einfachen Darstellung der Viertel-Wellenlängen-Resonator ist. Auf eine detaillierte Beschreibung soll hier jedoch verzichtet werden.

Folgendes Blockschaltbild veranschaulicht zusammenfassend den Ablauf der Sprachgenerierung [FEL84]:

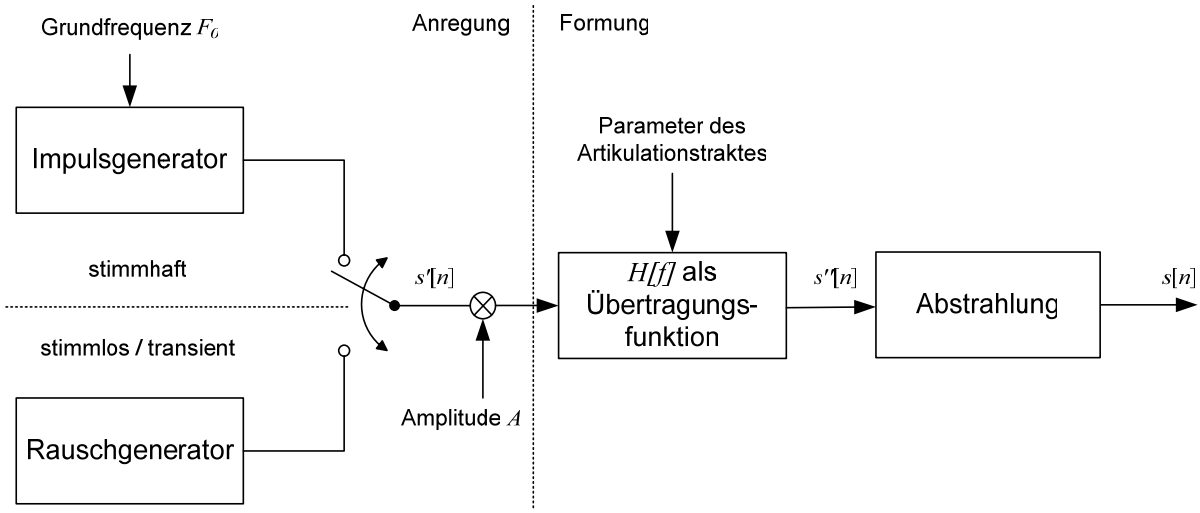


Abb. 3.2: Lineares Modell zur menschlichen Spracherzeugung

Im Bereich der Anregung kann das dargestellte lineare Modell jeweils noch um den Korrekturblock Klangfarbe in Folge des Impulsgenerators, und einen weiteren Korrekturblock nach dem Rauschgenerator erweitert werden. Der letzte Block repräsentiert die Abstrahlung über Nase und Mund, die für die letztliche Prägung des Signals ausschlaggebend ist.

Neben der menschlichen Produktion von Sprachlauten spielt für die Betrachtungen in Kap. 3 vor Allem auch die Physiologie des Gehörs eine wichtige Rolle, um sich bei der angestrebten automatischen Analyse an natürlichen Prozessen orientieren zu können. Abb. 3.3 gibt einen schematischen Überblick über den Aufbau des Ohrs [ZWI90]. Die Verarbeitung des akustischen Sprachsignals erfolgt dabei beim Mensch durch eine multiple Wandlung in den Bereichen Außen-, Mittel-, und Innenohr. Dabei erstreckt sich das Außenohr von der Ohrmuschel mit Gehörgang bis zum Trommelfell und das Mittelohr über die drei Gehörknöchelchen Hammer, Amboss und Steigbügel bis zum ovalen Fenster. Das Innenohr schließlich setzt sich aus der *Schnecke* genannten *Cochlea* und dem *Cortischen Organ* zusammen, welches für die Weiterleitung der Empfindung an den Hörnerv verantwortlich ist. Die Form des Gehörgangs ist ausschlaggebend für die Anhebung bestimmter Frequenzen. Die Übertragung von Luftschwankungen findet ausgehend vom

[29] Es gibt eine Reihe unterschiedlicher Kategorisierungen bezüglich der Artikulatoren. Unter Anderem werden auch die Abschnitte des Gaumens hinzugezählt.

Trommelfell über die Gehörknöchel auf das ovale Fenster statt, um diese dann in einem flüssigen Medium in Druckwellen zu wandeln. Letztere führen zu einer frequenzabhängigen Reizung von insgesamt 3.600 inneren Haarzellen im Abstand von 9 µm in der Cochlea [ZWI90]. Wahrnehmbare Frequenzen erstrecken sich dabei - personen- und altersbedingt - von 50 Hz bis ca. 16 kHz - 20 kHz. Aus der Unterteilung in 640 Stufen ergibt sich die Auflösung eines Reizes bis zu einer Genauigkeit von sechs Haarzellen. Diese Zellen senden nach ihrer Erregung elektrische Impulse, die letztlich für die weitere Verarbeitung ausschlaggebend sind.

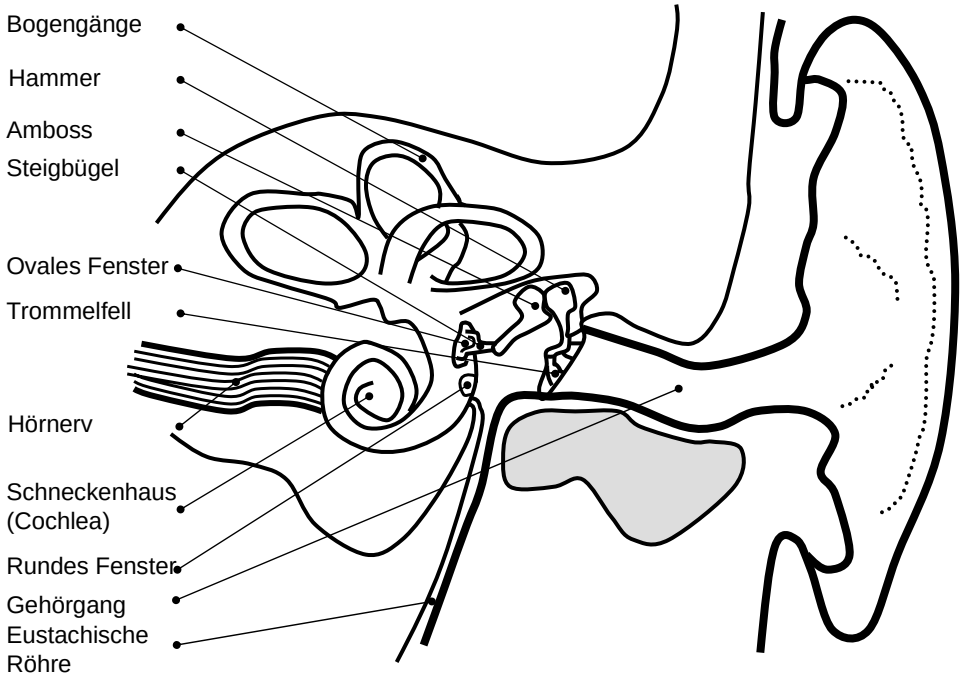


Abb. 3.3: Gehör als Schema im Schnitt

3.2 Vorverarbeitung

Wie in Kap. 2.6.1 beschrieben, ist der Startpunkt einer automatischen Klassifikation zunächst die Erfassung und Vorverarbeitung des zu analysierenden Sprachsignals $s(t)$. Speziell für die Aufzeichnung von Sprache wird hier im Echtzeitbetrieb als elektrischer Wandler für akustische Sprachsignale ein aktives Kondensatormikrophon der Firma AKG mit der Typenbezeichnung C1000-S MK II eingesetzt. Die Signalabtastung erfolgt mit einer Abtastfrequenz f_s von 16 kHz und die anschließende A/D-Wandlung mit einer Auflösung von 16 Bit[30]. Diese Zahlen sind für die Verarbeitung von Sprache typisch. Nach dieser Digitalisierung liegen somit zu diskreten Zeitpunkten n die Abtastwerte $s[n]$ vor. Der jeweils erste Wert einer Aufzeichnung sei bei $n = 0$ gewählt.

[30] In einzelnen Fällen wie bei der Anwendung vorab aufgezeichneter Datenbeispiele sind in Kap. 3.12.4 andere Abtastraten gewählt. Es erfolgt dort ein entsprechender Vermerk.

3.2.1 Segmentierung

Das Signal von Interesse muss zeitlich begrenzt werden. Ein bezüglich maximaler Länge limitierender Faktor ist dabei der Wunsch Abschnitte zu erhalten, die von ausschließlich einem Gefühlszustand geprägt sind. Zunächst ist somit eine Segmentierung in kleine Einheiten wie Einzelwörter oder Phoneme[31] als gesamter Betrachtungsrahmen bezüglich einer finalen Entscheidung über die aktuelle Emotion möglich. Jedoch wird eine robuste Schätzung durch eine zeitlich längere Beobachtung, welche supersegmentale Entwicklungen erfasst, positiv begünstigt. Es erscheint denkbar, dieses Verhalten mit der im Folgenden erläuterten Quasistationarität von Sprache auf höherer Ebene bezüglich dem zu Grunde liegenden Affekt zu vergleichen. In Summe ergibt sich somit, als Kompromiss, eine Zerlegung in sogenannte *Äußerungen*. Im Gegensatz zur später betrachteten geschriebenen Sprache, in der ein Satz anhand gegebener Interpunktion abgrenzbar ist, erfolgt die Limitierung hier in lautsprachlichen Einzeläußerungen innerhalb von Dialogakten[32]. Da es sich im Falle emotionaler Äußerungen nicht zwangsläufig um einen Dialog im engeren Sinne handeln muss, erfolgt die Untergliederung in erster Linie basierend auf ausreichend langen Sprechpausen. Die im Falle eines Echtzeitbetriebs erforderliche automatische Segmentierung wird in Kap. 4.3.2 beschrieben, da sie das Verständnis der im Folgenden beschriebenen Merkmale voraussetzt. Die Gefahr eines instantanen Gemütswechsels innerhalb der festgelegten Zeitspanne, ohne größere Pausen oder Hesitationen im Redefluss, erscheint vernachlässigbar. Es hat sich anderseits gezeigt, dass eine Betrachtung über mehrere Äußerungen der Quasistationarität bezüglich der Emotion nicht gerecht wird, da sich der emotionale Zustand bereits geändert haben kann.

3.2.2 Fensterung und Konturextraktion

Wie unter Kap. 2.6.1 beschrieben, ist eine Transformation des erfassten Signals in eine geeignetere parametrische Beschreibung in der Regel unerlässlich. Die dynamische Natur eines Sprachsignals bedingt als Ausgangsbasis der Betrachtung zunächst die Bildung von Wertereihen $x = (x_i)$ mit $i = 0,...,I-1$. Konkreter ergeben sich zwei- oder mehrdimensionale Entwicklungen über der Zeit. Man spricht daher auch von *Zeitreihen*, bzw. *Time Series*. Voraussetzung zur Beschreibung als Zeitreihe ist der hier durch die Abtastung realisierte Umstand einer Datenbetrachtung in diskreten, endlichen Zeitintervallen. In der Regel werden die Intervallgrenzen in Form äquidistanter Punkte gewählt. Allgemein kann dabei jedem solchen diskreten Zeitpunkt t[33] eine Zahl x_t oder ein Tupel $\underline{x_t}$ von Zahlenwerten zugeordnet werden. Man spricht entsprechend von *skalaren* beziehungsweise *vektoriellen* Werten und *uni-* beziehungsweise *multivariaten* Zeitreihen. Diese können als Funktionen im $\mathbb{R}^n$ dargestellt werden. Hier wird die Zeit über der Abszisse und die zugehörige Elongation, oder weitere abgeleitete charakteristische Signalverläufe wie Grundfrequenz oder

---

[31] Kleinste bedeutungsunterscheidende Einheit der Sprache.

[32] Auch als (Dialog-)*Turn* bezeichnet.

[33] Der Zeitindex wird hier zur im Folgenden anschaulicheren Notation als t für Tempus, lateinisch für Zeit, gewählt. Es sei jedoch wegen der Gefahr einer Irritation im Vergleich zur üblichen Bezeichnung bei kontinuierlicher Darstellung explizit auf die hier gewählte diskrete Natur hingewiesen.

Energie, zu jedem Zeitpunkt auf der Ordinate aufgetragen. Die Ableitung weiterer geeigneter Konturverläufe ist auf Grund der in Kap. 2.6.1 geforderten günstigeren Darstellung zur Repräsentation immanenter Emotion erforderlich. Idealerweise wird dabei der Anteil im Signal enthaltener, aus Sicht der Emotionserkennung redundanter Information, wie der gesprochene Inhalt, stimmliche Charakteristika des Sprechers, oder Umgebungsgeräusch, reduziert oder eliminiert.

Um nun die einzelnen Werte verschiedener Reihen zu berechnen, wird unter Berücksichtigung des quasistationären Verhaltens von Sprachsignalen das Signal $s[n]$ einer *Fensterung* unterzogen, wodurch es in sogenannte *Signalrahmen* s_t mit Zeitindex $t = 0,...,T-1$ aufgeteilt wird. Die gewählte Länge eines Fensters muss dabei eine Mindestdauer aufweisen, um eine ausreichende Werteauflösung transformierter Verläufe wie der Frequenz bei einem Leistungsspektrum zu erlauben. Eine zu große zeitliche Dauer hingegen würde den Grundsatz der Dynamik des Signals verletzen und zu starke Veränderungen innerhalb eines Rahmens erlauben. Somit würde sich beispielsweise die Grundfrequenz innerhalb eines Rahmens stark ändern können. Je Rahmen wird aber nur ein Wert abgeleiteter Signalkonturen bestimmt. Um Verzerrungen an den Fensterrändern zu vermeiden, werden allgemein üblich weiche Ausklänge mittels einer geeigneten Fensterfunktion $w[n]$ anstatt den steilen Flanken eines harten Ausschneidens entsprechend einer Fensterung mit einer Rechteckfunktion[34] gewählt. Um diese sogenannte *Fensterung* auszuführen, werden die Werte an der Position des aktuellen Rahmens mit der Fensterfunktion gewichtet. Eine Glättung des zeitlichen Verlaufs wird durch eine Überlappung von 50% der Rahmen gewährleistet. Für den gebildeten Signalverlauf $s_t[n]$ innerhalb eines Rahmens t ergibt sich somit:

$$s_t[n] = s\left[n + t \cdot \frac{N}{2}\right] \cdot w[n] \text{ und } N = T_w \cdot f_s \tag{3.1}$$

Hier wird die in folgender Gleichung gegebene Hanning-Funktion[35] $w_{Han}[n]$ mit einer Fensterbreite T_w=16 ms, was bei der gewählten Samplingfrequenz von f_s=16 kHz einer Zahl von $N = 256$ Werten je Rahmen entspricht, gewählt:

$$w_{Han}[n] = \begin{cases} a - a \cdot \cos\left(\frac{2 \cdot \pi \cdot n}{N}\right) & \text{für } 0 \le n \le N-1 \\ 0 & \text{sonst} \end{cases} \text{ und } a = \frac{1}{2} \tag{3.2}$$

Außerhalb des dargestellten Bereichs ist der Wert der Hanning-Funktion gleich Null. Sie wird hier für Betrachtungen im Zeitbereich bewusst angewandt, da sie die schmälste Fensterung aufweist, und somit am wenigsten gegenüber schnellen Signaländerungen empfindlich ist. Außerdem beträgt ihr Wert an der Fenstermitte exakt Eins und an den Rändern exakt Null. In [BOE93] wird gezeigt, dass sie im Vergleich zu den dort genannten Funktionen die exaktesten Werte für diverse hier betrachtete Größen ergibt. Für spektrale Transformationen wird die üblicher Weise verwendete

[34] Der Multiplikation mit einer Rechteckfunktion im Zeitbereich entspricht beispielsweise im spektralen Bereich eine Faltung mit einer Spaltfunktion. Diese weist eine hohe Welligkeit und unendliche Ausdehnung im Spektrum auf.

[35] Auch *von-Hann-Funktion* nach ihrem Namensvater *Julius von Hann*, einem österreichischen Meteorologen, benannt.

Hamming-Funktion w_{Ham} gewählt. Sie entspricht der Hanning-Funktion, nur mit entsprechend geändertem Wert $a = \frac{25}{46}$. Abb. 3.4 zeigt den Verlauf der Hanning-Funktion im Vergleich zur ähnlichen Hamming-Funktion. Es existiert daneben eine Reihe weiterer gebräuchlicher Fensterfunktionen wie *Bartlett*-, *Blackman*-, *Kaiser*- und *Welch*-Funktion, auf die hier nicht näher eingegangen werden soll.

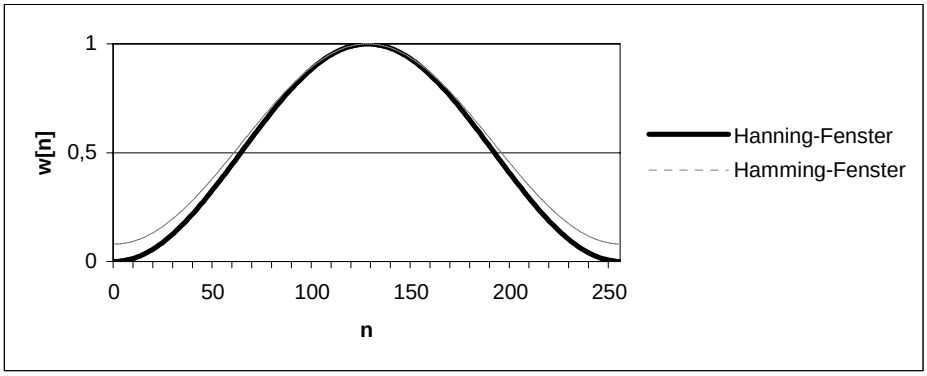


Abb. 3.4: Fensterfunktionen für die gewählte Breite von 256 Punkten

Aus der gewählten Abtastung und Fensterung ergeben sich äquidistante Zeitpunkte t, zu denen später Werte x_t abgeleiteter Zeitreihen $x = (x_t)$ mit $t = 0,...,T-1$ zugeordnet werden können. Es handelt sich dabei speziell um die leicht zu extrahierenden Verläufe der Nulldurchgänge und der Signalenergie sowie der Grundfrequenz, höherer Formanten, der harmonischen Ausgeprägtheit, des spektralen Roll-Off-Punkts, Flusses, Schwerpunkts sowie Verlaufs. Die Bestimmung dieser wird in den folgenden Abschnitten im Einzelnen behandelt. Eine signalangepasste, adaptive Distanz dieser Punkte ist denkbar, und unter Umständen vorteilhaft. Doch auf Grund des damit einhergehenden Synchronisierungsproblems unterschiedlicher Zeitreihen wird hierauf verzichtet. Es sei an dieser Stelle angemerkt, dass die angeführten Verläufe teilweise im aktuellen MPEG-7 Audiostandard als sogenannte *Low-Level-Deskriptoren (LLD)*[36] enthalten sind. Eine Emotionserkennung kann somit im Erfolgsfall der nachfolgend vorgestellten Betrachtungen unter Einschränkungen auch auf Basis von LLD nach dem MPEG-7 Standard durchgeführt werden.

Zusätzlich zu den beschriebenen Verläufen kann durch eine simple, jedoch gleichermaßen auch oft effektive Integration von diskreten Ableitungsfunktionen weitere Information über den zeitlichen Verlauf gebildet werden. Für die Deltawerte erster Ordnung entsprechend der Geschwindigkeit wird folgende Differenzbildung ab dem zweiten Wert der Reihe verwendet:

$$x'_t = x_t - x_{t-1} \tag{3.3}$$

Entsprechend ergibt sich für die jeweiligen Regressionskoeffizienten zweiter Ordnung im Sinne

[36] Im MPEG-7 Standard sind insgesamt 17 LLD definiert. Sie sollen eine Charakterisierung des Signaltyps erlauben. Im Einzelnen sind diese: Wellenform, Leistung, Grundfrequenz, harmonische Ausgeprägtheit, logarithmische Anstiegszeit sowie in Bezug auf das Audiospektrum die Einhüllende, die Basis, der Zentroid, die Ausweitung, die Flachheit, und die Projektion. Hinzu kommen für das harmonische Spektrum Schwerpunkt, Abweichung, Ausweitung und Variation. Schließlich sind allgemeiner spektraler und temporaler Schwerpunkt enthalten.

einer Beschleunigung ab dem dritten Signalwert:

$$x_t'' = x_t' - x_{t-1}' = x_t - 2 \cdot x_{t-1} + x_{t-2} \tag{3.4}$$

3.2.3 Konturglättung

Bei der Erfassung von Signalen und der Ableitung transformierter Verläufe sind verrauschte oder fehlerhafte Daten durch Quantisierung und Fensterung unumgänglich, wodurch sich eine Abweichung einzelner gemessener oder berechneter Werte von ihren realen Werten ergeben kann. Diesem Aspekt wird hier zunächst durch eine Glättung von gebildeten Konturverläufen Rechnung getragen. Allgemein wird darunter die Reduktion von irregulären Schwankungen einer Zeitreihe $x = (x_t)$ mit $t = 0,...,T-1$ verstanden. Es hat sich hier im Besonderen ein symmetrischer gleitender Mittelwertfilter, beziehungsweise *Simple-Moving-Average (SMA) Filter* bewährt. Im Gegensatz etwa zu Medianfiltern wird keine zu starke Kompensation von Ausreißern forciert, sondern eine gleichmäßige Glättung erzielt. Prinzipiell wird dabei ein gleitendes Fenster über den Signalverlauf verschoben, und über eine ungerade Anzahl B_f, *Filterbreite* genannt, von Werten des Eingangssignals x symmetrisch um jeden Wert x_t gemittelt, sodass sich der jeweilige gefilterte Wert $\bar{x}_t$ ergibt. Man spricht daher auch von einer lokalen Approximation.

$$\bar{x}_t = \frac{1}{B_f} \sum_{t=t+\frac{B_f-1}{2}}^{t-\frac{B_f-1}{2}} x_t \quad \text{mit } t = \frac{B_f-1}{2}, \cdots, N - \frac{B_f-1}{2} \tag{3.5}$$

Da für die ersten Werte nicht die volle Filterbreite an symmetrisch umliegenden Werten zur Verfügung steht, werden diese entsprechend der folgenden Vorschrift mit steigender Breite gefiltert:

$$\begin{aligned} \bar{x}_0 &= x_0 && B_f^* = 0 \\ \bar{x}_1 &= \tfrac{1}{3}(x_0 + x_1 + x_2) && B_f^* = 1 \\ &\vdots && \text{solange } B_f^* < \tfrac{B_f-1}{2} \end{aligned} \tag{3.6}$$

Die letzten Werte werden analog mit sinkender Filterbreite B_f^* behandelt. Wird jeder Wert x_t mit $t = 0,...,T-1$ durch den korrespondierenden gemittelten Wert $\bar{x}_t$ ersetzt, ergibt sich eine neue geglättete Wertereihe $\bar{x} = (\bar{x}_t)$. Die Übertragungsfunktion H_{SMA} im Frequenzbereich des SMA-Filters ergibt sich dabei zu [SMI99]:

$$H_{SMA}(f) = \frac{\sin(\pi \cdot f \cdot B_f)}{B_f \cdot \sin(\pi \cdot f)} \tag{3.7}$$

Diese weist ein deutliches Tiefpassverhalten mit hoher Welligkeit auf. Als Filterbreite B_f hat sich hier die geringe Breite von drei Werten bewährt, um eine Glättung ohne zu starke Verfälschung zu erreichen.

3.3 Statistische Kenngrößen

Um die Emotion eines segmentierten Abschnitts automatisch zuzuordnen, können Zeitreihen mit geeigneten Methoden, etwa der dynamischen Programmierung (vgl. Kap. 3.11), direkt klassifiziert werden. Nachteilig hierbei gestaltet sich jedoch eine starke Abhängigkeit vom gesprochenen Inhalt. Alternativ können aus der statistischen Datenanalyse bekannte Verfahren zur Ableitung von Funktionalen appliziert werden, um statische Größen auf einer makroskopischen zeitlichen Ebene zu konstruieren. Die hierzu erforderliche Abbildung eines Funktionenraums F wird als Funktional f bezeichnet:

$$f : F \to \mathbb{R} \tag{3.8}$$

Diese Funktionale können dann als Merkmale zur Klassifikation mit statischen Ansätzen verwendet werden. Die Tatsache, dass die Attribute innerhalb eines Merkmalsvektors $\underline{x}$ wie in Kap. 2.6.2 gezeigt keine Ordnung besitzen, ist für Funktionale dabei nicht relevant. Es bietet sich eine Reihe von statistischen Kenngrößen zur Analyse von Zeitreihen in Form von Funktionalen an, deren wichtigste Vertreter im Folgenden kurz vorgestellt werden. Dabei wird in der statistischen Datenanalyse zwischen modellabhängigen und –unabhängigen Methoden unterschieden. Nachfolgend wird sich auf die Anwendung von Verfahrensweisen der modellunabhängigen deskriptiven Statistik beschränkt.

3.3.1 Momente

Als die älteste Schätztechnik ist die *Momentenmethode*[37] aus dem Bereich der linearen Methoden bekannt. Den Momenten verschiedener Ordnung ist gemein, dass sie keine Information über die Dynamik enthalten. Speziell das Moment erster Ordnung ist der Erwartungswert beziehungsweise der durchschnittliche Wert einer gesamten Reihe. Dieser stellt eine der wesentlichsten Größen bei der Analyse von Wertereihen dar. Zur Bestimmung eines Mittelwerts werden vorrangig das Arithmetische Mittel und der Median angewandt. Das Arithmetische Mittel $\mu(x)$ einer Reihe $x = (x_i)$ mit $i = 0,...,I-1$ eignet sich besonders zur Beschreibung von intervallskalierten und normalverteilten Reihen:

$$\mu(x) = \frac{1}{I}\sum_{i=0}^{I-1} x_i \tag{3.9}$$

Bei ordinalskalierten und nicht normalverteilten Reihen hingegen bietet sich die Bestimmung des Medians an. Dieser ist als derjenige Wert einer Reihe definiert, bei dem höchstens die Hälfte der Werte der Reihe einen jeweils kleineren oder größeren Wert als der Median selbst besitzen. Ergeben sich zwei solcher Werte, wird das arithmetische Mittel dieser genommen. Ein Vorteil des Medians gegenüber dem arithmetischen Mittel ist seine geringere Empfindlichkeit gegenüber starken Ausreißern. Weitere Mittel wie harmonisches-, geometrisches-, quadratisches-, oder absolutes Mittel werden hier nicht betrachtet.

---

[37] Auch *Method-Of-Moments* nach Karl Pearson.

Neben dem Schwerpunkt entlang der Werteachse wie beim arithmetischen Mittel, kann auch jener entlang der Zeitachse von Interesse sein. Dieser sogenannte *Zentroid* c ergibt sich aus dem mit den Werten der Reihe gewichteten Mittelwert der Indexvariablen i:

$$c(x) = \frac{\sum_{i=0}^{I-1} x_i \cdot i}{\sum_{i=0}^{I-1} x_i} \tag{3.10}$$

Als weiteres wichtiges Charakteristikum besonders normalverteilter Werte wird die Standardabweichung als Moment zweiter Ordnung benutzt. Sie wird durch Ziehen der Quadratwurzel aus der Varianz berechnet, und trifft eine Aussage über die Dispersion um den Zentralwert.

$$\sigma(x) = \sqrt{\frac{1}{I}\sum_{i=0}^{I-1}(\mu - x_i)^2} \tag{3.11}$$

Neben diesen beiden betrachteten Momenten existieren noch höhere Momente wie die *Skewness*[38] und die *Kurtosis*[39], die für die angestrebte akustische Emotionserkennung geringeren Bezug zu psychoakustischen Größen aufweisen. Darüber hinaus werden hinsichtlich eines möglichst geringen Rechenbedarfs zu Gunsten einer angestrebten Echtzeitfähigkeit Momente geringerer Komplexität favorisiert.

3.3.2 Extrema

Neben der Bestimmung absoluter Extrema von Wertereihen bietet es sich unter Umständen an, diese zu relativieren. Dies geschieht hier mit Bezug auf den Mittelwert, und ist etwa bei der Erfassung des Intensitätsmaximums sinnvoll, um eine Normierung zu treffen.

Zusätzlich zu den minimalen oder maximalen Werten selbst ist oft auch die Position des Minimums $\min_{index}(x)$ beziehungsweise Maximums $\max_{index}(x)$ innerhalb einer Wertereihe $x = (x_i)$ mit $i = 1, ..., I$ von Interesse:

$$\begin{aligned} \min_{index}(x) &= \operatorname{argmin}_i(x_i) \\ \max_{index}(x) &= \operatorname{argmax}_i(x_i) \end{aligned} \tag{3.12}$$

Die Differenz des Maximums und des Minimums ergibt weiterhin den Wertebereich $\text{range}(x)$ der Verteilung an:

---

[38] Relativer Grad der Asymmetrie der Werteverteilung um den Zentralwert bezüglich der Normalverteilung.

[39] Angabe über den Grad der relativen Flachheit oder Steilheit einer Werteverteilung bezüglich der Normalverteilung. Es wird zwischen *platokurtischen*, respektive flachen, und dem entgegengesetzt *leptokurtischen* Verteilungen unterschieden.

$$\text{range}(x) = \max(x) - \min(x) \tag{3.13}$$

3.3.3 Normierung von Funktionalen

Einige der in den folgenden Kap. 3.7 und Kap. 3.11 vorgestellten Klassifikatoren, wie der Hauptteil der abstandsbasierten, reagieren sensibel hinsichtlich Erkennungsleistung auf verschiedene Zahlenbereiche unterschiedlicher Merkmale innerhalb eines Merkmalvektors. Hieraus ergibt sich der Bedarf einer Abbildung auf einen möglichst einheitlichen Bereich. Um gleichzeitig eine sinnvolle Nutzung des Zahlenbereichs eines verwendeten Rechners zu gewährleisten, erfolgt diese durch Befreiung der Merkmale vom Mittelwert $\mu(x)$ und Normierung auf die Standardabweichung $\sigma(x)$ der Verteilung $x = (x_i)$ mit $i = 0,...,I-1$ aller aufgetretener Werte x_i innerhalb der Lernmenge $\mathcal{L}$. Ein normierter Wert $\tilde{x}_i$ ergibt sich somit in folgender Weise:

$$\tilde{x}_i = \frac{x_i - \mu(x)}{\sigma(x)} \tag{3.14}$$

Die Standardabweichung der normierten Trainingsdaten beträgt somit Eins, und ihr Mittelwert Null. Für Daten, die nicht in $\mathcal{L}$ enthalten sind, gilt dies nur noch näherungsweise, da auch sie bezüglich $\mathcal{L}$ normiert werden. Alternativ kann auch eine Normierung auf den betragsmäßigen maximalen Wert erfolgen, um die Werte auf das Intervall $[-1,1]$ zu beschränken. Allerdings resultiert dies bei starken Ausreißern innerhalb der Wertereihe in einer suboptimalen Auflösung der Zahlen. Zur vereinfachten Notation wird im Weiteren von normierten Werten ausgegangen.

3.4 Prosodie

Die in Arbeiten zur Emotionserkennung am häufigsten betrachteten akustische Merkmale sind prosodische[40] Merkmale [PAN03]. Dies liegt unter Anderem daran, dass diese vergleichsweise einfach zu extrahieren sind. Unter ihnen werden suprasegmentale, also über mehrere Einzellaute verlaufende, Eigenschaften der Melodie und des Rhythmus in der Sprechweise zusammengefasst. Diese bleiben vom konkreten Lautinventar einer Sprache bis zu einem gewissen Grade unbeeinflusst, was vor dem Hintergrund einer inhaltsunabhängigen Erkennung der Emotion eine entscheidende Rolle spielt. Drei Hauptgruppen werden dabei in der Regel unterteilt: *Intensität*, *Intonation*, beziehungsweise Tonhöhenverlauf, und *Dauer*. Information, die durch diese Größen übertragen wird und weder syntaktischer noch semantischer Natur ist, wird als *nonverbal vokal* oder *paralinguistisch* bezeichnet. Im Folgenden soll die Berechnung konkreter prosodischer Merkmale näher beschrieben werden.

3.4.1 Elongation

Die Elongation beschreibt die unmittelbare Signalauslenkung über der Zeit. Der Mensch nimmt Änderungen in der Amplitude, dem jeweiligen Maximum einer Schwingung, im Mittel bei ungefähr

[40] Aus dem Griechischen: Prosodia, der Zugesang.

1 dB gerade noch wahr. Schwankungen werden dabei bei sinusoidalen Verläufen im Vergleich zu weißem Rauschen feiner aufgelöst [ZWI90].

Aus der Elongation wird zunächst die Rate der Nulldurchgänge *(ZCR)*[41] ermittelt. Nulldurchgänge lassen sich vergleichsweise einfach durch Finden von Stellen mit Vorzeichenwechsel im Verlauf der Elongation bestimmen und sind ein Standardverfahren bei der Analyse im Zeitbereich. Innerhalb eines Signalrahmens mit Index t ergibt sich für das Signal $s_t[n]$ die lokale Nulldurchgangsrate $ZCR_{s,t}$. Es wird zur übersichtlicheren Bezeichnung für diesen und weitere gebildete Verläufe vom Bezug zu einem Sprachsignal $s[n]$ ausgegangen. $ZCR_{s,t}$ vereinfacht sich hierdurch zu ZCR_t:

$$ZCR_t = \frac{1}{N} \cdot \sum_{n=1}^{N-1} \left| sign\left(s_t[n]\right) - sign\left(s_t[n-1]\right) \right| \tag{3.15}$$

Um durch verschiedene Aufzeichnungslängen keine Verzerrungen zu erleiden, wird die globale ZCR zusätzlich auf die Gesamtlänge T in Signalrahmen einer Phrase normiert. Trotz ihrer einfachen Berechnung erlaubt die ZCR Aussagen über die Stimmhaftigkeit eines Sprachsignals. So haben stimmhafte Laute eine deutlich niedrigere ZCR als stimmlose. Darüber hinaus wird sie generell zur Unterscheidung verschiedener Signaltypen verwandt [RIG04].

Die weitere Bestimmung des Mittelwerts und Medians ist ein Maß für den Offset einer Aufzeichnung, wie einem ungewollten Gleichanteil. Der Gebrauch sowohl von Median als auch Mittelwert erscheint zunächst redundant. Dies ist jedoch im Rahmen der später vollzogenen automatischen Generierung und Selektion von Merkmalen von Interesse (siehe Kap. 3.10 und Kap. 3.9). Nachfolgend findet sich eine Aufzählung von direkt aus dem Zeitsignal gebildeten Betrachtungen.

Funktionaltyp	**Notation**
Nulldurchgangsrate	*ZeroCrossingRate*
Langzeitmittelwert	*ElongMean*
Median	*ElongMed*

Tab. 3.5: Funktionale gebildet aus dem Verlauf des Zeitsignals

In Tab. 3.5 und den folgenden Tabellen zur Auflistung gebildeter Funktionale findet sich zusätzlich die in dieser Arbeit festgelegte Notation. Diese vereinfacht unter Anderem die Beschreibung relevanter Merkmale in Kap. 7.2.1 und Kap. A.2.

3.4.2 Intensität

Es besteht eine Reihe komplexer und umstrittener psychoakustischer Zusammenhänge zwischen der Lautstärkeempfindung eines Tones und der Frequenz und Dauer desselben. In dieser Arbeit wird

[41] Abkürzung für die im anglistischen Sprachraum gebräuchliche Bezeichnung *Zero-Crossing-Rate*.

jedoch, wie allgemein üblich, unter Verzicht auf eine nähere Betrachtung hierzu die in der digitalen Sprachverarbeitung weit verbreitete Kurzzeitenergie E_t eines Signalrahmens $s_t[n]$ mit Index t nach folgender Vorschrift abgeleitet:

$$E_t = \sum_{n=0}^{N-1} \left| s_t[n] \right|^2 \tag{3.16}$$

Dies entspricht einer physikalischen Betrachtung nach dem Schalldruck.

Problematisch an intensitätsbezogenen Größen ist der Umstand, dass bei der Sprachaufzeichnung die Aufnahmerichtung sowie die Distanz zu einem Mikrophon auf die Intensität einen großen Einfluss ausüben. In einigen Szenarien, wie im Fahrzeug, kann dies auf Grund konstanter Bedingungen unter Umständen vernachlässigt werden. Zur Kompensation werden aber in jedem Fall nur Änderungen der Energie betrachtet. Durch Bildung der ersten Ableitung erhält man schließlich Leistungswerte. Tab. 3.6 bietet einen Überblick über die abgeleiteten Größen aus diesem Verlauf. Zur Berechnung der An- und Abstiegszeiten werden die zeitlichen Abstände zwischen Extrema von Minimum zu Maximum und vice versa berechnet. Die Distanz der Wendepunkte ergibt sich aus der reziproken ZCR der Ableitung zweiter Ordnung.

Funktionaltyp	**Notation**
Langzeitmittelwert	*IntensMean*
Standardabweichung	*IntensStdDev*
Relatives Maximum	*IntensMax*
Relative Position des Maximums	*IntensPosMax*
Maximum der 1. Ableitung	*IntensDe1Max*
Mittelwert der Anstiegszeit	*IntensRisetimeMean*
Median der Anstiegszeit	*IntensRisetimeMed*
Mittelwert der Abstiegszeit	*IntensFalltimeMean*
Median der Abstiegszeit	*IntensFalltimeMed*
Mittelwert der Distanz der Wendepunkte	*IntensRePoDistMean*
Standardabweichung der Distanz der Wendepunkte	*IntensRePoDistStdDev*

Tab. 3.6: Funktionale gebildet aus dem Verlauf der Sprachsignalenergie

3.4.3 Intonation

In der Phonetik wird der Verlauf der Tonhöhe innerhalb eines Sprechaktes als Intonation oder Satzmelodie bezeichnet. Im Deutschen oder Englischen wird diese syntaktisch etwa zur Andeutung von Fragen benutzt. In tonalen Sprachen wie Mandarin hingegen wird sie zur Unterscheidung lexikalischen Gehalts verwendet. Darüber hinaus spielt sie eine entscheidende Rolle für den Ausdruck von Emotion. Die *Sprachgrundfrequenz (F_0)* ist das akustische Korrelat zu dieser perzeptiv wahrgenommenen Tonhöhe einer sprachlichen Äußerung. Zwischen den beiden genannten Größen der physikalischen Grundfrequenz[42], in der Regel in Schwingungen pro

[42] Im anglistischen Sprachgebrauch als *Fundamental Frequency* bezeichnet.

Sekunde, beziehungsweise Hz angegeben, und der wahrgenommenen Tonhöhe[43] besteht ein zwar monotoner, jedoch nichtdirekter und nichtlinearer Zusammenhang [ZWI90]. So löst das menschliche Gehör durch die logarithmische Frequenzwahrnehmung etwa tiefere Frequenzen feiner auf als höhere Frequenzen. Ferner kann die subjektive Tonhöhenwahrnehmung auch durch die Lautstärke und Dauer beeinflusst werden. Die komplexe Natur einer Verbindung diverser Frequenzanteile in der menschlichen Stimme begünstigt weiterhin die Divergenz zwischen perzipierter und physikalischer Grundfrequenz. Maßgebend hierfür ist der Einfluss der Amplitude von zugehörigen Obertönen zu F_0. Darüber hinaus kann auch ein Eindruck der Tonhöhenwahrnehmung in unharmonischen Signalen, wie etwa bandbeschränktem Rauschen, entstehen [ZWI90]. Dies beruht auf der frequenzanalytischen Arbeitsweise des Gehörs: Der größte gemeinsame Teiler einer Frequenzmenge wird vom Menschen als Grundfrequenz empfunden. Um generell eine Grundfrequenz wahrzunehmen, gelten 30 ms als ausreichend, wobei eine Unterscheidung zwischen verschiedenen Werten auch bei kürzeren Dauern möglich ist. Die in dieser Arbeit abgeleiteten Merkmale beziehen sich jedoch ausschließlich auf den aus dem Sprachsignal extrahierten Verlauf von F_0.

Zur Extraktion von F_0 aus dem Sprachsignal existieren eine Reihe sogenannter *Pitch Detection Algorithmen (PDA)*, die jedoch alle nur als Schätzung betrachtet werden können. Die tatsächliche Grundfrequenz wäre nur durch direkte Messung an der Glottis zu bestimmen. Dabei ist die Entstehung der Sprache im Menschen noch nicht vollständig geklärt. Im Allgemeinen kann eine Unterscheidung zwischen auf dem Zeitsignal, oder dem spektralen Verlauf basierenden Verfahren sowie Mischformen dieser beiden unterschieden werden, ohne dass sich dabei ein Ansatz als besonders geeignet erweist [MOU96]. Eine sinnvoll messbare Größe der Grundfrequenz tritt nur bei stimmhaften Artikulationen auf. Daher wird vor oder während der F_0-Schätzung eine stimmhaft/stimmlos Entscheidung getroffen. Alternativ kann, ergänzend zum Grundfrequenzverlauf, die Wahrscheinlichkeit der Stimmhaftigkeit für jeden Rahmen mit angegeben werden [WAN00]. Grundsätzlich ist die spektrale Energie der Grundfrequenz im Quellenspektrum des Vokaltrakts bei stimmhaften Lauten höher als die der zugehörigen Harmonischen, da sie die Anregung des Vokaltraktes darstellt. Resonanzen im Vokaltrakt, etwa der Formanten, können sich jedoch wie eine Reihe von Bandpassfiltern auswirken und somit F_0 im Verhältnis abschwächen. Dies führt im Allgemeinen zur sogenannten Grundfrequenzdopplung auf Grund der Selektion von höher harmonischen Anteilen bei der Maximumssuche, und erfordert die Betrachtung von F_0 im zeitlichen Verlauf über einen einzelnen Rahmen hinaus, um Konfusionen mit höheren Formanten zu vermeiden [YIN96].

Im Rahmen dieser Arbeit wurde ein auf dem Zeitsignal aufsetzendes Verfahren gewählt, welches auf der Autokorrelationsfunktion *(AKF)* beruht. Allgemein lässt sich sagen, dass auf der AKF basierende Ansätze mit die höchste Verbreitung besitzen. Nach [BOE93] sind sie grundsätzlich am besten geeignet und erzielen die größte Präzision. Die AKF_s des Signals s kann dabei als jene Transformation aufgefasst werden, die ein Maß der Ähnlichkeit des Signals bei einer Verschiebung um k zu sich selbst darstellt:

---

[43] In der anglistischen Literatur als *Pitch* beschrieben.

$$AKF_s[k] = \sum_{n=0}^{N-1-k} s[n] \cdot s[n+k] \text{ für } k = 0,...,N-1 \tag{3.17}$$

Bei der F_0-Berechnung macht man sich die Tatsache zu Nutze, dass für periodische Signale mit der Periodendauer T an den ganzzahligen Vielfachen $n \cdot T$ mit $n \in \mathbb{N}$ in der AKF des Signals $s[n]$ globale Maxima entstehen. Das erste solche globale Maximum findet sich im Ursprung der AKF eines Signals, und sein Wert entspricht grundsätzlich der Signalleistung von $s[n]$. Bei aperiodischen Signalen hingegen fällt die AKF steil und ohne ausgeprägte Maxima ab. Da sich die AKF zweier aufsummierter Signale ebenfalls addieren, finden sich an den Stellen der Vielfachen der Periodendauer T eines Signals s, welches aus additiver Überlagerung eines periodischen Signals mit Rauschen erhalten wird, lokale Maxima. Das globale Maximum ist in diesem Fall weiterhin im Ursprung zu finden, da dieser Wert nun der Gesamtleistung beider Signale entspricht. Das erste globale Maximum, mit Ausnahme des Maximums im Ursprung, befinde sich nun an der Stelle T_0. Aus dem Verhältnis von $AKF_s[T_0]$ zu $AKF_s[0]$ ergibt sich dann ein Wert für die harmonische Ausgeprägtheit im Intervall $[0,1]$. Zur Entscheidung über die Stimmhaftigkeit eines Sprachsignals kann daher ein Schwellwert gesetzt werden, welcher durch dieses Verhältnis von $AKF_s[T_0]$ zu $AKF_s[0]$ überschritten werden muss. Der Schwellwert wird hier auf 30% festgelegt. Zur zusätzlichen Sicherheit wird auch die ZCR des Signals unter Ausnutzung der Tatsache, dass stimmhafte Laute eine deutlich niedrigere ZCR aufweisen als stimmlose[44] (vgl. Kap. 3.4.1), betrachtet. Im Fall der Entscheidung eines stimmhaften Lauts entspricht der Ort des Maximums T_0 dem Reziprokwert der gesuchten Grundfrequenz F_0 [LAN95].

Zur näherungsweisen Kompensation von Verzerrungen durch die vorab erfolgte Fensterung wird die AKF des Sprachsignals $AKF_s[k]$ zusätzlich durch die leistungsnormierte AKF der Fensterfunktion $AKF_w[k]$ dividiert. Diese lautet für die gewählte Hanning-Funktion [BOE93]:

$$AKF_{w_{Han}}[k] = \left(1 - \frac{|k|}{N}\right) \cdot \left(\frac{2}{3} + \frac{1}{3} \cdot \cos\frac{2 \cdot \pi \cdot k}{N}\right) + \frac{1}{2 \cdot \pi} \sin\frac{2 \cdot \pi \cdot |k|}{N} \tag{3.18}$$

Damit lässt sich die Kurzzeit-Grundfrequenz $F_{0,t}$ innerhalb eines Signalrahmens mit Index t folgendermaßen bestimmen:

$$F_{0,t} = \frac{f_s}{N} \cdot \underset{k,k \neq 0}{\operatorname{argmax}} \frac{AKF_{s,t}[k]}{AKF_{w,t}[k]} \tag{3.19}$$

Eine häufig angewandte Variante der AKF die weniger Rechenkapazität benötigt, stellt die sogenannte *Average Magnitude Difference Funktion*, kurz *AMDF,* dar [ROS74]. Sie ist gegeben als:

---

[44] Generell weisen stimmhafte Laute auch eine höhere Energie und einen niedrigeren Frequenzbereich mit 800 Hz - 1600 Hz auf als stimmlose, die einen Bereich von 2.400 Hz - 3.200 Hz besitzen.

$$AMDF_s[k] = \sum_{n=0}^{N-1-k} |s[n] - s[n+k]|^j \text{ für } k = 1,...,N \tag{3.20}$$

Wird der Exponent *j* zu Eins gesetzt, ergibt sich eine Beschränkung auf Additions-Operationen an Stelle von Multiplikationen. Dies führt zu einer schnelleren Berechenbarkeit im Vergleich zur Anwendung der AKF. An den Vielfachen der Periodendauer T entstehen nun ausgeprägte Minima. Für die Umrechnung der gefundenen Periodendauer in die Grundfrequenz $F_{0,t}$ im Signalrahmen mit Index t ergibt sich entsprechend folgender Gesamtausdruck:

$$F_{0,t} = \frac{f_s}{N} \cdot \arg\min_{k,k\neq 0} \frac{1}{N} AMDF_t[k] \tag{3.21}$$

Die Vereinfachung der AKF zur AMDF erbringt jedoch in der Regel schlechtere Werte für die Grundfrequenz und die Berechnung der AKF wird im Weiteren ohnehin benötigt. Zusammenfassend lässt sich sagen, dass sich das gewählte Verfahren zur Grundfrequenzbestimmung robust gegenüber Rauschen erweist, jedoch eine Konfusion mit höheren Formanten (siehe Kap. 3.5.1) durch fehleranfällige Extremasuche möglich ist.

Um der eingangs erwähnten logarithmischen Hörempfindung bezüglich der Frequenzauflösung des Menschen gerecht zu werden, kann alternativ eine Transformation auf Halbtöne oder prozentuale Angabe erfolgen. Zur Übertragung in Halbtonschritte werden rektanguläre Semitonfilter gewählt. Sie reichen von der jeweiligen arithmetischen Mittelfrequenz eines Intervalls zu derjenigen adjazenter Mittelfrequenzen. In der nachfolgenden Gleichung beschreibt $f(x)$ die Frequenz einer musikalischen Note *x* und $f(x^{\#})$ diejenige des im Semitonintervall erhöhten Pendants $x^{\#}$.

$$f(x^{\#}) = f(x) \cdot \sqrt[12]{2} \tag{3.22}$$

Der Bereich der Grundfrequenz wird anschließend auf 47 Intervalle von *D* (73,416 Hz) bis *c'''* (1.046,502 Hz) auf den menschlichen Gesangsbereich[45] beschränkt.

Abschließend sind in Tab. 3.7 die berechneten Funktionale vorgestellt, um eine größere Unabhängigkeit vom gesprochenen Inhalt zu erzielen.

Das Merkmal *PitchArea* sei als Sonderfall einzeln beschrieben: Es wird die betragsmäßige Fläche des Grundfrequenzverlaufs um die mittlere Grundfrequenz gebildet und auf die Äußerungslänge normiert. Es ergibt sich so ein Maß für die gesamte Aktivität bezüglich starker Auslenkungen der Grundfrequenz:

$$PitchArea = \frac{1}{T} \sum_{t=0}^{T-1} \left| F_{0,s,t} - PitchMean \right| \tag{3.23}$$

Es ist zu erwarten, dass sich für Zustände geringer Aktivität ein niedriger Wert und vice versa

[45] Dieser ist für die klassischen Gesangsstimmen folgendermaßen beschrieben: Bass *D'-e*, Tenor *c-c''*, Alt *d-e''*, Sopran *c'-c'''*.

ergeben.

Funktionaltyp	**Notation**
Langzeitmittelwert	*PitchMean*
Standardabweichung	*PitchStdDev*
Relatives Maximum	*PitchMax*
Relatives Minimum	*PitchMin*
Relative Position des Maximums	*PitchPosMax*
Relative Position des Minimums	*PitchPosMin*
Wertebereich	*PitchRange*
Mittlere Steigung	*PitchDe1Mean*
Maximale Steigung	*PitchDe1Max*
Mittlere Distanz der Wendepunkte	*PitchRePoDistMean*
Standardabweichung der Distanz der Wendepunkte	*PitchRePoDistStdDev*
Normierte absolute Fläche	*PitchArea*

Tab. 3.7: Funktionale gebildet aus der Sprachgrundfrequenz

3.4.4 Dauer

Um das zeitliche Verhalten in den abgeleiteten Funktionalen über die Positionen von Extrema hinaus besser zu modellieren, werden ebenfalls Merkmale über die Dauer berechnet. Die Dauer stimmhafter Laute wird dabei in einer Näherung aus der Länge stimmhafter Abschnitte geschätzt. Die Entscheidung über einen stimmhaften Signalrahmen erfolgt wie in Kap. 3.4.3 beschrieben. Für den Zweck der Emotionserkennung erscheint diese Approximation durchaus adäquat.

Die Dauer von Pausen hingegen ergibt sich aus dem Verlauf der Signalenergie. Im Gegensatz zur Segmentierung des gesamten Signals in Einzelphrasen wie in Kap. 4.3.2 beschrieben, wird hier ein absoluter Schwellwert festgelegt. Die gesamte Dauer der Unterschreitung bis zur nächsten Überschreitung wird als Pause gewertet. Dabei kann diese maximal so lange andauern, bis durch die übergeordnete Segmentierung das Ende einer Phrase festgelegt wird. Tab. 3.8 zeigt die abgeleiteten Größen aus diesen beiden Dauern.

Funktionaltyp	**Notation**
Mittlere Dauer stimmhafter Laute	*DurationVoiSoMean*
Standardabweichung der Dauer stimmhafter Laute	*DurationVoiSoStdDev*
Rate stimmhafter Laute	*RateVoiSo*
Mittlere Pausendauer	*DurationSilMean*
Median der Pausendauer	*DurationSilMed*

Tab. 3.8: Funktionale gebildet aus dem zeitlichen Verhalten des Sprachsignals

Es sei erwähnt, dass das Zusammenspiel von Intonation, Intensität und Dauer eine entscheidende Rolle für die Betonung spielen. Generell sind Merkmale einer starken Betonung erhöhte Lautheit, gehobene Sprachgrundfrequenz sowie eine gestreckte Silbenlänge. Die sogenannte *Hauptbetonung* einer Phrase weist aus semantischer Sicht auf das zentrale Element dieser hin. Daneben existieren *Wortbetonungen* an der Stelle der meistbetonten Silbe sowie unbetonte Silben. Die Positionen des Phrasenmaximums bezüglich der Intensität und der Intonation spielen daher auch für die

Interpretation natürlicher Sprache eine wichtige Rolle.

Weiterhin sei angemerkt, dass ein klare Abgrenzung des Merkmaltyps Dauer sich als diffizil erweist: Zur Dauer können ebenfalls die ZCR, die An- und Abstiegszeiten der Intensität sowie die zeitlichen Distanzen von Wendepunkten gezählt werden. Die hier gewählte Einteilung beruht jedoch auf der Sichtweise der Merkmalsextraktion, und zählt nur Merkmale auf höherer Ebene wie Pausen- oder Vokallängen explizit zur Dauer.

3.5 Stimmqualität

Das *Timbre* einer Stimme, in der Musik als *Klangfarbe* bezeichnet, wird durch die Phonation und Artikulation beeinflusst. Es entspricht der individuellen Perzeption der Stimmqualität. Diese ist kein Träger linguistischer Information und daher unmittelbar mit dem emotionalen Charakter korreliert. Darüber hinaus hängt von ihr in erhöhtem Maße die Verständlichkeit des gesprochenen Inhalts in einer Äußerung ab. Als Beispiel sei Flüstern im Gegensatz zu geschrieener oder normal artikulierter Sprache genannt. Da das Verhältnis spektraler Bänder [KLA00A] und die Lage und Bandbreite von Formanten maßgeblich vom Timbre geprägt sind, sollen diese durch die folgenden Größen berücksichtigt werden.

3.5.1 Formanten

Formanten[46] sind Maxima im Spektrum, die durch ihre Mittenfrequenz und Bandbreite definiert sind. Ihre spektrale Lage ist dabei unabhängig von der wahrgenommenen Tonhöhe. Während sie im Normalfall oberhalb der Grundfrequenz zu finden sind, kann bei angehobener Grundfrequenz der erste Formant unterhalb dieser zu liegen kommen. Diejenigen Formanten, die oberhalb von ihr liegen, haben direkten Einfluss auf die beschriebene Klangfarbe eines Tonsignals. Induziert werden sie durch Resonanzen, die von der aktuellen Form des Vokaltraktes bei der Lautbildung abhängen. Sie treten in erster Linie bei stimmhaften Lauten, insbesondere bei Vokalen auf. Letztere prägen sie dabei unmittelbar: Obwohl sich ihre Lage individuell etwa nach Geschlecht und Alter stark unterscheidet[47], da ihre Position auch von der Länge des Ansatzrohres abhängt, kann ein bestimmter Vokal durch den ersten und zweiten über das sogenannte *Vokaltrapez* festgestellt werden. Auch der dritte Formant beeinflusst die Lautqualität, wohingegen höhere die Klangwahrnehmung prägen. Über den gesprochenen Vokal hinaus besteht jedoch ein Toleranzbereich, der etwa bei Sprechen unter Lächeln durch eine Verkürzung des Sprechtraktes zu einer Anhebung der Formantenlage führt. Zur menschlichen Wahrnehmung von Formanten ist anzumerken, dass eine Änderung der Mittenfrequenz ab 3%, eine der Bandbreite erst ab 40% Abweichung wahrgenommen wird [CAR79].

Die Bestimmung der Formanten erweist sich vergleichbar der Sprachgrundfrequenz als schwierig. Nicht zuletzt auch wegen der Gefahr von Konfusionen mit dieser. Sie werden hier unter zu

---

[46] Von lateinisch *formare* für *formen*, da sie bildlich den Klang formen. Eingeführt wurde der Begriff 1929 in einer Habilitationsschrift von Erich Schuhmann.

[47] Diese Eigenschaft verleiht ihnen auch in der forensischen Sprecherverifikation hohen Stellenwert.

Hilfenahme der *linearen Prädiktion* berechnet. Prinzip der linearen Prädiktion, kurz *LPC*[48], ist es, den aktuellen Signalwert $s[n]$ aus den bereits beobachteten p vorhergehenden Werten $s[n-k]$ mit $k = 1,...,p$ zu schätzen. Der Schätzwert $\hat{s}[n]$ ergibt sich dabei unter Verwendung der Prädiktionskoeffizienten a_k zu:

$$\hat{s}[n] = \sum_{k=1}^{p} a_k \cdot s[n-k] \tag{3.24}$$

Der Fehler $\varepsilon[n]$ zwischen Schätzwert und tatsächlichem Signalwert zu jedem Zeitpunkt n resultiert somit aus:

$$\varepsilon[n] = s[n] - \hat{s}[n] \tag{3.25}$$

Dieser Ansatz erlaubt eine Darstellung als rekursives Filter, welches $\varepsilon[n]$ als Eingangs- und $s[n]$ als Ausgangssignal besitzt. Die Übertragungsfunktion $H_{LPC}(z)$, die sich dabei im z-Bereich ergibt, lautet unter Verwendung der z-Transformierten $S(z)$ und $\mathrm{E}(z)$[49]:

$$H_{LPC}(z) = \frac{S(z)}{\mathrm{E}(z)} = \frac{1}{1 - \sum_{k=1}^{p} a_k \cdot z^{-k}} \tag{3.26}$$

$H_{LPC}(z)$ kann somit als lineares Filtermodell des Vokaltrakts interpretiert werden. In diesem Fall sollte die Anregungsfunktion, hier die Fehlerfunktion, der Anregung durch die Grundfrequenzimpulse entsprechen. Da das beobachtete Sprachsignal zum Zeitpunkt der Bestimmung der Formanten bekannt ist, besteht die Aufgabe der LPC-Analyse in der Bestimmung der einzelnen Prädiktionskoeffizienten, deren maximal mögliche Zahl durch die Gesamtzahl in einem Signalrahmen enthaltener Werte limitiert ist. Die Koeffizienten müssen möglichst so bestimmt werden, dass die folgende Summe E des Fehlerquadrats über die letzten p Abtastwerte minimal wird:

$$\mathrm{E} = \sum_{n=1}^{p} \varepsilon^2[n] \tag{3.27}$$

Diese Forderung ist bei stimmhaften Lauten leichter zu erfüllen, da der Signalverlauf bei stimmlosen Lauten stochastischer Natur ist und geringere Stationarität aufweist. Aus der Tatsache, dass E eine Linearkombination von quadratischen Funktionen ist, geht hervor, dass es ein eindeutiges Minimum für E gibt. Für dieses Optimierungsproblem erhält man bei Nullsetzung der

---

[48] Kurzform für die anglistische Bezeichnung *Linear Prediction Coding*. Sie beruht auf dem ursprünglichen Einsatz zur Datenreduktion in der Telekommunikation.

[49] Eine Eigenschaft die sich aus der Übertragungsfunktion ergibt, ist dass nur Polstellen, jedoch keine Nullstellen realisiert werden können. Dies spielt zwar in der Spracherkennung eine Rolle, ist jedoch für die Schätzung der Formanten irrelevant.

partiellen Ableitung von E nach den Koeffizienten a_k für $k = 1,...,p$ insgesamt p Gleichungen, aus denen sich die Koeffizienten bestimmen lassen. Es ist empfehlenswert, zu diesem Zweck die auch hier angewandte leichtere und vorteilhaftere Berechnung über die Autokorrelationsmethode oder alternativ die *Kovarianzmethode* zu verwenden. Dabei kann durch eine geschickte Umformung das Gleichungssystem in eine Toeplitz-Matrix transformiert werden. Diese erlaubt eine schnelle Lösung mit Hilfe des Levinson-Durbin-Algorithmus. Auf weitere Einzelheiten zur Realisierung sei hier unter Verweis auf die Literatur [RAB93] verzichtet.

Um nun von der LPC-Analyse ausgehend die Formantenlagen zu bestimmen, können zunächst lokale Maxima im Spektrum des LPC-Filters gesucht werden. Hierbei können ungewollt schwach ausgeprägte Formanten, die dicht neben stark ausgeprägten liegen, leicht übergangen werden. Als Abhilfe kann nach Sattelpunkten gesucht, oder eine Auflösung anhand des Phasenspektrums vollzogen werden. Eine genauere Lösung mit besserer Frequenzauflösung bietet hier die *Root-Solving*-Methode zur Bestimmung von Polen von $H_{LPC}(z)$. Diese ergeben sich aus komplexen Nullstellen $z_{o,k}$ mit $k = 1,...,p$ des Nennerpolynoms der Ordnung p und lassen sich wie folgt als Funktion der Formantenmittenfrequenz F_x und -bandbreite B_x darstellen:

$$z_{o,k} = f\left(-\pi B_x + j2\pi F_x\right) \tag{3.28}$$

Bei diesem Vorgang wird das LPC-Filter in Filter zweiter Ordnung zerlegt, wobei jedes Teilfilter einen Formanten modellieren kann. Für eine genauere Beschreibung wird hier auf [ACE99] verwiesen.

In einer zweiten Stufe der Bestimmung werden den gefundenen Resonanzfrequenzen die konkreten Formanten zugeordnet. Hierzu wird eine Tabelle mit den Frequenzbereichen der Formanten angewandt, um Fehlerkosten einzelner Zuordnungsmöglichkeiten innerhalb eines Signalrahmens festzulegen. Mittels dynamischer Programmierung (siehe Kap. 3.11.1) erfolgt schließlich eine Rückverfolgung der Trajektorien, um dadurch die Fehlerkosten bei der Verbindung benachbarter Rahmen durch Auswahl der jeweiligen lokalen Zuordnungsmöglichkeit global minimieren zu können und somit Kontinuität im Frequenzverlauf sicher zu stellen. Im Anschluss daran werden in gleicher Weise die Formantenbandbreiten bestimmt.

Die LPC-Ordnung hat starken Einfluss auf die Güte der Schätzung und sollte idealer Weise an den Sprecher angepasst sein. Bei zu geringer Ordnung kann es vorkommen, dass eine Resonanz ungewünschter Weise mehrere Formanten modelliert. Eine Auflösung ist dann nicht mehr möglich. Bei zu hoher Ordnung treten Frequenzverschiebungen in Richtung von Obertönen auf. Für das verwendete Wurzellösungsverfahren muss eine gerade Kardinalität gewählt sein, die hier zu 18 festgelegt wird. Da mittels LPC-Analyse keine Antiresonanzen modelliert werden, kommt es bei nasaler Artikulation unter Umständen zu Fehlern, weil im Bereich des dritten Formantens eine zusätzliche Resonanz auftreten kann. Um generell falsche Werte zu vermeiden, erfolgt eine Bestimmung nur in stimmhaften Intervallen.

Durch die in Kap. 3.1 beschriebene Abstrahlung des Sprachsignals an Mund und Nase bedingt, sind bei stimmhaften Lauten höhere Formanten schwächer ausgeprägt als niedrigere. Um dies zu

kompensieren, wird das Signal $s[n]$ wie folgt bereits vor der Fensterung einer Präemphase mit einem Filter mit der Übertragungsfunktion $H_{pre}(z)$ im z-Bereich mit dem *Präemphasefaktor* $\alpha_{pre} = 0,7$ unterzogen:

$$S^*(z) = S(z) \cdot H_{pre}(z) \text{ mit } H_{pre}(z) = 1 - \alpha_{pre} \cdot z^{-1} \tag{3.29}$$

Mit $s^*[n]$ als gefiltertes Signal entspricht dies im Zeitbereich:

$$s^*[n] = s[n] - \alpha_{pre} \cdot s[n-1] \tag{3.30}$$

Es sei angemerkt, dass für stimmlose Segmente, die hier aus genannten Gründen nicht betrachtet werden, auf Grund in der Regel stark ausgeprägter hoher Frequenzanteile umgekehrt vorgegangen werden sollte.

Für eine statische Modellierung werden aus den jeweiligen *X*-ten Formanten *FX* höherer Ordnung mit $X = 1,...,7$ eine Reihe Merkmale über eine gesamte Sprachsegmentdauer abgeleitet. Diese beziehen sich zunächst auf die spektrale Lage der Formanten:

Funktionaltyp	**Notation**
Langzeitmittelwert	*F1Mean, ..., F7Mean*
Standardabweichung	*F1StdDev, ..., F7StdDev*
Relatives Maximum	*F1Max, ..., F7Max*
Relatives Minimum	*F1Min, ..., F7Min*
Wertebereich	*F1Range, …, F7Range*
Mittlere spektrale Distanz zu F0	*F1ReDiF0, ..., F7ReDiF0*
Mittlere betragsmäßige Steigung	*F1De1Mean, …, F7De1Mean*
Maximale betragsmäßige Steigung	*F1De1Max, …, F7De1Max*

Tab. 3.9: Funktionale gebildet aus den Verläufen der Formanten höherer Ordnung

Die Ermittlung der mittleren relativen spektralen Distanz zur Grundfrequenz $F_{0,t}$ im Signalrahmen mit Index t erfolgt auf folgender Grundlage:

$$FXReDiF0 = \frac{1}{T} \sum_{t=0}^{T-1} \left(\frac{FX_t - F_{0,t}}{F_{0,t}} \right) \tag{3.31}$$

Die mittlere betragsmäßige Steigung ergibt sich aus:

$$FXDe1Mean = \frac{1}{T} \sum_{t=1}^{T-1} \left| FX_t - FX_{t-1} \right| \tag{3.32}$$

Die maximale betragsmäßige Steigung schließlich wird durch folgende Gleichung bestimmt:

$$FXDe1Max = \max\left(\left| FX_t - FX_{t-1} \right| \right) \tag{3.33}$$

Im Folgenden sind Merkmale beruhend auf der statistischen Analyse der spektralen Breite *BX* der Formanten aufgelistet.

Funktionaltyp	Notation
Langzeitmittelwert	*B1Mean, …, B7Mean*
Standardabweichung	*B1StdDev, ..., B7StdDev*
Relatives Maximum	*B1Max, …, B7Max*
Relatives Minimum	*B1Min, ..., B7Min*
Wertebereich	*B1Range, …, B7Range*
Mittlere betragsmäßige Steigung	*B1De1Mean, …, B7De1Mean*
Maximale betragsmäßige Steigung	*B1De1Max, …, B7De1Max*

Tab. 3.10: Funktionale gebildet aus den Verläufen der Bandbreiten von Formanten höherer Ordnung

3.5.2 Harmonische Ausgeprägtheit

Die Ausgeprägtheit von Harmonischen einer Schwingung *(HNR)*[50] wird ebenfalls nur in stimmhaften Anteilen geschätzt. Es wird die in periodischen Anteilen enthaltene Signalleistung im Verhältnis zu der des umgebenden Rauschens bestimmt. Dieses Merkmal wird gerne im Bereich der Musikverarbeitung verwendet, um die Musikalität eines Abschnitts zu beurteilen. Dies soll hier als weitere Größe zur Erfassung der Emotionalität eingehen. Darüber hinaus ist die HNR gut geeignet, um verschiedene Phonationsarten (vgl. Kap. 3.1) wie zum Beispiel Hauchen zu erkennen. Ihre Berechnung erfolgt unter Ausnutzung der in Kap. 3.4.3 beschriebenen Verhältnisse der AKF eines akustischen Signals. Während sie sich wie die Grundfrequenz alternativ aus dem spektralen Verlauf bestimmen lässt, ergibt der im Folgenden beschriebene Weg nach [BOE93] exaktere Werte: Ausgehend von einem periodischen Signal, das additiv durch Rauschen überlagert ist, wird zunächst das lokale Maximum, welches nicht dem Ursprung entspricht, gesucht. Dieses sei an der Stelle T_0. Die $AKF[T_0]$ entspricht dann der Gesamtleistung des periodischen Signalanteils. Ähnlich der Entscheidung zur Stimmhaftigkeit wird die $AKF[T_0]$ ins Verhältnis zur verbleibenden Leistung des Rauschanteils, die sich aus der Differenz der gesamten Signalleistung entsprechend $AKF[0]$ und der des periodischen Anteils gemäß $AKF[T_0]$ ergibt, gesetzt. Die Bestimmung der Kurzzeit-HNR HNR_t des Signals $s_t[n]$ innerhalb eines Rahmens t, üblicher Weise in dB angegeben, da so die menschliche Wahrnehmung besser gespiegelt wird, ergibt sich somit in folgender Weise:

$$HNR_t = 10 \cdot \log \frac{AKF_t[T_0]}{AKF_t[0] - AKF_t[T_0]} \tag{3.34}$$

Aus dieser Definition geht eine unendliche HNR für ideal periodische Signale hervor. Da dieser Fall in der Praxis für Sprachsignale nicht eintritt, wird der dann gegebenen Division mit Null als

[50] Kurz für *Harmonics-To-Noise-Ratio*.

Wert im Nenner nicht Rechnung getragen.

Aus dem Verlauf der HNR werden drei Merkmale analog zu vorangehenden Betrachtungen gebildet. Sie finden sich in Tab. 3.11.

Funktionaltyp	**Notation**
Langzeitmittelwert	*HNRMean*
Standardabweichung	*HNRStdDev*
Maximum	*HNRMax*

Tab. 3.11: Funktionale gebildet aus dem Verlauf der HNR

3.5.3 Spektrale Charakteristika

Zur Berechnung spektraler Merkmale wird eine *Fast-Fourier-Transformation (FFT)* mit linearer Frequenzaufteilung und einer Bandbreite von 20 Hz durchgeführt. Die FFT ist eine Variante der *Diskreten Fourier-Transformation (DFT)*, bei der die Zahl der Rechenschritte für N Punkte von der Kardinalität $2 \cdot N^2$ auf $2 \cdot N \cdot \mathrm{ld}(N)$ reduziert wird. Hierzu sei zunächst die DFT $S_t[f]$ mit der diskreten Frequenz f [51] des Signals $s_t[n]$ im Signalrahmen t gegeben:

$$\mathrm{DFT}\{s_t[n]\} = S_t[f] = \sum_{n=0}^{N-1} s_t[n] \cdot e^{-\frac{2 \cdot \pi \cdot j}{N} \cdot f \cdot n} \tag{3.35}$$

Das aus $s_t[n]$ zu bestimmende Spektrum $S_t[f]$ reicht von 0 Hz bis $f_{\max}$ entsprechend der Nyquist-Frequenz f_s. Für die folgenden Betrachtungen sei die vektorielle Schreibweise $\underline{S}_t = \{S_t[f]\}^T$ mit $f = 0,...,f_{\max}$ eingeführt. Um die DFT mittels FFT zu bestimmen, wird hier der weit verbreitete *Cooley-Tukey-Algorithmus*[52] angewandt. Dabei wird rekursiv nach dem Prinzip Teilen und Vereinen die Berechnung auf die Bestimmung mehrerer kleinerer DFT verlagert. Auf die Beschreibung von Details zu seinem Ablauf wird hier verzichtet.

Zur gehörgerechten Anpassung entsprechend der menschlichen Lautstärkenempfindung in Abhängigkeit von der Frequenz wird das Spektrum durch eine Annäherung der in ISO 226 definierten dB(A)-Bewertungskurve nichtlinear gewichtet. Die gewählte Approximation ist ein sich durch den *Least-Square-Algorithmus* (vgl. Kap. 3.8.2) ergebendes Polynom 2. Ordnung:

$$F[f] = -10{,}31524199 \cdot \log^2[f] + 70{,}0396577 \cdot \log[f] - 117{,}426918 \tag{3.36}$$

Dabei bezeichnet $F[f]$ den von der Frequenz f abhängigen dB-Wert welcher im Zuge der dB(A)-

---

[51] Obwohl f üblicherweise eine kontinuierliche Frequenz bezeichnet, ist f zur vereinfachten Notation hier auch für die diskretisierte Frequenz gewählt.

[52] Benannt nach einer Veröffentlichung von J. W. Cooley und J. W. Tukey aus dem Jahr 1965, obwohl dieser Algorithmus bereits Carl Friedrich Gauß bekannt war. Weitere Verfahren zur Bestimmung der FFT sind beispielsweise der *Primfaktor*-, der *Bruun*-, der *Rader*-, oder der *Bluestein*-Algorithmus.

Bewertung additiv dem entsprechenden Wert des Leistungsspektrums zu überlagern ist. Folgende Abb. 3.12 zeigt hierzu den Verlauf der Polynomnäherung im Vergleich zur dB(A)-Bewertungskurve:

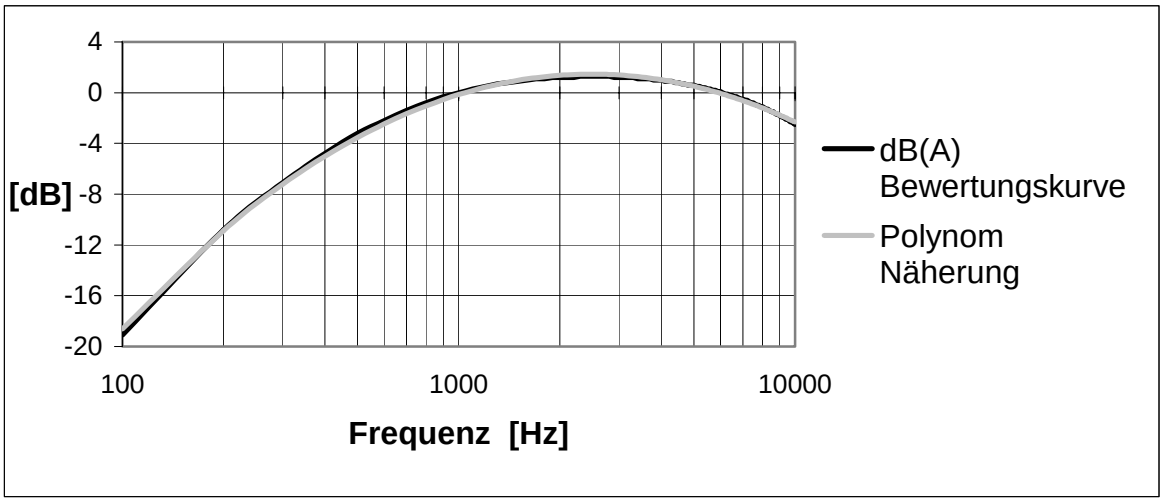


Abb. 3.12: dB(A) Bewertungskurve und Polynomnäherung

Einzelne Energiebänder neigen dazu, stark vom gesprochenen phonetischen Inhalt abzuhängen. Nichtsdestotrotz können ihr Verhältnis zueinander sowie Langzeitbetrachtungen für die Erkennung von Emotion von Interesse sein.

Funktionaltyp	Notation
Langzeitmittelwert 0-250 Hz	*SpecPo250Mean*
Langzeitmittelwert 0-650 Hz	*SpecPo650Mean*

Tab. 3.13: Funktionale gebildet aus dem Verlauf niedriger Energiebänder

Zunächst wird die kumulative spektrale Leistung im unteren Frequenzbereich äquivalent einem idealen Tiefpassfilter berechnet. Dabei wird einmal diejenige von 0 Hz - 250 Hz, und einmal diejenige von 0 Hz - 650 Hz bestimmt. Aus diesen Niederbandenergien werden jeweils die in Tab. 3.13 gezeigten Langzeitmittelwerte gebildet.

Der *Roll-Off-Punkt* ist ein weiterer aus dem DFT-Spektrum abgeleiteter Verlauf. Er bezeichnet die Frequenz, bei der von 0 Hz ab betrachtet 95%[53] der gesamten Signalenergie enthalten sind. Die Berechnung erfolgt durch Summation über die Frequenzbänder $S_t[f]$ beginnend mit $f = 0$ bis zur Frequenz $f_{rop,t}$, bei der der gewünschte prozentuale Anteil erreicht ist. Die Frequenz $f_{rop,t}$ wird per Definition zu Null gesetzt, falls die kummulative Energie innerhalb $s_t[n]$ Null entspricht. Die hierzu berechneten Funktionale sind im Einzelnen:

[53] Es existieren auch andere Definitionen zur oberen Eckfrequenz, beispielsweise 98% der Gesamtenergie.

Funktionaltyp	Notation
Langzeitmittelwert	*RollOffMean*
Standardabweichung	*RollOffStdDev*
Relatives Maximum	*RollOffMax*
Mittlere betragsmäßige Steigung	*RollOffDe1Mean*
Standardabweichung betragsmäßige Steigung	*RollOffDe1StdDev*
Maximale betragsmäßige Steigung	*RollOffDe1Max*

Tab. 3.14: Funktionale gebildet aus dem Verlauf des spektralen Roll-Off Punkts

Als letzte Kontur wird der spektrale Fluss[54] $S_{flux,t}$ betrachtet. Er ergibt sich aus der zeitlichen Verschiebung des FFT-Spektrums als Maß für die globale Dynamik. Die Berechnung ist erst ab dem zweiten Signalrahmen möglich:

$$S_{flux,t} = \left\| \underline{S}_t - \underline{S}_{t-1} \right\| \text{ mit } t = 1,...,T-1 \tag{3.37}$$

Die sich aus dem Fluss ergebenden Merkmale sind im Einzelnen:

Funktionaltyp	Notation
Langzeitmittelwert	*SpecFluxMean*
Standardabweichung	*SpecFluxStdDev*
Relatives Maximum	*SpecFluxMax*

Tab. 3.15: Funktionale gebildet aus dem Verlauf des spektralen Flusses

3.5.4 Mel-Frequenz-Cepstral-Koeffizienten

Mel-Frequenz-Cepstral-Koeffizienten, kurz *MFCC*, gehören zu den am häufigsten verwendeten Merkmalen in der automatischen Spracherkennung. Der Vollständigkeit halber sollen diese hier kurz erläutert werden: Es handelt sich bei MFCC um eine *homomorphe* spektrale Transformation. Hierunter ist eine strukturerhaltende Transformation zu verstehen, deren Aufgabe es hier ist, die Faltung eines Signals mit einer Übertragungsfunktion auf eine additive Superposition zurückzuführen [RIG04]. Der Zweck ist dabei in der Sprachverarbeitung für das beobachtete abgestrahlte Sprachsignal $s[n]$ das ursprüngliche Anregungssignal $s'[n]$ möglichst von den sprecherspezifischen Eigenschaften $H[f]$ des Vokaltrakts zu trennen (vgl. Kap. 3.1). Das nachfolgende Blockschaltbild soll das Prinzip veranschaulichen: Zunächst erfolgt eine DFT des Eingangssignals. Der Faltungsoperation im Zeitbereich entspricht dabei eine Multiplikation im Frequenzbereich. Eine anschließende Logarithmierung des Signals bringt eine Verlagerung auf eine Summation der beiden Anteile. Zusätzlich erfolgt in Anpassung an die menschliche Wahrnehmung eine nichtlineare spektrale Transformation. Dies wird durch eine trianguläre Filterbank mit N Kanälen realisiert, die auf der Mel-Skala $Mel[f]$ äquidistant angeordnet sind. Die Skala ist

[54] Im Anglistischen als *Spectral Flux* bezeichnet.

gegeben als:

$$Mel[f] = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right) \tag{3.38}$$

Im Anschluss an die Filterung der Bänder werden mittels einer inversen DFT die MFCC Koeffizienten berechnet. An Stelle einer inversen DFT werden in der Praxis zur Dekorrelation gerne wie hier auch eine Diskrete Cosinus-Transformation *(DCT)*, oder eine Hauptachsentransformation (vgl. Kap. 3.9.1) verwendet. Die DCT, eine Entwicklung nach einer gewichteten Summe von Kosinusfunktionen, lässt sich als Vektor-Matrix-Multiplikation ausführen, und die Berechnung erfolgt dabei im Gegensatz zur DFT ausschließlich mit reellen Koeffizienten, was ihre Bestimmung vereinfacht. Das Signal $s_t[n]$ im Rahmen t wird in folgender Weise zu $S_t[u]$ in den u-Bereich der DCT transformiert:

$$S_t[u] = \sqrt{\frac{2}{N}} \cdot C[u] \cdot \sum_{n=0}^{N-1} s_t[n] \cdot \cos\frac{(2 \cdot n + 1) \cdot u \cdot \pi}{2 \cdot N} \text{ mit } C[u] = \begin{cases} \frac{1}{\sqrt{2}} & \text{für } u = 0 \\ 1 & \text{sonst} \end{cases} \tag{3.39}$$

Über diese Schritte zur Bestimmung von MFCC hinaus erfolgt in der Regel, wie in Kap. 3.5.1 geschildert, eine Präemphase des Signals. Entsprechend dem allgemein üblichen Vorgehen wird der Präemphasefaktor dabei zu $\alpha_{pre} = 0,9$ gewählt.

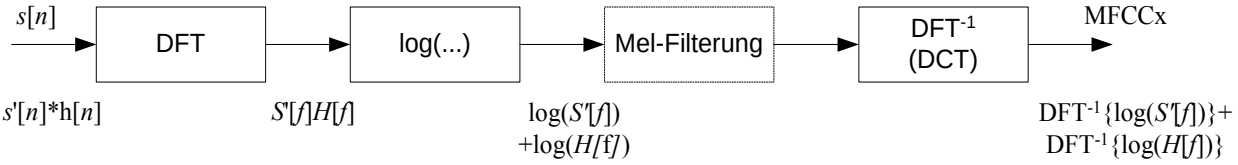


Abb. 3.16: Ablauf der Berechnung von MFCC-Koeffizienten

Es erscheint fraglich, inwiefern MFCC sich für die Emotionserkennung eignen, da insbesondere niedrige Koeffizienten sehr stark den gesprochenen Inhalt, und höhere sprechercharakteristische Eigenschaften repräsentieren. In höheren ist dabei auch die Grundfrequenz prinzipiell enthalten. Einzelne Bänder können jedoch, wie auch die berechneten Tiefbandenergieverläufe, von Interesse sein, was die in Kap. 3.9 beschriebene Selektion von Merkmalen zeigen wird. Des Weiteren sind sie die Vorraussetzung für die geschilderte Betrachtung der Verhältnisse spektraler Bänder zueinander, was durch genetische Generierung in Kap. 3.10 realisiert wird. Eine Reihe von Merkmalen wird im Zusammenhang mit MFCC analog der vorherigen Analyse betrachtet. Dabei bezeichnet *MFCCX* jeweils den *X*-ten Koeffizienten. Ferner wird auch der Betrag der Differenz benachbarter Koeffizienten *MFCCXDe1* betrachtet. Er kann erst ab dem zweiten Rahmen ermittelt werden. Es gilt:

$$MFCCXDe1_t = \left| MFCCX_t - MFCCX_{t-1} \right| \text{ mit } t = 1, \ldots, N-1 \tag{3.40}$$

Für die betragsmäßige Differenz zweiter Ordnung steht weiterhin als Abkürzung *MFCCXDe2*. Sie

wird erst ab dem dritten Signalrahmen berechnet:

$$MFCCXDe2_t=|MFCCDe1X_t- MFCCDe1X_{t-1}| \text{ mit } t = 2,...,N-1 \tag{3.41}$$

Es werden folgende charakteristische Größen entsprechend abgeleitet.

Funktionaltyp	**Notation**
Langzeitmittelwert	*MFCC1Mean, …, MFCC15Mean*
Standardabweichung	*MFCC1StdDev, …, MFCC15StdDev*
Relatives Maximum	*MFCC1Max, …, MFCC15Max*
Relatives Minimum	*MFCC1Min, …, MFCC15Min*
Mittlere Steigung	*MFCC1De1Mean, …, MFCC15De1Mean*
Relatives Maximum der 1. Ableitung	*MFCC1De1Max, …, MFCC15De1Max*
Langzeitmittelwert der 2. Ableitung	*MFCC1De2Mean, …, MFCC15De2Mean*
Relatives Maximum der 2. Ableitung	*MFCC1De2Max, …, MFCC15De2Max*

Tab. 3.17: Funktionale gebildet aus den Verläufen der MFCC

Speziell der Langzeitmittelwert wird üblicherweise unter Ausnutzung der homomorphen Charakteristik subtraktiv zur Kanalkompensation verwandt, da MFCC als sehr anfällig gegenüber Kanalstörungen und Hintergrundgeräuschen gelten. Es ist von daher davon auszugehen, dass diese Merkmale stark durch die Aufnahme geprägt sind, wobei der Einfluss als zeitlich unabhängig gesehen werden kann.

3.6 Artikulation

Die Artikulation beschreibt die Lautbildung je nach stimmhafter, oder –loser Artikulationsart. Merkmale hierzu werden daher auf segmentaler Ebene betrachtet. Im direkten Zusammenhang mit Ihnen steht der von den Muskeln geleistete Aufwand, weswegen Ihnen eine starke Korrelation mit der in Kap. 2.3.1 beschriebenen Aktivitätsdimension zugeschrieben wird [BUR00]. Als erste typische Größe sei der artikulatorische Aufwand bei der Lautbildung von Konsonanten genannt. Hierzu wird gerne der spektrale Schwerpunkt[55] mit der diskreten Frequenz f und der spektralen Energie $S_t[f]$ im Signalrahmen t betrachtet [VAN99]:

$$SpecPoCen_i = \frac{\sum_{f=0}^{(N/2)-20} S_t[f] \cdot f}{\sum_{f=0}^{(N/2)-20} S_t[f]} \tag{3.42}$$

Sollte sich die Leistung in einem Signalrahmen zu Null ergeben, wird der spektrale Schwerpunkt per Definition zu 0 Hz gesetzt, um eine Division durch Null zu umgehen.

[55] Im anglistischen Sprachgebrauch unter *Spectral Centroid* oder *Spectral Balance* bekannt.

Es werden aus dem Verlauf des Schwerpunkts ermittelt:

Funktionaltyp	Notation
Langzeitmittelwert	*SpecPoCenMean*
Standardabweichung	*SpecPoCenStdDev*
Relatives Maximum	*SpecPoCenMax*
Mittlere betragsmäßige Steigung	*SpecPoCenDe1Mean*
Standardabweichung der betragsmäßigen Steigung	*SpecPoCenDe1StdDev*
Maximale betragsmäßige Steigung	*SpecPoCenDe1Max*

Tab. 3.18: Funktionale gebildet aus dem Verlauf des spektralen Zentroids

Eine weitere Größe resultiert aus der Analyse der Artikulation von Vokalen. Hier ist besonders eine Bewegung innerhalb des gegebenen Spielraums bezüglich der Lage der ersten beiden Formanten bei der Bildung eines bestimmten Vokals von Interesse. Man spricht dabei von der *Zentralisierung* der Formanten. Eine zentrale Lage entspricht einer *Reduktion* oder dem *Vowel-Target-Undershoot*. Kommen die ersten beiden Formanten über den bei neutraler Sprechweise üblichen Frequenzen zum Liegen, wird dies als *Elaboration* oder *Vowel-Target-Overshoot* bezeichnet [KIE00]. Zwar werden hier die Formantenlagen, wie in Kap. 3.5.1 beschrieben, in die Betrachtung mit einbezogen - die exakte und automatische Berechnung der Zentralisierung setzt jedoch eine Kenntnis des aktuellen Vokals voraus, was eine zusätzliche Spracherkennung zur phonetischen Segmentierung erfordert. Weiterhin erschwert wird diese durch den Umstand der hohen Fehlerhaftigkeit sowohl der phonetischen Verschriftung auf Basis einer automatischen Erkennung, als auch der Bestimmung der Formanten selbst. Da, bedingt durch den Stand der Technik, hieraus eine zu geringe Genauigkeit bei der Schätzung der Zentralisierung zu erwarten ist, wird auf diese Berechnung verzichtet.

Ebenfalls auf einer automatischen Spracherkennung basierend können Auslassungen und Einschübe von Lauten, sogenannte *Elisionen* und *Epenthesen* sowie der generelle Grad der Dissimilierung im Vergleich zu einer normierten phonetischen Verschriftung eines neutral gesprochenen Wortes geprüft werden. Dies ist jedoch für eine automatische Extraktion besonders schwierig, da es nicht nur eine automatische Erkennung von Sprache, sondern auch eine exakte phonetische Verschriftung im Falle von ausgelassenen oder eingeschobenen Lauten verlangt.

Fasst man die Betrachtungen zu akustischen Merkmalen zusammen, ergibt sich abschließend die in Tab. 3.19 abgebildete Verteilung bestimmter Funktionale im Vergleich zur Zahl ihrer Basiskonturen:

Anzahl [#]	Basis-Verläufe	Verläufe	Funktionale
Zeitsignal	1	1	3
Intensität	1	3	11
Grundfrequenz	1	3	12
Dauer	(2)	(2)	5
Formanten (Lage, Bandbreite)	14	28	105
MFCC	15	45	120
HNR	1	1	3
FFT Spektrum	5	7	17
Gesamt	**38**	**88**	**276**

Tab. 3.19: Verteilung der Basiskonturen ohne Differenzbildung erster und zweiter Ordnung, der Gesamtkonturen und der gebildeten Funktionale geordnet nach Gruppen. Geklammerte Zahlen bei den Konturen der Dauer weisen auf die Berechnung aus anderen vorhandenen Verläufen hin.

3.7 Automatische Klassifikation von Mustern

Im Folgenden werden eine Reihe von Verfahren zur automatischen Klassifikation von Mustern vorgestellt. Der Einsatz unterschiedlicher Methoden wird durch die in Kap. 2.6.2 beschriebenen unterschiedlichen Stärken und Schwächen diverser Ansätze begründet. Im anschließenden Abschnitt wird ferner gezeigt, wie mehrere unterschiedliche Verfahren zu einem mächtigeren Klassifikator vereint werden können.

3.7.1 Abstandsklassifikatoren

Abstandsklassifikatoren werden auch *instanzbasierte Klassifikatoren* genannt. Grundgedanke ist es, den nächstgelegenen Referenzvektor im Musterraum mittels einer linearen *Entscheidungsfunktion* aufzusuchen und dessen Klasse als erkannte Klasse auszugeben. Aus dieser Vorgehensweise wird auch die Bezeichnung *Nächster-Nachbar-Klassifikator*[56], kurz *1NN*, abgeleitet. Ein Lernen vorab ist nicht erforderlich, da die Verarbeitung auf den Moment der Klassifikationsanfrage verschoben wird. Man spricht daher auch von *Lazy Learning* oder *Memory-based Learning*, also speicherbasiertem Lernen [ATK97]. Als Vorteil erweist sich somit, dass ein Lernprozess entfallen kann, da neue Referenzen jederzeit dem Referenzschatz zugefügt werden können. In der Erkennungsphase ergeben sich dabei sehr lange Verarbeitungszeiten, insbesondere bei einer hohen Zahl an Referenzmustern, da der beobachtete Vektor mit jedem gespeicherten Beispiel verglichen werden muss. Um dies zu vermeiden, können ein oder mehrere Schwerpunkte je Klasse gebildet werden. In diesem Fall wird jedoch das Hinzufügen neuer Referenzen erschwert, und die Leistung bezüglich maximaler Diskriminanz erleidet Einbußen.

Es wird eine Reihe von unterschiedlichen Maßen zur Berechnung der Distanz $d\left(\underline{x},\underline{x}_l\right)$ zwischen

[56] Englisch: *Nearest Neighbor Classifier.*

einem Mustervektor $\underline{x}$ und einem Referenzvektor $\underline{x}_l$ aus der Lernmenge $\mathcal{L}$ angewandt [FUK90][57]. Die beiden Vektoren seien dabei von der gleichen Dimensionalität *n*. Die höchste Verbreitung hat dabei die *Minkowski-Metrik* d_r:

$$d_r\left(\underline{x},\underline{x}_l\right)=\sqrt[r]{\sum_{i=1}^{N}\left|x_i-x_{l,i}\right|^r} \tag{3.43}$$

Für den Sonderfall $r=2$ wird diese auch *Euklidsches Abstandsmaß* genannt. Ebenfalls gerne verwendet wird der sogenannte *City-Block-* oder *Manhatten-Abstand* für $r=1$. Sehr populär ist ferner die *Cosinus-Distanzfunktion* $d_{\cos}$, bei der der Winkel zwischen den Vektoren $\underline{x}$ und $\underline{x}_l$ als Maß angewandt wird:

$$d_{\cos}\left(\underline{x},\underline{x}_l\right)=\cos\left(\underline{x},\underline{x}_l\right)=\frac{\sum_{i=1}^{n}x_i\cdot x_{l,i}}{\sqrt{\sum_{i=1}^{n}x_i^2}\cdot\sqrt{\sum_{i=1}^{n}x_{l,i}^2}} \tag{3.44}$$

Basierend auf dem Distanzmaß ergibt sich die erkannte Klasse κ_e schließlich entsprechend der Klasse κ desjenigen Vektors $\underline{x}_{l,\kappa}$ aus $\mathcal{L}$, für den sich der minimale Abstand für die gewählte Funktion ergibt:

$$\kappa_e=\underset{\kappa}{\operatorname{argmin}}\,d\left(\underline{x},\underline{x}_{l,\kappa}\right) \tag{3.45}$$

Alternativ kann auch ein Mehrheitsentscheid zwischen den Klassen der *k* nächsten Nachbarn des zuzuordnenden Musters unter den gelernten Referenzen erfolgen. Man spricht dann vom *k-Nächste-Nachbarn-Klassifikator (kNN)*. Die erkannte Klasse κ_e erhält man somit unter Einführung der Zahl k_κ der nächsten Nachbarn der Klasse Ω_κ innerhalb der k nächstliegenden Instanzen zu:

$$\kappa_e=\underset{\kappa}{\operatorname{argmax}}\,k_\kappa \tag{3.46}$$

Der kNN-Ansatz mit $k>1$ kann auch als Schätzung der a-posteriori Wahrscheinlichkeit $P\left(\Omega_\kappa\left|\underline{x}\right.\right)$ der Klasse Ω_κ für die Beobachtung des Vektors $\underline{x}$ interpretiert werden[58]. Damit stellt dieses Verfahren bereits ein statistisches Lernverfahren dar. Die Induktion probabilistischer Beschreibungen aus einer Trainingsmenge stellt eine beliebte Alternative zu instanzbasierten oder neuronalen Verfahren des maschinellen Lernens dar. Es folgen in Kap. 3.7.5, Kap. 3.7.6 und Kap. 3.11.2 eine Reihe weiterer statistischer Klassifikationsansätze.

---

[57] Beispiele weiterer, nicht betrachteter Abstandsmaße sind die *Dice-*, die *Jaccard-* und die *Mahalanobis*-Distanz, oder entropiebasierte Maße [CLE95], [DAY05].

[58] Für die bedingte Wahrscheinlichkeit P*(A=a|B=b)* mit den Zufallsvariablen *A* und *B* und den konkreten Werten *a* und *b* wird innerhalb dieser Arbeit soweit sinnvoll möglich vereinfachend P*(a|b)* gewählt.

3.7.2 Support-Vektor-Maschinen

Support-Vektor-Maschinen (SVM)[59] wurden in [VAP95] zur Klassifizierung von Daten eingeführt. Sie basieren auf der statistischen Lerntheorie und ihre theoretische Grundlage kann als Analogon zur Elektrostatik gesehen werden: Dabei korrespondiert ein Trainingsbeispiel zu einem geladenen Leiter an einem bestimmten Ort im Raum, die Entscheidungsfunktion zur elektrostatischen Potentialfunktion und die Lernzielfunktion zur Coulombschen Energie [HOC02]. In der aktuellen Mustererkennung erfreuen sich SVM steigender Beliebtheit auf Grund ihrer hohen Generalisierungsfähigkeit. Besonders zu erwähnen ist ferner die weitgehende Unabhängigkeit von der Komplexität bezüglich der Anzahl betrachteter Merkmale - beispielsweise im Vergleich zum beschriebenen 1NN. Im Wesentlichen werden vier Basiskonzepte in der Theorie der SVM vereint:

- Strukturelle Minimierung des Fehlerrisikos bei der Klassifikation,
- Abbildung in einen neuen Merkmalsraum, *Kerneltrick* genannt,
- Lineare Entscheidung im neuen Merkmalsraum und
- Optimierung eines dualen Problems.

Grundsätzlich sind SVM somit in der Lage, *zwei* Klassen voneinander zu trennen. Im Folgenden wird zunächst das grundlegende Verfahren zur Klassifikation vorgestellt. Anschließend werden verwendete Ansätze zur Trennung multipler Klassen erläutert.

Gegeben sei eine binäre Lernaufgabe mit der endlichen Trainingsmenge $\mathcal{L}$:

$$\mathcal{L} = \left\{ \left(\underline{x}_l, y_l \right) \mid l = 1, ..., L \right\} \text{ mit } \underline{x}_l \in \mathbb{R}^n \text{ und } y_l \in \left\{ +1, -1 \right\} \tag{3.47}$$

Die insgesamt zwei zu trennenden Klassen werden dabei als positive und negative Instanzen gesehen. Für die positiven Instanzen $\underline{x}_l \in \Omega_1$ gilt dabei per Definition $y_l = +1$, beziehungsweise $y_l = -1$ für die Menge der negativen Instanzen $\underline{x}_l \in \Omega_2$. Um im Weiteren die jeweiligen Instanzen voneinander strikt trennen zu können, wird eine durch den Vektor $\underline{w}$ und den Bias b definierte Hyperebene $H\left(\underline{w}, b\right)$ der Form

$$H\left(\underline{w}, b\right) = \left\{ \underline{x} \in \mathbb{R}^n \mid \underline{w}^T \underline{x} + b = 0 \right\} \text{ mit } \underline{w} \in \mathbb{R}^n \text{ und } b \in \mathbb{R} \tag{3.48}$$

gesucht, welche in der Lage ist, diese Forderung zu erfüllen:

$$y_l = \pm 1 \Rightarrow \qquad \underline{w}^T \underline{x}_l + b = \pm 1 \tag{3.49}$$

Unter der Voraussetzung, dass eine ideale lineare Diskrimination tatsächlich möglich sei, ist eine Normierung der Randbedingungen aus Gl. 3.49 durch geeignete Skalierung von $\underline{w}$ und b erreichbar [CHR00]:

---

[59] Auch als *Stützvektormaschinen* bezeichnet.

$$\begin{aligned} y_l = +1 \Rightarrow \quad & \underline{w}^T \underline{x}_l + b \geq +1 \\ y_l = -1 \Rightarrow \quad & \underline{w}^T \underline{x}_l + b \leq -1 \end{aligned} \tag{3.50}$$

Es ergibt sich nun unter Verwendung des folgenden vorzeichenbehafteten Abstands $r(\underline{x})$ eines Punktes $\underline{x}$ von der Hyperebene H

$$r(\underline{x}) = \frac{\underline{w}^T \underline{x} + b}{\|\underline{w}\|} \tag{3.51}$$

die Trennbreite[60] $\mu_{\mathcal{L}}$ als das Minimum der Beträge der Abstände aller Punkte $\underline{x}_1, ..., \underline{x}_l$ in $\mathcal{L}$ von H:

$$\mu_{\mathcal{L}}(\underline{w}, b) = \min_{l=1,...,L} \left(r(\underline{x}_l) \right) \tag{3.52}$$

Um eine maximale Diskriminativität zwischen den beiden Klassen zu gewährleisten, gilt es, diese Trennbreite zu maximieren. Hierzu wird eine optimale, die Trainingsmenge $\mathcal{L}$ separierende Hyperebene $H^* = H(\underline{w}^*, b^*)$ mit maximalem Wert $\mu_{\mathcal{L}}^*(\underline{w}^*, b^*)$ gesucht. Die jeweiligen Instanzen $\underline{x}_l^{sv} \in \mathcal{L}$, die Gl. 3.11 Genüge leisten, besitzen dabei die größte Nähe zur Hyperebene H^* und werden Support- oder Stützvektoren von H^* bezüglich $\mathcal{L}$ genannt. Ihr Abstand $r^*(\underline{x}_l^{sv})$ zur Hyperebene H^* beträgt auf Grund der Normierung der Trennbedingung:

$$r^*(\underline{x}_l^{sv}) = \frac{\pm 1}{\|\underline{w}\|} \text{ für } y_l = \pm 1 \tag{3.53}$$

In der Konsequenz ergibt sich ein Korridor zwischen den positiven und negativen Instanzen mit der Breite $2 \cdot \|\underline{w}\|^{-1}$. Sein Rand wird von den Stützvektoren gebildet, die in zugehöriger Abb. 3.20 im Drucksatz hervorgehoben sind. Anstelle der Maximierung der Korridorbreite kann auch eine Minimierung des Ausdrucks $0{,}5 \cdot \underline{w}^T \underline{w}$ vollzogen werden. Die sich ergebende zu minimierende Funktion ist eine streng konvexe Funktion von $\mathbb{R}^n$ in $\mathbb{R}$ und besitzt ein eindeutiges Minimum $\underline{w}$. Aus Gl. 3.50 ergeben sich hierfür die linearen Randbedingungen:

$$y_l \cdot \left(\underline{w}^T \underline{x}_l + b \right) - 1 \geq 0 \text{ mit } l = 1, ..., L \tag{3.54}$$

Zur Lösung dieses Randwertproblems können unter Anderem Lagrangesche Multiplikatoren genutzt werden, worauf hier unter Verweis auf [VAP95] nicht näher eingegangen wird.

---

[60] In der anglistischen Literatur als *Margin-Of-Separation* bezeichnet.

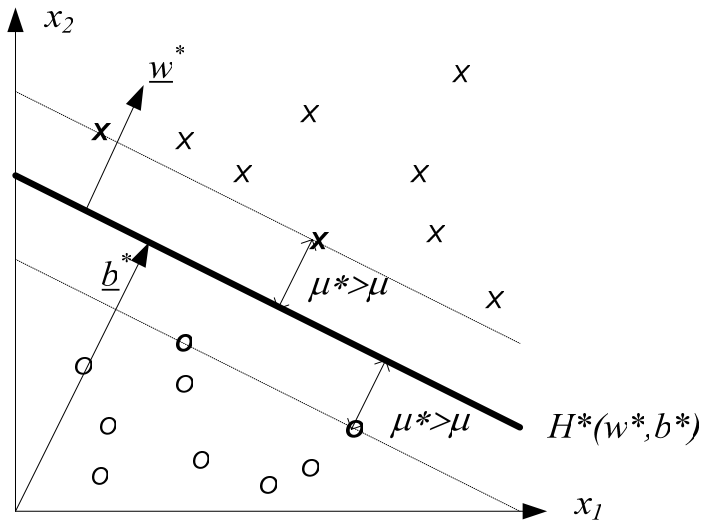


Abb. 3.20: Optimale Hyperebene $H^*\left(\underline{w},b\right)$ mit maximaler Trennbreite μ^*, Darstellung im zweidimensionalen Merkmalsraum

Im Allgemeinen, nicht trivialen Fall, ist eine lineare Trennbarkeit der Trainingsmenge $\mathcal{L}$ entgegen der vorab angenommenen Voraussetzung nicht gegeben. In diesem Fall werden die Gleichungen Gl. 3.50 um eine sogenannte Schlupfvariable $\xi \geq 0$ erweitert[61]. Es ist somit erlaubt, den Ansatz aufrecht zu erhalten, weil Vektoren, die die Trennebene durchdringen, auf der falschen Seite platziert werden dürfen:

$$\begin{aligned} y_l = +1 \Rightarrow \quad & \underline{w}^T \underline{x}_l + b \geq +1 - \xi_l \\ y_l = -1 \Rightarrow \quad & \underline{w}^T \underline{x}_l + b \leq -1 + \xi_l \end{aligned} \tag{3.55}$$

Somit ist der Ausdruck $\left\|\underline{w}\right\| + G \cdot \sum_{l=1}^{L} \xi_l$ zu minimieren, wobei *G* ein Fehlergewichtungsfaktor ist.

Als Ansatz zur Lösung nicht linear trennbarer Probleme wird eine Transformation des Merkmalsraums der Dimension *n* in einen in der Regel höher dimensionalen Raum der Dimension $\tilde{n}$ gemäß der Abbildungsvorschrift Φ mit

$$\Phi : \mathbb{R}^n \rightarrow \mathbb{R}^{\tilde{n}} \tag{3.56}$$

durchgeführt. Der Raum $\mathbb{R}^n$ wird dabei als Eingaberaum, der Raum $\mathbb{R}^{\tilde{n}}$ als Merkmalsraum[62] bezeichnet. Die Transformation wird mit Hilfe einer sogenannten *Kernel-Funktion* $K^{\Phi}\left(\underline{x},\underline{y}\right)$ der Form

$$K^{\Phi}\left(\underline{x},\underline{y}\right) = \Phi\left(\underline{x}\right)^T \cdot \Phi\left(\underline{y}\right) \tag{3.57}$$

---

[61] Man spricht nun von einem *Soft-Margin-Binary-Classifier*.

[62] Die Begriffe entsprechen den anglistischen Bezeichnungen *Input-Space* und *Feature-Space.*

vollzogen. Die Kernel-Funktion muss Symmetrie aufweisen, der Einhaltung der Cauchy-Schwarz Bedingung Genüge leisten und eine positiv semi-definite Matrix besitzen. Eine optimale Kernel-Funktion kann nur empirisch ermittelt werden. Ziel ist dabei neben einer optimalen Trennungsleistung ein geringer Rechenaufwand. In der Praxis finden vor allem folgende Funktionen Anwendung:

- Polynomial-Kernel mit $K^{\Phi}\left(\underline{x},\underline{y}\right)=\left(\underline{x}^{T}\cdot\underline{y}+1\right)^{p}$, wobei p die Polynomordnung bezeichnet,
- Gauß-Kernel[63] mit $K^{\Phi}\left(\underline{x},\underline{y}\right)=e^{-\frac{\left\|\underline{x}-\underline{y}\right\|^{2}}{2\sigma^{2}}}$, wobei σ die Standardabweichung für die Gaußsche Glockenkurve angibt, und
- Sigmoid-Kernel mit $K^{\Phi}\left(\underline{x},\underline{y}\right)=\tanh\left(k\left(\underline{x}^{T}\cdot\underline{y}\right)+\Theta\right)$, wobei k die Verstärkung, und Θ den Offset bestimmen.

Als Entscheidungsfunktion $d_{\underline{w},b}\left(\underline{x}\right)$ erhält man unter Verwendung der Abbildungsvorschrift aus Gl. 3.56:

$$d_{\underline{w},b}\left(\underline{x}\right)=\underline{w}^{T}\Phi\left(\underline{x}\right)+b \tag{3.58}$$

Somit kann folgende Entscheidungsregel verwendet werden:

$$\begin{aligned} d_{\underline{w},b}\left(\underline{x}\right)\geq 0 \quad &\Rightarrow \quad x\in\Omega_{1} \\ d_{\underline{w},b}\left(\underline{x}\right)<0 \quad &\Rightarrow \quad x\in\Omega_{2} \end{aligned} \tag{3.59}$$

Es existiert eine Reihe von Möglichkeiten, um nun die binäre Problemstellung auf multiple Klassen Ω_{κ} mit $\kappa=1,...,k$ und $k>2$ zu erweitern. Hier sollen die vier wichtigsten vorgestellt werden:

- *Eine-gegen-alle* Klassen Entscheidung[64], mit k binären Entscheidungen $\Omega_{1}=\Omega_{\kappa}$ und $\Omega_{2}=\left\{\Omega\setminus\Omega_{\kappa}\right\}$ für $\kappa=1,...,k$. Es wird diejenige Klasse gewählt, für die $d_{\underline{w},b}\left(\underline{x}\right)$ maximal ist.
- Paarweise *eine-gegen-eine (1-vs-1)* Klasse Entscheidung[65], wie in [HAS98] vorgestellt, mit $0,5\cdot k\cdot\left(k-1\right)$ binären Entscheidungen. Ein Mehrheitsentscheid der Einzelsiege jeder Klasse führt zur finalen Klasse. Nach [NIE03] ist der höhere Rechenaufwand gegenüber der Strategie eine-gegen-alle auf Grund höherer Leistung gerechtfertigt. Im Folgenden wird diese Methode angewandt, falls nicht explizit eine andere Angabe getroffen wird.

---

[63] Auch als radiale Basisfunktion *(RBF)* Kernel-Funktion bezeichnet.

[64] Im anglistischen Sprachgebrauch als *One-Against-All* Verfahren bekannt.

[65] In der anglistischen Literatur als *One-Against-One* Methode bezeichnet.

- SVM-Bäume: Diese Variante ist nicht vollständig erforscht. Ein Vorteil ist eine verkürzte Trainingszeit beruhend auf der Tatsache, dass für k Klassen nur $(k-1)$ SVM Modelle gebildet werden müssen. Vor allem aber beschleunigt sich der Ablauf in der Erkennungsphase, da nur SVM entlang des Astes der finalen Klasse bewertet werden müssen. Je nach Anordnung sind dies im günstigsten Fall nur eine, und im Mittel $ld(k)$ Einzelklassifikationen. Allerdings ist eine Parallelisierbarkeit somit nicht mehr gewährleistet. Durchgeführte Analysen im Laufe dieser Arbeit haben eine deutliche Abhängigkeit der Erkennungsleistung von der Anordnung der Klassen gezeigt. Es kann als optimale Regel empfohlen werden, schwer diskriminierbare Klassen zuletzt zu trennen. Dies kann aus der Konfusionsmatrix einer vorab vollzogenen eine-gegen-alle Entscheidung abgeleitet werden, wobei der Prozess auch automatisiert werden kann [SCH04A]. Weiterhin kann beobachtet werden, dass bei optimaler Anordnung eine Verbesserung in der Leistung gegenüber den beiden vorhergehenden Varianten erzielbar ist. Die nachfolgende Darstellung zeigt exemplarisch die für die konkreten Emotionen des MPEG-4 Sets mit ergänzender Neutralität gefundene günstigste Aufteilung auf die Ebenen des Baumes:

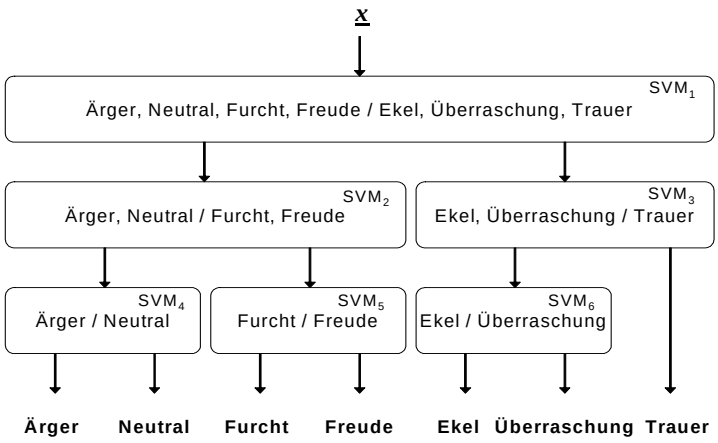


Abb. 3.21: Optimale Aufteilung auf die Ebenen bei Einsatz von Multi-Layer SVM für MPEG-4 Emotionen mit ergänzender Neutralität

- Mehrschicht- oder Multi-Layer-SVM (*ML-SVM*)[66] - eine leichte Abwandlung der SVM-Bäume, indem hier auf unterschiedlichen Ebenen zusätzlich diverse Merkmalsets zum Einsatz kommen können [XUC03], [SCH04D]. In speziellen Fällen kann dies von Vorteil sein, wie in Kap. 4.3.2 gezeigt wird.

Mit der Vorstellung der Grundzüge der Theorie zu SVM wird dieses Thema hier ohne weitere Details und Beweisführungen abgeschlossen. Zusammenfassend kann gesagt werden, dass die Klassifikation mit SVM bei einer Vielzahl von Support-Vektoren eine hohe Erkennungsdauer mit

[66] Die Abkürzung ML-SVM wird im Folgenden auch für SVM-Bäume verwendet. Werden auf den verschiedenen Ebenen diverse Merkmalssets angewandt, erfolgt die Angabe gesondert.

sich bringt[67]. Durch die Suche nach Stützvektoren ergibt sich aber eine Datenreduktion. Darüber hinaus zeigen SVM eine besondere Effektivität bezüglich Daten mit vielen Merkmalen auf Grund ihrer selbstständigen Suche nach der Komplexität. Ihre hohe Generalisierungsfähigkeit, durch strukturelle Risikominimierung mit gegebener Fehlerobergrenze bedingt, vermeidet im Besonderen eine Überanpassung.

Nachteilig zeigt sich jedoch ein hoher Zeitaufwand beim Trainieren neuer Lernbeispiele auf Grund der wiederholt notwendigen Optimierung. Es können auch keine allgemeinen Aussagen über die Klassifikationssicherheit oder zur Wahl der richtigen Kernel-Funktion getroffen werden. Eine richtige Datenverarbeitung und -Skalierung, sowie eine Wahl günstiger Attribute sind ferner essentiell. Trotz dieser Nachteile sind SVM dank ihrer sehr hohen Präzision für reale Problemstellungen sehr gut geeignet.

3.7.3 Künstliche Neuronale Netze

Künstliche Neuronale Netze (*ANN*[68]) sind in der Lage, praktisch beliebige Funktionen nachzubilden [NIE03], wodurch sie sich im besonderen Maße zur Bildung von Klassifikatoren eignen und sich einer sehr großen Verbreitung erfreuen. Die ersten mathematischen Modelle gehen dabei auf die Arbeiten von McCulloch und Pitts im Jahre 1943 zurück [MCC43], auf denen auch heutige ANN maßgeblich basieren. Die grundlegenden Eigenschaften lehnen sich an natürliche neuronale Netze an, wie sie in zentralen Nervensystemen vor allem von Wirbeltieren vorliegen. Die elementare Einheit zur Verarbeitung von Information ist dabei das *Neuron*, welches über sein zugehöriges *Axon* eine bestimmte Aktivität in Form elektrischer Impulse sendet [RIG99]. Diese Impulse werden über ein verzweigtes Netzwerk zur synaptischen Verbindung anderer Neuronen propagiert. In Abhängigkeit von der kumulativen Eingangserregung ändert sich jeweils die Aktivität eines Neurons, was sich in der Regel durch eine erhöhte Impulsfrequenz äußert. Maßgeblich hierfür ist allgemein eine Schwellwertentscheidung, die meist durch eine nichtlineare Übertragungsfunktion approximiert ist. Ein Neuronales Netz besteht somit aus Neuronen und ihren gerichteten Verbindungen. Von Interesse für den hier verfolgten technischen Einsatz sind dabei in erster Linie die Lernfähigkeit und Parallelität der Neuronalen Netze, welche eine Klassifikation in hoher Geschwindigkeit ermöglichen. Beschreibende Parameter sind:

- **Die Netzwerktopologie** mit der zugehörigen Menge der verarbeitenden Einheiten und Verbindungen.
- **Der Berechnungstyp der Einheiten**. Oft ist dabei eine homogene Anordnung gewählt.
- **Die Kodierung des Ausgangs**.

Abb. 3.22 zeigt ein einfaches Neuron, wie es in den im Rahmen dieser Arbeit angewandten Netzen

---

[67] Es existieren zur Lösung dieses Problems Ansätze, die Zahl der Support-Vektoren zu minimieren. Man spricht dann von *Reduced SVM* oder *RSVM*.

[68] Die gewählte Kurzform bezieht sich auf die anglistische Namensgebung *Artificial Neural Network*, und dient der besseren Unterscheidbarkeit zur Abkürzung kNN für den k-Nächste-Nachbarn Klassifikator.

verwendet wird. An den Eingängen liegen die Werte x_i mit $i = 1,...,N$ des Eingangsvektors $\underline{x} = \{x_i\}$ an - im hier betrachteten Fall der Merkmalsvektor. Die Eingangswerte erfahren eine Modulation mit den Gewichten w_i, wobei $i = 0,...,N$, zusammengefasst im Vektor $\underline{w} = \{w_i\}$. Als Sonderfall ist dabei das *Bias* genannte Gewicht w_0 zu nennen, mit welchem ein permanenter additiver Offset erzielt werden kann. Im Anschluss erfolgt eine Summation der gewichteten Anteile, deren Ergebnis u mit der, in der Regel nichtlinearen, Transferfunktion $T(u)$ verarbeitet wird. Das Ergebnis dieser Transformation liegt schließlich am Ausgang v an. Auf Grund der verwendeten Transferfunktionsarten wird dieser Neurontyp auch als *Schwellwertelement* bezeichnet. Die Schwellwertfunktion bildet dabei die Entscheidungsgrundlage und sollte idealer Weise hohe Flankensteilheiten aufweisen.

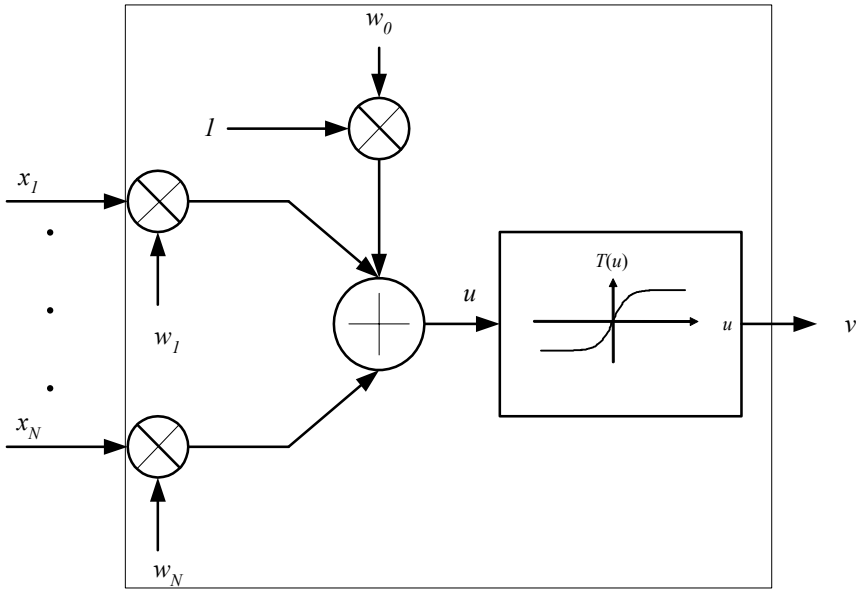


Abb. 3.22: Beispielhaftes Neuron

Beliebte Transferfunktionen sind dabei:

- Schwellwertfunktion: $T(u) = \begin{cases} 1 : u \geq 0 \\ 0 : u < 0 \end{cases}$
- Sigmoid-Funktion: $T(u) = \frac{1}{1 + e^{-\alpha u}}$, mit α als Steigungsparameter.
- Tangens-Hyperbolicus-Funktion, als Sonderfall der Sigmoid-Funktion mit additivem Offset.

In dieser Arbeit wird die oft eingesetzte Sigmoid-Funktion mit $\alpha = 1$ genutzt, wegen der durch sie vollzogenen Approximation eines idealen Schwellwertentscheiders mit dem Vorteil der Differenzierbarkeit, welche für das später gezeigte Lernen von Nöten ist. Abb. 3.23 zeigt zur Veranschaulichung den Verlauf ausgewählter Transferfunktionen.

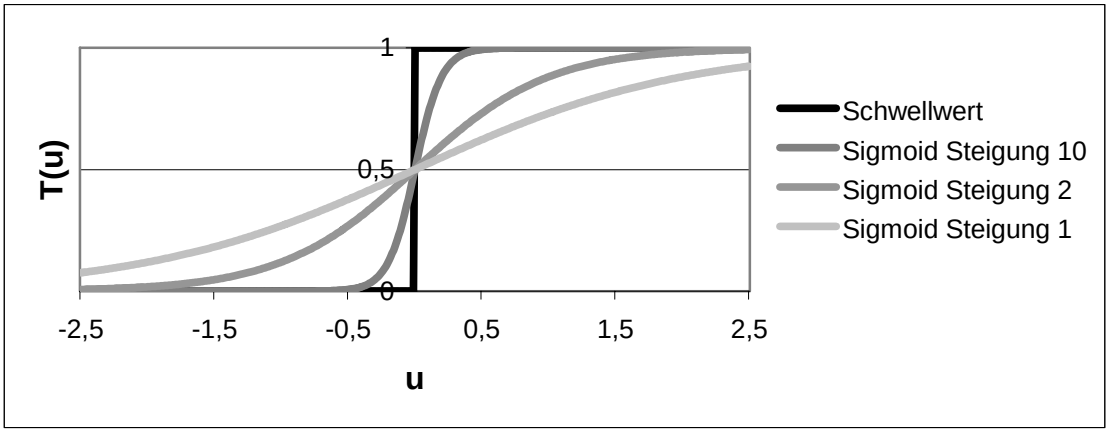


Abb. 3.23: Beispiele verwendeter nichtlinearer Transferfunktionen in Neuronen

Als Vertreter einer Vielzahl existierender Realisierungen von ANN wird hier das *Mehrschicht-Perzeptron*, kurz *MLP*[69], vorgestellt und eingesetzt - eine Standardvariante Neuronaler Netze, die sich großer Verbreitung zu Mustererkennungsaufgaben erfreut. Die Knoten sind im MLP in mehreren Schichten geordnet. Verbindungen bestehen nur zwischen adjazenten Schichten, und der Informationsfluss ist vorwärts gerichtet[70]. In der gewählten Topologie besteht eine vollständige Verbindung der Knoten, was jedoch nicht zwingend notwendig ist. Die erste Schicht ist die Eingangsschicht und besteht aus N Knoten für die Eingabewerte x_i mit $i = 1,...,N$ des Merkmalsvektors $\underline{x}$. In ihr erfolgt keine Verarbeitung, weswegen sie oft nicht als gesonderte Schicht gezählt wird. Anschließend können mehrere verborgene Schichten folgen. Für die Definition beliebiger Klassengrenzen im $\mathbb{R}^N$ ist jedoch nachweisbar *eine* verborgene Schicht ausreichend.

Die letzte Schicht ist schließlich die Ausgangsschicht. An den Ausgängen ihrer Neuronen y_j mit $j = 1,...,M$, die im Ausgangsvektor $\underline{y} = \{y_j\}$ zusammengefasst werden, findet sich das Klassifikationsergebnis des MLP. Abb. 3.24 zeigt ein MLP mit einer verborgenen Schicht.

Es ist eine entsprechende Kodierung des Ausgangsvektors erforderlich, um alle Klassen abzubilden. Im später betrachteten Fall von $k = |\Omega|$ Emotionsklassen Ω_κ werden k Ausgangsneuronen gewählt. Jede Zielvariable korrespondiert dabei direkt mit einer Klasse mit den binären Werten Null oder Eins, falls die zugehörige Klasse vorliegt. Man spricht auch von *1-aus-k* Kodierung. Vorteilig dabei ist die Möglichkeit zur Bewertung der Sicherheit der gewählten Klasse. Im Besonderen bietet sich die sogenannte *Softmax-Funktion* als Transferfunktion in den Ausgangsneuronen an, die die Summe der Ausgänge zu Eins normiert, um diese als a-posteriori Klassenwahrscheinlichkeiten

[69] Die Abkürzung leitet sich aus der anglistischen Bezeichnung *Multi-Layer-Perceptron* ab.

[70] Dieses Verhalten kennzeichnet der Begriff des *Feed-Forward-Netzes* im Gegensatz zu rekurrenten Netzen, die Rückkopplungen enthalten können. Letztere werden in Kap. 4.3.1 vorgestellt.

$P\left(\Omega_\kappa \left| \underline{x} \right.\right)$ interpretierbar zu gestalten[71]:

$$P\left(\Omega_\kappa \left| \underline{x} \right.\right) = y_\kappa = \frac{e^u}{\sum_{j=1}^{M} e^u} \text{ mit } M = k = \left|\Omega\right| \tag{3.60}$$

In der Erkennungsphase erfolgt die Berechnung des Ausgangsvektors $\underline{y}$ schichtenweise von der Eingangs- bis zur Ausgangsschicht. Je Schicht wird dabei für jeden Knoten die gewichtete Summe der Eingänge der vorhergehenden Schicht gebildet und mit der Nichtlinearität $T\left(u\right)$ gewichtet. Die erkannte Klasse κ_e wird bei Wahl der Softmax-Funktion und der beschriebenen Kodierung folgendermaßen bestimmt:

$$\kappa_e = \underset{\kappa}{\text{argmax}}\, P\left(\Omega_\kappa \left| \underline{x} \right.\right) \tag{3.61}$$

Alternativ kann beispielsweise eine binäre Kodierung, mit entsprechend $\text{ld}\left(k\right)$ Ausgangsneuronen, gewählt werden.

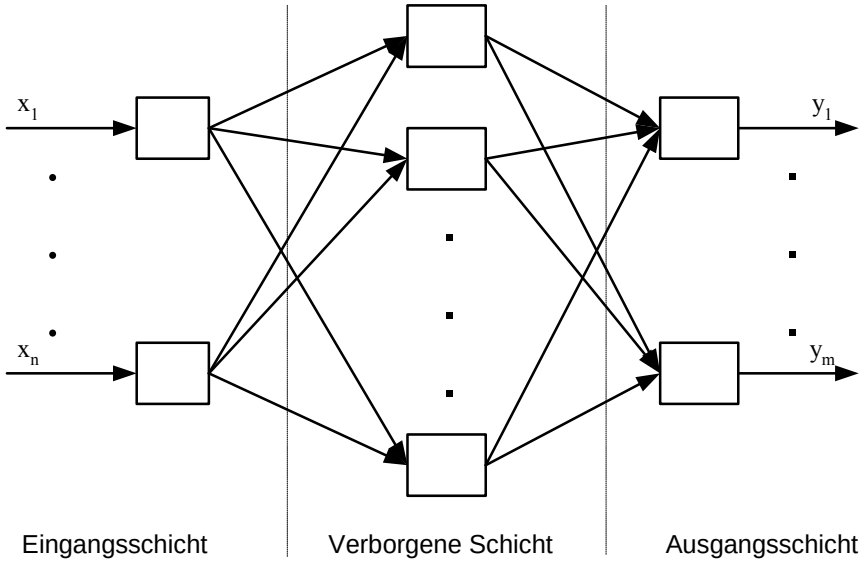


Abb. 3.24: Struktur eines Mehrschicht-Perzeptrons mit einer verborgenen Schicht

Es existiert eine Vielzahl von Lernverfahren, um die Gewichte des Netzes zur Klassifikation zu trainieren. Hier wird die weit verbreitete *Backpropagation*[72], ein Gradientenabstiegsverfahren,

[71] Es werden somit die Stochastizitätsbedingungen $\sum_{\kappa=1}^{k} P\left(\Omega_\kappa \left| \underline{x} \right.\right) = 1$ und $0 \le P\left(\Omega_\kappa \left| \underline{x} \right.\right) \le 1$ erfüllt.

[72] Auch als *allgemeine Delta-Regel* bekannt. Modernere, auf Geschwindigkeit oder Fehlerminimierung optimierte Varianten hierzu sind *Quickprob* oder *RProb (Resilient Propagation)*.

vorgestellt [RUM87], welches den Durchbruch für den erfolgreichen Einsatz von ANN mit sich brachte. Die Gesamtheit der Gewichte $\underline{w}_j$ einer Schicht mit $j = 0,...,I$ und I als Zahl der Neuronen in dieser Schicht seien in der Matrix $\underline{\underline{W}} = \left\{\underline{w}_j\right\}$ für diese Schicht zusammengefasst. Als Gütefunktion $F\left(\underline{x},\underline{\underline{W}}\right)$, zur Kontrolle des Lernfortschritts im Sinne eines überwachten Lernens, wird der mittlere quadratische Fehler *MSE*[73] zwischen der Zielvorgabe des bekannten Musters $f\left(\underline{x}\right)$ eines Merkmalsvektors $\underline{x}$ aus der Lernmenge $\mathcal{L}$ und dem korrespondierenden Ausgang des Netzes $y_j\left(\underline{x},\underline{\underline{W}}\right)$ bestimmt:

$$F\left(\underline{x},\underline{\underline{W}}\right) = \left|f\left(\underline{x}\right) - y\left(\underline{x},\underline{\underline{W}}\right)\right|^2 \tag{3.62}$$

Das Gradientenabstiegsverfahren rückwärts gerichteter Propagation zur Netzoptimierung lässt sich nach einer Initialisierung, die hier zufällig erfolgt[74], in drei sich jeweils wiederholende Schritte einteilen:

- Vorwärtspass als normaler Durchlauf
- Bestimmung des MSE nach Gl. 3.62
- Rückwärtspass mit Gewichtsanpassung durch den Korrekturterm Δw_{ij}:

$$w_{ij} \rightarrow w_{ij} + \Delta w_{ij} = w_{ij} - \beta \cdot \frac{\partial F\left(\underline{x},\underline{\underline{W}}\right)}{\partial w_{ij}} \tag{3.63}$$

mit der empirisch festzulegenden Schrittweite β und der Abstiegsrichtung $\frac{\partial F\left(\underline{x},\underline{\underline{W}}\right)}{\partial w_{ij}}$.

Abbruchkriterium dieses iterativen Ablaufs ist entweder eine maximale Schleifenzahl, oder eine minimale Änderung im Fehler. Besonders bei hoher Dimensionalität ist einem geeigneten Kriterium sorgfältig Rechnung zu tragen. Eine ausführliche Darstellung hierzu findet sich in [SAK94].

Günstige Parameter bei der Gestaltung des Netzes können leider nur empirisch ermittelt werden, was einen entscheidenden Nachteil von ANN darstellt. Es gibt jedoch Ansätze zur Automatisierung beim Netzwerkdesign. Um einem sogenannten *Overfitting* – einer Überadaption des Netzes – vorzubeugen, muss die Zahl der Parameter im Netz der Zahl der Lernbeispiele und der Dimensionalität gerecht werden. Sieht man zu viele Parameter, durch die Anzahl der Neuronen in der verborgenen Schicht, oder mehrere verborgene Schichten vor, geht die

---

[73] Abkürzung für *Mean Square Error*. In der Praxis werden auch weitere Fehlermaße wie der *McClelland-Fehler* oder die Kreuzentropie verwendet.

[74] Dieses Verfahren hat sich in der Praxis bewährt. Die Biasgewichte sollten dem Trainingsverfahren eine möglichst optimale Ausnutzung des Dynamikbereichs der Netzausgänge erlauben. Ergänzend können diese üblicherweise so gewählt werden, dass die a-priori Klassenwahrscheinlichkeiten an den Ausgängen des Netzes vorliegen, insofern die Softmax-Funktion am Ausgang gewählt wurde. Dabei wird davon ausgegangen, dass sich die Ausgangssumme bei der Initialisierung ohne Bias nahe Null ergibt.

Generalisierungsfunktion durch zu niedrige Bündelung verloren. Hierzu kann als Gütemaß im Training die Erkennungsleistung auf einem weiteren, disjunkten Set verwendet werden. Ein klarer Vorteil Neuronaler Netze ist ihre Fähigkeit mit redundanter Eingangsinformation umzugehen, da entsprechende Gewichte nahe Null gelernt werden. Außerdem erfolgt das Lernen diskriminativ, das heißt es wird die Information über alle Klassen gleichzeitig gelernt, und somit eine maximale Abgrenzung zwischen den Klassen erzielt. [RIG99]

3.7.4 Entscheidungsbäume

Als weitere Klassifikatoren sollen Entscheidungsbäume, abgekürzt *DT*[75], kurz vorgestellt werden. Ausgehend von einem *Wurzelknoten* verzweigt sich ein DT in Analogie zu einem realen Baum in mehrere *Äste* und *Knoten*. Die Knoten korrespondieren dabei zu Attributen, beziehungsweise Merkmalen. Die Äste hingegen repräsentieren Werte der jeweiligen Attribute. In der Erkennungsphase wird der beobachtete Attributswert mit den Werten die von einem Knoten zum jeweiligen Nachbarn reichen verglichen, und der entsprechende Ast gewählt. In den Terminalknoten, auch als *Blätter* bezeichnet, findet sich dem Pfad von der Wurzel aus folgend das Ergebnis der Klassifikation, respektive die zugehörige Klasse. Ein guter DT ist dadurch definiert, dass aus dem Set verbleibender Attribute dem Pfad der Wurzel entlang jeweils dasjenige Merkmal mit dem höchsten Informationsgehalt gemäß Shannon folgt [QUI93]. Das informationstheoretische Maß hierfür ist die *Shannon-Entropie* [SHA63]. Hierzu wird folgende allgemeine Definition der Entopie $H(P)$ einer Klassenverteilung $P(p_1,...,p_k)$ verwendet[76]:

$$H(P) = -\sum_{i=1}^{k} p_i \cdot ld(p_i) \tag{3.64}$$

Die notwendige mittlere Informationsmenge $H(\mathcal{L})$ um ein Element einer Datenmenge $\mathcal{L}$ einer Klasse $\Omega_\kappa \in \Omega$ zuordnen zu können, wird, unter Einführung der Menge $\mathcal{L}_\kappa \in \mathcal{L}$ aller Lernbeispiele der Klasse mit Index κ, bestimmt gemäß

$$H(\mathcal{L}) = -\sum_{\kappa=1}^{k} p_\kappa \cdot ld(p_\kappa) \text{ mit } p_\kappa = \frac{|\mathcal{L}_\kappa|}{|\mathcal{L}|} \tag{3.65}$$

Um nun den Beitrag eines einzelnen Merkmals x zur angestrebten Klassenzuordnung herauszufinden, wird zunächst die Menge $\mathcal{L}$ auf Basis der unterschiedlichen Werte des betrachteten Attributs in die Untermengen $\mathcal{L}_{x,j}$ mit $j = 1,...,J$ unterteilt. Dabei ist J die Zahl unterschiedlicher Attributwerte unter der Annahme diskreter Werte. Im Falle eines kontinuierlichen Wertebereichs erfolgt unter gesteigertem Rechenaufwand eine Quantisierung in Intervalle. Die nach Beobachtung des Merkmals x noch zur Klassenzuordnung benötigte mittlere Information $H(\mathcal{L}|x)$ ergibt sich als

---

[75] Diese Kurzform entspringt der anglistischen Bezeichnung *Decision Tree*.

[76] Bei k gleichwahrscheinlichen Nachrichten besitzt die Wahrscheinlichkeit einer Nachricht den Wert $1/k$ und die durch sie erbrachte Information den Wert ld(k). Weiterhin sei angemerkt, dass der Informationsgehalt einer sicheren Nachricht zu Null definiert ist.

gewichteter Mittelwert der benötigten Information $H\left(\mathcal{L}_{x,j}\right)$. Letztere Information ist dabei diejenige benötigte mittlere Information um ein Element der Untermenge $\mathcal{L}_{x,j}$ klassifizieren zu können.

$$H(\mathcal{L}|x) = \sum_{j=1}^{J} \frac{\left|\mathcal{L}_{x,j}\right|}{\left|\mathcal{L}\right|} \cdot H\left(\mathcal{L}_{x,j}\right) \tag{3.66}$$

Der Begriff *Information-Gain*, kurz *IG*, beschreibt, welcher Gewinn in Bezug auf die benötigte Information zur Zuordnung durch ein Merkmal *x* erzielt wird:

$$IG\left(\mathcal{L},x\right) = H\left(\mathcal{L}\right) - H\left(\mathcal{L}|x\right) \tag{3.67}$$

Diese Definition neigt jedoch dazu, Attribute mit einer hohen Zahl unterschiedlicher Werte *J* zu bevorzugen. Sollte ein Merkmal für jedes Datum der Menge $\mathcal{L}$ einen unterschiedlichen Wert aufweisen, ergibt sich $H(\mathcal{L}|x)$ zu Null, und damit ein maximales $IG\left(\mathcal{L},x\right)$. Dies kann durch die Einführung der *Information-Gain-Ratio IGR* kompensiert werden:

$$IGR\left(x,\mathcal{L}\right) = \frac{IG\left(\mathcal{L},x\right)}{H\left(\frac{\left|\mathcal{L}_{x,1}\right|}{\left|\mathcal{L}\right|},\ldots,\frac{\left|\mathcal{L}_{x,j}\right|}{\left|\mathcal{L}\right|},\ldots,\frac{\left|\mathcal{L}_{x,J}\right|}{\left|\mathcal{L}\right|}\right)} \tag{3.68}$$

Der Term im Nenner wird auch als *Splitinformation* bezeichnet, und ist diejenige Information, die man aus der beschriebenen Teilung der Menge $\mathcal{L}$ durch die Werte des Attributs *x* erhält.

Um einen Entscheidungsbaum zu erzeugen, kann das grundlegende *ID3* Verfahren nach [QUI83], wie in Abb. 3.25 gezeigt, verwendet werden. Der im Verlauf der Experimente angewandte C4.5 Algorithmus beinhaltet Erweiterungen dieses Basisalgorithmus zur Erstellung von DT: Durch ein sogenanntes *Pruning* kann ein Abschneiden von Teilbäumen erfolgen [QUI87], bei welchem ganze Unterbäume durch ein Blatt ersetzt werden, falls die Fehlerrate für einen Unterbaum höher als für das Blatt ist. Hierdurch reduziert sich einerseits die Komplexität des Baumes, andererseits aber auch die Zahl zu extrahierender Merkmale. Sowohl im Training, als auch in der Klassifikation können DT mit unvollständigen Merkmalsvektoren umgehen. In der Lernphase können die IG oder IGR Werte der Attribute aus Daten, in denen diese besetzt sind, berechnet werden. In der Erkennungsphase kann bei Fehlen eines Merkmals eine Schätzung über die Ausgangswahrscheinlichkeit anhand der gegebenen Werte geleistet werden.

ID3 (*X*: Merkmalsset, κ: Klasse, $\mathcal{L}$: Lernvektormenge**)** {

FALLS $\mathcal{L}$ leer **RÜCKGABE** singulärer Knoten mit Wert Fehler;
FALLS κ gleich für alle Daten in $\mathcal{L}$ **RÜCKGABE** singulärer Knoten mit Klasse κ*;*
FALLS *X* leer **RÜCKGABE** singulärer Knoten mit häufigster Klasse in $\mathcal{L}$;
// hierdurch können Fehler entstehen
SEI $x' = \arg\max_x IG(\mathcal{L}, x)$;
SEI $\{x_i' | i = 1,...,I\}$ Menge der in $\mathcal{L}$ enthaltenen Werte von *x';*
SEI $\{\mathcal{L}_i' | i = 1,...,I\}$ Menge der Subsets von $\mathcal{L}$,
die für x' nur die korrespondierenden Werte x_i' enthalten;
FÜR x_i' = 1 bis I **TUE** {
FÜR $\mathcal{L}_i'$ = 1 bis I **TUE** {
RÜCKGABE ID3 (*X*-{*x'*}, κ, $\mathcal{L}_i$**)**
// Gibt einen Baum mit *x'* in der Wurzel und den Ästen x_i' die zu den
// Bäumen ID3 (*X*-{*x'*}, κ, $\mathcal{L}_i$*)* führen aus
}
}
}

Abb. 3.25: Rekursiver ID3 Lernalgorithmus für Entscheidungsbäume in Pseudo-Code Darstellung

Abb. 3.26 zeigt einen durch C4.5 erhaltenen Entscheidungsbaum der Größe elf, wobei jeder Knoten inklusive der sechs Blätter gezählt wird. Der Baum erlaubt die Erkennung von Freude mit sechs akustischen Merkmalen abgeleitet aus der Sprachgrundfrequenz[77].

Das vorgestellte Beispiel zeigt auch, dass durch aggressives Pruning die Zahl der Merkmale reduziert werden kann. Im Baum sind die Merkmale *relative Position des Maximums der Grundfrequenz* sowie *Wertebereich der Grundfrequenz* nicht mehr enthalten. Wird nur die Wurzel des Baumes beibehalten, etwa durch extremes Pruning, spricht man vom sogenannten *Decision-Stump*. Obwohl dieser alleine selten eine vernünftige Erkennung erlaubt, spielt er dennoch eine bedeutende Rolle im Rahmen der *Ensembleklassifikation* (vgl. Kap. 3.8), bei der schwache Lernverfahren zu starken vereint werden. Generell können aus DT Regeln für eine regelbasierte Klassifikation abgeleitet werden. Dies erlaubt eine sehr anschauliche Darstellung des Ablaufs einer maschinellen Entscheidungsfindung.

[77] Für diesen Entscheidungsbaum mit starkem Pruning wurden 609 Lernbeispiele von 11 Sprechern aus dem in Kap. 3.12.1 vorgestellten Korpus EA-ACT verwendet. Die Erkennungsrate belief sich auf 92,1% in einer 10-fach stratifizierten Kreuzvalidierung.

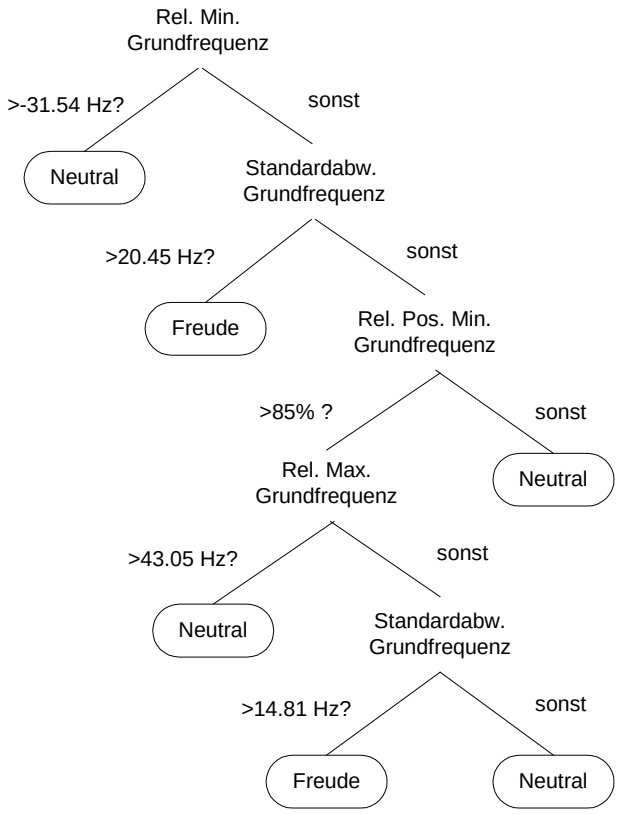


Abb. 3.26: Beispiel eines Entscheidungsbaumes

3.7.5 Naive-Bayes-Klassifikator

Die Namensgebung des statistischen *Naive-Bayes- (NB)* Klassifikators beruht auf der Anwendung des Bayesschen Satzes und der stark naiven Annahme, dass beobachtete Attribute bei gegebener Klasse statistisch bedingt voneinander unabhängig sind und keine versteckten oder latenten Attribute die Schätzung beeinflussen. Numerische Attribute werden dabei als vielen Problemen der realen Welt zugrunde liegende einfache Normalverteilung modelliert [JOH95]. Leider erweist sich die getroffene Beschränkung oft als nicht ausreichend. Als Vorteile des NB sind jedoch eine hohe Effizienz für Trainings- und Testläufe sowie hohe Transparenz zu nennen.

Ziel ist es, für einen Vektor $\underline{x}$ der Klasse Ω_κ die a-posteriori Wahrscheinlichkeit $P\left(\Omega_\kappa \middle| \underline{x}\right)$ zu maximieren. Anstelle dieser schwer bestimmbaren bedingten Wahrscheinlichkeit kann mit dem Satz von Bayes gefolgert werden:

$$P\left(\Omega_\kappa \mid \underline{x}\right) = \frac{P\left(\underline{x} \middle| \Omega_\kappa\right) \cdot P\left(\Omega_\kappa\right)}{P\left(\underline{x}\right)} \tag{3.69}$$

Die Maximierung von $P\left(\Omega_\kappa \middle| \underline{x}\right)$ unter Anwendung des Bayesschen Satzes wird als *Maximum-*

Likelihood[78] Methode bezeichnet. Der Term $P(\underline{x})$ im Nenner der rechten Seite von Gl. 3.69 wird hier als konstant betrachtet, und daher zu Gunsten einer Normierung der Summe der Wahrscheinlichkeiten $P(\Omega_\kappa|\underline{x})$ über alle Klassen zu Eins vernachlässigt. Für $P(\Omega_\kappa)$ wird entweder die reale Verteilung der Klassen κ, oder unter der Annahme einer Gleichverteilung vereinfachend $P(\Omega_\kappa)=k^{-1}$ gewählt. Ausgehend von der Hypothese der statistischen Unabhängigkeit der Attribute erhält man für $P(\underline{x}|\Omega_\kappa)$ mit den Komponenten x_i des Vektors $\underline{x}$ und $i=1,...,N$:

$$P(\underline{x}|\Omega_\kappa)=\prod_{i=1}^{N}P(x_i|\Omega_\kappa) \tag{3.70}$$

Die bedingten Wahrscheinlichkeiten $P(x_i|\Omega_\kappa)$ ergeben sich im diskreten Fall direkt aus der Auszählung der Fälle, in denen das Attribut x_i einen speziellen Wert bei gegebener Klasse Ω_κ besitzt. Für kontinuierlich dargestellte Merkmalsgrößen kann eine ebenfalls kontinuierliche Wahrscheinlichkeitsdichtefunktion *(WDF)*[79] $p(x_i=x|\Omega_\kappa)$ zur Modellierung angewandt werden. Im allgemeinen Fall des NB handelt es sich um eine einfache Normalverteilung $g(x,\mu_\kappa,\sigma_\kappa)$, die durch die beiden leicht bestimmbaren Parameter Mittelpunkt μ_κ und Standardabweichung σ_κ für die jeweilige Klasse Ω_κ definiert[80] ist:

$$p(x_i=x|\Omega_\kappa)=g(x,\mu_\kappa,\sigma_\kappa)=\frac{1}{\sqrt{2\pi}\sigma_\kappa}\cdot e^{-\frac{(x-\mu_\kappa)^2}{2\sigma_\kappa^2}} \tag{3.71}$$

Die Maximum-Likelihood Schätzwerte für diese beiden Parameter entsprechen dabei direkt den korrespondierenden Größen der Stichproben aus der Lernmenge $\mathcal{L}$.

Im Entscheidungsprozess wird die Klasse κ_e mit der höchsten Wahrscheinlichkeit für die Beobachtung $\underline{x}$ gewählt:

$$\kappa_e=\underset{\kappa}{\operatorname{argmax}}\left\{p(\underline{x}\,|\,\Omega_\kappa)\cdot P(\Omega_\kappa)\right\} \tag{3.72}$$

---

[78] Der Begriff der *Likelihood* ist die gebräuchliche Abkürzung für eine *Likelihood-Funktion L*. Diese ist proportional zu einer bedingten Wahrscheinlichkeit *P*(*A*|*B*), wird aber wie hier gezeigt unter Einbezug des Bayesschen Satzes umgekehrt als *L*(B|*A*)α*P*(*A*|B) gesehen. Dies bedeutet, dass das erste Argument als fixiert betrachtet wird.

[79] Im anglistischen Sprachgebrauch als *PDF* für *Probability Density Function* abgekürzt.

[80] Unter Berücksichtigung der Tatsache, dass eine reelle Zufallsvariable nur mit der Wahrscheinlichkeit Null einen konkreten Wert annimmt, wäre eine mathematisch korrekte Schreibweise der Gl. 3.71 nur in Intervallnotation möglich. Auf Grund der gewählten Normierung kann hierauf jedoch verzichtet werden.

3.7.6 Bayessche Netze

Bayessche Netze[81], kurz *BN*, haben sich zu einem beliebten mathematischen Beschreibungsmodell für wahrscheinlichkeitsbasiertes Behandeln von Unsicherheiten entwickelt. Sie erlauben als Besonderheit auch den Umgang mit unvollständigem Wissen. Mit BN kann sowohl Experten- und Domänenwissen modelliert werden, als auch ein datengetriebenes statistisches Lernen der Netztopologien sowie der bedingten Wahrscheinlichkeiten innerhalb des Netzes erfolgen. Der ihnen zugrunde liegende grafische Modellierungsansatz bringt zusätzlich den Vorteil einer anschaulichen Repräsentation von gelernten oder zu beschreibenden Zusammenhängen mit sich.

Ein BN besteht aus einem gerichteten azyklischen Graph[82], der sich aus Knoten, die Zufallsvariablen X_n mit $n = 1,...,N$ repräsentieren, und gerichteten Kanten, die bestehende statistische Abhängigkeiten zwischen jeweils einem Knotenpaar beschreiben, zusammensetzt [JEN96]. Den einzelnen Knoten des Netzes ist eine Verteilung der bedingten Wahrscheinlichkeiten vertretener Zufallsvariablen zugeordnet. Man nennt diese Verteilung auch *Conditional-Probability-Tables*. Kanten werden so angeordnet, dass der Kopf zu derjenigen Zufallsvariablen, deren Dichte angegeben wird, und das Ende zu derjenigen, die in der Dichte als Bedingung auftritt, zeigt. Man unterscheidet entsprechend zwischen *Eltern-* und *Kindknoten* eines Knotens n, respektive $E_{n,i}$ mit $i = 1,...,I$ und $K_{n,j}$ mit $j = 1,...,J$. Knoten ohne Eltern besitzen als assoziierte Wahrscheinlichkeitsverteilung unbedingte Verteilungen. Während die gesamten Verteilungen eines BN prinzipiell beliebiger Natur und sowohl kontinuierlich als auch diskret sein dürfen, werden in der Praxis Normalverteilungen (vgl. Kap. 3.7.5) oder diskrete Verteilungen bevorzugt. Die grundlegende Modellierungsaufgabe eines BN ist somit eine optimal kompakte Repräsentation der gemeinsamen Wahrscheinlichkeitsverteilung $P(X_1,...,X_N)$ beteiligter Variablen unter Anwendung bekannter Unabhängigkeiten. Sie wird *Joint-Probability-Distribution (JPD)* genannt und ist von exponentieller Größe. Jeder Eintrag in der JPD kann aus der im Netzwerk enthaltenen Information mittels folgender Kettenregel bestimmt werden:

$$P(X_1,...,X_N) = \prod_{n=1}^{N} P\left(X_n \middle| Eltern(X_n)\right) \tag{3.73}$$

In Abb. 3.27 ist als Beispiel der in Kap. 3.7.5 vorgestellte NB Klassifikator als triviales BN dargestellt:

---

[81] Auch als *Belief-*, *Probabilistic-* oder *Causal Networks* bezeichnet. In jüngerer Zeit scheinen sich jedoch der Begriff Bayesian Network und die hier gewählte Abkürzung BN als geläufigste Form heraus zu kristallisieren.

[82] In der anglistischen Literatur als *Directed-Acyclic-Graph*, kurz *DAG*, bezeichnet.

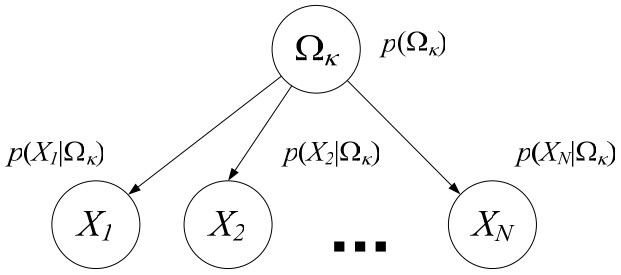


Abb. 3.27: Darstellung eines NB Klassifikators als Beispiel eines Bayesschen Netzes

Durch Einbringen neuen Wissens in ein Netz, *Evidenz* genannt, werden einzelne Knoten instantiiert [PEA88]. Die Evidenz wird von den instantiierten Knoten ausgehend durch das gesamte Netz propagiert. Hierfür existieren mehrere Propagationsalgorithmen, die sich in erster Linie in ihrer Effizienz unterscheiden [GUO02]. Für Graphen mit singulärer Verbindung der Knoten, das heißt jeder Knoten kann höchstens auf einem Pfad von einem anderen aus erreicht werden, sind Propagationsalgorithmen von polynomialer Ordnung in Bezug auf die Gesamtzahl von Knoten. Eine exakte Aktualisierung des Netzes ist jedoch oft NP-hart [GUO02]. Im Folgenden werden Baumstrukturen für die Netze gewählt, bei denen die Propagation mittels einer Nachricht π, die das Netz entlang der Pfeilrichtung durchläuft, und einer entgegengesetzt laufenden Nachricht λ, erfolgt [PEA88]. π-Nachrichten modellieren dabei *kausalen*[83] Einfluss im Netz, da durch Wissen über Knoten weiter oben im Netz Aussagen über weiter unten liegende getroffen werden können. λ-Nachrichten hingegen vertreten den *diagnostischen*[84] Informationsanteil, da durch Information über hierarchisch tiefer gelegene Knoten über höher angeordnete schlussgefolgert werden kann. Der Wert $\pi(X_n)$ des Knotens n ergibt sich mit den π-Nachrichten $\pi_n(E_{n,i})$, die von seinen Elternknoten $E_{n,i}$ gesendet wurden, zu:

$$\pi(X_n) = \sum_{i=1}^{I} P(X_n | E_{n,i}) \cdot \pi_n(E_{n,i}) \tag{3.74}$$

Dabei ist $P(X_n | E_{n,i})$ die bedingte Wahrscheinlichkeit der Kante $E_{n,i} \to X_n$. Auf Grund der hier gewählten Baumstruktur existiert jeweils nur ein Elternknoten E_n zu einem Knoten n. Der Wert $\lambda(X_n)$ des Knotens n ergibt sich mit den λ-Nachrichten $\lambda_{K_{n,j}}(X_n)$ seiner Kinder $K_{n,j}(X)$ zu:

$$\lambda(X) = \prod_{j=1}^{J} \lambda_{K_{n,j}}(X) \tag{3.75}$$

Somit ergibt sich die als $BEL(X_n)$[85] bezeichnete Wahrscheinlichkeit eines Knotens n mit einem

[83] Man spricht auch von *Predictive Support*.

[84] Auch als *Retrospective Support* bezeichnet.

[85] Diese Bezeichnung kommt von der Abkürzung für englisch *Belief*. Hieraus ergibt sich auch die erwähnte alternative Bezeichnung von BN als Belief Networks.

Normalisierungsfaktor α, der die Summe von $BEL(X_n)$ über alle Knoten auf Eins korrigiert, zu:

$$BEL(X_n) = \alpha \cdot \lambda(X_n) \cdot \pi(X_n) \tag{3.76}$$

Hierbei handelt es sich um folgende verallgemeinerte Form des Satz von Bayes:

$$P(X|E) = \alpha \cdot P(E|X) \cdot P(X) \tag{3.77}$$

Die von einem Knoten weiterzuleitenden Nachrichten ergeben sich folgendermaßen:

$$\pi_{K_j}(X) = \alpha \cdot \pi(X) \cdot \prod_{k \neq j} \lambda_{K_k}(X) \tag{3.78}$$

$$\lambda_X(E) = \sum_X \lambda(X) \cdot P(X|E) \tag{3.79}$$

Der Ablauf der Propagation verläuft insgesamt je Knoten n des Netzes in folgenden Schritten: Berechnung von $\lambda(X_n)$, $\pi(X_n)$ und $BEL(X_n)$, Senden der λ-Nachricht an alle Eltern und Senden der π-Nachricht an alle Kinder. In Abb. 3.28 wird beispielhaft der Ablauf einer Propagation dargestellt.

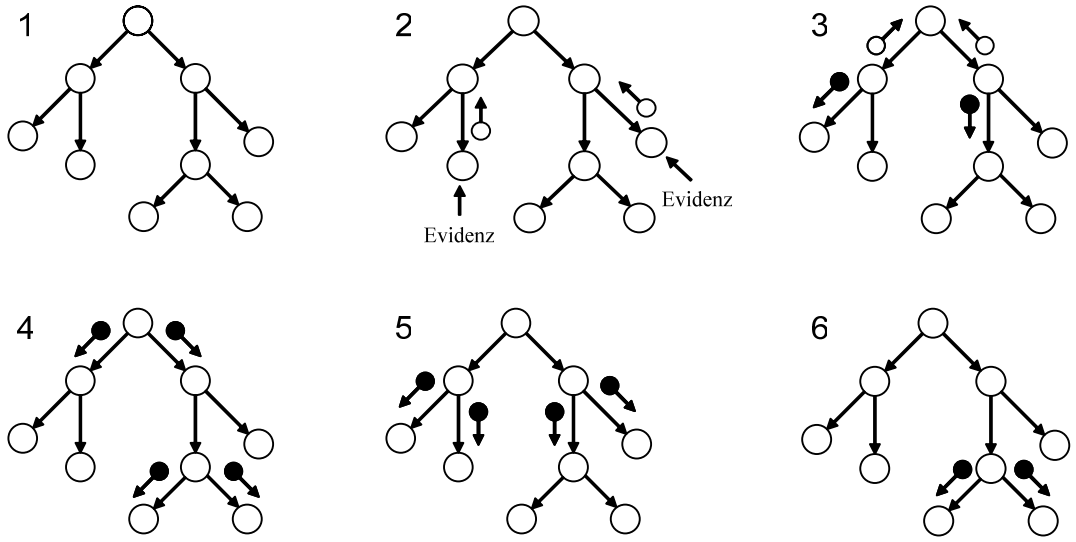


Abb. 3.28: Propagation von Nachrichten in einem BN. Ausgefüllte Kreise mit Pfeil zeigen π-Nachrichten, unausgefüllte λ-Nachrichten.

Nach Einbringung und Propagation des Wissens kann bei den gewählten baumförmigen Netztopologien die Wahrscheinlichkeit für eine Klasse im zugehörigen Wurzelknoten abgelesen werden. Dabei wird hier entweder für jede Klasse ein eigenes Netz, oder für jede Klasse innerhalb des gleichen Netzes ein eigener Wurzelknoten vergeben. Es wird dann die Klasse κ_e mit zugehöriger höchster Wahrscheinlichkeit $P(\Omega_\kappa)$ gewählt:

$$\kappa_e = \underset{\kappa}{\operatorname{argmax}}\, P(\Omega_\kappa) \tag{3.80}$$

In der Trainingsphase müssen die entsprechenden Wahrscheinlichkeiten, beziehungsweise WDF approximiert werden. Dies geschieht analog zu den in Kap. 3.7.5 und Kap. 3.11.2 vorgestellten Verfahren[86]. Wie auch bei der Wahl der Topologie von BN, kann hierbei ebenfalls Expertenwissen einfließen.

3.8 Ensemble-Klassifikation

Im Folgenden wird versucht die Erkennungsleistung, die mit bisher vorgestellten Klassifikationsverfahren erzielt werden kann, durch eine Kombination dieser in multiplen Instanzen zu steigern. Man spricht dann von sogenannten *Ensembles* oder *Committees* von Klassifikatoren. Ziel ist es, den mittleren quadratischen Fehler ε eines Klassifikators zu reduzieren. Dieser lässt sich als Komposition von Bias und Varianz darstellen, wobei $\hat{\theta}$ den Schätz- und θ den tatsächlichen Wert repräsentiert:

$$\varepsilon = E\left\{\left(\hat{\theta}-\theta\right)^2\right\} = \underbrace{\left[E\left(\hat{\theta}\right)-\theta\right]^2}_{\text{Quadratischer Bias}} + \underbrace{E\left\{\left[\hat{\theta}-E\left(\hat{\theta}\right)\right]^2\right\}}_{\text{Varianz}} \tag{3.81}$$

Idealerweise sollten beide Anteile reduziert werden. In der Regel wird jedoch entweder der Bias oder die Varianz maßgeblich beeinflusst. Instanzen von Klassifikatoren werden hierbei als Modelle $\mathcal{M}_m \in \mathcal{M}$ mit der Menge der Modelle $\mathcal{M}$ bezeichnet. Je nach Ensemblemethode werden unter Umständen mehrere Modelle des gleichen Verfahrentyps, basierend auf verschiedenen Trainingssets, zur Klassifikation erzeugt. Eine Entscheidung erfolgt dann meist auf Basis eines Mehrheitsentscheids, wobei Einzelergebnisse auch gewichtet werden können. Grundsätzlich stellt sich dabei somit die Aufgabe der Konstruktion von Ensembles und der Kombination aller Modelle. In [BRE96] wird gezeigt, dass sich für Klassifikatoren, die instabil auf Änderungen des Trainingssatzes reagieren, signifikante Steigerungen in der Leistung durch eine Aggregation erzielen lassen. Beispiele solcher Klassifikatoren sind Neuronale Netze, Entscheidungsbäume oder regelbasierte Modelle. Hingegen erweisen sich Nächste-Nachbarn-Verfahren, Naive-Bayes oder lineare Regression als stabil in dieser Hinsicht und können durch Ensemblebildung sogar negativ beeinflusst werden. Auf Grund des hohen Rechenaufwands der durch das erforderliche mehrfache Trainieren desselben Klassifikators aufkommt, werden vorzugsweise sogenannte *Weak Classifiers*, also einfache Klassifikatoren, verwendet.

Es werden zunächst zwei geläufige einfachere Varianten vorgestellt: *Bagging*[87], welches vorrangig die Varianz minimiert, und *Boosting*, dass Bias und Varianz senkt, letztere aber nachweislich schwächer als beim Bagging [WEB00]. Im Anschluss daran wird eine komplexere Alternative, das

[86] NB, BN und HMM lassen sich gemeinsam mit der Theorie der *Grafischen Modelle* (vgl. Kap. 4.2.2) beschreiben.

[87] Abkürzung für *Bootstrap Aggregating*.

Stacking, beschrieben.

3.8.1 Bagging, Boosting und MultiBoosting

Beim Einsatz von Bagging [BRE96] werden die M Modelle mit $M = |\mathcal{M}|$ eines Ensembles durch Lernen eines Klassifikatortyps mit unterschiedlichen Variationen[88] des Trainingssatzes gebildet. Diese Variationen werden durch Sub-Sampling mittels Ziehen mit Zurücklegen aus dem gesamten Lernsatz erzielt. Üblicherweise und hier wird dabei eine der Gesamtheit des ursprünglichen Satzes entsprechende Kardinalität gewählt, wodurch im statistischen Mittel 63,2% der Trainingsdaten je Set vertreten sind, und der verbleibende Anteil aus Duplikaten besteht. Die finale Klassifikation erfolgt dann anhand eines ungewichteten Mehrheitsentscheids über die Ergebnisse der Modelle. Neben der Anzahl der gewählten Variationen spielt auch die Anzahl der Iterationen, in denen mehrfach unterschiedliche Sets gewählt werden, eine Rolle.

Bei der Durchführung von Boosting[89] [VAL84] wird zur Leistungsoptimierung ein gewichteter Mehrheitsentscheid anstelle des ungewichteten eingesetzt. Dabei werden Gewichte umgekehrt proportional zur Fehlerwahrscheinlichkeit gewählt, um damit eine Konzentration auf die am diffizilsten zu diskriminierenden Klassen zu erzwingen [FRE96]. In der Konsequenz muss die Konstruktion der Ensembles iterativ verlaufen. Bei diesem Verfahren erreicht man in der Regel eine Verbesserung gegenüber Bagging, in Einzelfällen können sich jedoch auch Verschlechterungen ergeben [QUI96].

In Abb. 3.29 wird der hier gewählte *AdaBoost*[90]-Algorithmus gezeigt. Es sind $L = |\mathcal{L}|$ die Zahl der Lernbeispiele des Trainingssets $\mathcal{L}$ und $\dim(\underline{x}_l)$ die Zahl der Merkmale eines Vektors $\underline{x}_l \in \mathcal{L}$, mit zugehörigem kategorialem Attribut y_l. Durch den Algorithmus werden darüber hinaus Gewichte w_l zu jedem Lernbeispiel $\underline{x}_l$ vergeben, die einheitlich mit $w_l = L^{-1}$ initialisiert werden. Der durch Boosting zu optimierende Klassifikator muss in der Lage sein, die Gewichtung der Lernbeispiele zu berücksichtigen. Dies erfolgt hier durch Sampling des Lernsets mit Vervielfachung einzelner Beispiele nach ihrem Gewicht. AdaBoost ist nur zur Diskrimination zwischen zwei Klassen mit $y_l \in \{+1, -1\}$ geeignet. Eine Behandlung von Mehrklassenproblemen erfolgt, wie in Kap. 3.7.2 für SVM gezeigt, durch Aufteilung des Problems in mehrere Zweiklassenentscheidungen.

Besonders für Klassifikatoren mit einer Leistung knapp über der Ratewahrscheinlichkeit werden durch Boosting sehr hohe Gewinne an Leistung erzielt. Dem gegenüber weist Boosting durch die Vergabe hoher Gewichte für fehl klassifizierte Instanzen, die eine zufällig richtige Erkennung für Rauschen leisten, eine Anfälligkeit für verrauschte Daten auf. Darüber hinaus besteht die Gefahr einer Überadaption und es ist eine hohe Zahl an Lernbeispielen erforderlich.

---

[88] Man spricht von *Bootsrap Replikationen*.

[89] Dies ist die ursprüngliche Bezeichnung des Verfahrens, die vom Autor bevorzugt wird. Später ist dieses auch unter *Arcing* bekannt geworden.

[90] Es handelt sich um die Variante *AdaBoostM1*.

AdaBoost ($\mathcal{L}$: Lernvektormenge, I : Iterationszahl**)** {
// Elemente von $\mathcal{L}$ sind Tupel $(\underline{x}_l, y_l)$ aus Lernvektor $\underline{x}_l$ und Klasse y_l
SEI $L = |\mathcal{L}|$**; SEI** $\mathcal{L}' = \mathcal{L}$ und $w_l = 1$ mit $l = 1,...,L$; // Initialisierung
FÜR $m = 0$ bis M **TUE** {
MARKE *Lernen*;
SEI $\mathcal{M}_m = \text{Lerne}(\mathcal{L}')$;
SEI $\varepsilon_m = \frac{\sum_{\underline{x}_l \in \mathcal{L}' : \mathcal{M}_m(\underline{x}_l) \neq y_l} w_l}{L}$**;** // Fehler: Summe Gewichte falsch klassifizierter Samples
FALLS $\varepsilon_m > 0,5$ {
SEI $\mathcal{L}'$ =BootstrapSample$(\mathcal{L})$**; SPRINGE** *Lernen*;
}
FALLS $\varepsilon_m = 0$ {
SEI $\beta_m = 10^{-10}$; **SEI** $\mathcal{L}'$ =BootstrapSample$(\mathcal{L})$**;**
} **SONST** {
SEI $\beta_m = \frac{\varepsilon_m}{1-\varepsilon_m}$**;**
FÜR JEDES $\underline{x}_l \in \mathcal{L}'$ {
FALLS $\mathcal{M}_m(x_l) \neq y_l$ { **SEI** $w_l = \frac{w_l}{2 \cdot \varepsilon_m}$**;** // Fehlklassifikation
} **SONST SEI** $w_l = \frac{w_l}{2 \cdot (1-\varepsilon_m)}$**;**
FALLS $w_l < 10^{-8}$ **SEI** $w_l = 10^{-8}$**;**
}
}
AUSGABE $\mathcal{M}_* = \underset{y}{\text{argmax}} \sum_{m:\mathcal{M}_m(x)=y} \log \frac{1}{\beta_m}$**;** // Finaler Klassifikator
}

Lerne ($\mathcal{L}$: Lernvektormenge**)** {... // Trainiert den Basisklassifikator auf der Lernmenge }

BootstrapSample ($\mathcal{L}$: Lernvektormenge**)** {
... // Bildet ein Bootstrap Sample entsprechend den Gewichten aus der Lernmenge
... // Setzt $w_l = 1$ mit $l = 1,...,L$;
RÜCKGABE $\mathcal{L}'$
}

Abb. 3.29: AdaBoostM1 Algorithmus als Pseudocode dargestellt

Um die Stärken der höheren Varianzminimierung des Bagging und die Biasreduktion des Boosting zu vereinen, können diese Verfahren sequentiell kombiniert werden. Eine parallele Erzeugung von Instanzen durch Bagging und Boosting hingegen ist auf Grund der unterschiedlichen Gewichtungsstrategien ausgeschlossen. In der Praxis werden so durch AdaBoost gebildete Subensembles anschließend durch Bagging erweitert, sodass Ensembles aus Subensembles entstehen. Anstelle von Bagging wird jedoch bevorzugt *Wagging* gewählt - eine Variante bei der sichergestellt ist, dass alle Lernbeispiele verwendet werden. Diese Verkettung nennt sich *MultiBoosting* und stellt in der Regel die effizienteste Variante dar [WEB00]. Als Parameter ergeben sich die Zahl und Größe der Subensembles. Standardgemäß werden hier $\sqrt{M}$ Subensembles der Größe $\sqrt{M}$ gebildet, woraus sich gesamt M Modelle ergeben.

3.8.2 Stacking und Voting

Über eine Leistungssteigerung eines einzelnen Klassifikatortyps hinaus führt das sogenannte *Stacking* [WOL92], dessen Prinzip es ist mehrere diverse Klassifikatoren auf den Lernproben zu trainieren, um ihre inhärenten Stärken zu vereinen und Schwächen zu kompensieren. Ein übergeordneter Klassifikator lernt dann aus den Ergebnissen der einzelnen Modelle welches Verfahren in welcher Entscheidungssituation zu bevorzugen ist. Aus dieser hierarchischen Gliederung erfolgen die Bezeichnungen *Base-Level* oder *Level-0-Klassifikator* für Einheiten, die direkt mit Daten in Berührung kommen, und *Meta-* oder *Level-1-Klassifikator* [TIN99] für die beschriebene übergeordnete finale Instanz, die nur Vorentscheidungen sieht. Analog entspricht die Benennung der jeweiligen Lerndaten ihrem Level. Um den Metaklassifikator mit Level-1-Daten trainieren zu können, wird eine stratifizierte j-fache Kreuzvalidierung (vgl. Kap. 3.12.3) vollzogen. Somit wird sichergestellt, dass jeweils disjunkte Datensets auf Level Null für die Konstruktion der Modelle und die nachfolgende Erzeugung der Level-1-Trainingsdaten verwendet werden. Die Wahl der Level-0-, als auch die des Level-1-Klassifikators, erfolgt in der Literatur bisher weitgehend explorativ und nach Erfahrungswerten, da eine vollständige Beschreibung der Zusammenhänge noch nicht vollzogen worden ist. Bewährt haben sich auf niederer Ebene vor allem Kombinationen von: NB, kNN, C4.5 Entscheidungsbäumen [TIN99] und Kernelmaschinen [SEE03], hier SVM.

Zur finalen Entscheidung haben sich diese Verfahren jedoch als wenig geeignet gezeigt. Vorrangig wird statt dessen, wie auch hier, Multiple Lineare Regression *(MLR)* gewählt. Diese unterscheidet sich von einer simplen linearen Regression nur durch die Verwendung mehrerer Eingangsvariablen. Im Falle einer Angabe der jeweils gewählten Siegesklasse auf niederer Ebene, ohne Bewertung über die Sicherheit, wird diese Entscheidung für die Regression numerisch in eine harte Konfidenz $P_{m,\kappa}\left(\underline{x}\right) \in \{0,1\}$ je Basisklassifikator m von M, mit dem Wert Null oder Eins je Klasse Ω_κ, umgesetzt. Folgende Gleichung zeigt die Berechnung der MLR je Klasse:

$$MLR_\kappa\left(\underline{x}\right) = \sum_{m=1}^{M} \alpha_{m,\kappa} P_{m,\kappa}\left(\underline{x}\right) \tag{3.82}$$

In der Erkennungsphase wird für eine neue Beobachtung $\underline{x}$ die Klasse κ_e mit dem höchsten Wert für $MLR_\kappa\left(\underline{x}\right)$ gewählt.

$$\kappa_e = \arg\max_{\kappa} MLR_{\kappa}\left(\underline{x}\right) \tag{3.83}$$

Die Koeffizienten $\alpha_{m,\kappa}$ sind als Gewichte je Klasse und Instanz interpretierbar und auf nicht-negative Werte beschränkt[91] [BRE96]. Daraus ergibt sich, dass ein hohes $\alpha_{m,\kappa}$ ein hohes Vertrauen in die Leistung des Klassifikators m zur Bestimmung der Klasse Ω_{κ} ausdrückt. Zu ihrer Bestimmung wird die Lawson und Hansonsche *Least-Square* Kalkulation angewandt, die hier nicht näher ausgeführt wird. Das zu lösende Minimierungsproblem ergibt sich dabei aus der Minimierung des folgenden Ausdrucks, in dem j den Koeffizient des Trainingssetteils der j-fach stratifizierten Partitionierung angibt:

$$\sum_{j=1}^{J} \sum_{(\underline{x}_l, y_l)\in\mathcal{L}} \left(y_l - \sum_{\kappa=1}^{k} \alpha_{m,\kappa} \cdot P_{m,\kappa,j}\left(\underline{x}_l\right) \right)^2 \tag{3.84}$$

Es wird in [TIN99] darüber hinaus gezeigt, dass eine Metaklassifikation auf Basis von echten Werten über die Konfidenz $P_{m,\kappa}\left(\underline{x}\right) \in [0;1]$ der Level-0-Klassifikatoren eine weitere Verbesserung im Gegensatz zur Bewertung beruhend auf der harten Entscheidung bringt. Diese Variante ist auch als *StackingC*[92] [SEE03] bekannt. Auf eine Beschreibung der Konfidenzberechnung für die jeweiligen Klassifikatoren wird hier im Einzelnen verzichtet[93].

Wird zur Klassifikation auf Metaebene lediglich ein ungewichteter Mittelwert der Wahrscheinlichkeiten zur finalen Entscheidung vollzogen, spricht man von *Voting*.

Zusammenfassend lässt sich sagen, dass Ensemble-Verfahren auf Grund des mehrfachen Trainingsaufwands den höchsten Rechenbedarf zur statischen Klassifikation mit sich bringen. Stacking und Bagging sind dabei rechentechnisch voll parallelisierbar, wohingegen im Falle von Boosting diese Möglichkeit der Zeitoptimierung auf Grund des iterativen Charakters entfällt. Die niedrigste Fehlerrate erhält man in der Regel mit StackingC. Hier ist jedoch ein höherer Lernschatz durch die benötigte stratifizierte Aufteilung erforderlich. Ist dies nicht gewährleistet, können sich in Einzelfällen Bagging oder Boosting als leistungsstärker erweisen. Im Gegensatz dazu kann Stacking auch zur Optimierung stabiler Klassifikatoren angewandt werden. Abschließend sei erwähnt, dass Bagging und Boosting in Stacking integrierbar sind, was unter Umständen zu einer weiteren Leistungssteigerung führen kann.

3.9 Reduktion und Selektion von Merkmalen

Mittels geeigneter heuristischer Verfahren ist eine große Anzahl an Merkmalen erzielbar, die in ihrer Gesamtheit den Merkmalsraum bilden. Um jedoch den Aufwand bei der Merkmalsextraktion und Klassifikation akzeptabel zu halten, erscheint es sinnvoll, die Zahl der verwendeten Attribute

---

[91] Es kann gezeigt werden, dass diese Limitierung für Klassifikationsprobleme entfallen kann. Sie ist aber trotzdem üblich.

[92] *StackingC* steht dabei für *Stacking With Confidences*.

[93] Diese ist unter Anderem in [TIN99] für einige Klassifikatoren zu finden.

auf diejenigen mit hoher Leistung bezüglich der Diskriminanz zu beschränken, da damit stets eine Ersparnis an Rechenzeit im realen Einsatz verbunden ist. Darüber hinaus können, abhängig von der Art des Klassifikators, irrelevante Merkmale sogar negativen Einfluss ausüben. Unter stark negativ beeinflussbaren Lernverfahren befindet sich beispielsweise der beschriebene kNN Klassifikator. In [WIT00] wird weiterhin gezeigt, dass das Hinzufügen einer binären Zufallsvariable beim Einsatz von Entscheidungsbäumen die Erkennungsleistung um bis zu 10% gesenkt hat. Ferner erhöht eine überschaubare Zahl von Attributen auch das Verständnis der Aufgabe. Eine zu starke Reduktion führt jedoch im Allgemeinen zu Verlusten in der Klassifikationsleistung.

3.9.1 Principal-Component-Analysis

Die Hauptachsentransformation, oder *Principal-Component-Analysis*[94] (*PCA*), transformiert vorhandene Muster in Richtung maximaler Korrelation [JOL86]. Dies wird unter Anderem zur Normalisierung von Mustern, oder, wie hier zur Datenkomprimierung, beziehungsweise Dimensionsreduktion, angewandt. Hierzu wird für die Merkmalsvektoren $\underline{x}_i \in \mathcal{L}$ aus der Menge der Lernbeispiele $\mathcal{L}$ der Mittelpunktvektor $\underline{m}_x$ und die Kovarianzmatrix $\underline{C}_x$ berechnet, wobei $L = |\mathcal{L}|$ ihre Anzahl sei:

$$\underline{m}_x = E\{\underline{x}\} = \frac{1}{L}\sum_{i=1}^{L}\underline{x}_i \tag{3.85}$$

$$\underline{C}_x = E\left\{\left(\underline{x}-\underline{m}_x\right)\cdot\left(\underline{x}-\underline{m}_x\right)^T\right\} \tag{3.86}$$

Im Anschluss werden die zugehörigen Eigenwerte λ_n und Eigenvektoren $\underline{e}_n$ der Kovarianzmatrix bestimmt, die folgender Gleichung genüge leisten:

$$\underline{C}_x \cdot \underline{e}_n = \lambda_n \cdot \underline{e}_n \text{ mit } n = 1,...,N \text{ und } N = \dim\left(\underline{x}\right) \tag{3.87}$$

Nach Lösung dieses Eigenwertproblems wird die Hauptachsentransformation durch eine translatorische Verschiebung der Vektoren $\underline{x}_i$ um den Mittelpunktvektor $\underline{m}_x$ und eine anschließende Rotation durch eine Matrix der Eigenvektoren mit $\underline{E} = \{\underline{e}_n\}^T$ und $n = 1,...,N$ vollzogen, so dass man die neuen Vektoren $\tilde{\underline{x}}_i$ erhält:

$$\tilde{\underline{x}}_i = \underline{E}\cdot\left[\underline{x}_i - \underline{m}_x\right] \tag{3.88}$$

Hierdurch ergeben sich nachweislich folgende Bedingungen für die so transformierten Vektoren mit Einführung der diagonalen Matrix der Eigenwerte $\underline{\Lambda}$, die in ihrer Diagonalen durch die Eigenwerte λ_n besetzt ist [RIG04]:

---

[94] Auch unter den Bezeichnungen *Karhunen-Loeve-, Hotelling-*, oder *Eigenvektor-Transformation* geläufig.

$$\underline{m}_{\tilde{x}} = 0 \text{ und } \underline{C}_{\tilde{x}} = \underline{\Lambda} \tag{3.89}$$

Die Vektoren $\underline{\tilde{x}}_i$ sind somit in ihren Komponenten unkorreliert und mittelwertbefreit. Bei den Komponenten selbst handelt es sich um Linearkombinationen der ursprünglichen Merkmale, die als neue Attribute in einer günstigeren Repräsentationsform betrachtet werden können, wobei ihre Anzahl zunächst der ursprünglichen Merkmalszahl entspricht. Auf Grund der beschriebenen Tatsache, dass der zugehörige Eigenwert zu jeder neuen Größe seiner Varianz entspricht, lassen sich nun Merkmale mit kleinem Eigenwert streichen, ohne somit zu starke Leistungseinbußen zu riskieren. Hieraus ergibt sich ein mittlerer quadratischer Fehler ε zwischen einem reduzierten Vektor, der sich nur aus den Komponenten der $M < N$ höchsten Eigenwerte zusammensetzt, und dem ursprünglichen Vektor zu:

$$\varepsilon = \sum_{n=M+1}^{N} \lambda_n \tag{3.90}$$

Eine Reduktion der Dimensionalität nach dieser Vorgehensweise ist weit verbreitet, und wird im Folgenden als *PCA-FS*[95] abgekürzt. Nachteilig dabei zeigt sich allerdings, dass die gewonnenen künstlichen Merkmale geringe Transparenz aufweisen. Ferner erscheint die Anwendung dieser Methode nur unter Annahme einer gaußförmigen Verteilung sinnvoll, da im Vergleich zu GMM (siehe Kap. 3.11.2) hier nur eine Gaußverteilung und eine Kovarianzmatrix für alle Klassen gemeinsam verwendet wird[96]. PCA-FS wird daher im Folgenden in erster Linie als Vergleichsbasis zu bestehenden Arbeiten verwendet.

Abschließend ist in Abb. 3.30 als Beispiel zur PCA-FS die Projektion eines hochdimensionalen akustischen Merkmalsraums in den zweidimensionalen Raum gezeigt. Er wird von den auch als *Principal-Components (PC)* bezeichneten Eigenvektoren mit höchstem zugehörigem Eigenwert aufgespannt. Die Darstellung basiert auf der Datenbank EMO-DB (siehe Kap. 3.12.1). Es ist deutlich zu sehen, dass die Reduktion auf zwei Dimensionen keine Trennung aller Klassen sinnvoll zulässt. Mit einem einfachen distanzbasierten Klassifikator wie 1NN lassen sich jedoch Emotionen teilweise paarweise trennen[97].

---

[95] Kurz für PCA basierte Feature-Selection.

[96] Eine Verbesserung in dieser Hinsicht bietet die sogenannte *Lineare Diskriminanz Analyse* (*LDA*). In dieser werden auch Klassenverteilungen berücksichtigt. Sie findet jedoch im betrachteten Szenario der Reduktion von akustischen Merkmalen geringere Anwendung. Es werden daher generell günstigere Verfahren vorgestellt.

[97] So sind etwa Neutralität und Ärger in einer 10-fach SCV (siehe Kap. 3.12.3) noch mit 98,1% trennbar. Langweile und Neutralität, wie auch aus Sicht des dimensionalen Ansatzes zu erwarten, sind dies hingegen nicht mehr. Hier ergibt sich eine Erkennungsleistung von 53,5%, die nahe der Ratewahrscheinlichkeit von 50% liegt.

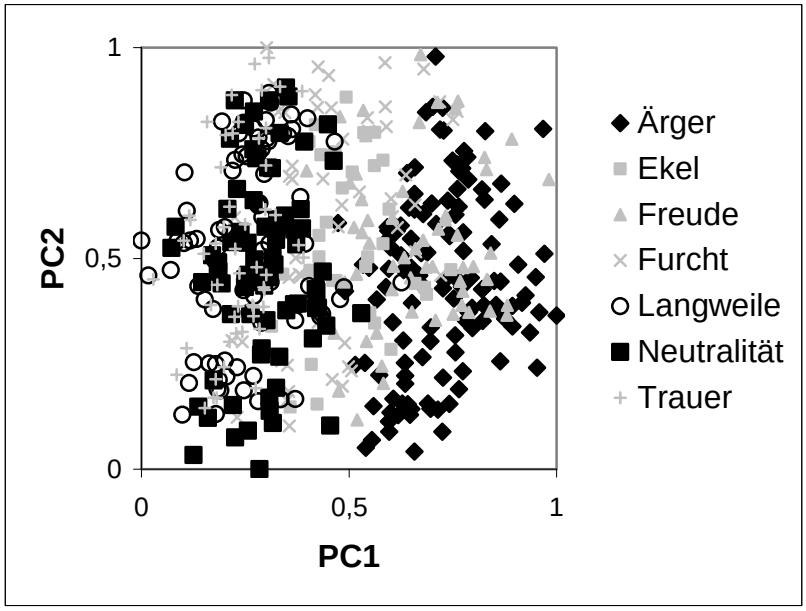


Abb. 3.30: Beispielhafte Projektion des 276 dimensionalen Merkmalsraums auf zwei Dimensionen mittels PCA-FS, Datenbank EMO-DB. Die Emotionen Ärger und Neutralität sind hervorgehoben, da sie in der Projektion noch trennbar sind. Weitere Abgrenzungen sind nicht mehr möglich.

3.9.2 Information-Gain und Gleitende Suchverfahren

Ziel einer wertenden Auswahl optimaler Merkmale ist eine echte Reduktion des zu extrahierenden Merkmalssets der Größe N auf ein Set der Größe $M < N$ (vgl. Kap. 3.9.1) hinsichtlich maximaler Leistung. Im Vergleich hierzu wird bei einer Reduktion im Rahmen etwa der vorgestellten Hauptachsentransformation prinzipiell nur die Zahl der zu klassifizierenden Merkmale reduziert, nicht aber die Zahl der ursprünglich zu extrahierenden. Es existiert hierzu eine Reihe automatischer Verfahren zur Merkmalsselektion, als *Feature-Selection* bezeichnet, die jedoch in der Regel nur suboptimale Lösungen erzielen. Sie werden generell in *filterbasierte*[98] und w*rapperbasierte Ansätze*[99] unterteilt [SCA03].

Bei Ersteren werden einfache Maße zur isolierten Bewertung der Relevanz singulärer Attribute unabhängig vom angestrebten Klassifikator verwendet. Häufig angewandte Beispiele hierfür sind die *Korrelation* und die aus der Informationstheorie bekannte *Transinformation* zur Berechnung der Stärke der statistischen Abhängigkeit zweier Zufallsgrößen, was zunächst den Vorteil einer gewissen Allgemeingültigkeit und geringen Rechenaufwands bringt. Somit eignen sich die Methoden der Filteransätze besonders, um eine Auswahl in hochdimensionalen Räumen zu ermöglichen, oder in diesen eine Vorreduktion zu erreichen.

Um bei insgesamt maximaler Klassifikationsleistung die geringste Merkmalszahl zu erreichen, empfiehlt es sich hingegen, ein Set von Attributen gemeinsam zu reduzieren, und den Bias der

[98] Auch als *Open-Loop FS* bezeichnet.

[99] In Bezug auf die einhüllende Funktion der Bewertungsfunktion. Diese Rückkopplung beschreibt auch der alternative Terminus *Closed-Loop FS*.

Bewertungsfunktion mit dem des Modellierungsansatzes in Übereinstimmung zu bringen. Dies wird in den wrapperbasierten Ansätzen realisiert, die diesen Verbund der Attribute unter rechenintensiver Verwendung des angestrebten Induktionsalgorithmus im Merkmalsraum optimieren. Anstelle aufwändiger Klassifikatoren im späteren Einsatz kann alternativ ein weniger kostenintensives Verfahren als Bewertungsfunktion eingesetzt werden[100], was jedoch in der Regel in einer leichten, negativen Abweichung des finalen Sets resultiert. Da sich ferner eine Vollsuche meist als NP-hart erweist [JOH94], ist ein geeignetes Verfahren wie *gleitende* oder *genetische Suche* erforderlich. Es werden im Folgenden entsprechende Methoden vorgestellt und in Kap. 3.12.4 angewandt:

Der sogenannte **Information-Gain** *(IG)* wird als Vertreter schneller filterbasierter FS-Methoden gewählt, da er oft zur Bewertung der Güte einzelner Merkmale eingesetzt wird [YAG97]. Dabei wird der Gewinn an Information in Bits bewertet, der für die korrekte Zuordnung eines Musters durch Einsatz eines singulären Merkmals erzielt wird. Eine ausführliche Beschreibung des IG findet sich in der Abhandlung zu Entscheidungsbäumen (vgl. Kap. 3.7.4.). Da Attribute mit geringer Streuung innerhalb des Wertebereichs bei einer Selektion nach höchstem IG benachteiligt werden, kann auch hier zur Vermeidung IGR verwendet werden. Eine Suche nach Attributen mit hohem IG oder hoher IGR wird im Folgenden als *IG-FS* oder *IGR-FS* abgekürzt.

Sequentielle Gleitende Suchverfahren erfreuen sich bei der Merkmalsauswahl unter den wrapperbasierten FS-Methoden größter Popularität. Es handelt sich dabei um die von Pudil et al. in [PUD94] vorgestellten *Sequential-Floating-Search-Methoden (SFSM)*[101]. Im direkten Vergleich mit anderen weit verbreiteten Suchverfahren wie *Branch-And-Bound* [ZON97], oder Genetische Algorithmen (vgl. Kap. 3.10) [HAO03] erweisen sie sich mit als herausragendes Verfahren[102] bezüglich Rechenzeit und Reduktionsleistung ohne Beschränkung auf Monotonie bezüglich des Merkmalsets. Es existieren zwei gegenläufige iterative Suchrichtungen im Rahmen der SFSM: Die Vorwärtssuche, *Sequential-Forward-Search (SFS)* genannt, sowie die Rückwärtssuche, *Sequential-Backward-Search (SBS)* genannt. Die Initialisierung der SFS geschieht dabei mit einem leeren Merkmalsset. In jedem folgenden Iterationsschritt wird das jeweils optimale Merkmalsset hinsichtlich der gegebenen Gütefunktion unter Ergänzung eines weiteren Attributs ausgewählt. Bei SBS hingegen wird vom maximalen Merkmalsset ausgegangen, um in jedem Schritt das Attribut zu eliminieren, welches den geringsten Verlust hinsichtlich der Gütefunktion aufweist. Diese Vorgehensweise erlaubt jedoch keine Korrektur von bereits gelöschten oder hinzugefügten Größen. Es kann somit zu sogenannten *Nesting*-Effekten kommen, da durch die Wahl eines Merkmales entscheidend die nachfolgenden Selektionen beeinflusst werden. Abhilfe schafft hier die *„Plus l – Take Away r“* Selektion, bei der jeweils *l* Merkmale hinzugefügt und *r* gelöscht werden. Diese Parameter lassen sich jedoch im Hinblick auf ein Optimum nicht statisch festlegen. Im Gegensatz

[100] Ein Beispiel wäre der Einsatz eines kNN-Klassifikators statt SVM, da letztere jeweils neu trainiert werden müssen.

[101] SFSM gehören zur Gruppe der als *Hill Climbing* bezeichneten Suchverfahren.

[102] Pudil et al. stellten über die vorgestellten SFS Methoden hinaus auch eine adaptive Variante, kurz als AFSM tituliert [SOM99], vor. Diese erzielt bezüglich der Optimierung mindestens die gleiche Güte, ist aber dem Optimum unter Umständen näher. Auf Grund des deutlich höheren erforderlichen Rechenaufwands bei dieser Methode wird auf ihre Anwendung hier verzichtet.

zur reinen SFS wird bei der erweiterten *Sequential-Floating-Forward-Search (SFFS)*[103] in jedem Iterationsschritt ergänzend dynamisch geprüft, ob durch Elimination von Merkmalen aus dem aktuellen Set eine Verbesserung in der Leistung zu beobachten ist. Diese schwebende Kombination aus Ergänzung und Streichung, beziehungsweise SFS- und SBS-Schritten wird als *Floating-Search* bezeichnet. Abb. 3.31 zeigt hierzu den hier verwendeten SFFS-Algorithmus in Form von Pseudocode.

SFFS (*X*: Merkmalsset, *n*: Zieldimension**)** {

// acc_{wrap} repräsentiert die Erkennungsleistung des gewählten Klassifikators

// $X = \{x_i \,|\, i = 1,...,N\}$, wobei N der Dimensionalität des Merkmalsvektors entspricht

$Y_0 = \{\ \}$; // Initialisierung

FÜR *k*=0 bis n **TUE** {

MARKE *Inklusion;*

SEI $x^+ = \arg\max_{x \in X - Y_k} acc_{wrap}\left(Y_k + x\right)$; // relevantestes Merkmal

SEI $Y_{k+1} = Y_k + x^+$;

SEI $k = k + 1$;

MARKE *Konditionelle_Exklusion;*

SEI $x^- = \arg\max_{x \in Y_k} acc_{wrap}\left(Y_k - x\right)$; // unrelevantestes Merkmal

FALLS $acc_{wrap}\left(Y_k - x^-\right) > acc_{wrap}\left(Y_{k-1}\right)$ {

SEI $Y_{k-1} = Y_k - x^-$**;**

SEI $k = k - 1$;

SPRINGE *Konditionelle_Exklusion*;

} **SONST SPRINGE** *Inklusion;*

}

AUSGABE Y_k; // Gibt das geordnete Merkmalsset des Betrags n aus

}

Abb. 3.31: SFFS-Algorithmus als Pseudocode dargestellt

Alternativ wird bei der *Sequential-Floating-Backward-Search (SFBS)* von einem komplett besetzten Merkmalsset ausgehend, in umgekehrter Weise, durch SBS- und folgende SFS-Schritte vorgegangen. Die jeweils zu favorisierende Richtung ergibt sich aus den Rahmenbedingungen der ursprünglichen und angestrebten Setgrößen. Für SFS und SFFS ergibt sich eine von der Bewertungsfunktion sowie der angestrebten Setgröße abhängige Komplexität der Ordnung $O(M)$.

[103] Der als Wrapper angewandte Klassifikator wird führend angegeben. So steht beispielsweise *SVM-SFFS* für eine SFFS mit SVM als Zielfunktion.

Im Gegensatz dazu ergibt sich bei SBS und SFBS die Komplexität in der Ordnung $O(N-M)$.

3.10 Genetische Generierung von Merkmalen

Um künstlich eine Vielzahl weiterer Merkmale zur Leistungssteigerung zu erzeugen, können solche aus bereits extrahierten durch eine automatische Generierung gebildet werden. Konkret kann dies durch Veränderung einzelner Merkmale mittels mathematischer Operationen wie Potenz-, Exponential-, Logarithmus-, oder trigonometrische Funktionen und Betragsbildung erzielt werden (vgl. [SCH05G]). Zur Kombinationen von Merkmalen bieten sich etwa Additions-, Multiplikations-, Potenzierungs- oder Vergleichsoperatoren an [ZHA05]. Als Beispiele einer sinnvollen Kombination können der Quotient aus Dauer stimmhafter Laute und Pausen [SCE82] sowie das Verhältnis einzelner Frequenzbänder zueinander genannt werden. Die Generierung muss dabei im Zusammenhang mit einer Bewertung über zusätzlich erzielte Güte erfolgen, um Generierung redundanter Merkmale, welche den Rechenaufwand und die Komplexität für den Klassifikator erhöht, zu vermeiden. Auch hier ist als Maß der im späteren Einsatz verwendete Klassifikator als ideal anzusehen (vgl. Kap. 3.9). Eine vollständige Suche durch systematische Bildung neuer Größen aus allen bestehenden ist jedoch für einen größeren Merkmalsraum in der Regel ebenfalls NP-hart. Somit sind automatische Selektion und Generierung als zwei Komponenten des gleichen Problems zu sehen: Wird eine Lernaufgabe durch zu wenige Merkmale beschrieben, kann versucht werden neue zu generieren - im umgekehrten Fall können überflüssige eliminiert werden. Auf diese Weise kann auch eine weitere Steigerung der Leistung gegenüber der reinen Selektion erreicht werden, ohne dabei wie bei Kernel-Funktionen von SVM (vgl. Kap. 3.7.2) eine günstige Abbildung empirisch ermitteln zu müssen. Weiterhin positiv an künstlich generierten Merkmalen erweist sich, dass zur Berechnung dieser in einem späteren Einsatz ein deutlich geringerer zeitlicher Aufwand im Vergleich zur ursprünglichen Extraktion neuer Attribute erforderlich ist.

Um den Suchaufwand zu verringern, bieten sich bei der kombinierten Generierung und Selektion unter Anderem Genetische Algorithmen, kurz *GA* [GOL89] an. Diese stellen ein weiteres bioanaloges Verfahren neben den ANN dar, und beruhen auf dem darwinistischen *Survival-Of-The-Fittest*-Prinzip von Mutation und Selektion in Anlehnung an die Evolution. Sie eignen sich im besonderen Maße zur sukzessiven Optimierung in großen Suchräumen, die eine Vielzahl lokaler Maxima und Minima aufweisen[104]. Neo-Darwinisten fassen die Überkreuzung elterlicher Erbinformation als zusätzliche Varietätsgröße zur zufälligen Mutation auf. Zwar benötigen GA große Rechenleistung - sie lassen sich aber stark parallelisieren.

Abb. 3.32 zeigt den gesamten iterativen Ablauf einer genetischen Programmierung zur hier realisierten gemeinsamen Generierung und Selektion von Merkmalen mittels GA, wobei $\underline{x}$ den Ausgangsmerkmalsvektor, und $\underline{x}'$ den finalen bezeichnen:

[104] Obwohl GA mit zu den stabilsten Optimierungsverfahren gehören, begleitet sie oft die Schwierigkeit, eine realitätsnahe Beschreibung zu finden, die zur Lösung eines bestimmten Problems führt.

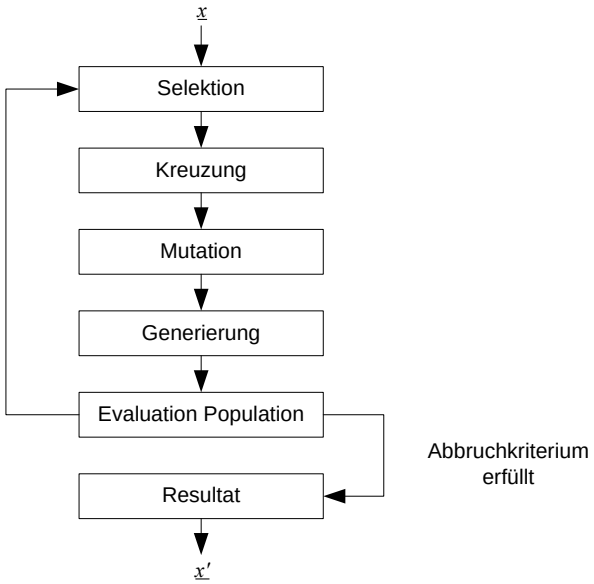


Abb. 3.32: Ablauf der hybriden Generierung und Selektion mit GA

Zur Initialisierung des Verfahrens muss zunächst ein *Population* genannter Bestand deutlich unterscheidbarer simultan verwalteter *Individuen*, auch als *Suchpunkte* bezeichnet, geschaffen werden, die mögliche Lösungen der Suchaufgabe repräsentieren. Im betrachteten Fall repräsentieren diese Individuen Merkmalssets in Form binär kodierter Chiffre-Ketten, *Chromosomen* genannt. Ein Chromosom besitzt jeweils n *Gene*, entsprechend der Zahl der Größen im aktuellen Set. Jedes dieser Gene schließlich steht für ein Attribut aus einer Liste, in der diese beschrieben sind, und gibt binär kodiert seinen Aktivitätszustand an. Die in Frage kommenden Merkmale des Ausgangsvektors $\underline{x}$ sind dabei auf numerische Größen beschränkt. Zur Initialisierung werden die Attribute auf $N = \dim(\underline{x})/n$ Individuen zufällig aufgeteilt. Durch den iterativen Ablauf von GA wird die Population dann über multiple *Generationen* verfolgt.

Im Einzelnen werden folgende fünf Punkte iterativ abgearbeitet, bis ein Abbruchkriterium erfüllt ist:

- Die **Selektion** erfolgt bezogen auf die *Fitness* eines Individuums. Die zugehörige *Fitnessfunktion* bestimmt als Optimierungskriterium entscheidend das globale Verhalten der Suche. Als Funktion wird hier entsprechend den in Kap. 3.9.2 vorgestellten wrapperbasierten Verfahren die Klassifikationsleistung mit SVM auf einem Kreuzvalidierungsset (vgl. Kap.3.12.3) verwendet. Als einfacher Selektionsalgorithmus wird das *Roulette-Wheel*-Verfahren gewählt: Die 360° eines Rouletterades werden proportional zur Fitness anteilig auf die Individuen aufgeteilt. Im Anschluss erfolgt die Selektion entsprechend mehrmaligem Drehen des Rades. Es wird so aus einer Population mit N Individuen N mal mit Zurücklegen gezogen. Die auserwählten Individuen werden im sogenannten *Mating-Pool* gesammelt. Die Wahrscheinlichkeit für die Reproduktion auserwählt zu werden ist für Individuen mit besserer Fitness somit höher. Zusätzlich wird

hier sichergestellt, dass das Individuum höchster Fitness auf jeden Fall selektiert wird[105].

- Zur **Kreuzung** werden aus dem Mating-Pool *N/2* mal ohne Zurücklegen Individuen mit der Wahrscheinlichkeit *1/N* gezogen. Die Kreuzung selbst wird dann mit einer gewissen Wahrscheinlichkeit an jedem Paar vollzogen. Sie muss, im Gegensatz zu normalen GA, angepasst werden, um, durch die später vollzogene Generierung bedingt, Chromosomen mit variabler Länge zuzulassen. Zunächst müssen jeweils zwei *Eltern* ausgesucht werden. Die einzelnen Elternteile werden dabei als *Elter* bezeichnet. Eine sinnvolle Rahmenbedingung bei der Wahl des Verfahrens zur Kreuzung ist, dass der Abstand zwischen Eltern und Kindern nicht größer sein sollte, als zwischen den Eltern selbst. Dieses gewährt das gewählte *Single-Point-Crossing*-Verfahren, bei dem jedes Elterchromosom an einer Stelle um die Mitte herum geteilt wird. Die jeweiligen Hälften werden im Anschluss über Kreuz zu neuen, *Kind* genannten, Individuen zusammengefügt. Die Fitness bestimmt weiterhin, wie viele Kinder Eltern haben dürfen. Abb. 3.33 veranschaulicht den Informationsaustausch zweier Individuen durch Single-Point-Kreuzung.

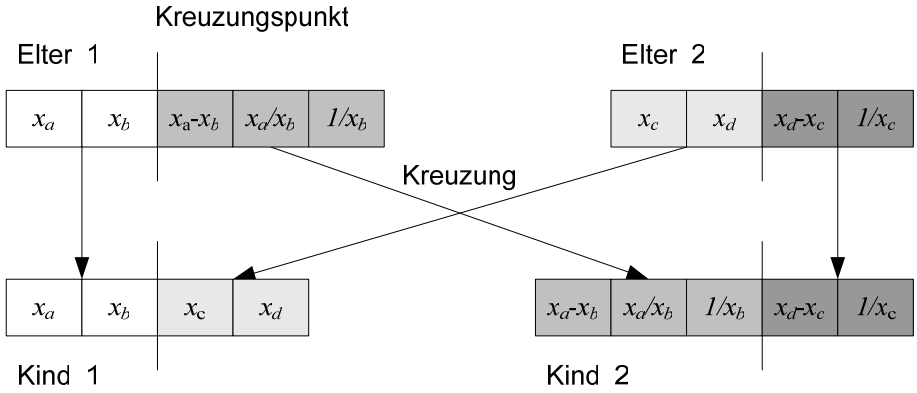


Abb. 3.33: Single-Point-Kreuzung mit variabler Länge bei GA. Die Chromosomhälften sind zur Veranschaulichung der Kreuzung in diversen Graustufen hinterlegt.

- Während der **Mutation** von Individuen wird der Zustand aktiv oder inaktiv bei jedem Gen eines Individuums, respektive Merkmal eines Sets, mit einer bestimmten Wahrscheinlichkeit geändert, wie in Abb. 3.34 beispielhaft dargestellt.

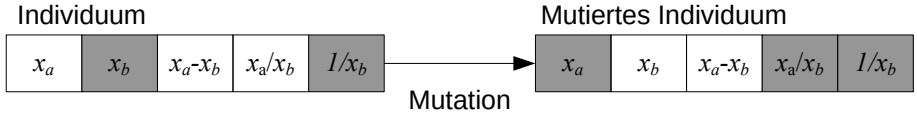


Abb. 3.34: Mutation bei GA. Grau hinterlegte Gene eines Chromosoms entsprechen deaktivierten Merkmalen.

[105] Die sichergestellte Wahl des elitären Individuums wird als *Elitist Selection* bezeichnet.

- Zur **Generierung** kommen hier die Operationen *Addition*, *Subtraktion*, *Multiplikation*, *Division* und *Kehrwertbildung* zum Einsatz. Diese werden als *Generator* bezeichnet, und zufällig ausgewählt. Je nach Generatortyp wird im Anschluss die notwendige Zahl von Merkmalen im Sinne von Operanden zur Generierung ausgewählt. Das generierte Merkmal wird dann dem Individuum angehängt. Da sich die Generierung nicht nur auf ursprüngliche Attribute beschränkt, können auch komplexere Verkettungen von Operationen auftreten. Die Zahl der insgesamt je Schritt zu erzeugenden Größen wird dabei vorab festgelegt. Die folgende Abbildung zeigt das Vorgehen anhand eines Beispiels.

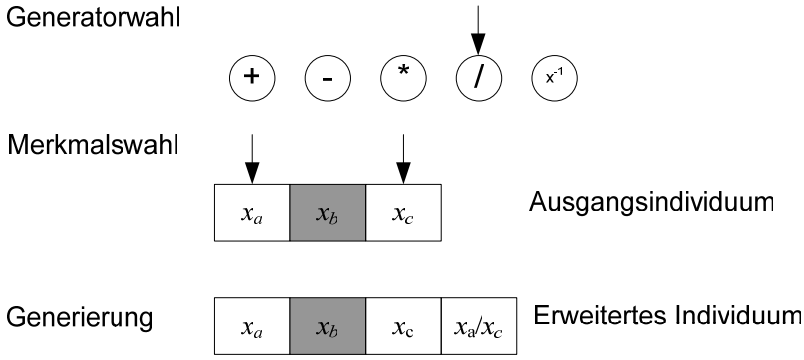


Abb. 3.35: Mutation bei GA. Grau hinterlegte Gene eines Chromosoms entsprechen deaktivierten Merkmalen.

- Die **Evaluation** der Population entspricht der Durchführung des Fitnesstests.

Der dargestellte Prozess wird, wie eingangs erwähnt, iterativ so lange wiederholt, bis ein Abbruchkriterium erfüllt ist, welches die entstandene Lösung als ausreichend optimiert klassifiziert. Im Folgenden wird als Abbruchkriterium eine maximale Zahl von 50 Generationen mit einer Beschränkung auf maximal 25 Generationen ohne Verbesserung festgelegt. Die Initialisierungs-, Kreuzungs- und Mutationswahrscheinlichkeiten von Attributen werden jeweils auf 50% festgesetzt. Um Rechenzeit zu sparen, wird vor der genetischen Generierung des Merkmalssets vorab eine Selektion mit SVM-SFFS vollzogen. Hierdurch wird ein Satz an geeigneten Individuen zur Initialisierung gewählt.

3.11 Automatische Klassifikation von Zeitreihen

Ziel der folgenden Verfahren ist es, an Stelle der Ableitung von Funktionalen und nachfolgender statischer Klassifikation von Zeitreihen $\underline{x} = (\underline{x}_t)$ mit $t = 0,...,T-1$ diese direkt zu klassifizieren. Hierzu ist eine dynamische Komponente erforderlich, die eine Verzerrung entlang der Zeitachse erlaubt.

3.11.1 Dynamic-Time-Warping

Ein sehr einfaches Verfahren zur Klassifikation von Zeitreihen ist das sogenannte *Dynamic-Time-Warping (DTW)*. Ähnlich wie bei den unter Kap. 3.7.1 vorgestellten Klassifikatoren handelt es sich um eine instanzbasierte Methode, bei der die Distanz zwischen zwei Zeitreihen berechnet wird,

wobei zusätzlich einer dynamischen zeitlichen Verzerrung Rechnung getragen wird. Der möglichst optimale zeitliche Ausgleich erfolgt durch Minimierung der kumulativen Distanz lokaler Abstände zwischen zugeordneten Punkten. Der Name DTW bezieht sich auf die Tatsache, dass die Zeitachsen zweier zu vergleichender Reihen dergestalt verzerrt werden, dass korrespondierende Abschnitte aufeinander abgebildet werden.

Zum Vorgehen werden ausgehend von zwei multivariaten Zeitreihen $\underline{x} = \left(\underline{x}_t\right)$ mit $t = 0,...,T-1$ und $\underline{x}_l = \left(\underline{x}_{l,z}\right)$ mit $z = 0,...,Z-1$ zunächst die Distanzen zwischen je zwei Vektoren $\underline{x}_t$ und $\underline{x}_{l,z}$ zu den Zeitpunkten t, z mit einer Distanzfunktion $d\left(t,z\right)$ bestimmt. Hierzu wird oft der Euklidsche Abstand, wie in Kap. 3.7.1 vorgestellt, verwendet:

$$d\left(t,z\right) = \left[\underline{x}_t - \underline{x}_{l,z}\right]^T \cdot \left[\underline{x}_t - \underline{x}_{l,z}\right] \tag{3.91}$$

Es sind jedoch auch weitere dort diskutierte Abstandsmaße einsetzbar. Die Abstände aller Vektoren untereinander werden in der Distanzmatrix $\underline{D} = \left\{d\left(t,z\right)\right\}$ mit $t = 0,...,T-1$ und $z = \left(0,...,Z-1\right)$ zusammengefasst. Im Weiteren wird mittels Backtracking (vgl. Kap. 3.11.2) der optimale Pfad durch die Distanzmatrix bestimmt, so dass der aufsummierte Gesamtabstand minimal wird. Dabei werden sowohl lokale Einschränkungen an das Wegediagramm, als auch globale in Form eines Suchkorridors, wie in Abb. 3.36 dargestellt, getroffen. Damit werden zu starke Verzerrungen ausgeschlossen und das Verfahren beschleunigt, da Pfade zu starker Verzerrung nicht berechnet werden müssen. In der Erkennungsphase wird die Klasse desjenigen Referenzmusters gewählt, das den geringsten Abstand zum zu klassifizierenden aufweist (vgl. Kap. 3.7.1).

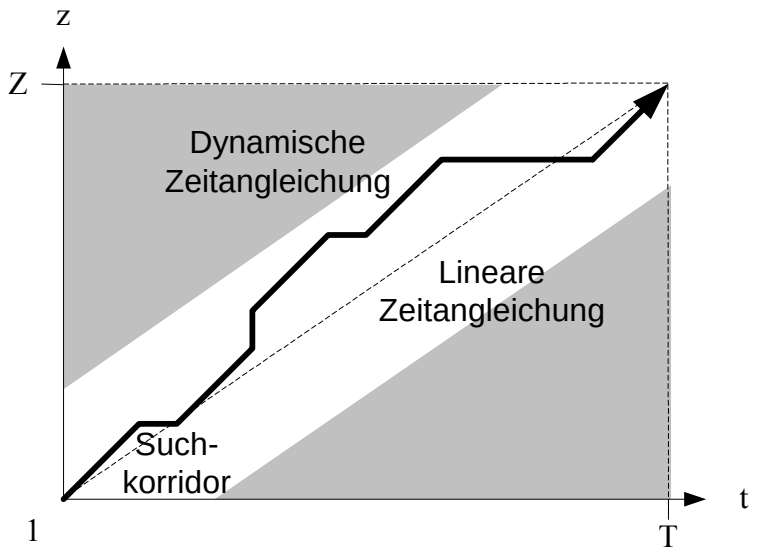


Abb. 3.36: Verzerrte Zeitangleichung beim DTW-Verfahren

Nachteile von DTW sind der hohe Speicher- und Rechenbedarf während der Erkennungsphase auf Grund seines instanzbasierten Charakters. Andererseits ist kein explizites Training erforderlich, da neue Lerninstanzen einfach abgespeichert werden können. Da kein Modell gebildet werden muss, kann bereits mit sehr geringem Trainingsschatz eine Erkennung stattfinden. Für die

sprecherabhängige Emotionserkennung kann dies interessant sein, weil hier keine zu große Zahl an Lernbeispielen des potenziellen Nutzers zu erwarten sind. Schließlich ist DTW sehr leicht zu implementieren, und findet daher in kommerziellen Produkten durchaus Einsatz. Im Allgemeinen werden jedoch die nachfolgend vorgestellten HMM auf Grund der meist deutlich höheren Leistung im Einsatz bevorzugt, obwohl durch neue Ansätze, wie hierarchisches DTW, ein weiterhin bestehendes Interesse an DTW bezeugt wird.

3.11.2 Hidden-Markov-Modelle

Hidden-Markov-Modelle erlauben eine probabilistische Modellierung von multivariaten Zeitreihen und sind ein Standardverfahren im Bereich der Spracherkennung, wo sie erstmals in [BAK75] vorgestellt wurden. In [RAB89] findet sich eine detaillierte Beschreibung der hier angewandten Prinzipien.

Als Ausgangsbasis sollen zunächst diskrete Markov[106]-Prozesse vorgestellt werden [RAB89]. In Abb. 3.37 ist hierzu ein exemplarisches System gezeigt, das sich zu jedem Zeitpunkt in einem der hier gewählten Zustände $S_1,...,S_3$ befindet.

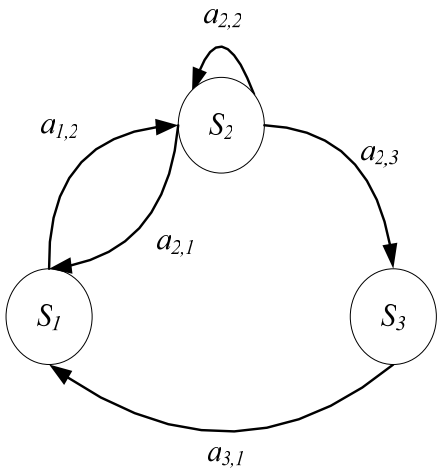


Abb. 3.37: Beispielhafte Markov-Kette mit drei Zuständen und ausgewählten Übergängen

Mit jedem fortschreitenden Zeitindex findet ein Zustandsübergang in einen neuen Zustand oder ein Selbstübergang mit der Wahrscheinlichkeit $a_{i,j}$ für einen Übergang vom Zustand *i* in den Zustand *j* mit $i = 1,...,N$ und $j = 1,...,N$ und N als Gesamtzahl der Zustände statt. Folgende Stochastizitätsbedingungen gelten:

$$a_{i,j} > 0 \text{ und } \sum_{j=1}^{N} a_{i,j} = 1 \tag{3.92}$$

Aus der Tatsache, dass zu jedem Zeitpunkt nachvollziehbar ist in welchem Zustand sich der

[106] Benannt nach dem russischen Mathematiker Andrei Andrejewitsch Markov. Aus dem Kyrillischen oft auch als *Markow* in lateinische Lettern übertragen.

beschriebene stochastische Prozess befindet, folgt die Bezeichnung *beobachtbar*. In einer Erweiterung wird nun in jedem Zustand eine Wahrscheinlichkeit für eine Beobachtung $\underline{x}_t$ in einer weiteren Wahrscheinlichkeitsfunktion dargestellt. Es ergibt sich hieraus eine Beobachtungsfolge[107] $\underline{x} = \{\underline{x}_t\}$ mit $t = 0,...,T-1$ der Länge T. Sie korrespondiert in der Anwendung mit der Merkmalsvektorfolge $\underline{x}$, deren Modellierung hier das Ziel ist. In der Konsequenz erhält man einen doppelt stochastischen Prozess, in dem rein anhand beobachteter Emissionen der Zustände der unterlagerte Prozess der Übergänge nicht mehr nachvollziehbar, beziehungsweise *verborgen*[108] ist. Man spricht daher von einem *Hidden-Markov-Modell*, welches kausale, stationäre Prozesse abbilden kann, die nicht von der Vergangenheit abhängen[109]. Eine Reihe von Parametern erlaubt die vollständige Beschreibung eines im Folgenden als λ notierten HMM:

- Die **Anzahl der Zustände** N: Diese kann im Zusammenhang mit einer physikalischen Bedeutung stehen. In der Spracherkennung modellieren die Zustände meist Phoneme, wodurch ihre Gesamtzahl in direkter Verbindung mit der Wortlänge steht. Als Bezeichnung für die Zustände sei $S = \{S_1,...,S_N\}$ gewählt. Ferner gibt q_t den zum Zeitpunkt *t* aktuellen Zustand an.

- Die **Übergangswahrscheinlichkeiten** $a_{i,j}$**,** zusammengefasst in der Matrix $\underline{A} = \{a_{i,j}\}$**:** Diese entsprechen den beschriebenen Wahrscheinlichkeiten der Markov-Ketten, wobei ein zusätzlicher Einsprungszustand S_0 und Aussprungszustand S_{N+1} eingeführt werden, in denen das HMM beginnen und enden muss [YOU02][110]. Gilt für alle Übergangswahrscheinlichkeiten $a_{i,j} > 0$, was bedeutet dass jeder Zustand von jedem Zustand aus erreichbar ist, so spricht man von einem *ergodischen* HMM. Es wird eine Reihe weiterer Topologien unterschieden, bei denen Einträge in $\underline{A}$ mit Nullen besetzt sind. Die in der Sprachverarbeitung wichtigste Form, und auch im Folgenden angewandte, sind die sogenannten *Links-Rechts*-Modelle[111]. Bei diesen führt ein Inkrement im Zeitindex zu einer Stagnation oder ebenfalls zu einem Inkrement im Zustandsindex, wodurch retrovertierte Transitionen ausgeschlossen sind. Zusätzlich wird oft eine Beschränkung an die Sprungweite gestellt[112], wobei nur Übergänge von einem Zustand *i* zu einem Zustand $j \le i + \Delta, \quad 1 \le \Delta \le N - i$ erlaubt sind[113]. Unter Anwendung der Definition der Zustände lässt sich zusammenfassen:

---

[107] Englisch: *Observation*.

[108] Englisch: *Hidden*.

[109] Sprachsignale sind nicht stationär (vgl. Kap. 3.2.2). Auch die Einschränkung, dass nur unmittelbar vorhergehende Zustände betrachtet werden ist nicht ausreichend. Der Einsatz von HMM ist somit hier als Kompromiss zwischen Modellierungsgenauigkeit und Berechenbarkeit zu sehen.

[110] Diese Notation erspart eine zusätzliche Festlegung eines Einsprungvektors.

[111] Für diese ergibt sich die Matrix $\underline{A}$ in Dreiecksform.

[112] In der Literatur als *Constrained-Jump-HMM* bezeichnet.

[113] Eine typische maximale Sprungweite ist etwa bis zum übernächsten Zustand, wie bei den sogenannten *Bakis* HMM. Eine noch stärkere Einschränkung besitzen *lineare* HMM, bei welchen nur Autotransitionen und Sprünge in den jeweils nächsten Zustand erlaubt sind.

$$a_{i,j} = P\left(q_{t+1} = S_j \,\middle|\, q_t = S_i\right) \text{ mit } 0 \le i, j \le N+1 \tag{3.93}$$

- Die **Ausgabedichten** $b_i\left(\underline{x}_t\right)$**.** Sie geben die Wahrscheinlichkeit an, dass im Zustand i die Beobachtung $\underline{x}_t$ erfolgt:

$$b_i\left(\underline{x}_t\right) = p\left(\underline{x}_t \middle| q_t = S_i\right) \tag{3.94}$$

Im Fall sogenannter *kontinuierlicher* HMM werden als Ausgabedichten meist kontinuierliche Gaußfunktionen zur Approximation der originalen WDF verwendet[114]. Gegeben sei hierzu die n-variate Gaußsche Dichte $g\left(\underline{x}, \underline{m}_i, \underline{C}_i\right)$ mit der Kovarianzmatrix $\underline{C}_i$[115], dem Mittelpunktvektor $\underline{m}_i$ und $\underline{x} \in \mathbb{R}^n$:

$$g\left(\underline{x}, \underline{m}_i, \underline{C}_i\right) = \frac{1}{\sqrt{\left(2\pi\right)^n} \cdot \left|\underline{C}_i\right|^{\frac{1}{2}}} \cdot \mathrm{e}^{\left(-\frac{1}{2}(\underline{x}-\underline{m}_i)^T \cdot \underline{C}_i^{-1} \cdot (\underline{x}-\underline{m}_i)\right)} \tag{3.95}$$

Da eine einzelne Gaußkurve keine multimodalen und nicht linear korrelierten Dichten modellieren kann, wird eine gewichtete additive Überlagerung von M Kurven betrachtet. Dabei werden die einzelnen, als Mixturen bezeichneten Gaußschen Dichten, mit zugehörigen Faktoren $c_{i,m}$ mit $m = 1,...,M$ gewichtet, die wegen der Stochastizitätsbedingung in ihrer Summe Eins ergeben müssen:

$$b_i\left(\underline{x}_t\right) = \sum_{i=1}^{M} c_{i,m} \cdot g\left(\underline{x}_t, \underline{m}_{i,m}, \underline{C}_{i,m}\right) \text{ und } \sum_{m=1}^{M} c_{i,m} = 1 \tag{3.96}$$

Auf diese Weise erhält man eine hohe Anzahl an Parametern zur Modellierung der Ausgabedichten. Dies setzt jedoch auch eine große Menge von Lernbeispielen in $\mathcal{L}$ voraus. Alternativ können die Komponenten der WDF für alle Zustände identisch gewählt werden[116]. Ersetzt man die Ausgabedichten durch eine endliche Zahl verschiedener diskreter Wahrscheinlichkeiten, erhält man *diskrete* HMM. Diese erfordern den Einsatz von Verfahren der *Vektorquantisierung*[117] *(VQ)* zur Erstellung von *Codebüchern*[118], um die Eingangsvektoren $\underline{x}_t$ zu jedem Zeitpunkt $t = 0,...,T-1$ einer endlichen Zahl C von

---

[114] Beispielsweise können auch Polynomapproximationen eingesetzt werden. Anstelle einer Zustandsdichte kann generell auch ein Klassifikator wie ANN oder SVM zur Schätzung der Wahrscheinlichkeiten eingesetzt werden (vgl. Kap. 3.7), was teilweise zu Leistungssteigerungen führt.

[115] Bei annähernd unkorrelierten Merkmalen in $\underline{x}$ kann von einer nur in der Hauptdiagonalen besetzten Kovarianzmatrix $\underline{C}_i$ ausgegangen werden. Auf die Invertierung dieser kann dann zu Gunsten einer einfachen Division verzichtet werden.

[116] Da die Zustände auf die gleichen Mixturen zugreifen, spricht man von *verbundenen Mixturen*, oder englisch von *Tied Mixtures*.

[117] Verlustbehaftete Kompressionsmethode basierend auf dem Prinzip der Blockcodierung.

[118] Es existiert eine Reihe von Verfahren, um das Codebuch optimal an die beobachteten Verteilungen aus der Lernmenge anzupassen. Zwei der beliebtesten Methoden der Codebuchoptimierung sind *Linde-Buzo-Gray*-VQ *(LBG-VQ)* und der Einsatz von *K-Means* Klustering.

Codebucheinträgen zuzuordnen. Jeder Eintrag repräsentiert dabei einen Codevektor, welcher alle Vektoren in seiner Codierungsregion, ähnlich einem Runden im eindimensionalen Raum, approximiert. Besonders bei geringer Codebuchgröße C kann es jedoch zu starken Quantisierungsfehlern bei hoher Dimensionalität von $\underline{x}_t$ kommen. Diskrete HMM benötigen vergleichsweise wenig Trainingsmaterial - ihre Erkennungsleistung ist aber entscheidend geringer.

Zusammengefasst wird ein HMM somit vollständig durch das Set seiner freien Parameter beschrieben: $\lambda = (\underline{A}, b_1, ..., b_N)$. Abb. 3.38 zeigt ein exemplarisches Links-Rechts-HMM, welches hier eine Klasse $\Omega_\kappa \in \Omega$ repräsentiert, und damit zur Notation λ_κ mit $\kappa = 1, ..., k$ und $k = |\Omega|$ führt. In der Erkennungsphase konkurrieren die Modelle untereinander: Das Modell, das die höchste Generierungswahrscheinlichkeit für die beobachtete Folge $\underline{x}$ besitzt, gibt schließlich die erkannte Klasse κ_e an (vgl. Kap.3.7.5):

$$\kappa_e = \underset{\kappa}{\operatorname{argmax}} \left\{ p(\underline{x} \mid \lambda_\kappa) \cdot P(\Omega_\kappa) \right\} \tag{3.97}$$

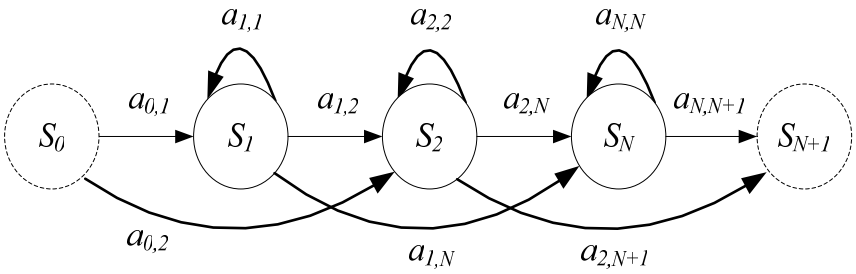


Abb. 3.38: Beispielhaftes Links-Rechts-HMM vom Typ Bakis mit drei Zuständen

Die Aufgabe der *Evaluation* ist nun die effiziente Berechnung der maximalen Emissionswahrscheinlichkeit der Beobachtungsfolge $\underline{x} = (\underline{x}_t)$ mit $t = 0, ..., T-1$ bezüglich der Dichte $p(\underline{x} \mid \lambda_\kappa)$ für ein HMM λ_κ. Eine direkte und vollständige Berechnung aller Pfade mit zugehörigen Emissionen durch das Modell ist prinzipiell möglich. Dies erfordert jedoch die erneute Berechnung aller Wahrscheinlichkeiten für jeden Zeitpunkt, was zu einer größenordnungsmäßigen Gesamtzahl von $2 \cdot T \cdot N^T$ Rechenschritten führt. Mit Hilfe des sogenannten *Vorwärtsalgorithmus* kann diese Zahl auf die Größenordnung $T \cdot N^2$ reduziert werden. Hierzu werden partielle Wahrscheinlichkeiten $\alpha_t(j)$ eingeführt, die die Wahrscheinlichkeit angeben, dass bis zum Zeitpunkt t der Zustand j unter korrekter Erzeugung des Merkmalsverlaufs $\underline{x} = (\underline{x}_t)$ von $t = 0$ aus erreicht wurde [SCT95]. In einer Initialisierung werden zuerst die partiellen Wahrscheinlichkeiten $\alpha_0(j)$, auch Vorwärtswahrscheinlichkeiten genannt, für jeden Zustand wie folgt bestimmt:

$$\alpha_0(j) = a_{o,j} \cdot b_j(\underline{x}_0) \text{ mit } 1 \le j \le N \tag{3.98}$$

Die Wahrscheinlichkeiten α_t für jeden höheren Zeitschritt mit $t = 1, ..., T-1$ ergeben sich im Anschluss rekursiv gemäß:

$$\alpha_{t+1}(j) = \left(\sum_{i=1}^{N} \alpha_t(i) \cdot a_{i,j} \right) \cdot b_j\left(\underline{x}_{t+1}\right) \text{ mit } 1 \le t \le T-2,\ 1 \le j \le N \tag{3.99}$$

Abb. 3.39 zeigt zur Veranschaulichung dieses Prinzips ein sogenanntes *Trellis-Diagramm*. Dabei wird für jeden Zeitpunkt t jeder potenziell mögliche Zustand eines HMM gezeigt:

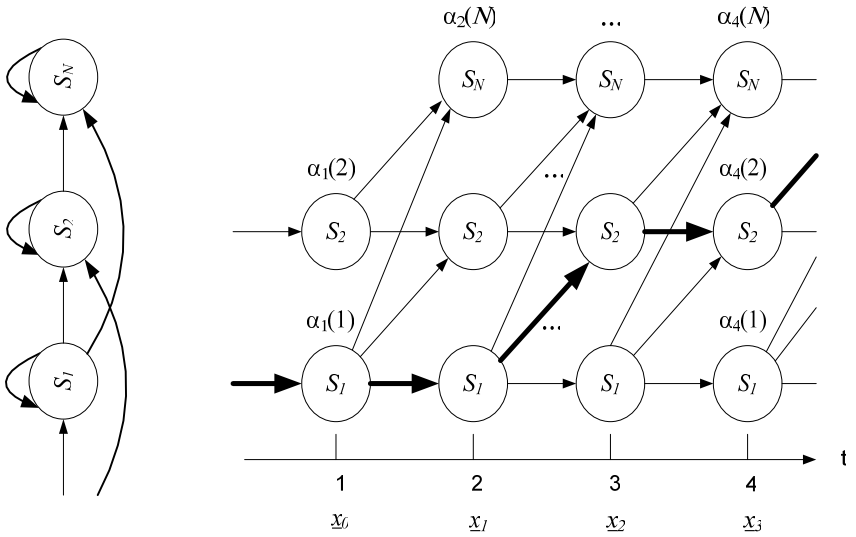


Abb. 3.39: Trellis-Diagramm für ein Bakis HMM mit drei Zuständen und Beispiel eines optimalen Pfads (hervorgehoben)

Um die Wahrscheinlichkeit der Beobachtung $\underline{x}$ bei gegebenem HMM λ_κ zu berechnen, wird der Vorwärtsalgorithmus mit einer Summation aller Teilwahrscheinlichkeiten abgeschlossen:

$$p\left(\underline{x} \mid \lambda_\kappa\right) = \sum_{i=1}^{N} \alpha_{T-1}(i) \tag{3.100}$$

Wünscht man nur den optimalen Pfad durch das Modell, kann alternativ auch der *Viterbi-Algorithmus* angewandt werden. Mit diesem ist auch eine Dekodierung möglich, die die optimale Hintergrundsequenz aufschlüsselt. Damit die Abfolge der Zustände erhalten bleibt, wird eine Reihe mit Elementen $\psi_t(j)$ zur Speicherung der besten Teilpfade eingeführt. Analog zum Vorwärtsalgorithmus wird zunächst eine Initialisierung vollzogen:

$$\delta_1(j) = a_{0,j} \cdot b_j\left(\underline{x}_1\right) \text{ und } \psi_1(j) = 0 \text{ mit } 1 \le j \le N \tag{3.101}$$

Anschließend erfolgt auch hier für höhere Zeitschritte eine rekursive Bestimmung:

$$\begin{aligned} \delta_t(j) &= \max_{1 \le i \le N} \left(\delta_{t-1}(i) \cdot a_{i,j} \right) \cdot b_j\left(\underline{x}_t\right) \\ \psi_t(j) &= \operatorname*{argmax}_{1 \le i \le N} \left(\delta_{t-1}(i) \cdot a_{i,j} \right) \text{ mit } 1 \le t \le T-1,\ 1 \le j \le N \end{aligned} \tag{3.102}$$

Der Abschluss wird durch die Berechnung der gesamten Pfadwahrscheinlichkeit vollzogen:

$$\begin{aligned} p^* &= \max_{1\le i\le N}\left(\delta_T(i)\right) \\ q_T^* &= \operatorname*{argmax}_{1\le i\le N}\left(\delta_T(i)\right) \end{aligned} \tag{3.103}$$

Ist auch die Abfolge der Zustände selbst von Interesse, erfolgt zusätzlich zu den bereits beim Vorwärtsalgorithmus vorgestellten Schritten Initialisierung, Rekursion und Abschluss ein sogenanntes *Backtracking* zu deren Dekodierung:

$$q_t^* = \psi_{t+1}\left(q_{t+1}^*\right) \text{ mit } t = T-2, T-3, \cdots, 0 \tag{3.104}$$

Der Hauptunterschied zwischen Vorwärts- und Viterbi-Algorithmus liegt somit in der Betrachtung der Generierungswahrscheinlichkeit des optimalen Pfads durch das Modell, wie in Abb. 3.39 dargestellt, anstelle der Summation über alle Pfade beim jeweiligen Abschluss, wodurch sich ein Geschwindigkeitsvorteil bei der Berechnung zu Lasten einer suboptimalen Lösung ergibt.

Abschließend stellt sich die Aufgabe des Trainings der Modelle. Die Bestimmung der freien Parameter erfolgt hier mit dem Ziel der Maximierung der Wahrscheinlichkeitsdichte $p\left(\underline{x}\,|\,\lambda\right)$ [119] für die Generierung einer Folge $\underline{x}$:

$$\lambda^* = \operatorname*{argmax}_{\lambda} p\left(\underline{x}|\lambda\right) \tag{3.105}$$

Dieser Ansatz erfordert, dass die Parameter des Modells λ zur Berechnung des optimalen Modells λ^* bereits vorliegen, was mit Hilfe des *Expectation-Maximization- (EM)* Algorithmus[120] für HMM[121] gelöst werden kann. Hierbei wird iterativ alternierend im sogenannten *E-Schritt* der Erwartungswert aus den aktuellen Parametern der zu optimierenden Größe bestimmt und im *M-Schritt* eine bessere Schätzung der Modellparameter hinsichtlich der angestrebten Maximierung erzielt. Auf eine Darstellung des exakten Ablaufs wird zu Gunsten einer Skizzierung verzichtet: In Analogie zur Vorwärtswahrscheinlichkeit $\alpha_t(i)$ wird eine Rückwärtswahrscheinlichkeit $\beta_t(i)$ bestimmt:

$$\begin{aligned} &\beta_T(i) = 1,\ 1 \le i \le N \\ &\beta_t(i) = \sum_{j=1}^{N} a_{i,j} \cdot b_j\left(\underline{x}_{t+1}\right) \cdot \beta_{t+1}(j) \text{ mit } t = T-2, T-3, \cdots, 0;\ 1 \le i \le N \end{aligned} \tag{3.106}$$

Mit $\beta_t(i)$ lässt sich die Wahrscheinlichkeit γ, die das Auftreten des Zustands i zum Zeitpunkt t

---

[119] In der Praxis wird die logarithmierte Dichte $L = \log p\left(\underline{x}\,|\,\lambda\right)$ verwendet, da der Logarithmus sich streng monoton verhält, und die Berechnung vereinfacht. Weiterhin sei angemerkt, dass das hier beschriebene Maximum-Likelihood-Training nur für unendlich viele Beispiele optimale Modelle ergibt, da die Transinformation zwischen Modellzustandsfolge und Merkmalsfolge nicht optimiert wird. Dies leistet das deutlich aufwändigere *diskriminative Training*, für das im betrachteten Szenario jedoch kein signifikanter Zugewinn an Leistung zu erwarten ist.

[120] Die Bezeichnung Algorithmus wird der strengen Definition auf Grund mangelnder Instruktionen für konkrete Rechenschritte des allgemeinen EM-Algorithmus nicht gerecht. Sie hat sich jedoch als beliebte Bezeichnung etabliert.

[121] Dieser ist für HMM auch als *Baum-Welch-Algorithmus* bekannt.

bei gegebener Beobachtungsfolge beschreibt, bestimmen:

$$\gamma_t(i) = \frac{\alpha_t(i) \cdot \beta_t(i)}{p(\underline{x} \mid \lambda)} \tag{3.107}$$

Weiterhin führt $\beta_t(i)$ zur Wahrscheinlichkeit ξ, dass zum Zeitpunkt t der Zustand i vorgelegen hat und in den Zustand j übergegangen wurde:

$$\xi_t(i,j) = \frac{\alpha_t(i) \cdot a_{i,j} \cdot b_j(x_{t+1}) \cdot \beta_{t+1}(j)}{p(\underline{x} \mid \lambda)} \tag{3.108}$$

Mit der Einführung dieser Hilfsgrößen kann schließlich eine Lernvorschrift für die Parameter eines HMM angegeben werden:

$$a_{0,j} = \gamma_1(j),\ a_{i,j} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \tag{3.109}$$

$$b_j(\underline{x}_t) = \frac{\sum_{t=0,\underline{x}=\underline{x}_t}^{T-1} \gamma_t(j)}{\sum_{t=0}^{T-1} \gamma_t(j)} \tag{3.110}$$

Das Ziel, die Wahrscheinlichkeiten an das Trainingsmaterial anzupassen, wird, durch den rekursiven Charakter bedingt, nur asymptotisch approximiert. Als hierdurch erforderliches Abbruchkriterium werden in der Praxis sowohl eine maximale Iterationszahl als auch eine minimale Änderung in Modellparametern vorgegeben.

Ein genereller Nachteil von HMM ist, dass optimale Parameter nur heuristisch zu finden sind. Eine Abschätzung kann in gewissen Grenzen aus der Charakteristik der zu modellierenden Muster getroffen werden. Im Allgemeinen werden sehr viele Trainingsmuster benötigt, um die partiellen Wahrscheinlichkeiten in einem zeitintensiven Trainingsvorgang sinnvoll zu berechnen. Dies gilt im Besonderen bei kontinuierlichen Modellen. Darüber hinaus ist die mangelnde Modellierung der zeitlichen Abhängigkeit beim Zustandsübergang in vielen Fällen nicht adäquat. Hinzu kommt ein negativ exponentieller Wahrscheinlichkeitsverlauf für die Verweildauer innerhalb eines Zustandes [RAB89]. Für die Modellierung der meisten physikalischen Zeitreihen wäre aber vielmehr eine Verweildauer in analytischer Form wünschenswert. Bei den hier gewählten Links-Rechts-Modellen, in denen sich die Anzahl der Zustände proportional zur durchschnittlichen Dauer orientiert, ist eine explizite Modellierung der Verweildauerdichte oft weniger relevant.

Abschließend sei zu HMM angemerkt, dass der Sonderfall eines kontinuierlichen HMM mit einem einzigen Zustand ohne Selbstübergang einem sogenannten *Gaussian-Mixture-Modell* λ, kurz *GMM*, entspricht. Auf Grund so gegebener mangelnder Berücksichtigung der Zeitkomponente handelt es sich nicht mehr um eine dynamische Modellierung. GMM stellen somit gleichzeitig

einen Grenzfall statischer BN dar, und werden daher im Weiteren nicht gesondert behandelt.

3.12 Experimente und Ergebnisse

Im Folgenden wird eine Reihe von im Rahmen dieser Arbeit durchgeführten Versuchen zur Beurteilung der Qualität des vorgestellten ganzheitlichen Konzepts zur akustischen Emotionserkennung mittels Merkmalsextraktion, hybrider Generierung und Reduktion sowie Klassifikation vorgestellt. Die Angaben zur Erkennungsleistung erfolgen dabei in Prozent richtig zugeordneter Daten und die Genauigkeit wird jeweils im sinnvollen Verhältnis zur Zahl der Stichproben gewählt.

3.12.1 Verwendete Datenbanken

Im Vergleich zu verwandten Gebieten wie der automatischen Sprachverarbeitung existieren erst seit jüngerer Zeit Datenbanken mit emotionalen Sprachbeispielen [DOU00], [CAM02], [DOU04]. Sie unterscheiden sich stark in Zahl und Art bezogen auf Emotionen, Personenzahl und Umfang. Tab. 3.40 und Tab. 3.41 geben einen informativen Überblick über bedeutende Korpora natürlicher und gespielter Daten. Darüber hinaus bestehen unter Anderem weitere Datenbanken in finnischer, französischer, hebräischer, russischer und slowenischer Sprache. Aus Gründen der Vergleichbarkeit mit anderen Forschungsergebnissen werden im Weiteren öffentliche und frei zugängliche Datensammlungen bevorzugt.

In [VER03] werden 31 akustische Datenbanken zur Emotionserkennung verglichen. In [VER04A] verwenden dieselben Autoren in Bezug auf diesen Vergleich die *Danish Emotional Speech*-Datenbank, kurz *DES*[122] [ENG96]. Im Sinne eines potenziellen Vergleichs mit anderen Arbeiten soll diese auch hier eingesetzt werden. In ihr sind vier Emotionen des MPEG-4-Basissets, namentlich *Ärger, Freude, Trauer, Überraschung*, und ergänzend *Neutralität* enthalten. Zwei weibliche Sprecher, mit den Schlüsseln *DHC* (34 a) und *KLA* (52 a) sowie zwei männliche Sprecher mit den Schlüsseln *JZB* (38 a) und *HO* (52 a) sind darin enthalten. Die Sprecher sind professionelle Schauspieler mit mehrjähriger Berufspraxis, und sprechen jeweils in ihrer Muttersprache Dänisch die Wörter *ja* und *nein*, neun weitere Sätze und zwei Passagen fließenden Textes in jeder der fünf Emotionsarten. Aus Konsistenzgründen wurden die beiden Textpassagen in vier und fünf einzelne Phrasen aufgeteilt, da die Erkennung hier grundsätzlich je Einzelphrase erfolgt. Es ergeben sich somit insgesamt 414 Äußerungen. Die Aufnahme erfolgte in einem schallgedämpften Tonstudio unter akustischen Idealbedingungen mit 16 Bit, 20 kHz und PCM-Kodierung. Ein Operator hat jeweils den Ablauf überwacht und ein weiterer die Qualität der Aufzeichnungen.

[122] Datenbank der Aalborg Universität, Dänemark, die ursprünglich zum Vergleich realer mit synthetisch erzeugten Emotionssamples erstellt wurde.

Bezeichnung	Sprache	Emotionen	Herkunft	Umfang
AEC [BAT04]	Deutsch, Englisch	Ärger, Bemutterung, Betonung, Empfindlichkeit, Freude, Hilflosigkeit, Langweile, Maßregelung, Neutralität, Überraschung	Kinder interagieren mit Spielzeug-Roboter	51 deutsche Kinder, 48.412 Wörter 30 englische Kinder, 5.822 Wörter
CREST [CAM02]	Englisch Japanisch Chinesisch	Breites Spektrum	Freiwillige nehmen ihre alltäglichen Konversationen auf	Ziel: 1.000 h über 5 a
READING-LEEDS [STI00]	Englisch	Breites Spektrum	Personen schildern im Radio ein emotionales Erlebnis	~ 4,5 h
SYMPAFLY [STE04]	Deutsch	Ärger, Betonung, Empfindlichkeit, Freude, Hilflosigkeit, Ironie, Neutralität, Panik, Überraschung	Flugbuchung über Dialogsystem, Nutzer nicht eingeweiht	110 Dialoge, 29.200 Wörter
Fernandez und Picard [FER00]	Englisch	Neutralität, Stress	Numerische Antworten auf mathematische Fragen während Fahrsituation	4 Personen
Mozziconacci [MOZ98]	Nieder-ländisch	Ärger, Ekel, Freude, Furcht, Glücklichkeit, Hochmut, Indignation, Langweile, Neutralität, Schuldgefühl, Sorge, Trauer, Wut	Vorlesen emotional neutraler Sätze nach Einstimmung durch emotionale Texte	3 Schauspieler lesen 8 Sätze je dreimal

Tab. 3.40: Auswahl bedeutender Sprachdatenbanken natürlicher Emotionen

Bezeichnung	Sprache	Emotionen	Text	Umfang
Abelin [ABE00]	Schwedisch	Ärger, Dominanz, Ekel, Freude, Furcht, Schüchternheit, Trauer, Überraschung	emotional neutral	1 Person
EMO-DB [EMO05]	Deutsch	Ärger, Ekel, Freude, Furcht, Langweile, Neutralität, Trauer	emotional neutral	10 Personen, 5 weiblich, lesen je 10 Sätze je Emotion
DES [ENG96]	Dänisch	Ärger, Freude, Neutralität, Trauer, Überraschung	emotional neutral	4 Personen, 2 weiblich lesen je 2 Wörter, 9 Sätze und 2 Passagen je Emotion
Pereira [PER00]	Englisch	Ärger, Freude, Neutralität, Trauer, Wut	emotional neutral	2 Personen lesen je 2 mal je Emotion einen Satz und vier Ziffern
van Bezooijen et al. [BEZ84]	Nieder-ländisch	Ärger, Ekel, Freude, Furcht, Interesse, Missachtung, Neutralität, Scham, Trauer, Überraschung	emotional neutral	8 Personen, 4 weiblich, lesen 4 Phrasen je Emotion
Yacoub et al. [YAC03]	Englisch	Ärger, Ekel, Elation, Freude, Furcht, Interesse, Langweile, Missachtung, Neutralität, Panik, Scham, Stolz, Trauer, Verzweiflung, Wut	emotional	8 Schauspieler 2.433 Phrasen

Tab. 3.41: Auswahl bedeutender Sprachdatenbanken gespielter Emotionen durch Vorlesen

Als weitere Datenbank für den Test akustischer Erkennungsleistung wird die öffentliche *Berlin*

Database Of Emotional Speech, kurz *EMO-DB*[123] verwendet [KIE00], [EMO05], die einen Vergleich zu deutscher Sprache erlaubt. Sie besitzt ein leicht abgewandeltes Emotionsset des MPEG-4-Sets, in dem *Langweile* an Stelle von Überraschung enthalten ist. Zehn emotional neutrale Sätze (siehe Kap. A.3, Tab. A.7) wurden von zehn ebenfalls professionellen Schauspielern gesprochen, von denen fünf weiblich sind. Die Aufnahme erfolgte mit 16 Bit, 16 kHz und PCM kodiert, unter Studiobedingungen. Annotiert wurden die Sätze von 20 Personen hinsichtlich der Emotion, der Ausgeprägtheit und der Natürlichkeit der Aufzeichnungen. Die Datenbank verfügt über insgesamt 816 Sätze, wobei nur ein kleinerer, allgemein verwendeter Satz von insgesamt 488 Beispielen mindestens 80% Erkennungsleistung und mindestens 60% Natürlichkeit im menschlichen Perzeptionstest aufweist. Dieser besitzt folgende Verteilung unter den Emotionen:

Ärger	Ekel	Furcht	Langweile	Freude	Neutral.	Trauer	Summe
127	38	55	79	58	78	53	**488**

Tab. 3.42: Übersicht Verteilung der Sprachsamples unter den Emotionen, Datenbank EMO-DB

Schließlich wird hier, um auch spontane Sprachbeispiele zu berücksichtigen, die Datenbank *Aibo Emotion Corpus 1.0 (AEC)*[124] der Universität Erlangen-Nürnberg [BAT04] verwendet. Es handelt sich dabei um Sprachaufzeichnungen von insgesamt 51 Kindern zweier deutscher Schulen, davon 30 weiblich, im Alter von 10 a bis 13 a. Diese spielen mit dem Roboter *Aibo* der Firma Sony in einer WOO-Studie: Es wird per Fernsteuerung durch einen versteckten Versuchsleiter vorgetäuscht, dass der Roboter Sprache versteht. Die Aufnahme erfolgte mittels Headset und DAT-Rekorder. 9 h aufgezeichnetes Sprachmaterial wurden automatisch auf Wortebene segmentiert, und auf 16 Bit und 16 kHz abwärts konvertiert. Es ergaben sich dabei 48.412 Wörter in 13.588 Phrasen. Diese wurden von fünf Annotatoren auf Wortebene in zehn Kategorien (vgl. Tab. 3.40) eingeordnet, wobei nur vier davon mit einer ausreichenden Zahl von Daten vertreten sind. Diese sind *Ärger, Bemutterung, Betonung* und *Neutralität*. *Bemutterung* und speziell *Betonung* sind dabei eher als Sprechstil zu werten. Tab. 3.43 zeigt die Verteilung der Beispiele nach Klasse und Übereinstimmung der Annotatoren. Es zeigt sich ein deutliches Missverhältnis zu Gunsten neutraler Beispiele. Mit $\frac{3}{5}$ Übereinstimmung ergaben sich weitere 14.146 Beispiele. Mit $\frac{2}{5}$ Konsonanz schließlich noch 3.664.

[#]	Ärger	Bemutter.	Betonung	Neutralität	Rest	Summe
5/5 Übereinstimmung	191	16	67	12.646	2	12.922
4/5 Übereinstimmung	473	509	550	16.104	24	17.660
Summe	**664**	**525**	**617**	**28.750**	**26**	**30.582**

Tab. 3.43: Übersicht Verteilung der Sprachsamples nach Emotion, Datenbank AEC

Zu den Aufzeichnungen existiert eine Transkription des gesprochenen Inhalts inklusive nonverbaler Abschnitte, und zwar *Atmen, Lachen, Husten* und *Geräusche* sowie *nasale, vokale* und *gemischte*

[123] Datenbank der TU Berlin.

[124] Dieser wird auch *CEICES*-Korpus genannt. CEICES steht dabei für „*Combining Efforts for Improving Automatic Classification of Emotional User States*", und bezeichnet eine Forschungskooperation zwischen FAU Erlangen, TAU, QUB, LIMSI-CNRS, ITC-IRST, UA, interACT, der Universität Karlsruhe sowie der TU München.

Hesitationen. Hierzu sind weitere Details in Kap. 4.7.1, Tab. 4.22 ersichtlich. Für die Versuche mit dieser Datenbank werden hier ähnlich wie beim Vorgehen bei der Datenbank EMO-DB nur Beispiele mit mindestens 80% Sicherheit bezüglich der Annotation gewählt[125]. Es ergeben sich somit 2.606 Samples auf Wortbasis und 1.480 auf Phrasenbasis, von denen 352 volle Übereinstimmung in der Annotation besitzen[126].

In weiteren hier gewünschten Betrachtungen treten jedoch Problemstellungen auf, zu denen keine konkreten und allgemein verfügbaren Datenbanken vorhanden sind. Hierin begründet sich die Motivation zur Erstellung vier eigener Korpora akustischer Emotionsbeispiele, die in den folgenden Experimenten verwendet werden: Die erste ist die Datenbank *EA-WSJ*[127] mit emotional gespielten Sätzen aus dem Schatz der *Wall Street Journal-Datenbank*[128] [PAU92]. Je Proband wurden 26 Sätze je zweimal aufeinander folgend in zufälliger Reihenfolge ärgerlich oder neutral gesprochen. Die Zuordenbarkeit zu einer der beiden gespielten Emotionen neutral oder ärgerlich wurde durch zwei Probanden gewährleistet. Nur im Falle einer Übereinstimmung beider Annotatoren wurde ein Sample in die Datenbank aufgenommen, wodurch die menschliche Erkennungsleistung als vollständig korrekt zuordenbar angesehen werden kann. Zu diesen Sätzen existiert auch eine phonetische Annotation. Im Rahmen der Aufzeichnung wurde jeder Proband gebeten, seine Englischkenntnisse und schauspielerischen Leistungen auf einer Skala von Eins, entsprechend sehr gut, bis Sechs für ungenügend, selbst einzuschätzen. Englischkenntnisse wurden dabei von Eins bis Drei und im Mittel mit 2.0 - das Schauspieltalent von Eins bis Vier und mit 2,6 im Mittel bewertet.

Um eine aussagekräftige Datenbank mit möglichst vielen Beispielen verschiedener Sprecher im Vergleich zu den vorgestellten öffentlichen Datenbanken unter akustischen Idealbedingungen zu erhalten, wurden weiterhin Sprachbeispiele der Emotionen des MPEG-4 Basissets mit ergänzender Neutralität von 39 Personen in einem reflexionsarmen Raum über einen Zeitraum von zwei Wochen aufgezeichnet. Die hohe Gesamtzahl unter konstanten akustischen Konditionen wurde dabei durch Schauspielen der Emotionen wie bei den Datenbanken EMO-DB und DES erzielt. Die Datenbank trägt im Weiteren daher die Bezeichnung EA-ACT. Der gesprochene Inhalt war für die Probanden zu jeder Zeit frei wählbar, damit diese einen für sie persönlich besonders passenden aussuchen konnten. Weiterhin sollten die seit mindestens drei Jahren in Deutschland lebenden Sprecher unterschiedlicher Herkunft diejenige Sprache frei wählen, mit der sie sich am besten in den emotionalen Zustand versetzen konnten. Zwei weitere männliche Probanden im Alter von 23 a und 30 a wählten aus den so erhaltenen Äußerungen 2.720 aus, so dass sich je Sprecher und Emotion neun oder zehn Beispiele in adäquater Natürlichkeit und übereinstimmender Annotation zwischen dem Sprecher und den Annotatoren ergaben. Die intendierte Emotion war während der

---

[125] Um eine Gleichverteilung unter den Emotionen anzunähern, werden nur 800 neutrale Beispiele aus den vorhandenen mit Sicherheit von 100% selektiert. Hierzu werden von jedem Kind in gleicher Zahl die ersten Beispiele genommen.

[126] Da in dieser Arbeit eine phrasen- und keine wortweise Betrachtung vollzogen wird, wurden die Phrasen entsprechend eines Mehrheitsentscheids unter ihren Wörtern annotiert.

[127] Die Bezeichnung EA wurde hier zur Vereinfachung für *Emotionsdatenbank - akustisch* gewählt und wird für die weiteren Datenbanken entsprechend getroffen. WSJ kürzt die größte englisch sprachliche Tageszeitung der USA Wall Street Journal ab. Die Sätze stammen aus dem November 1993.

[128] Bekannte Datenbank für automatische Spracherkennung (vgl. Kap. 7.1.1).

Verschriftung unbekannt.

Beschreibung	**EA-ACT**	**EA-WSJ**	**EAL-F+W**	**EA-CAR**
Domäne	Versuch	Wall Street Journal	Spielfilm Webchat	Fahrzeug
Inhalt	Frei gewählt	26 Sätze aus WSJ1 S3 P0, je 2x	Drehbuchakte Webchat	MMI-Interaktion Fahrzeug
Emotionen	Basisset, Neutralität	Ärger, Neutralität	Basisset, Neutralität	Ärger, Freude, Irritation, Neutralität
Gesprochene Sprache	Deutsch, Englisch, Französisch, Mandarin	Englisch	Englisch	Deutsch
Art	Gespielt	Gespielt	Gespielt	Real
Sprecher	39, 5 weiblich 21 a - 30 a Ø 25,0 a	10, 1 weiblich 23 a - 33 a Ø 26,0 a	3, 1 weiblich 22 a - 30 a Ø 27,0 a	10, 2 weiblich 22 a - 26 a Ø 23,2 a
Muttersprache	28x Deutsch, 1x Englisch, 1x Französisch, 1x Mandarin, 3x Serbisch, 5x Türkisch	9x Deutsch, 1x Serbisch	2x Deutsch, 1x Portugies.	Deutsch
Phrasenzahl	2.720	520	3.125	2.022
Verteilung	gleich	gleich	ungleich	ungleich
Phrasenlänge	1-18 Wörter, Ø7,7 Wörter	3-18 Wörter, Ø10,9 Wörter	1-20 Wörter Ø7,0 Wörter	1-6 Wörter Ø3,2 Wörter
Sampling / Kodierung	16 kHz, 16 Bit, PCM	44,1 kHz, 16 Bit, PCM	44,1 kHz, 16 Bit, PCM	11 kHz, 16 Bit, PCM
Mikrophon	Kondensator Yoga EM240	Kondensator AKG-1000S	Kondensator AKG-1000S	Kondensator Yoga EM240
Umgebung	Usability Labor	Usability Labor	RAR	Fahrsimulator
Annotation	Emotion	Emotion, Inhalt, Phonetische Annotation	Emotion, Inhalt	Emotion
Annotatoren	2 plus Sprecher	2	-	3 plus Sprecher
Öffentlich	Nein	Ja	Nein	Nein

Tab. 3.44: Übersicht über die im Rahmen dieser Arbeit aufgezeichneten akustischen Datenbanken zur Emotionserkennung

Tab. 3.44 gibt einen Überblick über die diskutierten und weiteren Variablen der erstellten Datenbanken zur akustischen Emotionserkennung. Die Datenbanken EAL-F+W und EA-CAR werden in Kap. 4.7.1 und Kap. 7.3.2 vorgestellt. Sie sind zur besseren Übersicht in dieser Tabelle mit angeführt.

3.12.2 Menschliche Leistung in der Diskrimination

Als Richtlinie für die Bewertung der Leistung, die mit automatischen Verfahren erreicht werden kann, soll zunächst die menschliche Sicherheit in der Erkennung von Emotion aus dem akustischen Sprachsignal analysiert werden. Für die Interpretation von besonderer Relevanz ist dabei, dass ein menschlicher Entscheider grundsätzlich den gesprochenen Inhalt mitbewerten wird, insofern die Sätze nicht emotional neutral oder in einer Fremdsprache sind.

Zur Datenbank EMO-DB existiert hierzu ein Hörversuch mit 15 Versuchspersonen mit einem durchschnittlichen Alter von 24,8 a (Minimum 22 a, Maximum 28 a), davon vier weiblich [BUR00]. Diesen wurde der emotional unbestimmte Satz „*In sieben Stunden wird es soweit sein.*“ in sieben Emotionen von zwei der zehn enthaltenen Sprechern präsentiert. Die Probanden hörten die insgesamt 14 Beispiele in zufälliger emotionaler Reihenfolge und mussten jeden Satz einer der sieben Emotionskategorien zuordnen. Über die Anzahl der Beispiele je Emotion besaßen die Testpersonen keine Kenntnis. Es ergab sich folgende Verteilung in der gemittelten Erkennungsleistung:

Akk. [%]	Ärger	Ekel	Furcht	Freude	Neutral.	Langweile	Trauer
μ	73,3	63,3	96,7	93,3	99,9	90,0	73,3

Tab. 3.45: Menschliche Leistung in der Klassifikation akustischer Emotionsbeispiele, Datenbank EMO-DB, mittlere Leistung 84,3%

Ein ähnliches Experiment wurde auf Basis der Datenbank DES durchgeführt [ENG96], [VER04A]: Die Güte der Datenbank wurde von 20 Testhörern, davon 10 weiblich, der Altersstufen 19 a bis 59 a und 38 a im Mittel bewertet[129], wobei die Erkennungsleistung im Bereich von 59% bis 80% mit einem Mittelwert von 67% schwankte. Tab. 3.46 zeigt die Ergebnisse nach Emotion:

Akk. [%]	Ärger	Freude	Neutral.	Trauer	Überras.
μ	75,1	56,4	60,8	85,2	59,1

Tab. 3.46: Menschliche Leistung in der Klassifikation akustischer Emotionsbeispiele, Datenbank DES, mittlere Leistung 67,3%

Erscheint die Stichprobenzahl zu gering, um generelle Rückschlüsse über die interklassenspezifische Verteilung zu erlauben, so zeigt sich dennoch, dass rein auf Basis der Sprechweise Emotion transmittiert werden kann.

[129] Eine genaue Beschreibung der Testbedingungen ist in [ENG96] zu finden.

Da zur Datenbank AEC kein gesonderter Perzeptionstest existiert, wird für den späteren Vergleich die mittlere Übereinstimmung der Annotatoren von 84,8%[130] aus den im vorhergehenden Abschnitt gewählten Beispielen stellvertretend gewählt.

Auf der hier vorgestellten Datenbank EA-ACT soll darüber hinaus geprüft werden, inwiefern sich der Sprecher im Anschluss an die Aufnahme selbst einstufen kann. Diese Reklassifikation wurde von zehn Sprechern durchgeführt. Jeder Sprecher hat dabei nur seine selbst gespielten Emotionen nochmals validiert. Tab. 3.47 zeigt die erhaltene Konfusionsmatrix nach Wahrheit und Prädiktion sowie die mittlere Gesamtleistung.

[%] Wahr Prädiktion	Ärger	Ekel	Freude	Furcht	Neutral.	Trauer	Überras.
Ärger	**92,0**	3,3	0,0	0,8	0,0	0,0	4,0
Ekel	4,2	**80,3**	0,5	9,5	0,0	0,9	4,6
Freude	0,7	4,2	**85,3**	5,5	2,2	2,2	0,0
Furcht	1,3	6,9	3,5	**81,3**	0,0	3,1	0,0
Neutralität	0,3	1,0	2,9	1,9	**83,5**	10,4	0,0
Trauer	0,0	13,3	3,8	2,8	3,9	**76,3**	0,0
Überras.	1,7	6,8	2,5	1,0	0,0	0,4	**87,5**

Tab. 3.47: Menschliche Leistung in der Reklassifikation akustischer Emotionsbeispiele, mittlere Leistung 83,7% ± 2,1%. Überras. steht für Überraschung, Neutral. Für Neutralität.

Weiteren sechs Probanden wurden je 150 zufällig ausgewählte Sprachbeispiele fremder Sprecher vorgeführt. Dabei sank die Klassifikationsleistung auf 64,7% ± 9,6% im Mittel.

3.12.3 Evaluierungsstrategien

Als Messverfahren für die Güte der betrachteten Algorithmen wird eine j-fach stratifizierte Kreuzvalidierung *(SCV)*[131] gewählt [WIT00]. Dabei wird die gesamte, aus L Instanzen bestehende, Datenmenge $\mathcal{L}$ mit $L = |\mathcal{L}|$ in j getrennte Teilmengen mit $j \leq L$ partitioniert. In j Testdurchläufen werden jeweils die j-te Teilmenge als Testmenge und die $j-1$ verbleibenden Teilmengen als Trainingsmengen verwendet. Als Gesamterkennungsleistung wird der Mittelwert aus den Einzelerkennungsleistungen[132] der j Durchläufe berechnet. Mit steigendem j begünstigt sich dabei das Verhältnis zu Gunsten des Lernmaterials, was den Gesamtrechenbedarf und in der Regel auch die Erkennungsleistung steigert, während die Stichprobenzahl je Durchlauf sinkt. Als Modus Procedendi haben sich im Allgemeinen für j die Werte zwei, drei, fünf, zehn und zwanzig entsprechend einem Trainingsanteil von 50%, 66,7%, 80%, 90% und 95% durchgesetzt. Eine gleichmäßige Distribution unter den j Teilmengen, eine sogenannte *Stratifizierung*, erlaubt eine

[130] Diese Angabe bezieht sich auf die hier gewählte Phrasenbasis. Auf Wortbasis sind dies 85,2%.

[131] Abkürzung der anglistischen Bezeichnung *Stratified Cross Validation*.

[132] Die Begriffe *Erkennungsleitung*, *Erkennungsrate* und *Akkuratheit* werden hier synonym für das Verhältnis richtig zugeordneter Beispiele zur Gesamtzahl von Testbeispielen verwendet. Im Gegenzug hierzu ergibt sich die *Fehlerrate* aus der Gesamtzahl falsch zugeordneter Beispiele im Verhältnis zur Gesamtzahl.

Minimierung der Varianz der Fehlerraten. Die folgende Abb. 3.48 zeigt den Einfluss des Partitionierungsfaktors j auf die Erkennungsrate anhand der Datenbanken DES und EMO-DB. Erwartungsgemäß steigt diese mit steigendem j, jedoch nicht streng monoton.

Hauptvorteil und Begründung des Einsatzes dieses Verfahrens ist die Möglichkeit, einen trainingsdisjunkten Test auf allen vorhandenen Daten durchzuführen. Dies ist im hier betrachteten Einsatzgebiet entscheidend, da vergleichbar geringe Datenbankgrößen zu beklagen sind.

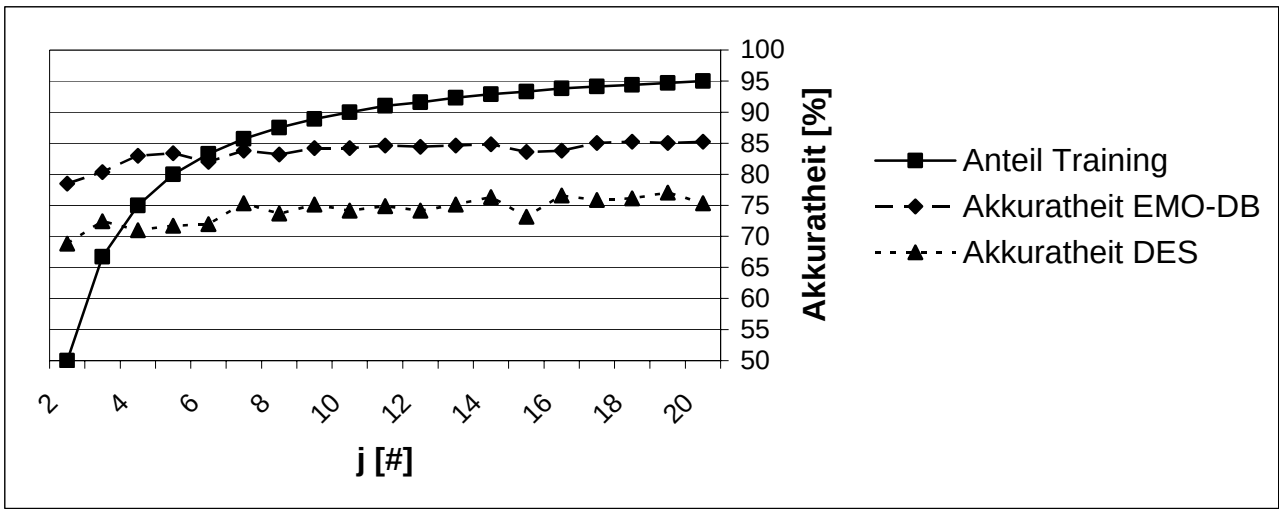


Abb. 3.48: Einfluss des Partitionierungsfaktors j in der Kreuzvalidierung auf die Erkennungsleistung, Klassifikator SVM

Einen geringeren Rechenaufwand erfordert die Variante, einen prozentualen Anteil der gesamten Datenmenge $\mathcal{L}$ als Testset festzulegen. Das Training erfolgt dann entsprechend mit den verbleibenden Daten der Menge $\mathcal{L}$. Da auf diese Art nur ein Trainings- und ein Testlauf erforderlich sind, bietet sich diese Möglichkeit im Besonderen zur Bewertung aufwändiger Verfahren an. Nachteilig ist die entsprechend geringere Aussagekraft der Evaluation.

Als eine weitere Evaluationsstrategie wird auch personenunabhängige Erkennungsleistung betrachtet. Dabei wird sichergestellt, dass keine Daten einer Testperson im Trainingsmaterial enthalten sind. Grundsätzlich ist zu erwarten, dass ein großer Schatz an Lernmaterial, ins Besondere vieler Personen, eine sinnvolle Basis zur personenunabhängigen Erkennung bildet. Auf Grund der in der Regel geringen Datenbankgrößen in der Emotionserkennung bietet sich die *Leave-One-Out (LOO)* Methode an, bei der das Trainingsmaterial optimal genutzt wird: In ihrer strengen Definition bezieht sich die Bezeichnung auf eine Kreuzvalidierung mit einer Zahl j von Testdurchläufen gleich der Gesamtheit L der Instanzen der Datenmenge $\mathcal{L}$. Speziell in der Sprachverarbeitung stellt das Testverfahren *Leave-One-Speaker-Out (LOSO)* eine Variante dar, bei der die Daten jeweils eines Sprechers im Training nicht verwendet werden. Diese werden im unmittelbaren Anschluss als Testmaterial genutzt, um sprecherunabhängige Erkennungsleistung zu bewerten. Die Angabe der Erkennungsrate bei einer Validierung mittels LOSO erfolgt dabei als Mittelwert über alle Sprecher. Da eine hohe Zahl an Durchläufen notwendig ist, um eine Aussage über mehrere Sprecher treffen zu können, werden alternativ Personen gruppenweise zusammengefasst.

Abschließend sei erwähnt, dass sich Aussagen zur Signifikanz eines Ergebnisses im Folgenden jeweils auf die Nullhypothese $H_0 = 0$ keiner Änderung und das weit verbreitete Signifikanzniveau von $\alpha = 0,05$ beziehen. Als Test wird ein zweiseitiger *Student-T-Test* gewählt. Auf explizite Angabe eines p-Werts wird verzichtet.

3.12.4 Evaluierung der akustischen Analyse

Als erstes werden hier statische und dynamische Modellierung miteinander verglichen. Hierzu werden als Basiskonturen 15 MFCC gewählt, welche die höchste Verbreitung in der Sprachverarbeitung aufweisen. Diese werden einmal alleine, und anschließend zusammengefasst mit Geschwindigkeits- und Beschleunigungswerten angewandt. Zusätzlich erfolgt ein Vergleich zur systematischen Generierung von Funktionalen aus ihnen. Tab. 3.49 zeigt die jeweils maximal erzielte Erkennungsleistung unter optimaler Parameterkonfiguration für Versuche mit der Datenbank EMO-DB. Als Klassifikatoren für dynamische Modellierung werden DTW und Links-Rechts-HMM eingesetzt, für statische Modellierung werden zum Vergleich leistungsstarke SVM gewählt. Zusätzlich ist in Klammern die Dimension des jeweiligen Merkmalvektors angegeben.

Akkuratheit [%]	Konturen, DTW	Konturen, HMM	Funktionale, SVM
MFCC	47,5 (15 dim)	55,1 (15 dim)	68,4 (60 dim)
$MFCC + \partial + \partial\partial$	50,8 (45 dim)	59,3 (45 dim)	**73,8 (120 dim)**

Tab. 3.49: Evaluierung der dynamischen und statischen Modellierung, Datenbank EMO-DB, 2-fach SCV

HMM zeigen sich erwartungsgemäß auf Grund der komplexeren Modellierung leistungsstärker als DTW, doch es ist deutlich zu sehen, dass der Ansatz der Funktionalbildung eine höhere Akkuratheit im Vergleich zur direkten Klassifikation anhand von Basiskonturen aufweist. Dies liegt in erster Linie daran, dass der gesprochene Inhalt dynamisch übermodelliert wird, obwohl die Datenbank EMO-DB nur zehn verschiedene Sätze beinhaltet. Für beliebigen Inhalt verstärkt sich diese Diskrepanz. Ein ausführlicher Vergleich hierzu findet sich in [SCH03A]. Die Integration der Ableitungen hingegen bringt in jedem Fall einen zusätzlichen Gewinn.

In Tab. 3.50 werden die in Kap. 3.7 vorgestellten statischen Klassifikatoren bezüglich erzielbarer maximaler Leistung unter optimaler Konfiguration miteinander verglichen. Als Datenbanken sind die drei öffentlichen Datenbanken EMO-DB, DES und AEC gewählt. Für kNN ergibt sich bei einer Analyse in Einzelschritten eine Zahl zwischen zehn (EMO-DB, AEC) und zwölf (EMO-DB) Nachbarn als Optimum. Bezüglich SVM sind hier polynomiale Kernel-Funktionen mit einem Grad von Eins (DES, AEC) oder Zwei (EMO-DB) im Vergleich zu einer RBF-Kernel-Funktion günstiger. Bei MLP hat sich eine versteckte Schicht mit der halben Zahl an Neuronen im Vergleich zur Eingangsschicht mit durchgängig Sigmoid-Transferfunktion als ideal erwiesen. Für alle Ensemblevarianten wurde eine Iterationszahl von je Zehn als Kompromiss zwischen Rechenzeit und Leistung gewählt. Eine höhere Zahl kann zu weiteren Verbesserungen führen, die sich jedoch als nicht signifikant erwiesen. Die Angabe der Leistung erfolgt in Tab. 3.50 jeweils einmal mit vollem Merkmalsset, und einmal mit durch SVM-SFFS auf optimale Größe verringertem. Durch die Reduktion ergibt sich für alle Klassifikatoren bis auf BN und für alle Datenbanken eine

deutliche Leistungssteigerung. Die Klassifikatoren DT und MLP, die bereits selbst Merkmale unterschiedlich gewichten, profitieren davon jedoch weniger. Die jeweiligen Sets und ihre Größe variieren stark in Abhängigkeit von der Datenbank. Sie sind in Kap. A.2 zu finden, wo neben der Angabe des SVM-SFFS Rangs auch die IGR zu jedem Attribut angegeben ist. Bemerkenswert dabei erscheint, dass die Reihenfolge der Merkmale bei Anwendung der IGR-FS stark von der bei SFFS abweicht. Dies begründet sich darin, dass einzelne, weniger relevante Merkmale bei der Setoptimierung ein Set komplettieren können.

Akkuratheit [%]	EMO-DB		DES		AEC	
	Top 75	Alle	Top 100	Alle	Top 107	Alle
1NN	75,8	63,5	31,2	29,5	73,5	70,7
kNN	78,9	67,6	50,7	39,4	74,5	70,7
MLP	86,5	84,8	72,0	65,2	78,7	76,5
DT C4.5	61,5	61,1	51,7	48,1	63,6	62,2
AdaBoosting C4.5	74,6	72,3	58,7	55,1	71,4	70,4
Bagging C4.5	74,8	70,7	58,9	51,9	72,4	71,5
MultiBoosting C4.5	74,6	72,5	58,7	55,1	71,4	70,7
NB	74,0	73,6	52,4	45,2	67,0	65,2
BN	74,4	72,1	51,7	52,9	68,7	67,6
SVM 1-vs-1	**87,5**	84,8	**74,2**	65,9	**79,1**	74,9
StackingC MLR kNN BN MBC4.5	83,2	78,1	60,1	56,3	77,0	73,5
StackingC MLR 1NN NB SVM C4.5	80,5	76,2	69,1	64,0	77,5	75,7
Voting 1NN NB SVM C4.5	79,9	76,0	59,4	49,5	76,0	74,4
StackingC MLR 1NN NB C4.5	79,9	75.4	54,1	51,7	74,7	71,9
Voting 1NN NB C4.5	78,5	73,2	51,7	46,9	74,7	72,5

Tab. 3.50: Akkuratheit unterschiedlicher Klassifikatoren zur akustischen Verarbeitung, alle Merkmale oder Top-N mit SVM-SFFS,10-fach SCV

Als stärkster Klassifikator zeigt sich generell SVM. Der in dieser Arbeit vorgestellte ML-SVM Ansatz weist die insgesamt maximale Leistung zur Behandlung multipler Klassen auf: Für die Datenbank EMO-DB ergeben sich 88,8%, für die Datenbank DES 76,2% bei optimiertem Merkmalssatz. Da ML-SVM aber aufwändiger zu trainieren sind, wird als Standardklassifikator im Weiteren 1-vs-1 SVM verwendet. Eine polynomiale Kernel-Funktion hat sich dabei stets als optimal erwiesen, wobei die ideale Polynormordnung schwankt. Dicht gefolgt werden SVM von ANN. Schwach hingegen sind die distanzbasierten und Bayesschen Ansätze. Ensembles zeigen sich in der Regel stärker als ihre Basisklassifikatoren - sobald jedoch SVM im Ensemble enthalten sind, sind diese schwächer als SVM alleine. MultiBoosting bringt im Vergleich zu AdaBoosting oder Bagging nur einen geringen Gewinn an Akkuratheit.

Abb. 3.51, Abb. 3.52 und Abb. 3.53 zeigen nun den unterschiedlichen Effekt der drei in Kap. 3.9 vorgestellten Verfahren zur Selektion von Merkmalen SVM-SFFS, IGR-FS und PCA-FS für die

drei öffentlichen Datenbanken EMO-DB, DES und AEC.

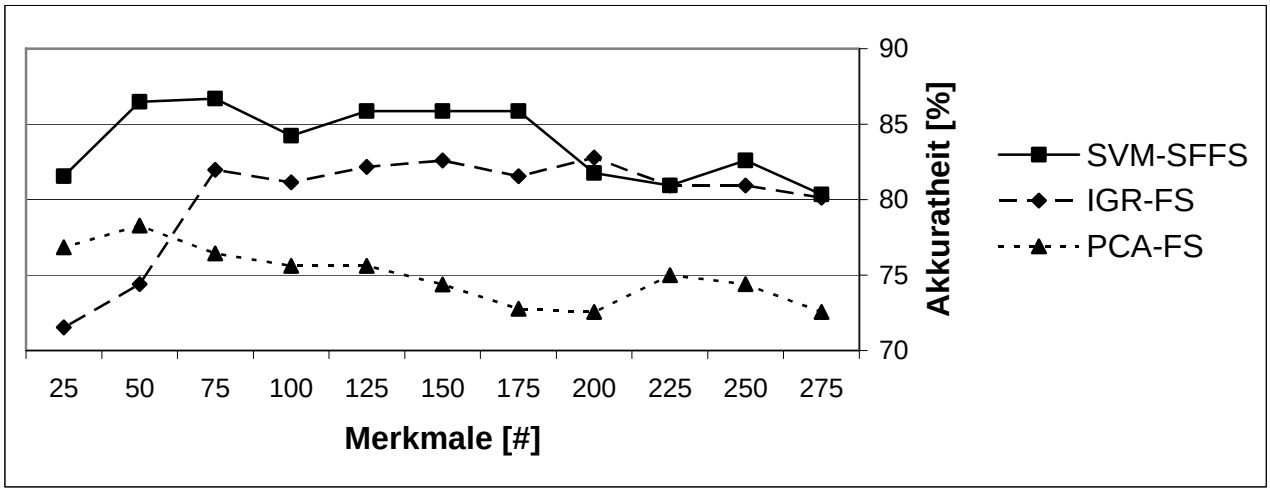


Abb. 3.51: Vergleich FS Verfahren, Datenbank EMO-DB, Klassifikator SVM, 3-fach SCV

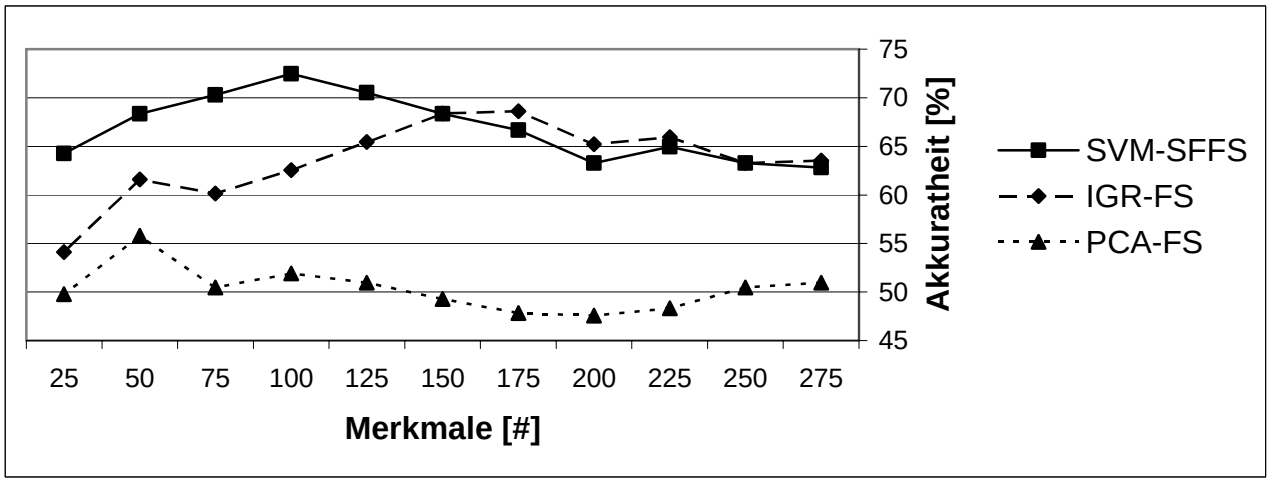


Abb. 3.52: Vergleich FS Verfahren, Datenbank DES, Klassifikator SVM, 3-fach SCV

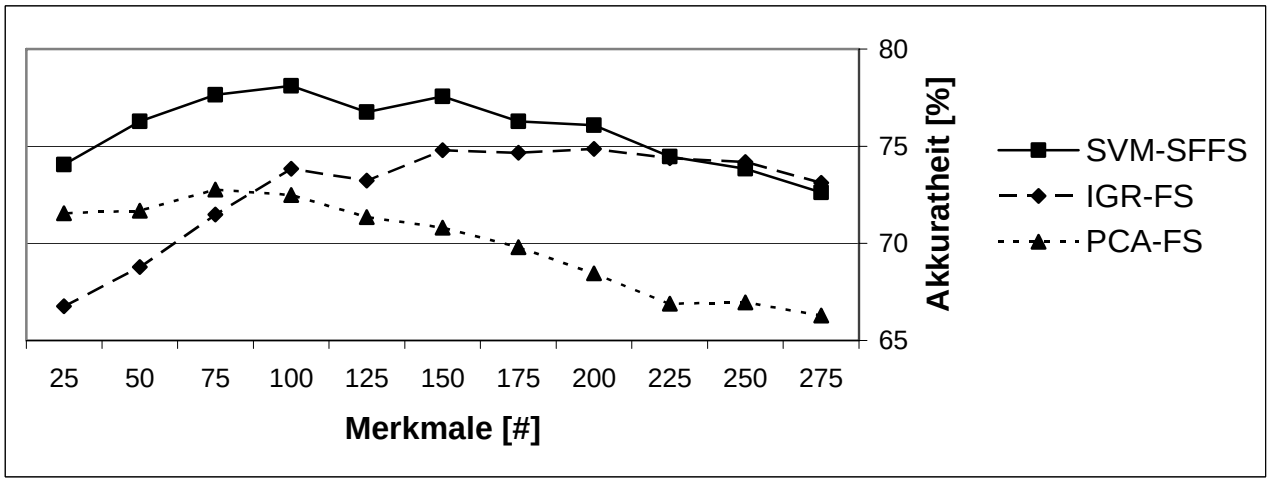


Abb. 3.53: Vergleich FS Verfahren, Datenbank AEC, Klassifikator SVM, 3-fach SCV

Die höchste Leistung insgesamt ergibt sich dabei jeweils mit SVM-SFFS bei deutlich reduzierter Setgröße, da nur hier ein Verbund von Merkmalen optimiert wird. PCA-FS hingegen ist die ungünstigste Wahl, nicht zuletzt auch auf Grund der Tatsache, dass der Extraktionsaufwand bei PCA-FS gesteigert wird, während er bei SFFS und IGR-FS gesenkt wird. Die vergleichsweise geringe Reduktionsleistung beim PCA Ansatz liegt auch in der Untermodellierung der Komplexität durch singuläre Normalverteilungen während der Selektion. Allen drei Verfahren ist gemein, dass sich ein Optimum bei einem reduzierten Set einfindet. Die Klassifikation bei voller Setgröße ist demnach die schlechteste Wahl.

Die folgende Darstellung zeigt das Verhalten diverser Klassifikatoren bei Reduktion der Dimensionalität von $\underline{x}$ für die Datenbank EMO-DB. Für diesen Vergleich wurde IGR-FS ausgewählt, da das Verfahren unabhängig vom Zielklassifikator ist. Der Entscheidungsbaum C4.5 zeigt dabei erwartungsgemäß die geringste Anfälligkeit, da dieser bereits selbst eine Priorisierung von Attributen auf Entropiebasis vollzieht. 1NN hingegen zeigt sich klar am empfindlichsten gegenüber hoher Dimensionalität.

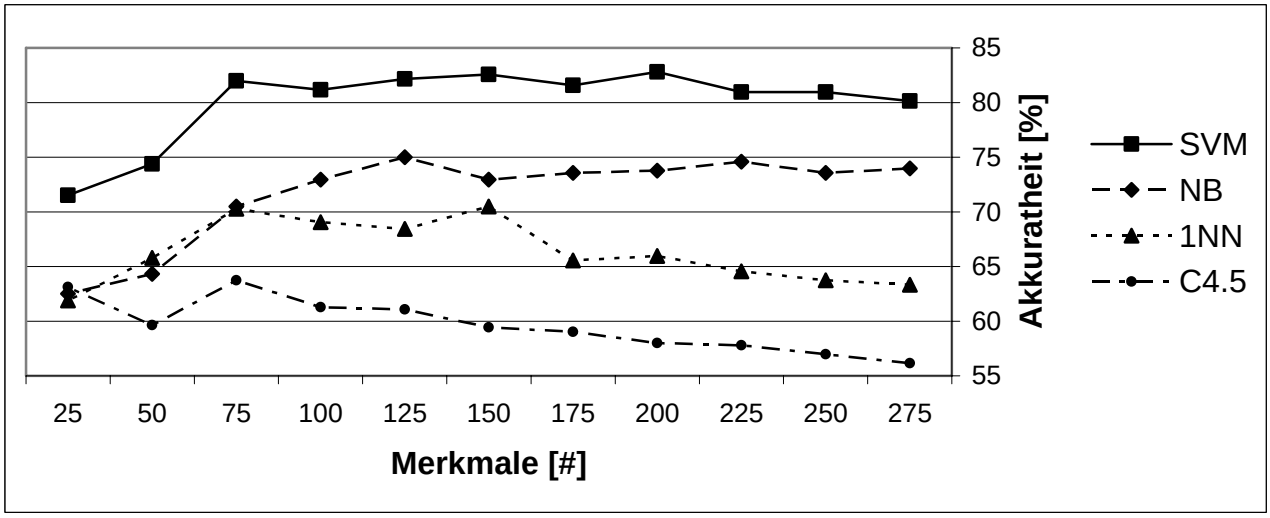


Abb. 3.54: Merkmalsreduktion mit IGR-FS, Verhalten unterschiedlicher Klassifikatoren, Datenbank EMO-DB, 3-fach SCV

Die Eignung einzelner Merkmalsgruppen zur Emotionserkennung ist in Tab. 3.55 auf unterschiedlichen Datenbanken zu sehen. Diese Merkmalsgruppen sind dabei aufsteigend nach der durchschnittlichen Leistung über alle Datenbanken sortiert. Es zeigt sich, dass MFCC basierte Größen die höchste Relevanz besitzen, obwohl diese den gesprochenen Inhalt stark mitmodellieren. Für die weiteren Gruppen ist die Reihenfolge datenbankübergreifend nicht mehr eindeutig. Die insgesamt zweitwichtigste Gruppe sind Merkmale aus dem FFT-Spektrum, was auch im Kontext der hohen Relevanz der ebenfalls spektrumsbasierten MFCC zu sehen ist. Relativ schwach erscheinen die Gruppen HNR, Dauer und Zeitsignal, was jedoch vor dem Hintergrund zu bewerten ist, dass diese auch diejenigen mit der geringsten Zahl von abgeleiteten Größen darstellen.

Akkuratheit [%]	EMO-DB	DES	AEC	EA-WSJ
HNR	30,3	25,6	38,5	57,9
Dauer	27,5	19,1	49,9	61,4
Zeitsignal	33,4	36,2	52,4	89,8
Grundfrequenz	62,1	32,9	47,8	74,6
Intensität	48,2	39,6	53,9	77,1
Formanten	63,7	41,8	65,7	82,9
FFT Spektrum	70,7	39,9	63,5	86,0
MFCC	77,3	58,0	73,9	98,5
Gesamt	84,8	65,9	74,9	98,1
Top N	**87,5**	**74,2**	**79,1**	**99,2**

Tab. 3.55: Vergleich von Leistung nach Merkmalsgruppen, Klassifikation mit SVM, 10-fach SCV

Durch genetische Generierung von Merkmalen kann die Erkennungsrate mit einem 1NN-Klassifikator von 75,8% auf 77,5% in einer 10-fach SCV auf der Datenbank EMO-DB zusätzlich gesteigert werden. Aus den ursprünglich mittels SVM-SFFS ausgewählten TOP 75 Größen wurde dabei als neues Merkmal *MFCC3Mean+HNRMean* generiert. Wird dieses den Top 75 Merkmalen hinzugefügt, steigert sich auch für eine Klassifikation mit SVM die Erkennungsleistung von 87,5% auf 87,7%. Auf der Datenbank DES wurde in analoger Weise das leistungsfördernde Merkmal *B2StdDev-MFCC3Mean* generiert. Tab. 3.56 zeigt hierzu die Ergebnisse. Für den Korpus EA-WSJ ergibt sich bei Einsatz von genetischer Generierung für einen 1NN eine Steigerung von 92,3% auf 93,3%. Insgesamt lassen sich also durch genetische Generierung weitere signifikante Zugewinne erzielen - diese sind allerdings deutlich unter der Größenordnung derer durch FS erzielter Gewinne.

Akkuratheit [%]	EMO-DB		DES		AEC	
	Top 75	+ Gen.	Top 100	+ Gen.	Top 107	+ Gen.
SVM	87,5	**87,7**	74,2	**74,4**	79,1	79,1

Tab. 3.56: Leistungssteigerung durch genetisch generierte Merkmale, Klassifikation mit SVM, 10-fach SCV, *+ Gen.* steht dabei für das Hinzufügen genetisch generierter Merkmale

Den Ablauf und die Leistungsfähigkeit des bisherigen Ansatzes der selber lernt, welche Information für ihn relevant ist und sich die günstigste Darstellung dieser sucht, soll ein Beispiel in Fortführung von Tab. 3.49 zusammenfassend verdeutlichen. Die Angaben zur Erkennungsleistung beziehen sich auch hier auf die Datenbank EMO-DB in einer zweifachen SCV mit HMM und SVM als Klassifikatoren. Als Standardmerkmale in der Sprachverarbeitung werden wieder MFCC gewählt. Zunächst werden die Verläufe dieser berechnet. Im Anschluss werden systematisch Ableitungen dieser gebildet, hier die Geschwindigkeit und die Beschleunigung, und Funktionale gebildet. Im nächsten Teilschritt erfolgt die Selektion relevanter Funktionale mittels SVM-SFFS um die Komplexität zu reduzieren und so eine kompakte Basis für die kombinierte genetische Generierung und Selektion von Merkmalen zu schaffen. Durch die Veränderung und Kombination von Attributen im letzten Schritt wird unter Einbringung von Zufall nach weiteren günstigeren Repräsentationsformen der Information gesucht. Tab. 3.57 zeigt hierzu die jeweilige Dimensionalität des Merkmalsvektors $\underline{x}$, den günstigsten Klassifikator, der sich aus der Modellierung ergibt sowie die Akkuratheit während des geschilderten Vorgehens. Wie zu sehen ist,

steigt die Erkennungsleistung mit jedem Teilschritt signifikant und streng monoton, wobei die grundlegende Information, 15 MFCC Verläufe, nur günstiger modelliert wird.

Teilschritt	Dimension [#]	Klassifikator	Akkuratheit [%]
Konturextraktion	15	HMM	55,1
Ableitung	45	HMM	59,3
Funktionalbildung	120	SVM	73,8
Reduktion, SFFS	98	SVM	74,8
Genetische Generation	**100**	**SVM**	**75,4**

Tab. 3.57: Beispiel Teilschritte bei der Optimierung des Merkmalsraumes, Basismerkmale 15 MFCC, Datenbank EMO-DB, 2-fach SCV

Für den praktischen Einsatz weiterhin von Interesse ist der Einfluss der Sprecherabhängigkeit, da in vielen Fällen keine Daten des aktuellen Sprechers vorliegen. Tab. 3.58 zeigt hierzu die sprecherabhängige Erkennungsleistung speziell für die Datenbank DES. Es werden jeweils exklusiv die Daten eines Sprechers in einer 10-fach SCV gelernt und getestet. Im Anschluss wird für jeden Sprecher das Set von Top 100 Attributen individuell mit SVM-SFFS optimiert. Hierbei zeigt sich eine interpersonelle Schwankung nahe 10%. Im Mittel ist die Erkennung mit 83,0% jedoch deutlich höher als bei einer 10-fach SCV über den gesamten Korpus, die bei Einsatz des gleichen Klassifikators und gemeinsamer Merkmalsselektion nur 74,2% beträgt. Dies liegt einerseits daran, dass Fremdsprecheräußerungen das Modell verfälschen können, und andererseits an der beschriebenen lokalen Optimierung der Merkmalssets von Partitionen des gesamten Datensatzes.

Akkuratheit [%]	Frauen		Männer		
	DHC	KLA	JZB	HO	Ø
276 Merkmale	71,6	64,9	72,3	70,0	69,7
Top 100/Sprecher	87,3	75,7	89,1	80,0	**83,0**

Tab. 3.58: Erkennungsleistung sprecherabhängig, Datenbank DES, Top 100 Merkmale einer personenbezogenen SVM-SFFS, Klassifikator SVM, 10-fach SCV

Als weiteres Experiment wurde die Datenbank AEC in einer 2-fach Kreuzvalidierung entsprechend der in [BAT05] festgelegten Aufteilung in die zwei Schulen, aus denen die Kinder stammen, die mit dem Aibo-Roboter spielen, evaluiert. Hierdurch sinkt die Erkennungsrate auf 70,6% im Mittel im Vergleich zu 79,1% bei einer 10-fach SCV über den gesamten Korpus. Dies begründet sich in dem Umstand, dass bei einer 10-fach SCV nahezu doppelt so viel Trainingsmaterial und im Normalfall auch Daten der Testpersonen je Durchlauf zur Verfügung stehen.

Um eine sprecherunabhängige Erkennungsrate auf der Datenbank EA-WSJ zu erzielen, wird diese in zwei Sets zu je fünf Sprechern für Training und Test in einer Kreuzvalidierung geteilt. Die mittlere Erkennungsleistung beträgt dabei 85,6% bei Verwendung des vollen Merkmalsvektors und SVM als Klassifikator. Im Vergleich zur Erkennungsleistung von 99,2% zeigt sich auch hier, dass eine sprecherunabhängige Erkennung sich ungleich schwieriger gestaltet.

Es soll nun in den folgenden drei Darstellungen betrachtet werden, inwiefern sich die Erkennungsrate unter Streichung von weniger relevanten Emotionen steigern lässt.

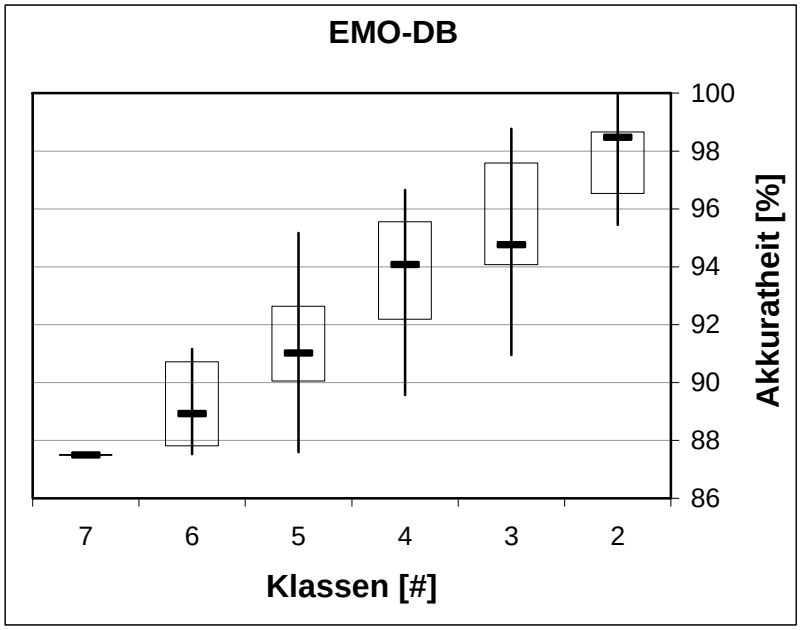
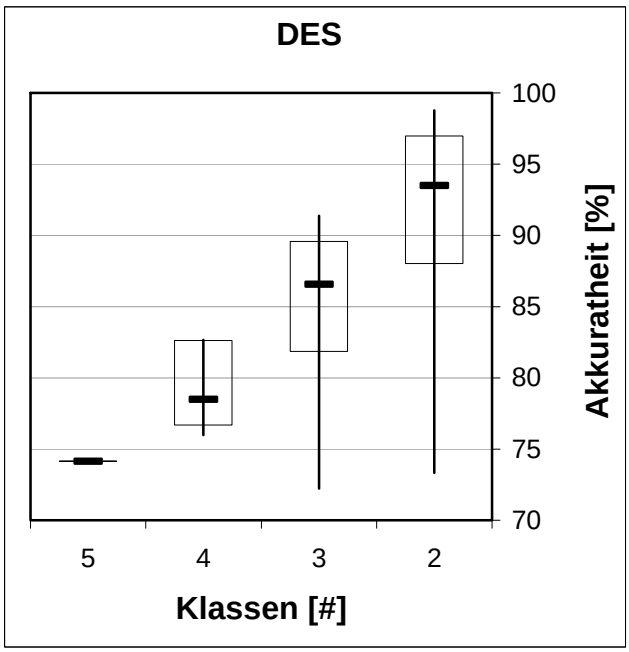


Abb. 3.59: Boxplot Quartile und Extrema des Einflusses der Reduktion des Emotionssets auf die Erkennungsleistung, Datenbank EMO-DB, Top 75 Merkmale, Klassifikation mit SVM, 10-fach SCV



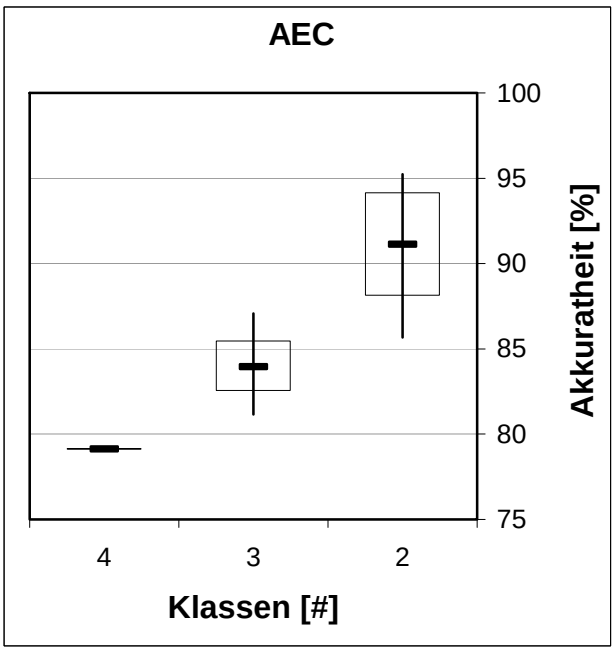


Abb. 3.60: Boxplot Quartile und Extrema des Einflusses der Reduktion des Emotionssets auf die Erkennungsleistung, Datenbank DES, Top 100 Merkmale (links), Datenbank AEC, Top 107 Merkmale (rechts), Klassifikation mit SVM, 10-fach SCV

In konkreten Szenarien wie der Erkennung eines verärgerten Call-Center-Anrufers können so beispielsweise nur zwei oder drei Emotionen, dafür aber robuster erkannt werden. Die drei Boxplots zeigen hierzu den Einfluss der Reduktion des Emotionssets auf die Erkennungsgüte für die Datenbanken EMO-DB[133], DES[134] und AEC[135]. In der Darstellung sind Extrema und Quartile enthalten. Die durchschnittliche Akkuratheit steigt wie erhofft streng monoton mit sinkender Klassenzahl. Die Streuung in Abhängigkeit von den im jeweiligen Set enthaltenen Emotionen ist dabei jedoch sehr hoch, so dass im Extremfall ein günstiges Set mit mehr Emotionen trotzdem besser erkannt werden kann, als eines mit weniger Emotionen. Besonders extrem ist dieses Verhalten für die Datenbank DES, bei der für bestimmte Kombinationen mit zwei oder drei affektiven Zuständen die Erkennungsrate geringer ausfällt als mit allen fünf Emotionen.

Im Gegenzug zur Streichung von Emotionen können diese auch zusammengefasst werden. Hier soll als interessantes Beispiel die Unterscheidung nur nach aktiv/passiv oder negativ/(neutral)/positiv anhand der Datenbanken EMO-DB und DES gezeigt werden. Die Klusterung der Emotionen erfolgt dabei nach der Aktivität a und Valenz v, wie in Kap. 2.3.1 eingeführt und in Tab. 3.61 dargestellt:

Emotion	EMO-DB	DES
Aktiv	Ärger, Freude, Furcht, Ekel	Ärger, Freude, Überraschung
Passiv	Langweile, Neutralität, Trauer	Neutralität, Trauer
Positiv	Freude	Freude
Negativ	Ärger, Ekel, Furcht, Trauer	Ärger, Trauer
Neutral	Langweile, Neutralität	Neutralität, Überraschung

Tab. 3.61: Zuordnung der Emotionen nach dimensionaler Aufteilung, Datenbanken EMO-DB und DES

In Tab. 3.62 werden die erzielten Erkennungsleistungen bei Einteilung von Emotionen nach dimensionaler Aufteilung gezeigt. Als Klassifikator wurde SVM in einer 10-fach SCV gewählt. Neben der Akkuratheit wird auch jeweils die je nach Problemstellung deutlich unterschiedliche optimale Zahl von Merkmalen gezeigt, wobei das ursprüngliche Ranking entsprechend der normalen Emotionsaufteilung und SVM-SFFS beibehalten wurde. Wie zu erkennen ist, lässt sich eine automatische Erkennung der Aktivität[136] deutlich leichter realisieren als eine der Valenz. Fasst man neutrale Äußerungen zu negativen hinzu, lässt sich die Leistung erwartungsgemäß steigern.

---

[133] Bei der Berechnung ergeben sich bei sieben Emotionen insgesamt 120 Permutationen für die vollständige Berechnung des beschriebenen Problems. Um den Aufwand für die Berechnung zu reduzieren, wurde eine Approximation mittels eines gleitenden Klassenfensters mit insgesamt 42 Durchläufen gewählt.

[134] Es wurden alle 26 Permutationen, die zur vollständigen Berechnung des Klassenreduktionsverhaltens bei fünf Klassen erforderlich sind, durchgeführt.

[135] Die für vier Klassen insgesamt elf Permutation zur vollständigen Berechnung des Reduktionsverhaltens wurden durchgeführt.

[136] Zum Vergleich mit Ergebnissen aus [HAM05] wurde auch eine 5-fach SCV für die Datenbank DES durchgeführt. Dabei ergibt sich eine Akkuratheit von 95,9%, die das beste dort genannte Ergebnis zur Unterscheidung aktiv/passiv von 88,1% deutlich übersteigt.

	EMO-DB		DES	
	Akk. [%]	**Top N [#]**	**Akk. [%]**	**Top N [#]**
Aktiv / Passiv	97,5	57	96,1	78
Positiv / Neutral / Negativ	90,6	62	63,0	100
Positiv / Negativ	92,4	59	82,6	96

Tab. 3.62: Erkennungsleistung der Emotionen nach dimensionaler Aufteilung und optimale Merkmalszahl N, Datenbanken EMO-DB und DES, Klassifikator SVM, 10-fach SCV

Abschließend soll ein Vergleich zur menschlichen Erkennungsleistung und zu anderen Arbeiten helfen, die in dieser Arbeit erzielten Leistungen einstufen zu können: Tab. 3.63 zeigt die maximale hier erzielte Leistung, aktuell berichtete maximale anderer Arbeiten und die mittlere menschliche Klassifikationsleistung (siehe Kap. 3.12.2). Als absolute Mindestanforderung sind zum Vergleich weiterhin ein rein zufälliges Raten, entsprechend der reziproken Klassenzahl, und die stete Wahl der Klasse mit maximaler a-priori Auftrittswahrscheinlichkeit angegeben. Letzterer Wert berücksichtigt, dass die Klassen nicht immer ganz gleich verteilt sind.

Akkuratheit [%]	**EMO-DB**	**DES**	**AEC**
Raten	14,3	20,0	25,0
Häufigste Klasse	26,0	31,1	37,2
Mensch	84,3	67,3	84,8
Andere Arbeiten	77,4 [VOG05]	50,6 (61,1) [VER04A]	70,0* [BAT05]
Maximum hier	**88,8**	**76,2 (83,0)**	**70,6*/ 79,1**

Tab. 3.63: Vergleich Erkennungsraten zufällige Entscheidung, stete Wahl der häufigsten Klasse und Mensch sowie andere Arbeiten und hier erzielte maximale Leistung unter gleichen Testbedingungen mit 10-fach SCV. In runden Klammern stehen Erkennungsraten für sprecherabhängige Erkennung. Die mit * gekennzeichneten Erkennungsraten stehen für eine sprecherunabhängige 2-fach Kreuzvalidierung.

Für gespielte Daten zeigt sich der automatische Ansatz dabei äußerst leistungsstark: Die Akkuratheit übersteigt sogar augenscheinlich die des Menschen. Hierzu ist allerdings anzumerken, dass es sich um eine Kreuzvalidierung handelt, was bedeutet, dass das Lernverfahren Gelegenheit hat Eigenschaften jedes Sprechers kennen zu lernen. Außerdem handelt es sich bei EMO-DB und DES jeweils um Studioaufnahmen unter Idealbedingungen. Schließlich ist auch die Angabe zur menschlichen Erkennungsleistung ein Mittelwert, wobei einzelne Testpersonen deutlich über diesem lagen, und man ferner davon ausgehen kann, dass es Menschen mit grundsätzlichen Zuordnungsschwierigkeiten gibt. Die spontane, im Feld aufgezeichnete Datenbank AEC zeigt außerdem, dass unter realistischen Bedingungen die menschliche Performanz obsiegt. Trotzdem ist auch hier der absolute Abstand von 5,6% nicht groß. Bei dieser Datenbank sei erwähnt, dass die menschliche Übereinstimmung auf der Grundlage akustischer und linguistischer Information beruht, während zur automatischen Erkennung hier nur die akustische genutzt wurde. Ein entsprechend höheres Ergebnis unter Verwendung beider Analysen findet sich in Kap. 6.1.

Tab. 3.64 und Tab. 3.65 zeigen noch differenziert den Vergleich Mensch und Maschine nach Emotionen. Die Raten einzelner Erkennung resultieren dabei stets aus der gleichzeitigen Erkennung des gesamten Sets. Es ergeben sich dabei deutliche Schwankungen, welche beweisen, dass einzelne Emotionen schwerer erkannt werden.

Akk. [%]	Ärger	Ekel	Furcht	Freude	Neutral.	Langweile	Trauer
Mensch	73,3	63,3	96,7	93,3	99,9	90,0	73,3
Hier	92,9	92,1	89,1	69,0	84,6	89,9	92,5

Tab. 3.64: Vergleich Erkennungsleistung Mensch-Maschine nach Emotionen, Datenbank EMO-DB, Klassifikation mit SVM, 10-fach SCV

Akk. [%]	Ärger	Freude	Neutralität	Trauer	Überraschung
Mensch	75,1	56,4	60,8	85,2	59,1
Hier	81,0	60,5	81,0	80,2	68,4

Tab. 3.65: Vergleich Erkennungsleistung Mensch-Maschine nach Emotionen, Datenbank DES, Klassifikation mit SVM, 10-fach SCV

Tab. 3.66 zeigt schließlich die Erkennungsleistung nach Emotion aufgeschlüsselt für die Datenbank AEC. Zu dieser existiert leider kein Perzeptionstest mit ausreichender Aussagekraft.

	Ärger	Bemutterung	Betonung	Neutralität
Akkuratheit [%]	82,4	70,3	82,6	66,9

Tab. 3.66: Erkennungsrate nach Emotion, Datenbank AEC, Klassifikation mit SVM, 10-fach SCV

Eine generelle Tendenz, welche Emotion leicht zu erkennen ist, lässt sich nicht angeben. Dies beruht nicht zuletzt auf der Diversität der Sets. Einzig Ärger scheint generell am besten erkennbar zu sein, was für viele Applikationen bereits ausreichend ist. Auf die Darstellung von Konfusionsmatrizen für die automatische Erkennung wird hier verzichtet, da sie im Wesentlichen von statistischem Rauschen geprägt sind.

4

Linguistische Emotionserkennung

„Verschiedene Menschen können mit Hilfe der Sprache ihre Erlebnisse bis zu einem gewissen Grade miteinander vergleichen.“

ALBERT EINSTEIN (1879-1955)

In diesem Kapitel werden Methoden vorgestellt, den gesprochenen oder geschriebenen *Inhalt* einer Äußerung hinsichtlich ausgedrückter Emotion zu analysieren. In [ARU01] wird gezeigt, dass der linguistische Anteil dies grundsätzlich erlaubt. Dies beruht im Allgemeinen auf der erlernten Verwendung spezieller Ausdrücke und grammatikalischer Veränderungen in Abhängigkeit vom aktuellen affektiven Zustand. Ein Beispiel hierfür sind Kraftausdrücke im Fall von Ärger oder unsinnige Wiederholungen im Fall von Irritation. Die Aufgabe der automatischen Emotionserkennung ist es somit, auf dieser Ebene einen Ansatz zu finden, der in einem Satz ausgedrückte Gefühle anhand semantischer Information zuordnet. Um dies zu ermöglichen, sollen semantische Einheiten wie Wörter in eine numerische Repräsentation umgesetzt werden. Bisher betrachtete Ansätze stützen sich dabei auf die Wortwahl einzelner oder zusammenhängender Einheiten. Die einfachste Methode ist dabei die reine Suche nach affektiven Schlüsselwörtern. Als komplexere Variante wird gerne die Modellierung der Emotion aus Wortfolgen mittels klassenbasierter Sprachmodelle verwendet. Sie wird daher im Verlauf des Kapitels als Vergleichsbasis angewandt. Darüber hinaus wird gezeigt, inwiefern sich Verfahren aus dem Bereich der Textsuche für die Erkennung von Emotion eignen. Als dritte Möglichkeit wird eine grafische Modellierung zur Erfassung des semantischen Zusammenhangs innerhalb einer gesamten Äußerung vorgestellt.

Um sowohl gesprochenen, als auch geschriebenen Äußerungen in der Vorverarbeitung gerecht werden zu können, werden auch innerhalb der Arbeit entwickelte Verfahren zur Sprach- und Handschrifterkennung sowie zur fehlertoleranten Behandlung digitalen Textes vorgestellt.

4.1 Vokabular

Die Ausgangsbasis für die Erfassung akustisch oder manuell eingegebenen Textes zur späteren Analyse ist ein sogenanntes *Vokabular*, welches alle zu betrachtenden lexikalischen Einheiten beinhaltet. Die kleinste dieser Einheiten sind die sogenannten *Semune,* eine Kurzform für *semantische Untereinheiten*. Semune sind dabei in einer strengen Definition nicht *Wörtern* gleichzusetzen, da sich emotionaler Ausdruck bereits in Silben oder Wortfragmenten, aber auch erst in zusammengesetzten Wörtern widerspiegeln kann. Im Folgenden wird jedoch - dem allgemeinen, vereinfachenden Gebrauch folgend - von Wörtern w gesprochen.

4.1.1 Generierung

Als globale Einheit der Betrachtung wird hier, wie in Kap. 3, eine Phrase $\mathcal{S}$ [137] mit $\mathcal{S} = \{w_1,...,w_S\}$ und zugehöriger Phrasenlänge $S = |\mathcal{S}|$ gewählt. Ausgehend von einem Korpus $\mathcal{L}$ solcher textueller Phrasen mit zugehöriger Klassenangabe wird die Gesamtheit unterschiedlicher Wörter gesucht. Diese ergibt das sogenannte *Vokabular* $\mathcal{V}$ der Wörter w_i mit $\mathcal{V} = \{w_1,...,w_V\}$. Die Vokabulargröße entspricht dabei $V = |\mathcal{V}|$. Dieses Vokabular bewegt sich in der Größenordnung mehrerer Tausend Einträge[138], wobei nur ein Bruchteil für die tatsächliche Emotionserkennung relevant ist. Zur Reduktion der Komplexität hinsichtlich der Erkennungsleistung, des Speicherbedarfs und der Rechenzeit werden daher hier Verfahren zur Beschränkung des Vokabulars auf relevante Termini und Zusammenfassung solcher in Gruppen eingesetzt.

Um einen generellen Eindruck zu erhalten, welcher Anteil an Wörtern für sich emotional erscheint, wurden von zwei Testpersonen (vgl. Kap. 4.7.1) die 10.000 häufigsten Wörter der englischen Sprache [QUA98] bezüglich ihres emotionalen Charakters in den sieben Emotionen des MPEG-4 Sets mit ergänzender Neutralität annotiert. Hierbei ergaben sich 815 bedeutungstragende Wörter. In [SIE95] wird außerdem ein Vokabular mit 277 emotionalen Termini, *Balanced Affective Wordlist* genannt, in den Emotionen *ängstlich, negativ, neutral* und *positiv* vorgestellt.

Im weiteren Verlauf wird das Vokabular jeweils in einer Kreuzvalidierung aus der betrachteten Datenbank erstellt.

4.1.2 Stopping

Die direkte Reduktion von Vokabeln wird in der linguistischen Analyse als *Stopping* bezeichnet. Beim traditionellen, expertenbasierten Stopping werden Funktionswörter ohne Bedeutung für die potentiellen Zielklassen, hier Emotionen, mittels einer *Stoppliste* händisch aus dem Vokabular gestrichen. Unter der Vorraussetzung eines hinreichend großen Trainingsschatzes mit annähernd gleicher Distribution unter den Klassen ist zu erwarten, dass sich für diese Termini ohnehin eine

---

[137] Für englisch *Sentence*.

[138] In der automatischen Spracherkennung hat es sich eingebürgert, die Vokabulargröße in Tausenden, mit *k* notiert, anzugeben. Man spricht dann von *geschlossenen* Vokabularen. Speziell die deutsche Sprache besitzt jedoch ein *offenes* Vokabular, da Wortverbindungen prinzipiell zu unendlich vielen neuen Wörtern führen können.

Gleichverteilung ergibt, was ein geeigneter Klassifikator lernen würde. In der Praxis wird dieser Fall jedoch selten auftreten, da Korpora auf Grund teuren, schwer erhältlichen Trainingsmaterials das adäquate Ausmaß kaum aufweisen. Ferner sollte redundante Information auch dann gestrichen werden, um die Komplexität für den Lernalgorithmus zu reduzieren.

Alternativ kann ein automatisches Stopping erfolgen. Die in Kap. 3.9.2 vorgestellten gleitenden Suchverfahren, oder ähnliche wrapperbasierte Verfahren, erweisen sich auf Grund der betrachteten Ursprungsgröße des Merkmalssets hierzu als nur bedingt praktikabel. Stattdessen werden oft einfache Bedingungen, wie eine Mindestanforderung an die Auftrittshäufigkeit innerhalb des Trainingsmaterials, als erster Ansatzpunkt gewählt. Vor allem eine filterbasierte Selektion auf Basis des Information-Gain (siehe Kap. 3.7.4) oder der *Salience* (siehe Kap. 4.4.2) gelten hier als effektiv [JOA97B], [YAG97][139]. Auch PCA (siehe Kap. 3.9.1) wird für diesen Zweck angewandt. Als Beispiel zum automatisierten Stopping zeigt Tab. 4.1 die Salience von ausgewählten Vokabulareinträgen aus dem in dieser Arbeit betrachteten Korpus EAL-F+W (siehe Kap. 4.7.1). Ein hoher Wert spricht dabei für große Relevanz.

Wort w_i	**Salience**
great	2,887
wonderful	2,686
horrible	1,897
air	1,649
car	1,337
raincoat	1,283
hell	0,907
depressing	0,872
of	0,104
that	0,081
on	0,074

Tab. 4.1: Salience ausgewählter Wörter der Filmdatenbank EAL-F+W

Wie bereits in diesem Auszug zu sehen ist, sind automatisierte Verfahren wegen der zu geringen Korpusgröße nur bedingt geeignet. Es ergibt sich zwar ein hohes Maß an Salience für die emotional gefärbten Ausdrücke *great, wonderful* oder *horrible*, und ein geringes für die Funktionswörter *of, that,* und *on.* Im mittleren Bereich jedoch ist eine Schranke sehr schwer applizierbar, ohne auch Wörter hohen emotionalen Gehalts zu verlieren. Im Beispiel gestaltet sich etwa die Durchführung einer Filterung der Termini *air, car* und *raincoat* als diffizil ohne auch *hell* oder *depressing* zu eliminieren.

Generell sollte von einem zu radikalen Ausschluss von Wörtern abgesehen werden, da es oft nicht direkt ersichtlich ist, ob ein Wort auch emotional konnotiert sein kann. So können gerade auch

[139] Weitere Verfahren sind etwa *DF-Thresholding,* χ^2*-test*, und *Term-Strength-Criterion,* auf die hier nicht näher eingegangen wird.

Possessivpronomina oder bestimmte Zahlen wie in den Äußerungen *„Meins!“* oder *„Ausgerechnet 13!“* durchaus vor dem kulturellen oder kontextuellen Hintergrund emotional behaftet sein. Im Korpus EAL-F+W wurden beispielsweise durch Expertenannotation gezielt 187 von 2.334 Vokabeln wie Artikel, Hilfsverben, Pronomen, Präpositionen, Konjunktionen, Personalpronomina, Eigennamen, Ortsangaben und Zahlen gestoppt. Dies entspricht einer absoluten Reduktion von 8%. Abb. 4.2 zeigt hierzu einen Auszug aus der Stoppliste. Es zeigt sich dabei auch, dass ein Korpus in der Realität nur ein sehr eingeschränktes Abbild erlaubt.

Alan, Alvy, Anna, Annie Ben, Bernie, Bianca, Billie, Bobby, Carl, David Elton, Harrison, Janice, Robert, Terry,…

four, 5, seven, 8, fifteen, 17, 30, 32, 2000, …

about, across, am, and, are, at, be, been, before, being, both, by, did, does, for, has, have, having, here, hers, he, him, his, hundred, it, its, of, off, she, the, them, there, these, they, this, those, to, went, were, …

Abb. 4.2: Auszug aus der Liste aufgetretener Stoppwörter in der Datenbank EAL-F+W

4.1.3 Stemming

Um bei der Reduktion von Vokabeln im Besonderen dem sprachlichen Hintergrund Rechnung zu tragen, bietet sich als natürliche und effiziente Form die Zusammenfassung morphologischer Varianten von Wörtern des gleichen Stamms an. Der Begriff *Wort* bezeichnet dabei die unterschiedlichen Wortformen innerhalb eines *Lexems* l_i. Lexeme sind Klassen lexikalisch äquivalenter Wortformen, welche dieses in unterschiedlicher Umgebung repräsentieren[140]. Der Begriff *Lexikon* beschreibt die Menge von Lexemen einer Sprache. Reduziert man also die im Vokabular erfassten Wörter auf ihre Lexeme, ergibt sich dadurch nicht nur, wie bei der üblichen Reduktion von Merkmalen, eine Einsparung an erforderlichem Speicherplatz, Rechenzeit und Trainingsmaterial, sondern es erhöht sich gleichzeitig die Leistung für die Klassifikation einer neuen Äußerung. Dies ergibt sich aus dem Umstand, dass die Häufigkeit $TF(w_i)$[141] für ein Lexem innerhalb des Lernmaterials deutlich höher sein wird als für einzelne Flexionsformen. Auf diese Weise kann ein aussagekräftigeres Modell für das Lexem gebildet, und ein bisher nicht gesehenes Wort, man spricht von *Out-Of-Vocabulary (OOV)* Wörtern, mit einem geeigneten Verfahren auf ein im Vokabular enthaltenes Lexem zurückgeführt werden. Eine Vielzahl von automatisierten Verfahren, als *Stemmer* bezeichnet, existiert hierzu: Vom einfachen N-Gramm Stemmer, der die ersten *n* Zeichen als Wortstamm annimmt, bis zu präziseren Varianten wie den Stemmern von Lovins [LOV68], Dawson [DAW74], Porter [POR80], Paice und Husk [PAI90] und Krovetz [KRO93],

[140] In der automatischen Spracherkennung existieren die analogen Bezeichnungen *Token* und *Typen*. Bei diesen Begriffen repräsentiert jedoch der Typ als Kluster von Token Aussprachevarianten anstelle morphologischer Varianten.

[141] Diese allgemein verwendete Notation bezieht sich auf *Term-Frequency*, englisch für Worthäufigkeit.

welche aus einer Suffixliste sowie Regeln zur Behandlung dieser bestehen. Natürliche Sprache beinhaltet jedoch immer Ausnahmen von starren Regelsets. Dies führt zwangsläufig zu Fehlern durch sogenanntes *Understemming*, falls zusammengehörige Varianten fälschlich nicht geklustert werden, und beispielsweise durch Anhängen eines Suffixes sich der Stamm ändert wie in *deceive* und *deception*[142]. Das entsprechende Pendant hierzu ist das *Overstemming*, bei dem Endungen irrtümlich als Suffix gehandelt werden, wie in *sad* und *sadist*[143].

Bei akzeptabler Korpusgröße ist daher Stemming durch einen Experten zu bevorzugen. Alternativ zu einer Behandlung während der Erkennung können Flexionsformen systematisch zugefügt werden, um die Lexeme und somit das Vokabular zu erweitern. Dabei muss jeweils auf eine äquivalente emotionale Konnotation geachtet werden. Im Vergleich zu Stemming mittels N-Grammen (vgl. Kap. 4.4) auf Buchstabenebene, wird bei dem später betrachteten Korpus EA-F+W das optimale Ergebnis hierdurch erzielt. Durch Berücksichtigung morphologischer Varianten wie *Deklination*, *Konjugation*, und *Komparation* können so 644 Wörter in 260 Lexemen zusammengefasst, und die Dimensionalität des Merkmalsvektors um 16%, also etwa um das Doppelte im Vergleich zu Stopping, reduziert werden. Tab. 4.3 zeigt hierzu als Beispiel einen Auszug aus der Stemmliste des Korpus EAL-F+W.

Lexem l_i	**Variante** $w_{i,1}$	**Variante** $w_{i,j}$	**Variante** $w_{i,J}$
claim	*claims*	*claiming*	*claimed*
damage	*damages*	*damaging*	*damaged*
feel	*feels*	*feeling*	*feelings*
hope	*hopes*	*hoping*	*hopeful*
kill	*killer*	*killing*	*killed*
weak	*weakness*	*weaknesses*	*weakened*
worry	*worries*	*worrying*	*worried*

Tab. 4.3: Auszug aus der Liste durch Stemming geklusterter morphologischer Varianten, Datenbank EAL-F+W

In der Praxis werden die beobachteten Wörter, die in einer durch Stemming erhaltenen Liste als morphologische Variante aufgeführt sind, durch den Klusternamen substituiert. Im Beispiel des Korpus EAL-F+W beträgt die Dimensionalität des finalen Vektors nach Stopping und Stemming 1.853. Auf Grund dieser gemeinhin noch zu hohen Komplexität, sind weitere Maßnahmen zur Reduktion des Merkmalsraums notwendig [RUE00]. Tab. 4.4 zeigt hierzu die höchsten Ränge bei einer Lexemselektion mittels IGR-FS (vgl. Kap.3.7.4). Einzig ungewöhnlich erscheint hier das Wort *face* auf Rang vier.

[142] Englisch für *betrügen* und *Betrug*.

[143] Englisch für *traurig* und *Sadist*.

Rang	Wort	IGR	Rang	Wort	IGR
1	*disgusting*	0,5583	13	*Jesus*	0,2221
2	*yuck*	0,4707	14	*beautiful*	0,2192
3	*dirty*	0,4506	15	*bitch*	0,2191
4	*face*	0,2824	16	*thank*	0,2073
5	*lucky*	0,2688	17	*happy*	0,2025
6	*perfect*	0,2609	18	*glad*	0,1886
7	*delighted*	0,2609	19	*sad*	0,1869
8	*afraid*	0,2578	20	*sorry*	0,1685
9	*great*	0,2405	21	*damn*	0,1627
10	*wonderful*	0,2364	22	*shit*	0,1605
11	*Christ*	0,2356	23	*love*	0,1548
12	*cool*	0,2281	24	*crap*	0,1426

Tab. 4.4: Ranking der emotionalsten Wörter mittels IGR-FS im Korpus EAL-F+W

4.2 Verarbeitung geschriebener Eingabe

4.2.1 Texterfassung

Zur Erfassung von textueller Eingabe bietet sich zunächst die konventionelle Eingabe über die Tastatur an.

Bei vielen potenziellen Anwendungen wie bei Palmtop-Geräten oder aktuellen öffentlichen Auskunftssystemen steht darüber hinaus ein Touchpad oder -screen zur Verfügung. Hierzu wurde hier eine Strichzugerkennung für Einzelbuchstaben mit Hilfe kontinuierlicher Links-Rechts-HMM mit je 16 Zuständen je Strichzug realisiert. Merkmale sind SMA-gefilterte planare Ortskoordinaten sowie Geschwindigkeits- und Beschleunigungsregressionskoeffizienten, die im Hintergrund zu einer überlagerten Applikation erfasst werden. Dabei erfolgt die Extraktion schritthaltend, was einen zeitlichen Bezug in die Betrachtung einbezieht. Verwendet werden dabei nur jeweils unterschiedliche Koordinatenpunkte. Die Koordinaten der Bewegung werden auf ein umschreibendes Rechteck sowie den Startpunkt der Bewegung normiert, um Invarianz gegenüber Größe und Ort sicherzustellen. Eine regelbasierte Nachbearbeitung auf Grundlage des Breiten-Höhenverhältnisses optimiert weiterhin die Erkennung der Strichzüge „|“ und „—“, und eine semantisch höhere Schicht auf Basis von BN erlaubt die Interpretation von zusammengesetzten Strichzügen (vgl. Kap. 4.6). So können Buchstaben in einem Zug, oder auch aus mehreren Einzelstrichen ausgeführt werden (vgl. [SCH02A]). Der Erkenner leistet, bei einem Inventar von 26 Buchstaben und zehn Ziffern, 98,4% Erkennungsleistung für Einzelzeichen bei Verwendung von 3.600 gleichverteilten Lern- und 1.980 disjunkten Testbeispielen.

4.2.2 Soft-String-Matching

Grundsätzlich muss digitaler Text zunächst einer adäquaten Vorverarbeitung unterzogen werden.

Hierzu gehört die Wandlung in ASCII-Code[144], das Trennen in einzelne Phrasen entsprechend der Interpunktion, die Umsetzung in ausschließlich kleine Lettern sowie das Entfernen von Sonderzeichen. Um sicherzustellen, dass falsche oder alternative Schreibweisen, wie etwa nach alter oder neuer deutscher Rechtschreibung, oder entsprechend britischem oder amerikanischem Englisch, ebenfalls behandelt werden können, wird in einer nächsten Stufe ein *Soft-String-Matching* Verfahren [NAV01] eingesetzt. Dieses erlaubt die Überführung zweier sich stark ähnelnder Zeichenfolgen[145], und auch die Durchführung eines automatisierten Stemmings (vgl. Kap. 4.1.3). Es existiert eine Reihe geeigneter Ansätze hierzu [COH03], wobei hier als Vergleichsbasis konventionelle *Levenshtein-Distanz (LD)* [LEV66] sowie ein hier vorgeschlagener Lösungsansatz (vgl. Kap. 4.6) vorgestellt werden.

Bei Matching mit LD wird mittels dynamischer Programmierung (vgl. Kap. 3.11.1) der minimale Abstand zwischen zwei Strings w_i und w_j durch die hier gleich gewichteten Editieroperationen *Ersetzung*, *Auslassung* und *Einfügung* von Einzelzeichen bestimmt. Es wird dann der String w_i mit insgesamt geringstem Abstand innerhalb eines Vokabulars $\boldsymbol{\mathcal{V}}$ dem Suchstring w_j zugeordnet.

Weiterhin wird ein Ansatz mit Bayesschen Netzen vorgeschlagen, welche um die Fähigkeit der Behandlung zeitlicher Abfolgen und weicher Evidenzen erweitert werden. Da es sich hierbei nicht mehr um BN im klassischen Sinne handelt, wird hier auch der allgemeine Begriff der *Grafischen Modellierung* [LAU96] verwendet. Als Ausgangsbasis sei zunächst ein BN der Topologie eines NB Klassifikators, wie in Abb. 3.27, Kap. 3.7.6 dargestellt, gewählt. Um Symbol- oder später Wortabfolgen zu berücksichtigen, wird Evidenz nur dann eingebracht, wenn die Reihenfolge von links nach rechts eingehalten wird. Zur verschiedenen Gewichtung unterschiedlicher Abstände zwischen Symbolen oder Wörtern bezüglich Test- und Referenzmustern, werden anstelle der bei BN üblichen harten Evidenzen, das heißt ein Ereignis tritt ein oder nicht, weiche Evidenzen, im Folgenden als *Konfidenzmaße* c mit $c \in [0;1]$ bezeichnet, eingeführt. Dabei werden die Zufallsvariablen an Knoten, an denen Wissen eingebracht wird, folgendermaßen bestimmt:

$$P^*(X) = P(X) + c \cdot (1 - P(X)) \tag{4.1}$$

Es wird vermieden, für jeden Vokabulareintrag $w_i \in \boldsymbol{\mathcal{V}}$ ein eigenes Modell zu generieren, um den Speicherbedarf zu reduzieren und die Rechenzeit zu verkürzen. Statt dessen wird für den Suchstring w_j ein generisches Modell zur Laufzeit erstellt und jeder Eintrag des Vokabulars in dieses eingebracht. Der String w_i, der dabei die höchste Wurzelwahrscheinlichkeit $P(w_j)$ erreicht, gilt dann als Matchstring.

---

[144] *American Standard Code for Information Interchange* – Erster 8 Bit Standardcode, in dem Lettern, Symbole, Interpunktion, Formatierungs- und Steuersymbole enthalten sind. International im Standard *ISO Latin 1* festgehalten, und im heute weit verbreiteten 16 Bit *Unicode* enthalten, der Alphabete diverser Nationen beinhaltet.

[145] Englisch: *String*. Soft-String-Matching bezieht sich ins Besondere auf einen intelligenten Vergleich von Zeichenfolgen in einem Vokabular mit dem aktuellen Vergleichsstring. Im Gegensatz hierzu können Listen von Stringgruppen als einfachere Lösung verwendet werden. Diese erfassen jedoch keine zum Zeitpunkt der Listenerstellung noch nicht gesehenen Stringvarianten.

Ergänzend werden Zeichensequenzen $z_{n,m}$ aus ein oder mehreren Einzelzeichen, die sich phonetisch oder orthografisch ähnlich sind, expertenbasiert zu Klustern $s_n = \{z_{n,1},...,z_{n,M}\}$ zusammengefasst (vgl. Kap. 4.1.3)[146]. Tab. 4.3 zeigt exemplarisch geklusterte Sequenzen.

Sequenz s_n	**Variante** $z_{n,1}$	**Variante** $z_{n,m}$	**Variante** $z_{n,M}$
ei	*ai*	*ey*	*ay*
c	*cc*	*kk*	*ck*
d	*dd*	*dt*	*tt*

Tab. 4.5: Auszug aus der Liste geklusterter Zeichensequenzen

Der Suchstring w_j wird nun in möglichst große zusammenhängende Zeichensequenzen s_n mit $n = 1,...,N$ eingeteilt, so dass $w_j = \{s_1,...,s_N\}$ gilt. Zu jedem $s_n \in w_j$ wird dann ein Knoten unter Berücksichtigung der Reihenfolge im generisch erzeugten Modell eingefügt. Jeder Knoten, der eine Sequenz s_n repräsentiert, hat, um ähnliche Schreibweisen zu modellieren, als seine Kindknoten die alternativen Zeichen $z_{n,m}$ mit $m = 1,...,M$ innerhalb des jeweiligen Klusters. Bei diesem Vorgehen muss sichergestellt werden, dass beim Vergleich nur jeweils eine Variante Evidenz erlangen kann. Die a-posteriori Wahrscheinlichkeit $P(z_{n,m}|s_n)$ einer Zeichensequenz bei auftretender Variante kann so je nach Grad der Ähnlichkeit unterschiedlich gewichtet und gelernt werden. Abb. 4.6 zeigt das gesamte Prinzip anhand eines allgemeinen Stringnetzes.

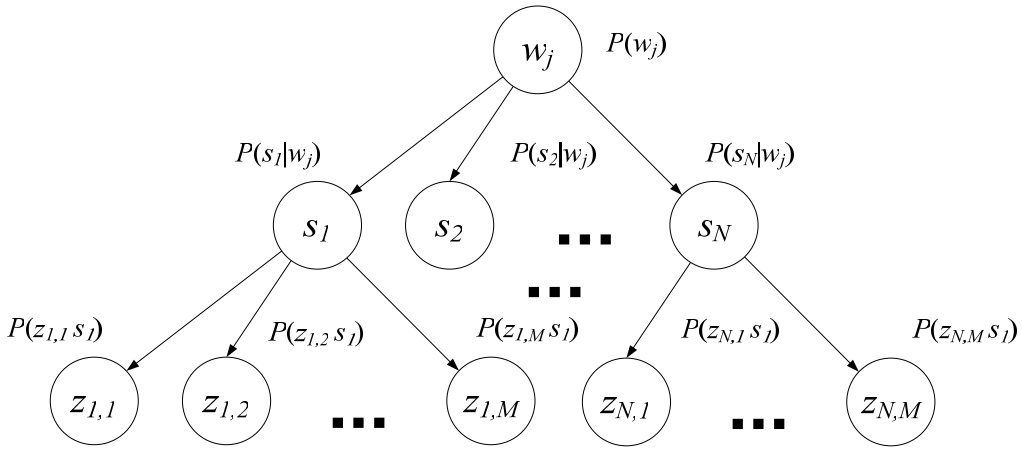


Abb. 4.6: Dynamisch generiertes BN zum Soft-String-Matching

Um eine Anpassung an die Länge eines Suchstrings zu vollziehen, wird eine Gleichverteilung unter den Zeichensequenzen s_n eines Suchstrings angenommen und folgende empirisch ermittelte vereinfachende Anpassung, in Abhängigkeit von der Stringlänge $|w_j|$, gewählt [SCH04B]:

[146] Das Erstellen von Zeichensequenzklustern kann bei ausreichend vorhandenem Lernmaterial auch automatisch erfolgen.

$$P(s_n \mid w_j) = \begin{cases} -0,008 \cdot |w_j| + 0,75 & \textit{für} \quad 1 \le |w_j| \le 15 \\ 0,63 & \textit{für} \quad |w_j| > 15 \end{cases} \tag{4.2}$$

Dieses Vorgehen trägt der in der Praxis bestätigten Annahme Rechnung, dass einzelne Zeichensequenzen in einem längeren String geringere Relevanz besitzen, als in einem kürzeren. Unterschiedliche Stringlängen werden darüber hinaus auch bei der Initialisierung der Modelle berücksichtigt:

$$P(w_j) = \begin{cases} -0,008 \cdot |w_j| + 0,25 & \textit{für} \quad 1 \le |w_j| \le 15 \\ 0,13 & \textit{für} \quad |w_j| > 15 \end{cases} \tag{4.3}$$

Um nun bei der Behandlung der Reihenfolge während des Vergleichs zweier Strings die Evidenzen je nach Zahl ausgelassener Symbole im Sinne weicher Evidenz unterschiedlich zu gewichten, wird die a-priori Wahrscheinlichkeit $P(z_{n,m}|s_n)$ dem Prinzip von Gl. 4.1 folgend angepasst: Mit der ersten Zeichensequenz des Suchstrings w_j beginnend, werden Treffer innerhalb des aktuellen Matchstrings $w_i \in \boldsymbol{\mathcal{V}}$ gesucht. Dabei werden jeweils die Sequenzen $z_{n,m}$ mit hoher Wahrscheinlichkeit $P(z_{n,m}|s_n)$ bevorzugt, und der Abstand $D_n(w_i, w_j)$ zwischen gültigen Treffern für die Zeichensequenzgruppe s_n gespeichert. Die adaptierte Wahrscheinlichkeit ergibt sich somit als:

$$P(z_{n,m}|s_n)^* = P(z_{n,m}|s_n) - 0,1 \cdot D_n(w_i, w_j) \tag{4.4}$$

Der Fall $|w_j| > |w_i|$, das heißt der Suchstring ist länger als der String des Vokabulars, wird dadurch erfasst, dass nicht alle Knoten Evidenz erfahren. Der umgekehrte Fall, $|w_i| > |w_j|$, ist jedoch noch nicht berücksichtigt, wodurch alle Varianten eines Strings, die Suffixe besitzen, nicht wünschenswerter Weise die gleiche Wahrscheinlichkeit erhalten. Um dies zu kompensieren, wird der letzte Knoten s_n welcher Evidenz erhält, entsprechend der Ausgeprägtheit der Längenübertretung abgeschwächt:

$$P(z_{n,m}|s_n)^* = P(z_{n,m}|s_n) - 0,05 \cdot (|w_i| - n) \tag{4.5}$$

Um die Leistungsfähigkeit dieses String-Matching Verfahrens und desjenigen nach LD zu evaluieren, werden insgesamt fünf verschiedene Vokabulare ausgewählt, von denen hier von besonderem Interesse vier sind, die nach Quasthoff et al. [QUA03] die jeweils am häufigsten verwendeten 10.000 Wörter vierer betrachteten Sprachen beinhalten. Zusätzlich wurde ein Vokabular mit 14k Musiktiteln für die Evaluierung verwendet, welches für ein später im Rahmen der Arbeit vorgestelltes Audiosuchsystem von Interesse ist (siehe Kap. 7.2.3). Tab. 4.7 beschreibt die Vokabulare mit der Gesamtzahl der Einträge $V = |\boldsymbol{\mathcal{V}}|$, der Verteilung der Wortlänge L_w und der durchschnittlichen wortübergreifenden Levenshtein-Distanz LD [LEV66].

	V	$\min(L_w)$	$\max(L_w)$	$\mu(L_w)$	$\sigma(L_w)$	$\mu(LD)$	$\sigma(LD)$
Deutsch	10.000	1	27	8,16	3,09	8,25	2,57
Englisch	10.000	1	18	7,11	2,52	7,33	2,02
Französisch	10.000	1	19	7,72	2,67	7,81	2,13
Niederländisch	10.000	1	26	7,60	2,89	7,79	2,36
Musiktitel	14.186	4	39	16,9	2,51	17,9	6,32

Tab. 4.7: Eigenschaften der betrachteten Sprach- und Musiktitelvokabulare

Um die Zuordnungsleistung beurteilen zu können, wurden je 1.000 zufällig gewählte Wörter jedes Vokabulars an ebenfalls zufälligen Stellen unter Einhaltung einer festgesetzten relativen Levenshtein-Distanz durch ausgewählte Editieroperationen verändert. Dies ist als die schwierigste Situation für einen String-Matching-Algorithmus anzusehen, da hierzu kein Fehlermodell erstellt werden kann. Nur wenn ein veränderter String zu seinem zugehörigen Ausgangsstring zugeordnet wird, gilt dies in der folgenden Evaluierung als richtig. Tab. 4.8 zeigt die mittlere Erkennungsleistung bei gleichverteilten Editieroperationen mit BN.

Akkuratheit [%]	**5%**	**10%**	**20%**	**30%**	**40%**	**50%**
Deutsch	98,8	95,6	89,5	74,3	68,9	51,7
Englisch	99,5	98,1	91,7	77,2	67,4	55,0
Französisch	97,1	93,4	89,1	70,8	62,6	50,2
Niederländisch	98,9	95,8	91,8	76,7	67,0	55,6
Liedtitel	99,7	99,2	97,6	95,7	92,6	89,3

Tab. 4.8: Zuordnungsleistung Soft-String-Matching mit BN bei gleichverteilten Editier-Operationen mit unterschiedlichen Vokabularen und Angabe in Rel. LD nach rechts

Abb. 4.9 und Abb. 4.10 zeigen schließlich einen Vergleich zwischen Modellierung mit BN und LD-basiertem String-Matching. Dabei wurde zur Veranschaulichung über alle Sprachen gemittelt.

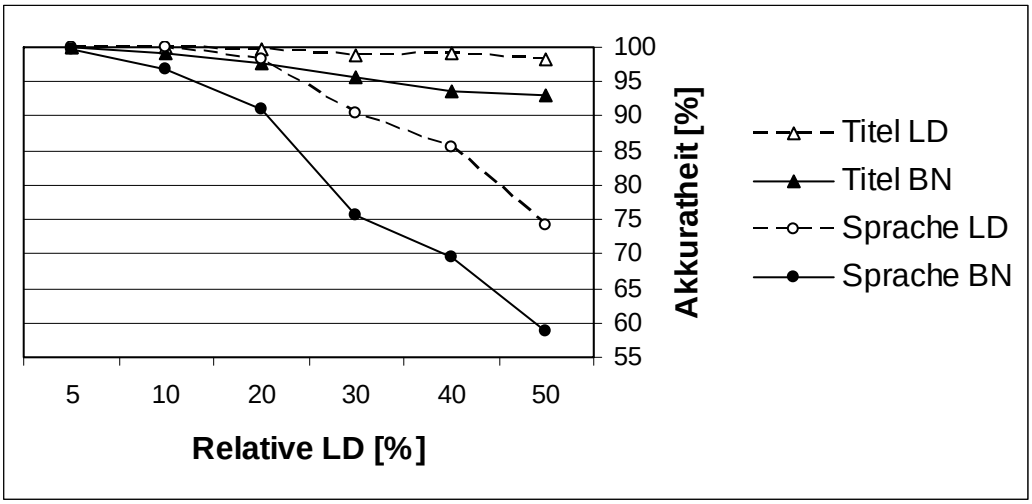


Abb. 4.9: Vergleich Zuordnungsleistung String-Matching mit LD und BN bei gleichverteilten Editieroperationen

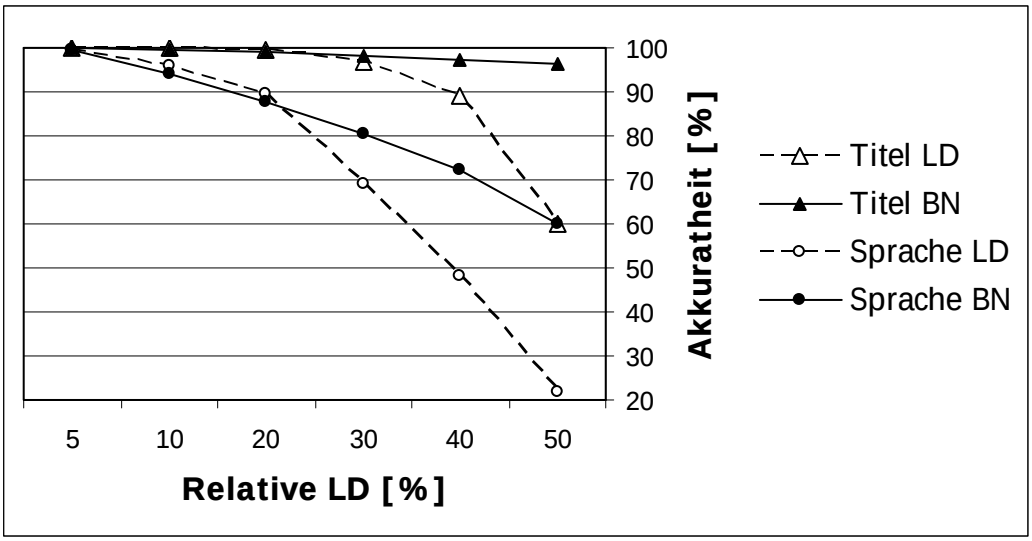


Abb. 4.10: Vergleich Zuordnungsleistung String-Matching mit LD und BN bei exklusiver Auslassung von Zeichen

Es zeigt sich, das der Ansatz der grafischen Modellierung mit BN speziell bei exklusiver Löschung von Zeichen deutlich stärker ist. Dies wird hier genutzt, wenn einzelne Zeichen etwa bei bildbasierter Schrifterfassung sehr unsicher erkannt und daher ausgelassen werden. Generell erlaubt der Ansatz mit BN ein Miteinbeziehen von Konfidenzwerten als Maß über die Sicherheit untergeordneter Erkenner.

4.3 Verarbeitung gesprochener Eingabe

4.3.1 Automatische Spracherkennung

Im Rahmen dieser Arbeit kommen zwei Lösungen für die automatische Spracherkennung, kurz ASR[147], alternativ zum Einsatz. Beide Varianten sind sowohl kontinuierlich online als auch offline für evaluative Zwecke anwendbar und verfügen über kein Sprachmodell. Die einzelnen Wörter werden demnach probabilistisch unabhängig voneinander erkannt, und zur Weiterverarbeitung auf linguistischer Ebene werden n-Best Listen mit Hypothesen und Konfidenzmaßen auf Wortebene weitergegeben.

Die erste eingesetzte ASR-Einheit ist ein kommerzieller sprecherunabhängiger Erkenner. Es handelt sich dabei um einen aktuellen HMM- und phonembasierten Erkenner der Firma Scansoft®[148] mit der Typenbezeichnung ASR 1602 GED ASRAPI V3.2.

Alternativ dazu wird im Folgenden ein phonembasierter Erkenner, beruhend auf einer hybriden ANN/BN Architektur, vorgestellt. Die Phonemerkennung erfolgt dabei mit einem Neuronalen Netz. Im Anschluss erfolgt ein String-Matching, wie in Kap 4.2.2 vorgestellt, auf Phonembasis. Hierzu

[147] Gebräuchliche Abkürzung für *Automatic Speech Recognition*.

[148] Vormals Lernout & Hauspie®.

muss das Vokabular, wie in der Spracherkennung üblich, ebenfalls in Phonemverschriftung vorliegen. Auf ein Sprachmodell wird vorerst verzichtet. Stattdessen werden ebenfalls n-Best Listen mit Hypothesen und Konfidenzmaßen auf Wortebene entsprechend der Wurzelwahrscheinlichkeit des BN an die semantisch höhere Verarbeitung weitergegeben. Soll LD für das Matching eingesetzt werden, so erfolgt eine Konfidenzangabe indirekt proportional zur LD.

Die Erkennung von Phonemen wird mit Hilfe von *Rekurrenten Neuronalen Netzen (RNN)* durchgeführt. Dies sind rückgekoppelte ANN, wodurch erreicht wird, das vergangene Muster und Entscheidungen zur aktuellen Klassifikation beitragen[149], und sich für längere Phonemsequenzen weniger Konfusionen, die für das String-Matching kritisch sind, ergeben. Im Gegenzug kann es jedoch für sehr kurze Phoneme zu Auslassungen kommen. In der hier verwendeten Architektur wird der Ausgang eines einschichtigen ANN teilweise an den Eingang rückgekoppelt[150], wodurch die Schichten zeitlich aneinander gefügt werden. Dies macht eine verborgene Schicht überflüssig, wodurch die Zahl der Parameter auf ca. 60% reduziert werden kann. Es wird hier die Sigmoid-Funktion für rückgekoppelte, und die Softmax-Funktion für Ausgangsknoten eingesetzt (vgl. Kap. 3.7.3). Zum Training wird *Backpropagation-Through-Time* angewandt, eine Variante der Backpropagation, bei der das Netz über der Zeit ausgefaltet wird. Weitere Details hierzu finden sich in [STA03].

Die Vorverarbeitung des Sprachsignals sowie die Extraktion von 12 MFCC und der logarithmierten Signalrahmenenergie, inklusive Geschwindigkeits- und Beschleunigungswerten, als Merkmale erfolgt analog zu Kap. 3.2 sowie Kap. 3.4.2 und Kap. 3.5.4. Es ergibt sich somit ein 39-dimensionaler Merkmalsvektor je Signalrahmen des Eingangssprachsignals.

47 Phoneme der englischen Sprache inklusive zweier Pausenmodelle[151] auf Basis der WSJ-Datenbank [PAU92] (vgl. Kap. 3.12.1), einer Standarddatenbank im Bereich der Spracherkennung, wurden trainiert. Die Akkuratheit beträgt dabei 83,24% auf Phonembasis [STA05] mit dem vorab beschriebenen RNN.

Für einen Leistungstest zum vorgeschlagenen Vorgehen wird ebenfalls die vorgestellte Datenbank WSJ verwendet. Tab. 4.11 zeigt die Eigenschaften des Vokabulars dieser Datenbank mit der Gesamtzahl der Einträge $V = |\mathcal{V}|$, der Verteilung der Vokabelwortlänge L_w und der durchschnittlichen wortübergreifenden Levenshtein-Distanz LD [LEV66].

	V	$\min(L_w)$	$\max(L_w)$	$\mu(L_w)$	$\sigma(L_w)$	$\mu(LD)$
WSJ, 25k	24.307	1	20	7,49	2,39	7,58
WSJ, 5k	4.985	1	18	7,03	2,38	7,19

Tab. 4.11: Eigenschaften der WSJ-Datenbank unterschiedlicher Vokabulargröße

[149] Der Einfluss nimmt dabei exponentiell mit der Zeit ab.

[150] Bei der alternativen *Jordan-Architektur* werden Neuronen der versteckten Schicht an den Eingang rückgekoppelt. Dies bringt einen vereinfachten Trainingsalgorithmus, bietet aber weniger Flexibilität.

[151] Dieses sind: 21 Vokale, 24 Konsonanten, kurze Pause und Stille.

Tab. 4.12 zeigt die mittels Phonemerkennung auf Basis von RNN und nachfolgendem String-Matching mit LD oder BN bei unterschiedlichen Vokabulargrößen (25k und 5k Wörter) erzielten Erkennungsraten auf Wortebene. Hierzu wurden 1.500 zum Phonemtraining disjunkte Testsätze gewählt. Der Ansatz mit grafischer Modellierung ist in jedem Fall zu bevorzugen, ins Besondere bei kleinerem Vokabular. Der Grund dafür ist die beschriebene Tatsache, dass die Phonemerkennung mit RNN nur eine geringe Phonemkonfusion zu Lasten von Auslassungen verursacht. Wie bereits in Kap. 4.2.2 gesehen, ist der BN-Ansatz hier LD überlegen, und die Berechnung erfolgte auf dem gleichen Rechner[152] 2,8 mal schneller mit BN als mit LD[153].

Akk. [%]	RNN / LD	RNN / BN
WSJ, 25k	54,86	**59,91**
WSJ, 5K	64,60	**64,92**

Tab. 4.12: Vergleich Erkennungsleistung Spracherkennung anhand phonembasiertem String-Matching mit LD und BN, Datenbank WSJ

Tab. 4.13 zeigt schließlich zum direkten Vergleich unter exakt gleichen Testbedingungen erzielte Ergebnisse mit traditionellen Methoden der Spracherkennung mit HMM. Zu diesem Zweck wurde zunächst eine hybride Architektur aus Phonemschätzung mit RNN und Warping mit einem HMM gewählt [STA03]. Dies erlaubt einen direkten Vergleich zwischen der Anwendung eines HMM und dem BN-Ansatz, der ersterem überlegen ist. Weiterhin wurde auf denselben Daten ein kontinuierliches HMM mit Gaußmixturen, wie standardmäßig in Spracherkennern verwendet, evaluiert. Auch dies zeigt sich als nicht konkurrenzfähig mit dem propagierten Ansatz.

Akk. [%]	RNN / HMM	Gauß HMM
WSJ, 5K	53,53	56,36

Tab. 4.13: Vergleich Erkennungsleistung Spracherkennung mit hybridem RNN / HMM und HMM mit Gaußmixturen, Datenbank WSJ

Zur Interpretation der Ergebnisse sei daran erinnert, dass an dieser Stelle explizit auf ein Sprachmodell verzichtet wurde. Durch ein solches ergeben sich deutliche Verbesserungen in der Wortfehlerrate[154]. Der Einsatz eines Sprachmodells folgt im weiteren Verlauf des Kapitels, unter Anderem mit BN, was eine nahtlose Integration der vorgestellten Methoden erlaubt.

4.3.2 Segmentierung

Um einzelne Phrasen aus einem kontinuierlichen Signalstrom herauszuschneiden, wird eine Stimmaktivitätserkennung realisiert. Dies ist hier erforderlich, wenn Audiodaten von Filmen oder

[152] Intel® Pentium 4®, 2,8 GHz Prozessortakt, 1 GB Hauptspeicher.

[153] Es handelt sich jeweils um Implementierungen in ANSI-C. Die Kompilierung erfolgte jeweils mit Microsoft® .NET® unter Microsoft® Windows 2000®. Dennoch tragen suboptimale Programmstrukturen zu Messverfälschungen bei.

[154] Diese wird allgemein *WER* für *Word Error Rate* abgekürzt.

Hörspielen automatisch nach Emotionen segmentiert werden sollen sowie im sogenannten *offenen Mikrofonbetrieb*, bei dem die erkennende Instanz grundsätzlich aktiv ist. Letzteres erlaubt einem Anwender zu jeder Zeit mit einer Benutzerschnittstelle zu sprechen, ohne dies beispielsweise manuell durch einen Taster initiieren zu müssen[155]. Dies bringt zwar eine Reihe von Problemen mit sich, wie das Herausfiltern von nicht sprachlichen Anteilen oder solcher fremder Sprecher (vgl. Kap. 6.2), ist aber für die Emotionserkennung nahezu unerlässlich. Grund hierfür ist, dass emotionale Reaktionen in der Regel spontan sind, und somit nicht vorausgesetzt werden kann, dass ein Nutzer entsprechend segmentiert. Darüber hinaus sind viele emotionale Ausrufe nicht direkt an ein System gerichtet, für dieses aber unter Umständen von Interesse. Bei einer emotionalen Observation ist ebenfalls nur ein permanent aktiver Erkenner möglich.

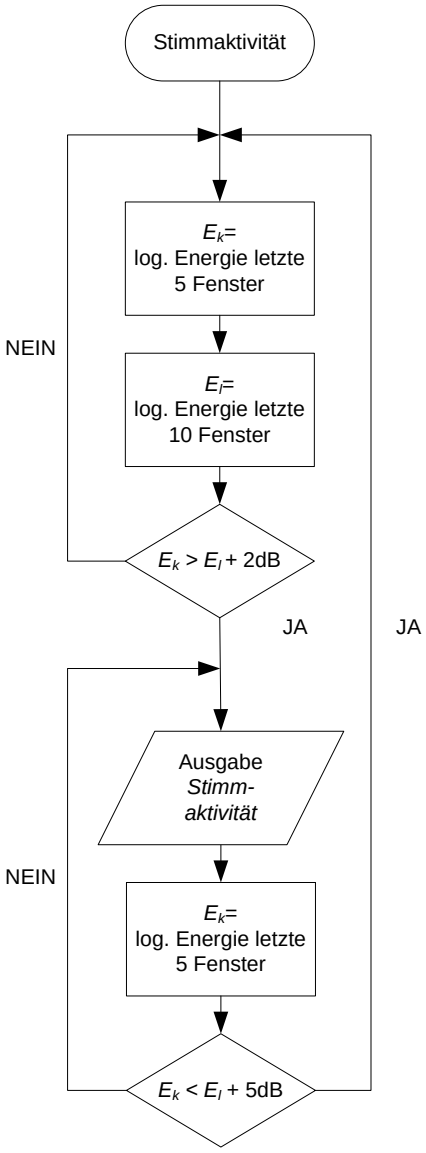


Abb. 4.14: Programmablaufplan zur Stimmaktivitätserkennung

[155] Man spricht dann von *Push-To-Talk* (*PTT)*.

Zur Erkennung von Stimmaktivität wird eine einfache Segmentierung auf Basis von Energieschwellwerten angewandt. Die Initialisierung muss dabei in einem sprachfreien Moment stattfinden. Zunächst wird in einer ersten Schleife jeweils die mittlere logarithmische Kurzzeitenergie E_k über fünf und die entsprechende Langzeitenergie E_l über zehn Siganlrahmen berechnet. Überschreitet die Kurzzeitenergie einen dynamischen Schwellwert gebildet aus der Langzeitenergie und einem addierten Erfahrungswert von 2 dB, wird ein sogenannter *Sprach-Onset* detektiert. Das System emittiert den Zustand *Stimmaktivität*. Sobald in einer zweiten Schleife eine statische Schwelle aus der letzten beobachteten Langzeitenergie und einem Erfahrungswert von 5 dB unterschritten wird, wird vom sogenannten *Sprach-Offset* ausgegangen, und das System springt in die erste Schleife ohne beobachtete Stimmaktivität zurück. Dieser Ablauf ist in Abb. 4.14 in Form eines Programmablaufplans dargestellt.

Da diese Methode grundsätzlich nur geeignet ist, Abschnitte höherer und geringerer Signalaktivität zu trennen, findet zusätzlich eine Klassifikation nach Störgeräusch und Sprache auf Basis von 12 MFCC mit Ableitungen erster und zweiter Ordnung als Merkmale und SVM als Klassifikator statt[156]. Hierzu wurden Aufzeichnungen mit einem Kondensatormikrophon Yoga EM 240 im offenen Betrieb mit der beschriebenen energiebasierten Segmentierung dahingehend ausgewertet, dass 1.000 Beispiele ausgeschnittener sprachfreier Anteile in 11 kHz und 16 Bit gesammelt werden konnten. Die Aufnahmen erfolgten dabei in den in Kap. 2.5.1 beschriebenen Räumen sowie in einer typischen Büroumgebung. In einer Diskrimination zwischen diesen und den Sprachsamples der Datenbank EA-ACT ergeben sich in einer 5-fach SCV 99,7% richtige Zuordnung. Diese nahezu fehlerfreie Klassifikation erweist sich für den in Kap. 7 beschriebenen Anwendungsfall als sehr geeignet.

4.4 Statistisches Sprachmodell

Um aus dem erfassten Text schließlich Emotion erkennen zu können, werden als Vergleichsbasis zunächst die aus der klassischen Spracherkennung bekannten statistischen Sprachmodelle in Form von *klassenbasierten N-Grammen* betrachtet. Allgemein geben N-Gramme die Auftrittswahrscheinlichkeit $P\left(w_j \middle| w_1,...,w_{j-1}\right)$ eines Wortes w_j unter der Bedingung vorangegangener Wörter mit Index $j=1,...,j-1$ von links nach rechts innerhalb einer Phrase $\mathcal{S}$ an.

Ein klassenbasiertes N-Gramm erlaubt darüber hinaus eine gemeinschaftlich probabilistische Modellierung von adjazenten Wörtern hinsichtlich einer Klasse, hier Emotion, Ω_κ: $P\left(\Omega_\kappa \middle| w_1,...,w_j\right)$. Zur vollständigen Schätzung dieser Wahrscheinlichkeit ergibt sich nach dem Zipfschen Gesetz[157] ein sehr hoher Bedarf an Trainingsmaterial. Dies wird durch die Erhöhung der

---

[156] Werden im Anschluss an diese Signaltypdiskrimination Emotionen aus akustischen Merkmalen mit SVM erkannt, lässt sich dieses Vorgehen auch als ML-SVM, wie in Kap. 3.7.2 vorgestellt, auffassen.

[157] Das *„Principle of Least Effort"* nach George Kingsley Zipf postuliert, dass in natürlicher Sprache wenige inhaltsleere Funktionswörter und Worttypen vergleichsweise häufig genutzt werden. Im Gegenzug existieren jedoch viele Worttypen die nur sehr selten auftreten, was die Erfassung dieser Fälle stark erschwert.

Emotionsklassenzahl weiterhin multiplikativ gesteigert. Man beschränkt sich daher aus Gründen der Praktikabilität auf Sequenzen von wenigen n Wörtern. Nachteilig hieran ist, dass unterschiedliche Wortfolgen einer Länge größer n gleich behandelt werden. Es ergibt sich mit der sogenannten *Markov-Annahme* $(n-1)$*-ter Ordnung* folgende Approximation:

$$P\left(\Omega_\kappa \middle| w_1,...,w_j\right) \approx P\left(\Omega_\kappa \middle| w_{j-(n-1)},...,w_{j-1},w_j\right) \tag{4.6}$$

Trotz geringer Anzahl modellierter Wörter bleibt für $n=1$ ein Bedarf an ca. $20 \cdot 10^3$ Wortfolgen je Klasse bestehen, für $n=2$ entsprechend bereits ca. $400 \cdot 10^6$, und für $n=3$ ca. $8 \cdot 10^{12}$. Dies kann jedoch, vor Allem durch Bildung von Wortgruppen, deutlich entschärft werden (vgl. Kap.4.1.3). Außerdem sollte ohnehin ein eher niedriger Koeffizient n gewählt werden, um speziell für die Emotionserkennung unabhängig vom weiteren Sinngehalt einer Äußerung zu bleiben. Dies beruht auf der Tatsache, dass für die emotionale Färbung einer Äußerung oft nur einzelne Wörter verantwortlich sind. Somit sind mindestens *Unigramme* entsprechend $n=1$ erforderlich, bei denen jedem Wort isoliert Wahrscheinlichkeiten bezüglich zugrunde liegender Gemütszustände zugeschrieben werden, wie beispielsweise in [DEV03] angewandt. *Bigramme* mit $n=2$ hingegen erfassen den Zusammenhang zweier benachbarter Wörter, wie in [POL00] betrachtet. In [ANG02] schließlich werden sogar *Trigramme* gemäß $n=3$, jedoch mit geringem Erfolg, eingesetzt. Tab. 4.15 zeigt die Anzahl der jeweiligen N-Gramme der in Kap. 4.7.1 betrachteten Datenbank EAL-F+W für linguistische Emotionserkennung nach Mindestauftrittshäufigkeit[158]. Ab einer solchen Mindestauftrittshäufigkeit von vier je N-Gramm sind Unigramme zahlenmäßig überlegen, jedoch ist auch hier die Gesamtzahl nur noch 366. Diese geringe Zahl spricht dafür, dass hier vorrangig Unigramme eine sinnvolle Modellierung erlauben.

Häufigkeit [#]	**Uni-gramme**	**Bi-gramme**	**Tri-gramme**	**Tetra-gramme**
> 1	1.656	6.550	8.789	9.293
> 2	751	1.355	780	507
> 3	489	570	167	52
> 4	366	323	50	6
> 5	300	217	24	3
> 10	162	57	4	0
> 20	86	11	0	0

Tab. 4.15: N-Gramm Häufigkeiten in der Datenbank EAL-F+W

Zur Veranschaulichung des grundsätzlichen Prinzips von N-Grammen sind in Tab. 4.16 die häufigsten zehn Uni- bis Tetragramme des Korpus EAL-F+W mit ihrer Gesamtauftrittshäufigkeit angeführt. Während mit hohem n die Aussagekraft sichtbar steigt, nimmt die Gesamtauftrittshäufigkeit rapide ab. Auch ein Verhalten nach dem Zipfschen Gesetz ist zu beobachten, da auf den ersten Rängen unter den Unigrammen vorrangig Funktionswörter

[158] *Tetragramme* stehen dabei für $n=4$.

auftauchen[159].

Unigramm	[#]	Bigramm	[#]	Trigramm	[#]	Tetragramm	[#]
I	495	*I don't*	62	*I don't know*	19	*I can't believe it*	6
you	425	*I can't*	44	*what are you*	11	*what do you mean*	6
to	247	*are you*	32	*I can't believe*	10	*what are you doing*	5
a	232	*I was*	30	*oh my god*	10	*you're not going to*	4
the	214	*I am*	30	*I have to*	9	*I don't know how*	4
it	170	*I have*	27	*what the hell*	8	*oh god oh god*	4
that	155	*do you*	27	*what do you*	8	*stop listening to him*	3
don't	149	*you know*	22	*why are you*	8	*hey don't do that*	3
me	146	*in the*	21	*I don't think*	7	*what the hell happened*	3
what	122	*don't know*	20	*I don't want*	7	*I never told you*	3

Tab. 4.16: Häufigste zehn N-Gramme in der Datenbank EAL-F+W

4.4.1 Schätzung der Wahrscheinlichkeiten

Ausgehend von einer Modellierung mit Unigrammen und einem Trainingskorpus $\mathcal{L}$, in dem jeder Phrase $\mathcal{S}$ mit den Wörtern w_j eine Emotion zugeordnet ist, werden die Wahrscheinlichkeiten $P(\Omega_\kappa \mid w_j)$ der im Vokabular $\mathcal{V}$ enthaltenen Wörter $w_i \in \mathcal{V}$ eine Emotion zu repräsentieren geschätzt. Sollte eine Äußerung prinzipiell verschiedene Gefühle beinhalten können, wird sie jeweils als neuer Trainingssatz mit entsprechenden weiteren emotionalen Zuordnungen behandelt. Jedem einzelnen Wort $w_j \in \mathcal{S}$ wird nun zur späteren Auszählung die Emotion Ω_κ des Satzes $\mathcal{S}$ zugewiesen.

Die einfachste Variante zur Schätzung der klassenbedingten Unigramm-Wahrscheinlichkeiten ist die sogenannte *Maximum-Likelihood-Estimation (MLE)*, bei der die Häufigkeit $TF(w_i, \Omega_\kappa)$, mit der das Wort w_i einer Emotion Ω_κ zugehört, ins Verhältnis zur Gesamtauftrittshäufigkeit in allen Klassen $TF(w_i, \Omega)$ gesetzt wird:

$$P_{MLE}(\Omega_\kappa \mid w_i) = \frac{TF(w_i, \Omega_\kappa)}{TF(w_i, \Omega)} \tag{4.7}$$

Als problematisch beim Vorgehen nach MLE erweist sich, dass die Wahrscheinlichkeit nicht auftretender Wort/Emotionspaare zu Null gesetzt wird. Dies setzt in einer produktbasierten Berechnung der Gesamtwahrscheinlichkeit einer Äußerung wie in Gl. 4.10 letztere ebenfalls zu Null, was auch bei großer zur Verfügung stehender Trainingsmenge nicht vermeidbar ist.

Alternativ kann ein *Maximum-A-Posteriori (MAP)* Schätzer angewandt werden. Das Vorwissen, dass fast jede Wortfolge prinzipiell über die insgesamt k Emotionen aus der Menge Ω verteilt auftreten kann, fließt als a-priori Wissen in Form einer initialen Gleichverteilung ein:

---

[159] Im Vergleich hierzu sind nach [QUA03] die häufigsten Wörter der englischen Sprache: *of, to, and, a, in, for, is, the, that,* und *on.* Vier dieser Begriffe treten auch hier unter den häufigsten auf.

$$P_{MAP}\left(\Omega_\kappa \mid w_i\right) = \frac{TF\left(w_i, \Omega_\kappa\right) + 1}{TF\left(w_i, \Omega\right) + k} \text{ mit } k = \left|\Omega\right| \tag{4.8}$$

Diese Vorgehensweise ist auch unter der Bezeichnung *Add-One-* oder *Laplace-Glättung* bekannt. Für kleine Korpora, wie in der Emotionserkennung gegeben, kann dies jedoch leicht zur Überschätzung nicht beobachteter Ereignisse führen.

Als Lösung wird eine Interpolation zwischen MLE und MAP Schätzung nach dem Lidstoneschen Gesetz [LID20] gewählt. Hierbei wird als Initialisierung für jedes Wort eine Mindestwahrscheinlichkeit für die Zugehörigkeit zu jeder Klasse gewählt. Dies geschieht durch Addition des Interpolationskoeffizienten λ anteiliger fiktiver Beobachtungen, wobei oft $\lambda = 0,5$ gewählt wird:

$$P_\lambda\left(\Omega_\kappa \mid w_i\right) = \frac{TF\left(w_i, \Omega_\kappa\right) + \lambda}{TF\left(w_i, \Omega\right) + k \bullet \lambda} \text{ mit } k = \left|\Omega\right| \tag{4.9}$$

Für den Korpus EAL-F+W (siehe Kap. 4.7.1) wurden die Werte $\lambda \in \left\{0, \frac{2}{25}, \frac{4}{25}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, 1\right\}$ evaluiert, wobei sich $\lambda = \frac{4}{25}$ als Optimum herauskristallisierte.

Tab. 4.17 zeigt zur abschließenden Veranschaulichung zur Schätzung der Wahrscheinlichkeiten drei Beispiele: die Unigramme *hell, cry* und *why*. Die Werte entsprechen einer Berechnung mit $\lambda = \frac{4}{25}$ bezüglich des Korpus EAL-F+W.

$P_\lambda\left(\Omega_\kappa \mid w_i\right)$	**Ärger**	**Ekel**	**Furcht**	**Freude**	**Neutral.**	**Trauer**	**Überr.**
hell	**0,394**	0,030	0,151	0,091	0,061	0,061	0,212
cry	0,154	0,077	0,077	0,077	0,077	**0,462**	0,077
why	0,280	0,020	0,040	0,020	0,040	0,160	**0,440**

Tab. 4.17: Auszug aus den klassenbedingten Auftrittswahrscheinlichkeiten von Unigrammen nach Lidstone-Glättung, Datenbank EAL-F+W. Die wahrscheinlichsten Klassen sind hervorgehoben.

Das Wort *hell* etwa spricht hier zunächst für die Emotion Ärger wie in *„Hell no!“*, in zweiter Linie kommt es aber auch in überraschten Ausrufen wie *„What the hell?“* vor. *Cry* hingegen spricht hier klar für Trauer. Man erkennt auch die Anwendung der Lidstone-Glättung, da das Wort nur in ärgerlichen oder traurigen Ausrufen beobachtet wurde, aber die Wahrscheinlichkeiten für die weiteren Emotionen gleichverteilt, jedoch von Null verschieden sind.

4.4.2 Probabilistische Modellierung

Der Fall der Erkennung wird hier anhand des im Weiteren verfolgten Ansatzes mit Unigrammen gezeigt. Eine Berechnung für N-Gramme mit $n > 1$ erfolgt in direkter Analogie hierzu.

Die erkannte Klasse κ_e für eine beobachtete Phrase $\mathcal{S}$ ergibt sich aus dem maximalen klassenbezogenen Produkt aller a-posteriori Wahrscheinlichkeiten $P\left(\Omega_\kappa \middle| w_j\right)$ der Wörter w_j die im

Vokabular enthalten sind:

$$\kappa_e = \arg\max_{\kappa} \prod_{(w_j \in \mathcal{S}) \wedge (w_j \in \mathcal{V})} P\left(\Omega_\kappa \middle| w_j\right) \tag{4.10}$$

Um nun einer ungleichen Verteilung der Beispiele in der Lernmenge gerecht zu werden, kann diese Gleichung erweitert werden zu:

$$\kappa_e = \arg\max_{\kappa} \prod_{(w_j \in \mathcal{S}) \wedge (w_j \in \mathcal{V})} \frac{P\left(\Omega_\kappa \middle| w_j\right)}{P\left(\Omega_\kappa\right)} \text{ mit } P\left(\Omega_\kappa\right) = \frac{\left|\mathcal{L}_\kappa\right|}{\left|\mathcal{L}\right|} \tag{4.11}$$

$P\left(\Omega_\kappa\right)$ ist dabei die a-priori Wahrscheinlichkeit der Klassen, gebildet aus dem Verhältnis der Lernbeispiele $\left|\mathcal{L}_\kappa\right|$ der Klasse κ und der Gesamtheit der Lernbeispiele $\left|\mathcal{L}\right|$.

Als Variante hierzu wird das informationstheoretische Maß der *Salience* $\text{sal}\left(w_j\right)$, beruhend auf der *Self-Mutual-Information (SMI)* $\text{smi}\left(\Omega_\kappa, w_j\right)$ [COV91], verwendet:

$$\text{sal}\left(w_j\right) = \sum_{\kappa=1}^{k} P\left(\Omega_\kappa \middle| w_j\right) \cdot \text{smi}\left(\Omega_\kappa, w_j\right) \text{ mit } \text{smi}\left(\Omega_\kappa, w_j\right) = \text{ld} \frac{P\left(\Omega_\kappa \middle| w_j\right)}{P\left(\Omega_\kappa\right)} \tag{4.12}$$

Die Salience ist dabei ein Maß dafür, wie bedeutend ein Wort zur Erkennung der Emotion ist (vgl. Kap. 3.7.4). Für einen Satz ergibt sich dann nach [LEE02A]:

$$\kappa_e = \arg\max_{\kappa} \sum_{(w_j \in \mathcal{S}) \wedge (w_j \in \mathcal{V})} P\left(\Omega_\kappa \middle| w_j\right) \cdot \text{sal}\left(w_j\right) \tag{4.13}$$

Hauptproblem bei der linguistischen Analyse sind sogenannte OOV-Ereignisse, dass heißt, für ein oder mehrere Wörter w_j der betrachteten Phrase $\mathcal{S}$ gilt $w_j \notin \mathcal{V}$. Je weniger Wörter enthalten sind, desto unsicherer wird im Allgemeinen die Aussage über die Emotion. Im schlimmsten Fall ist entsprechend $\mathcal{S} \cap \mathcal{V} = \varnothing$ kein Wort der Phrase enthalten und somit keine Aussage möglich. Hier wird dann, mit der Begründung, dass keine emotionalen Wörter beobachtet wurden, von einer neutralen Äußerung ausgegangen.

4.5 Vektorraummodell

Die sogenannte *Bag-Of-Words*[160] *(BOW)* Repräsentation von Text ist im Bereich der automatischen Textkategorisierung ein weit verbreiteter und erfolgreicher Ansatz. Vergleichbar dem in dieser Arbeit betrachteten Szenario werden digitale Texte in eine feste Anzahl vordefinierter Kategorien wie Wirtschaft, Sport und Kultur unterteilt. Es soll hier geprüft werden, inwiefern sich dieser Ansatz auch zur Modellierung von Emotion eignet.

---

[160] Auch als *Bunch-Of-Words* Repräsentation bekannt.

Für eine automatische Klassifikation wird der betrachtete Text, äquivalent zu den bereits vorgestellten akustischen Merkmalen, in Form von Attributen mit zugehörigem Wert repräsentiert. Im hier relevanten Fall der Abbildung einer Phrase $\mathcal{S}$ auf eine Emotionsklasse Ω_κ bedeutet dies, dass jedes verschiedene Wort w_j innerhalb der Äußerung mit einem Merkmal $x_i \in \underline{x}$ entsprechend seiner Häufigkeit korrespondiert [JOA97A]. Eine betrachtete Äußerung wird dann durch einen über den gesamten Vokabularraum reichenden Vektor repräsentiert, woraus sich die Bezeichnung *Vektorraummodell* ableitet. Eine Modellierung in Bezug auf Wortabhängigkeiten oder die relative Position eines Wortes im Satz erfolgt dabei nicht. Die Dimension des Merkmalsvektors ergibt sich äquivalent der Menge unterschiedlicher Termini $\mathcal{V} = \{w_1, ..., w_V\}$, die in der Lernprobe $\mathcal{L}$ von Texten oder Phrasen beobachtet wurden zu $V = |\mathcal{V}|$. Es ergibt sich in der Regel hieraus eine sehr hohe Dimensionalität, welche oft in keinem sinnvollen Verhältnis zur Zahl der zur Verfügung stehenden Lernbeispiele steht. Für den in Kap. 4.7.1 beschriebenen Korpus EAL-F+W etwa ergibt sich wie genannt eine Vokabulargröße von 2.334 Einträgen.

Das zunächst einfachste Abbildungsmaß $w_i \rightarrow x_i$ eines Terms des Vokabulars ist eine binäre Notation die aussagt, ob ein solcher in einer Äußerung enthalten ist. Geht die Häufigkeit eines Terms direkt ein, so ist die Ausgangsbasis die Term-Frequency $TF(w_i, \mathcal{S})$ (vgl. Kap. 4.4.1) [SAL88]. Sie spiegelt direkt die Häufigkeit eines Wortes $w_i \in \mathcal{V}$ des Vokabulars innerhalb der zu analysierenden Phrase $\mathcal{S}$ wieder. Eine Normierung auf die Gesamtlänge $S = |\mathcal{S}|$ der Phrase trägt dabei einer Varianz in dieser Größe Rechnung. Für eine Komponente $x_{TF,i}$ des zu bildenden Merkmalsvektors $\underline{x}$ ergibt sich somit:

$$x_{TF,i} = \frac{TF(w_i, \mathcal{S})}{S} \tag{4.14}$$

In [Sal88] wird darüber hinaus gezeigt, dass eine Normierung in Bezug auf die Auftrittshäufigkeit eines Wortes in verschiedenen Dokumenten eine Verbesserung ergibt. Hierzu sei die *Document Frequency* $DF(w_i)$ eingeführt die angibt, in wie vielen Dokumenten, in unserer Betrachtung Einzelphrasen, das Wort w_i beobachtet wird. Aus ihr ergibt sich die inverse Document Frequency, kurz *IDF*, in folgender Form, wobei $L = |\mathcal{L}|$ die Gesamtzahl der Lernphrasen bezeichnet:

$$IDF(w_i) = \frac{L}{DF(w_i)} \tag{4.15}$$

Unter Anwendung einer Skalierung gemäß IDF erhält man für die Komponenten x_i des Merkmalsvektors $\underline{x}$:

$$x_{TFIDF,i} = TF(w_i, \mathcal{S}) \cdot IDF(w_i) \tag{4.16}$$

Wird die TF zusätzlich logarithmiert, um Linearitäten zu kompensieren, ergibt sich der Merkmalstyp *logTF*:

$$x_{\log TF,i} = \log\left(o_{TF} + TF\left(w_i, \mathcal{S}\right)\right) \text{ mit } o_{TF} = \left\{\tfrac{1}{2}, 1\right\} \tag{4.17}$$

Die in der Gleichung enthaltene Offset-Konstante o_{TF} ist vor der Logarithmierung erforderlich, um einen Definitionsfehler bei nicht vorhandenen Termini zu umgehen. Zu ihrem Wert existiert eine Reihe unterschiedlicher Definitionen, wobei im Rahmen der folgenden Ausführungen o_{TF} zu Eins gesetzt wird. Entsprechend erhält man schließlich das Maß *logTFIDF*:

$$x_{\log TFIDF,i} = \log\left(o_{TF} + TF\left(w_i, \mathcal{S}\right) \cdot IDF\left(w_i\right)\right) \text{ mit } o_{TF} = \left\{\tfrac{1}{2}, 1\right\} \tag{4.18}$$

Zur Klassifikation dieser hochdimensionalen Merkmalsvektoren haben sich insbesondere SVM (siehe Kap. 3.7.2) als besonders geeignet erwiesen [DUM98]. Grund hierfür ist unter Anderem der Schutz vor Overfitting unabhängig von der Dimensionalität.

4.6 Grafische Modellierung

Als grundsätzlicher Nachteil des zuletzt betrachteten Bag-Of-Words Verfahrens erweist sich die fehlende Behandlung der Wortreihenfolge sowie die Sicht über bestehende Zusammenhänge. Während diese Punkte in der Klassifikation von Texten unter Umständen nur eine untergeordnete Rolle spielen können, ergibt sich eine Reihe mangelhaft repräsentierter Fälle bei der emotionalen Zuordnung singulärer Phrasen. Vorrangig sei hier eine sinngemäße Negierung wie in der Äußerung *„Ich bin nicht böse.“* genannt. Während bei klassenbasierten N-Grammen höherer Ordnung diese Reihenfolge mit berücksichtigt wird, sind sie zu stark auf die exakte Wortgruppe fixiert. Verändern sich etwa Wörter in der Mitte, wird das N-Gramm als solches nicht mehr erkannt. Das Ziel der folgenden Modellierung ist daher eine Suche über Einzelwörter hinaus nach Phrasen, die jedoch in ihrem Inneren durchbrochen sein dürfen. Zusätzlich wird es hierdurch möglich, Ausprägungen sensibler wahrzunehmen. Als Beispiel seien hierzu die Aussagen *„Ich bin leicht verärgert.“* und *„Ich bin extrem verärgert.“* genannt. Die Termini *leicht* oder *extrem* können von einem geeigneten Verfahren mit ausgewertet werden.

Auch bei dieser Modellierung werden, wie für das String-Matching (vgl. Kap. 4.2.2), erweiterte BN eingesetzt. Es erfolgt eine Klusterung auf vier Hierachieebenen: von Wort über Lexem zu Phrase hin zur Emotion [SCH04A], [MÜL05A]. Hierbei wird ein eigenes Modell je Phrase erzeugt, wobei in der Wurzel des BN die Emotion Ω_κ selbst steht, die diese Phrase ausdrückt. Die Evidenz wird dann in alle Phrasenmodelle, unter Berücksichtigung der richtigen Reihenfolge übereinstimmender Wörter der zu analysierenden Phrase, eingebracht. Dabei werden Konfidenzwerte der untergeordneten Schicht zur Erkennung auf Signalebene direkt in das Modell übernommen. Um die Berechnung zu beschleunigen, wird im Gegensatz zum String-Matching darauf verzichtet, verschiedene Wahrscheinlichkeiten auf unterster Ebene des Modells zuzulassen. Stattdessen wird ein Lexem l_i dann als evident gewertet, wenn eines der enthaltenen Wörter $w_{i,j}$ an der richtigen Stelle gefunden wird. Die Lexeme werden in einer nächsten Ebene zu N Teilphrasen $\mathcal{S}_l'$ mit je M Lexemen zusammengefasst. Diese Phrasen können Verneinungsphrasen, Phrasen der emotionalen Ausprägung oder emotionale Teilphrasen sein und setzen sich in der Regel aus ein bis drei Lexemen zusammen. Abb. 4.18 zeigt zur Veranschaulichung ein allgemeines Phrasennetz.

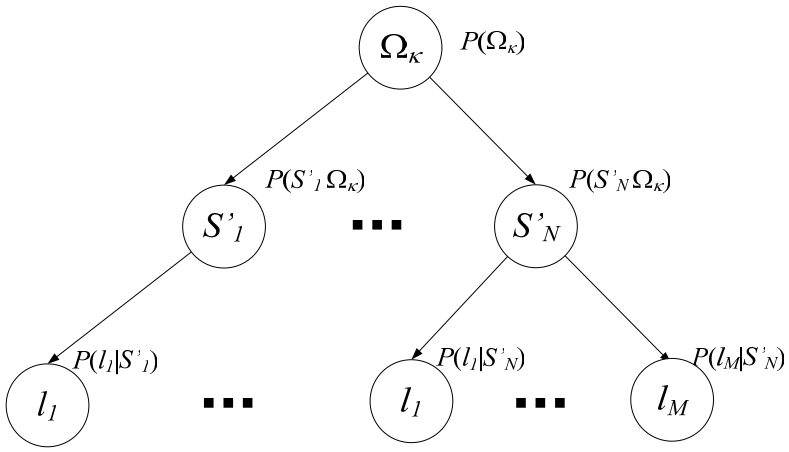


Abb. 4.18: Bayessches Phrasennetz zur linguistischen Emotionserkennung

Im Gegensatz zu dem in Kap. 4.2.2. vorgestellten String-Matching werden die einzelnen Modelle nicht dynamisch zur Laufzeit gebildet, da die Struktur der natürlichen Sprache im Vergleich zu String-Matching zu komplex ist. Vielmehr erfolgt hier ein Lernen anhand von Trainingsbeispielen bereits im Vorfeld. Um dieses zu ermöglichen, ist eine Verschriftung eines Datensatzes entsprechend der Lexeme und Teilphrasen erforderlich. Dazu werden die Phrasen des Lernmaterials als erstes nach Emotion sortiert, und anschließend, entsprechend einem semantischen Modell, gelabelt. Dieses grundlegende Modell muss durch einen Experten vorgegeben werden, und legt fest, welche Wörter welchen Lexemen zugehören und welche Teilphrasen existieren. Ferner können einzelne Lexeme als *Parameterlexeme* definiert werden, sodass bei Ihrer Erkennung automatisch das jeweilige Wort ausgegeben wird. Dies kann genutzt werden, um wie beschrieben die Ausprägung zusätzlich zu modellieren. Grundsätzlich lässt es sich nicht vermeiden, vor der Verschriftung ein semantisches Modell festzulegen. Es ist dabei im Besonderen zu erwähnen, dass die Güte dieses Modells im entscheidenden Maß für die generelle Erkennungsleistung verantwortlich ist. Ein zu fein untergliedertes Modell wird in der Anwendung wenige Treffer erzielen, ein zu grobes Modell hingegen führt zu Unschärfen bei der Abgrenzung einzelner Teilphrasen und letztlich auch einzelner Emotionen. Im Anschluss an die Verschriftung werden die jeweiligen Wahrscheinlichkeiten der BN automatisch durch Auszählen gebildet. Die BN werden schließlich in Form von XML gespeichert, was auch eine übersichtliche Nachbearbeitung und Ergänzung durch einen Experten ermöglicht.

4.7 Experimente und Ergebnisse

Im Folgenden werden die vorgestellten Verfahren statistisches Sprachmodell, Vektorraummodell und grafische Modellierung auf drei Datenbanken getestet. Es wird dabei grundsätzlich auf eine Kreuzvalidierung mittels SCV bei der Vokabularbildung, dem Lernen und der Evaluierung geachtet (vgl. Kap. 3.12.3).

4.7.1 Verwendete Datenbanken

Für das Training und die Beurteilung der Erkennungsgüte in der semantischen Emotionserkennung sind im Rahmen dieser Arbeit zunächst in der als *EL*[161]*-FILM* bezeichneten Datenbank Sätze aus Drehbüchern amerikanischer Filme in der original sprachlichen Fassung annotiert worden. Es handelt sich dabei um die in Tab. 4.19 angeführten insgesamt sieben U.S. amerikanischen Filme. Diese wurden nach Genre so ausgesucht, dass möglichst alle Basisemotionen vertreten sind, und auch möglichst oft vorkommen.

Filmtitel	**Jahr**	**Genre**	**Drehbuchautoren**
Alien	1986	Science-Fiction	Cameron, James
Annie Hall	1977	Komödie	Allen, Woody; Brickman, Marshall
Five Easy Pieces	1970	Drama	Joyce, Adrian
Notting Hill	1999	Komödie	Curtis, Richard
Scream	1996	Horror	Williamson, Kevin
Ten things I hate about you (The taming of the shrew)	1999	Komödie	Lutz, Karen; Smith, Kirsten; Shakespeare, William
Toy Story	1995	Animation, Fantasie	Whedon, Joss; Stanton, Andrew; Cohen, Joel; Sokolow, Alec

Tab. 4.19: Beschreibung der Filme für ausgesuchte Drehbuchpassagen in der Datenbank EL-FILM

Es werden jeweils einzelne zu spielende Phrasen des Drehbuchs satzweise betrachtet. Dabei ist eine gesamte Phrase einer oder mehreren Emotionen zugeordnet. Beispiele solcher Phrasen sind etwa für Freude: *„It's wonderful,"* oder für Ärger: *„Will you just do what the hell I tell you!"*[162]. Die Annotation erfolgte von zwei Personen verschiedenen Geschlechts, beide Studenten im Alter von 25 a mit langjährigen Aufenthalten in englischsprachigen Ländern. Um eine Behandlung aus dem Kontext heraus zu vermeiden, wurden die Sätze vor der Beurteilung zufällig gemischt. Eine Zuordnung zu einer Emotion erfolgte jeweils nur dann, wenn diese der Testperson sicher möglich erschien. Tab. 4.20 zeigt eine Übersicht über den gesamten Satz ausgewählter Phrasen, der sich aus der Vereinigungsmenge $\mathcal{L}_1 \cup \mathcal{L}_2$ der jeweiligen Annotatoren ergibt sowie den letztlich verwendeten Satz übereinstimmend zugeordneter Phrasen, entsprechend der Schnittmenge $\mathcal{L}_1 \cap \mathcal{L}_2$. Dabei entspricht L der jeweiligen Phrasenzahl und $W_{\mathcal{L}}$ der Gesamtzahl der Wörter in $\mathcal{L}$. S bezeichnet die Phrasenlänge und V die Vokabulargröße disjunkter Termini in $\mathcal{L}$.

[161] Entsprechend den in Kap. 3.12.1 gewählten abkürzenden Bezeichnungen kürzt *EL* für die Korpora dieses Abschnitts *Emotionsdatenbank - linguistisch* ab.

[162] In Kap. A.3.1 wird eine Reihe weiterer Beispiele aufgeführt.

Anzahl [#]	L	$W_{\mathcal{L}}$	$\min(S)$	$\max(S)$	$\mu(S)$	V
$\mathcal{L}=\mathcal{L}_1\cup\mathcal{L}_2$	1.312	9.217	1	20	7,0	-
$\mathcal{L}=\mathcal{L}_1\cap\mathcal{L}_2$	1.144	8.049	1	20	7,0	1.421

Tab. 4.20: Datenbank EL-FILM: Emotionale Passagen aus Drehbüchern, Statistik der Wörter

Tab. 4.21 schlüsselt die jeweiligen Mengen $\mathcal{L}_1, \mathcal{L}_2$ und $\mathcal{L}_1 \cap \mathcal{L}_2$ zusätzlich nach Anzahl von Phrasen je Emotion auf. Es zeigt sich dabei ein deutliches Ungleichgewicht zu Gunsten von Ärger, Neutralität und Trauer. Die Verteilung unter den Annotatoren hingegen ist relativ ähnlich.

Anzahl [#]	**Ärger**	**Ekel**	**Freude**	**Furcht**	**Neutral.**	**Trauer**	**Überra.**
$\mathcal{L}_1$	374	18	97	105	291	276	150
$\mathcal{L}_2$	348	29	110	106	288	255	176
$\mathcal{L}_1\cap\mathcal{L}_2$	329	18	89	96	226	241	145

Tab. 4.21: Datenbank EL-FILM: Emotionale Passagen aus Drehbüchern, Verteilung der Emotionen, Angabe in Anzahl der Sätze je Emotion

Mit dem Ziel hier eine gleichmäßigere Verteilung zu erhalten, wurden als weitere Quelle emotionale Textphrasen aus Internet-Diskussionsforen gesammelt, die mit einem Emoticon versehen waren. Die identischen Annotatoren, wie bei der Verschriftung der Datenbank EL-FILM, haben geprüft, ob Emoticons und Inhalt zusammenpassen, um beispielsweise Ironien zu vermeiden. Hierbei ergab sich der Datensatz EL-WWW[163] mit weiteren 363 Sätzen. EL-FILM und EL-WWW umfassen somit insgesamt 1.507 Äußerungen in der Datenbank EAL-F+W[164]. Um im später folgenden Kap. 6.1 auch die Integration von akustischer mit linguistischer Analyse testen zu können, wurden diese Phrasen auch von drei Laien über einen Zeitraum von sechs Monaten nachgespielt. Dieser lange Zeitraum soll Antizipationseffekten der Probanden vorbeugen. Die Emotionen wurden in zufälliger Reihenfolge simuliert. Die Neuaufzeichnung der emotionalen Passagen vermeidet die akustischen Überlagerungen der originalen Filmausschnitte bestehend aus zusätzlichen Hintergrundgeräuschen, Sprachanteilen anderer Sprecher und Filmmusik, die das Ergebnis stark verfälschen. In Tab. 3.44, Kap. 3.12.1 sind weitere Details der Datenbank hinsichtlich der akustischen Einspielung beschrieben worden.

Eine weitere gesammelte, rein textuelle Datenbank besteht aus einem Satz von Filmkritiken. Prinzipiell geht es hierbei um die Bewertung nach positivem oder negativem Eindruck des Autors in Schriftform [FIN03] und nicht um spontane Emotionen im engeren Sinne. Als Vorteil ergibt sich hier neben einer hohen verfügbaren Quantität von Beispielen eine Annotation dieser durch den Verfasser selbst. Letzteres beruht auf der Tatsache, dass ausschließlich Filmkritiken gewählt wurden, bei denen ebenfalls eine Bewertung in numerischer Repräsentation, etwa Noten, Sternchen, oder ähnliches, vorliegt. Der Internetdienst *Metacritic*® [MET05A] sammelt Kritiken zu Filmen,

[163] WWW wird als Abkürzung für *World-Wide-Web* verwendet, da die Phrasen aus Internetchat stammen.

[164] EAL steht für *Emotionsdatenbank - akustisch und linguistisch*.

Videospielen, Büchern und Musikalben, und projiziert deren Bewertung auf eine Skala von 0 bis 100 Punkte, die wiederum in die Kategorien *positiv*, *neutral* (50-74 Punkte) und *negativ* unterteilt wird. Darüber hinaus wird jeweils das Fazit als Exzerpt aus der Kritik bereitgestellt. Dieses beträgt in der Regel einen Satz. Der Höchstwert entspricht dabei der Bestbewertung [MET05B]. Ferner bietet dieser Dienst eine Sammlung der am besten und der am schlechtesten bewerteten Filme seit Beginn der Filmgeschichte an. Um die hier getroffene Auswahl nachvollziehbar zu gestalten, sind in der Datenbank *EL-META* die Kritiken der 60 best- und der 100 am schlechtesten beurteilten Filme bis zum Jahr 2005 ausgesucht worden. Das unterschiedliche Verhältnis ist durch die Tatsache bedingt, dass zu niedrig bewerteten Filmen meist weniger Kritiken existieren. Eine Gleichverteilung konnte somit nur durch entsprechende Wahl mehrerer schlecht abschneidender Filme erzielt werden. Kritiken im neutralen Bereich wurden dabei nicht betrachtet, da sie auf Grund der polarisierenden Auswahl zahlenmäßig zu gering vertreten waren.

Tab. 4.22 gibt einen Gesamtüberblick über die verwendeten Datenbanken zur linguistischen Analyse. Neben der Datenbank EAL-F+W existieren auch zur Datenbank AEC akustische Beispiele, wie bereits in Tab. 3.44, Kap. 3.12.1 gezeigt wurde.

Beschreibung	**EL-META**	**AEC**	**EAL-F+W**	
Domäne	Filmkritik	Kinder	Spielfilm	Internetchat
Inhalt	Schlüsselsatz aus Filmkritiken	Kinder spielen mit Roboter, WOO Versuch	Dialogakte aus Drehbüchern	Phrasen aus Chat-Konversation
Emotionen	Positiv, Negativ	Ärger, Bemutterung, Betonung Neutralität	MPEG-4, Neutral	
Sprache	Englisch	Deutsch	Englisch	
Art	Real	Real, spontan	Fiktiv	Real, spontan
Phrasenzahl	2.714	1.480	1.144+363=1.507	
Wortzahl	54.162	2.606	9.866	
Phrasenlänge	Ø20,0 Wörter	Ø 1,8 Wörter	Ø7,0 Wörter	
Vokabulargröße	9.735	269	2.334 (reduziert 1.853)	
Verteilung	gleich	ungleich	ungleich	ungleich
Annotation	Emotion, Inhalt	Emotion, Inhalt	Emotion, Inhalt	Emotion, Inhalt
Annotatoren	Autor	5	2	2 + Autor
Öffentlich	Ja	Ja	Nein	Nein

Tab. 4.22: Übersicht über die Eigenschaften verwendeter textueller Datenbanken zur Emotionserkennung. Zu diesen existieren teilweise auch akustische Beispiele.

4.7.2 Menschliche Leistung in der Diskrimination

Als Vergleichsbasis zur Einstufung der automatisch erzielten Akkuratheit soll, wie bereits in Kap. 3.12.2, die menschliche Erkennungsrate betrachtet werden. Hierzu wurden in einer Studie 50

Phrasen der Datenbank EL-FILM durch 15 Probanden den vorgegebenen sieben Emotionen des MPEG-4 Sets mit ergänzender Neutralität zugeordnet. Die Phrasen waren dabei nahezu gleichverteilt unter den Emotionen entsprechend der Notation der beiden ursprünglichen Annotatoren ausgewählt. Die Testpersonen waren alle Studenten im mittleren Alter von 24,8 a (Minimum 22 a, Maximum 28 a), davon vier weiblich. Mehrfachnennungen wurden indirekt proportional zu ihrer Gesamtzahl gewichtet. Betrachtet man beim Ergebnis jeweils nur die Emotion mit der höchsten Übereinstimmung pro Phrase, ergibt sich ein Mittel von 55,7% unter den Probanden und allen 50 Phrasen. Tab. 4.23 zeigt diese Übereinstimmung für alle sieben Emotionen aufgeschlüsselt. Das Versuchsergebnis ist detailliert in Tab. A.5, Kap. A.3 abgebildet.

Akk. [%]	Ärger	Ekel	Furcht	Freude	Neutral.	Trauer	Überras.
μ	55,9	59,1	36,8	68,2	45,6	52,9	65,5

Tab. 4.23: Menschliche Leistung in der Klassifikation linguistischer Emotionsbeispiele, Datenbank EL-FILM, mittlere Leistung 55,7%

Dieses Ergebnis sagt aus, dass durch eine linguistische Analyse einzelner Phrasen, im Vergleich zu einer rein akustischen Betrachtung, niedrigere Erkennungsleistungen zu erwarten sind (84,3% und 67,3% ergaben sich aus akustischer Analyse für die linguistisch emotional neutralen Datenbanken EMO-DB und DES, siehe Kap. 3.12.2).

4.7.3 Evaluierung der linguistischen Analyse

Grundsätzlich stellt sich bei der linguistischen Analyse die Frage, wie Phrasen behandelt werden sollen, die ausschließlich aus OOV-Wörtern bestehen. Obwohl dieser Fall in der Regel nur einem geringen Prozentsatz entspricht, hat er dennoch Auswirkungen auf die Beurteilung der Zuordnungsleistung. Es kann beispielsweise registriert werden, dass eine aktuelle Phrase nicht zuordenbar ist. Im Folgenden wird jedoch in solch einem Fall, wie in Kap. 4.4.2 beschrieben, zu Gunsten einer neutralen Äußerung entschieden, da aus der Sicht des Erkenners kein emotionaler Gehalt beobachtet wurde.

Tab. 4.24 zeigt anhand des Korpus EL-META die Erkennungsleistung mit verschiedenen Merkmalstypen bei der BOW-Repräsentation von Text.

Merkmalstyp	Binär	TF	TFIDF	logTF	logTFIDF
Akkuratheit [%]	81,9	82,6	**83,2**	**83,2**	82,8

Tab. 4.24: Vergleich verschiedener Merkmalstypen bei der BOW-Repräsentation, Datenbank EL-META, Klassifikation mit SVM, 3-fach SCV

Tab. 4.25 zeigt einen Vergleich zwischen den erzielten Leistungen mittels Unigrammen, Vektorraumdarstellung mit SVM, und grafischer Modellierung mit BN. Zum Vergleich sind die Erkennungsraten, die sich bei rein zufälliger Klassenwahl und steter Auswahl der häufigsten Klasse ergeben, gezeigt. Die Leistungsangaben beruhen auf reiner Textbasis, um das Ergebnis an dieser Stelle nicht durch Erkennungsfehler eines Spracherkenners zu verzerren. Eine Betrachtung der textuellen Analyse unter Einbezug der vorab nötigen Spracherkennung erfolgt in Kap. 6.1.

Akkuratheit [%]	EL-META	EAL-F+W	AEC
Raten	50,0	14,3	25,0
Häufigste Klasse	50,7	27,5	37,2
Unigramm	67,5	39,6	79,6
BOW SVM	**83,2**	42,2	**79,7**
BN	67,7	**64,8**	76,2

Tab. 4.25: Evaluierung der linguistischen Emotionserkennung, 10-fach SCV und Vergleich zu den rechnerischen Erkennungsraten zufällige Entscheidung und stete Wahl der häufigsten Klasse

Anhand der Tabelle ist zu sehen, dass Unigramme insgesamt die ungünstigste Wahl darstellen, und von den beiden hier vorgestellten Alternativverfahren bezüglich der Erkennungsleistung übertroffen werden. Die grafische Modellierung mit BN zeigt sich bei der Datenbank EAL-F+W als beste Lösung, während Vektorraumrepräsentation mit Worthäufigkeitsmerkmalen sonst die günstigste Wahl ist. Dies ist in erster Linie auf eine starke Abhängigkeit von der zugrunde liegenden Phrasenlänge innerhalb der Korpora zurückzuführen. Die Datenbank EL-META besitzt dabei die höchste Satzlänge, welcher der Ansatz grafischer Modellierung nicht gerecht wird. Bei den sehr kurzen Äußerungen der Datenbank AEC hingegen ergibt sich kein nennenswerter Vorteil eines phrasenorientierten Ansatzes mehr.

Abschließend zeigt Tab. 4.26 zur Datenbank AEC die Akkuratheit nach Emotion aufgeschlüsselt zum Vergleich zur akustischen Analyse in Tab. 3.66, Kap. 3.12.4.

	Ärger	Bemutterung	Betonung	Neutralität
Akkuratheit [%]	81,6	71,9	74,1	82,0

Tab. 4.26: Linguistische Analyse, Erkennungsrate nach Emotion, Datenbank AEC, Klassifikation mit SVM, 10-fach SCV

5

Manuelle Emotionserkennung

„Wirkliches Neuland in einer Wissenschaft kann wohl nur gewonnen werden, wenn man an einer entscheidenden Stelle bereit ist, den Grund zu verlassen, auf dem die bisherige Wissenschaft ruht, um gewissermaßen ins Leere zu springen.“

WERNER HEISENBERG (1901-1976)

Neben der bisher in Kap. 3 und Kap. 4 betrachteten Sprache, die sich auch als Eingabemedium etwa im Fahrzeugbereich oder bei Mobiltelefonen zunehmend ergänzend durchsetzen kann, ist in der Kommunikation mit Heim- und Bürocomputern vorrangig noch eine GUI basierte Bedienung Stand der Technik. In der somit hauptsächlich manuellen Interaktion werden im Folgenden zwei verwandte Eingabeparadigmen betrachtet: Die Eingabe über eine Maus wie in grafischen Benutzeroberflächen üblich, oder über eine berührungssensitive Oberfläche. Letztere Art von Eingabe wird im Zusammenhang mit der zunehmenden Verbreitung von stift- oder direktberührungsorientierten Szenarien in elektronischen Notizblöcken, PDA, Tablet PC, oder öffentlichen multimedialen Informationsständen betrachtet, um zu eruieren, inwiefern Erkennung einer Benutzeremotion direkt aus diesen Interaktionsdaten möglich ist. Dabei soll explizit auf den Einsatz ergänzender Hardware, wie oberflächlich angebrachte physiologische Sensoren, verzichtet werden. Ein Erfolg diesbezüglich würde einen großen Anwendungsbereich, etwa in Internetportalen, Lernprogrammen oder Ähnlichem, erschließen.

5.1 Verarbeitung von Mauszeigerbewegungen

Die Betrachtung von Mausbewegungen erfolgt im Folgenden unabhängig von der aktiven Applikation, mit der ein Nutzer interagiert. Dies ist insofern berechtigt, als es hier das Ziel ist, die Emotion während des Umgangs mit einer bestimmten, bekannten Anwendung zu erkennen. Die Maus als Eingabemedium wurde dabei repräsentativ für eine Reihe manueller Eingabegeräte gewählt, die nach ähnlichem Prinzip funktionieren. Abb. 5.1 zeigt neben dieser noch einen Trackball und einen analogen Joystick. Es handelt sich dabei um Eingabegeräte, die über mindestens einen Selektionsknopf verfügen, und die Bewegung der ganzen Hand oder einzelner

Finger in vier primäre Richtungen sowie in vier sich aus Superposition ergebende sekundäre Richtungen umsetzen.

Abb. 5.1: Beispiele manueller Eingabegeräte, von links: Maus, Trackball, analoger Joystick

5.2 Extraktion der Merkmale

Die Mausbewegungen werden als Hintergrundprozess kontinuierlich in Bildschirmkoordinaten umgesetzt. Entscheidend dabei ist in erster Linie die aktuelle Bildschirmauflösung, die bei der Erfassung maßgebend für die Registrierung feiner Bewegungsunterschiede ist. Auch die zeitliche Auflösung kann unter starker Systemlast limitierender Faktor für die Präzision werden[165]. Entsprechend hier getroffener Beobachtungen beeinflusst der unterschiedliche Einsatz optischer oder mechanischer Mäuse die Genauigkeit bezüglich zeitlicher und örtlicher Genauigkeit gering. Mechanische Mäuse mit Mikroschaltern und rotierender Kugel zur Bewegungserfassung tendieren jedoch eher zu Fehlern bei der Umsetzung der tatsächlichen Bewegung des Anwenders, bedingt durch die Oberflächengüte der Mausunterlage, oder durch Verunreinigungen innerhalb der Maus. Optische Mäuse indes sind hauptsächlich gegenüber spiegelnden Auflagen empfindlich. Im Falle stärkerer Reflektion ist jedoch eine Benutzung nicht mehr sinnvoll gewährleistet, weswegen dieser Fall von weiterer Betrachtung ausgeschlossen wird.

5.2.1 Segmentierung

Zur Verarbeitung von Mausinteraktionen wird als erstes eine Segmentierung einzelner Bedienschritte, ähnlich der Sprachsegmentierung (vgl. Kap. 3.2.1), vollzogen. Eine grobe Segmentierung unterscheidet zunächst zwischen Aktionen bei gedrückt gehaltener Maustaste und solchen ohne deren Benutzung. Das Bewegungssegment bei gedrückter Maustaste beginnt mit dem Andruck der jeweiligen Taste und endet mit Loslassen dieser Taste. Als Segmentgrenze dient auch der Wechsel zwischen diversen Tasten einer Maus. Diffiziler gestaltet sich die Unterteilung von Aktionen ohne Tastendruck. Bei kurzen Hesitationen während einer Bewegung soll diese nicht unterteilt werden. Es wird daher ein maximaler Zeitwert ohne Bewegung festgelegt, dessen Überschreitung zur Festlegung des Segmentendpunktes führt. Bei erneuter Positionsänderung wird dann ein neuer Startpunkt registriert.

[165] Weitere Abhängigkeiten bestehen bezüglich des verwendeten Maustreibers und Betriebssystems.

5.2.2 Extraktion der Basiskonturen

Um aus der Mausbewegung geeignete Merkmale für die Klassifikation extrahieren zu können, werden zunächst zwei Konturen abgeleitet:

- Der **Ortdeltaverlauf** beschreibt in Pixel die gerichtete Abweichung der Bewegung von der Ideallinie zwischen Startpunkt $\underline{k}_s$ und Endpunkt $\underline{k}_e$. Das kartesische Bildschirmkoordinatensystem mit den orthogonalen Bildschirmkoordinaten k_1 und k_2 sowie dem Ursprung $\underline{0}$ in der Bildmitte, wird zunächst ins Unendliche erweitert, um bei der Berechnung einen Überlauf zu vermeiden. Anschließend erfolgt eine Translation der Bewegungslinie mit dem Vektor $-\underline{k}_s$, so dass der Startpunkt der Mausbewegung im Ursprung $\underline{0}$ zu liegen kommt sowie eine Rotation um den Winkel $-\alpha$, um die Ideallinie mit der Abszisse in Deckung zu bringen, wie in Abb. 5.2 oben abgebildet. Für jeden Punkt $\underline{k} = (k_1, k_2)^T$ der Bewegungslinie ergibt sich somit ein transformierter Punkt $\underline{k}'$:

$$\underline{k}' = \begin{bmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{bmatrix} \cdot (\underline{k} - \underline{k}_s) \text{ mit } \alpha = \arctan \frac{k_{e,2} - k_{s,2}}{k_{e,1} - k_{s,1}} \tag{5.1}$$

  Die so transformierte Kurve ist in Abb. 5.2 unten links dargestellt. Die Auslenkung k_2' entlang der Ordinate entspricht so dem vorzeichenbehafteten Abstand eines Punktes von der Ideallinie. Dieser wird - mit in Bezug auf die Bildschirmauflösung minimal möglicher äquidistanter Abtastung - in den Ortdeltaverlauf χ mit $\chi = (\chi_t)$ und $t = 0,...,T-1$ umgesetzt, so dass jeder neu erfasste unterschiedliche Bildschirmpunkt einem Fortschreiten in der Zeit entspricht. Hierbei handelt es sich um eine Zeitreihe ohne zeitliche Äquidistanz in Bezug auf die tatsächliche Bewegung, da diese in der Regel mit variabler Geschwindigkeit erfolgt. Diesem Aspekt wird durch Integration eines Zeitverlaufs Rechnung getragen.

- Der **Zeitdeltaverlauf** erfasst die verstrichene Zeit zwischen zwei adjazenten Ortspunkten in ms. Ein neuer Datenpunkt wird stets nur dann erfasst, wenn sich eine Änderung in der k_1- oder k_2-Koordinate ergibt. Zu jedem Punkt o_t des Ortdeltaverlaufs korrespondiert somit ein Punkt des Zeitdeltaverlaufs τ mit $\tau = (\tau_t)$ und $t = 0,...,T-1$. Sei $t(\chi_t)$ die absolute Systemzeit zur Erfassung des χ_t zugehörigen Punktes, so gilt:

$$\tau_t = t(\chi_t) - t(\chi_{t-1}) \tag{5.2}$$

  Der erhaltene Verlauf ist in Abb. 5.2 unten rechts exemplarisch gezeigt.

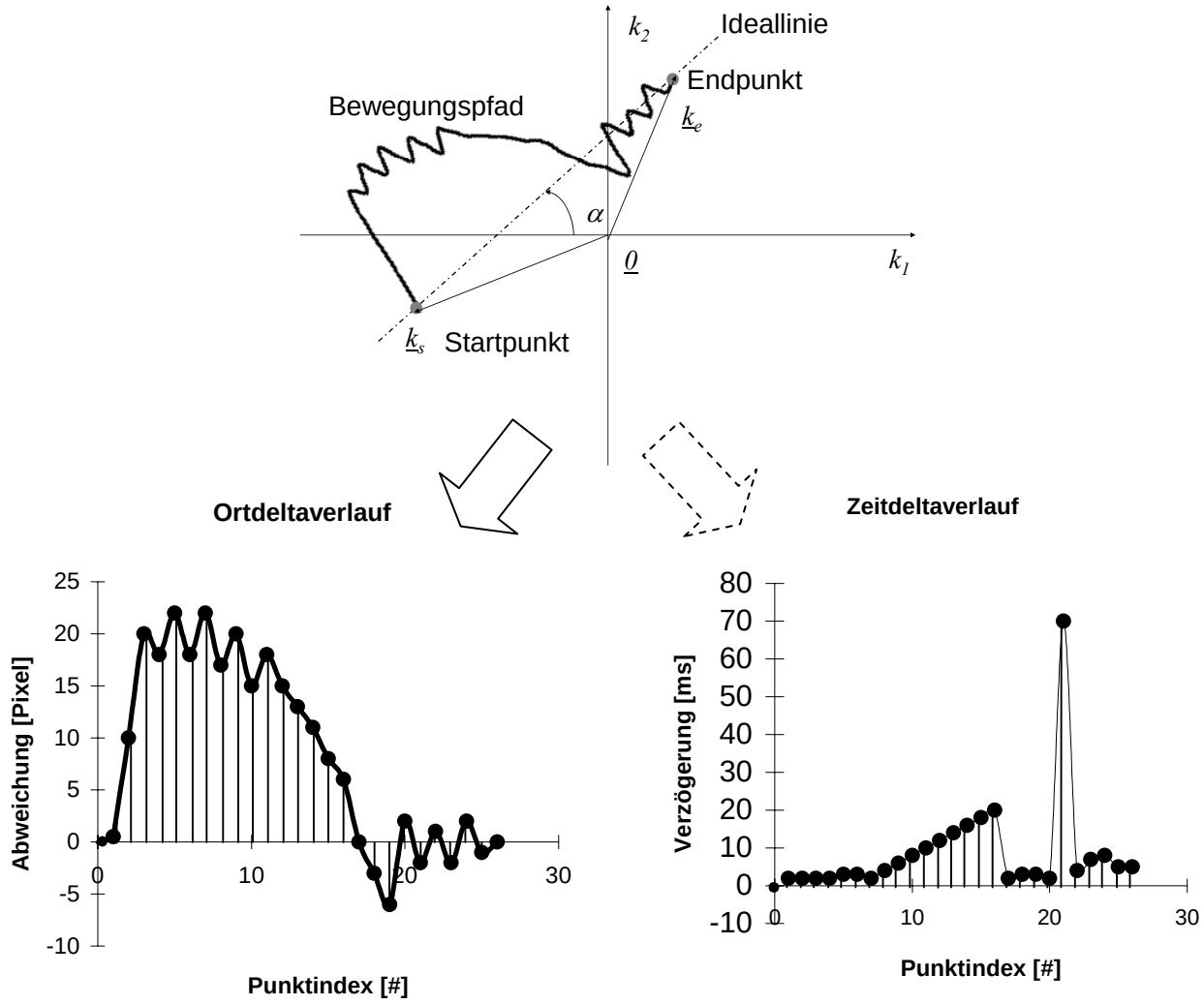


Abb. 5.2: Beispiel Bewegungspfad der Maus und Extraktion der Ort- und Zeitdeltaverläufe

5.2.3 Funktionalbildung

Ähnlich der akustischen Emotionserkennung würde eine direkte Betrachtung der Basiskonturen für die Erkennung des Affekts hier zu stark von der eigentlichen Bewegung abhängen. Um dies zu vermeiden, werden auch hier statistische Analysen über die Verläufe durchgeführt. Aus dem Ortdeltaverlauf werden als Merkmale berechnet:

- Länge der Ideallinie von Start- zu Endpunkt der Bewegung l:

$$l = \left\| \underline{k}_e - \underline{k}_s \right\| \tag{5.3}$$

- Gesamtabweichung von der Ideallinie χ_{ges}:

$$\chi_{ges} = \sum_{t=0}^{T-1} \left| \chi_t \right| \tag{5.4}$$

- Zahl der Nulldurchgänge

- Momente erster und zweiter Ordnung des betragsmäßigen Ortdeltaverlaufs
- Maximum des betragsmäßigen Ortdeltaverlaufs

Für den Zeitdeltaverlauf werden analog gebildet:

- Gesamtzeit der Bewegung τ_{ges}:

$$\tau_{ges} = \sum_{t=0}^{T-1} \tau_t \tag{5.5}$$

- Momente erster und zweiter Ordnung des betragsmäßigen Zeitdeltaverlaufs

Des Weiteren werden die Funktionale Extrema und Momente erster und zweiter Ordnung aus folgenden jeweils abgeleiteten Verläufen gebildet:

- Autokorrelationsfunktion
- Ableitungen erster und zweiter Ordnung
- Verteilungsfunktion

Ebenfalls ähnlich wie bei den akustischen Merkmalen (vgl. Kap. 3) ist auch hier das Ziel zunächst eine möglichst große Basis an Merkmalen zu schaffen, die prinzipiell geeignet erscheinen den emotionalen Benutzerzustand erkennen zu lassen. Die endgültige Wahl finaler Merkmale erfolgt analog mit Verfahren der Merkmalsselektion (vgl. Kap. 3.9).

5.3 Verarbeitung berührsensitiver Eingabe

In Abb. 5.3 sind exemplarisch Geräte gezeigt, bei denen eine berührsensitive Bedienung erfolgt. Die Bedienung kann dabei prinzipiell direkt mit einem Finger oder einem geeigneten Eingabegerät wie einem Stift erfolgen.

Abb. 5.3: Beispiele berührsensitiver Eingabegeräte, von links: PDA, Tablet PC, Infostand

Im Folgenden wird die Bedienung und Funktionsweise eines Touchscreens stellvertretend für weitere berührsensitive Geräte betrachtet.

5.3.1 Basistechnologie

Auch zur Erfassung berührsensitiver Eingaben per Hand direkt, oder mittels eines Stiftes, existiert eine Vielzahl von technischen Lösungen. In erster Linie sind dies optische, resistive, kapazitive, piezoelektrische und akustische Touchscreens oder –pads. Dabei sind bei einem Touchscreen im Gegensatz zu einem -Pad Eingabe- und Ausgabebereich in unmittelbare Deckung gebracht. Speziell akustische Oberflächenwellen und piezoelektrische Verfahren erlauben über die Aufnahme von parallel zur Eingabeebene verlaufenden Bewegungen hinaus auch die solcher senkrecht zu dieser. Diese Eigenschaft soll hier mit betrachtet werden. Im Folgenden wird auf Grund der hohen Verbreitung in öffentlichen Einrichtungen und der hohen Lichtdurchlässigkeit speziell ein akustischer Oberflächenwellen Touchscreen[166] gewählt und näher beschrieben: Signalgeber in den vier Bildschirmecken und Reflektoren an den seitlichen Rändern des Bildschirms erzeugen eine möglichst einheitliche Dichte von Wellenenergie. Hierzu werden elektrische Impulse mit der Frequenz 5 MHz durch Wandler in parallel zur Bildschirmoberfläche verlaufende akustische Oberflächenwellen umgesetzt, die sich mit 3 m/ms ausbreiten, und von Arrays aus piezoelektrischen Wandlern, die ebenfalls am Bildschirmrand positioniert sind, erfasst werden. Durch Aufdruck des Fingers wird Wellenenergie an der Berührstelle absorbiert, und die Position letzterer in k_1- und k_2-Richtung parallel zu den Bildschirmrändern (siehe Abb. 5.2) kann direkt durch die piezoelektrischen Wandler bestimmt werden. Zusätzlich kann durch den Grad absorbierter Energie die Andruckstärke in k_3-Richtung senkrecht zur Oberfläche des Schirms ermittelt werden. Dies wird durch die Tatsache ermöglicht, dass die Fingerspitze durch stärkeren Aufdruck ebenfalls stärker deformiert wird, und die Andruckfläche sich vergrößert. Die erfasste Intensität wird hier in 8 Bit kodiert, wodurch 256 Werte aufgelöst werden. Beim Einsatz eines anderen Selektionsmediums wie eines Stiftes, welcher grundsätzlich möglich ist, ist in der Regel die k_3-Koordinate nicht mehr sinnvoll messbar.

Prinzipiell kann diese Technologie auf diversen Oberflächen angebracht werden. Bei der Anwendung über einem Bildschirm ist eine Kalibrierung erforderlich, um Anzeige und Bedienfläche in Übereinstimmung zu bringen. Im Folgenden wird ein 12“-TFT-Monitor für die Anzeige verwendet. Die Auflösung in k_1- und k_2-Richtung wird hier entsprechend der dargestellten Auflösung angepasst.

5.3.2 Merkmalsextraktion

Die Segmentierung, wie in Kap. 5.2.1 beschrieben, wird bei der Erfassung berührsensitiver Interaktion im Vergleich zur Mausinteraktion wesentlich vereinfacht, da es keine Unterscheidung zwischen Aktionen bei gedrückter oder nicht gedrückter Maustaste gibt. Der Bewegungsanfang erfolgt mit Aufsetzen des Fingers oder Stifts und endet mit dem Abheben desselbigen. Hierzu werden die Werte der Andruckstärke $\zeta = (\zeta_t)$, mit $t = 0,...,T-1$, zwischen Null ohne Berührung, beziehungsweise Eins beim leichtesten noch erfassbaren Andruck, und 255 beim härtesten messbaren Andruck, gemessen. Während eines Interaktionsschritts muss dann $\zeta_t > 0$ gelten.

---

[166] Bei dem eingesetzten Touchscreen handelt es sich um den Typ Intellitouch 2500s der Firma Elo TouchSystems Inc.

Aus dem Bewegungsverlauf parallel zu den Bildschirmrändern werden dieselben Merkmale wie bei der Mausinteraktion abgeleitet. Als zusätzliche Größen werden bei Berührung durch einen Finger Attribute aus der Andruckstärke ζ_t durch systematische Bildung abgeleiteter Verläufe und Funktionale und anschließender Selektion (vgl. Kap. 3 und Kap. 5.2) betrachtet. Basiskonturen sind hier neben der Andruckstärke selbst ihre Geschwindigkeit und Beschleunigung sowie die jeweils zugehörige AKF. Aus diesen werden entsprechend dem bisherigen Vorgehen Momente, Extrema und die Zahl der Vorzeichenwechsel berechnet. Insgesamt ergeben sich so 220 Merkmale als Ausgangsbasis.

5.4 Experimente und Ergebnisse

Zur Eignungsprüfung der vorgestellten Merkmale aus der manuellen Interaktion bezüglich der Erkennung von Emotion werden rechnerische Tests, aber auch eine Benutzerstudie durchgeführt.

5.4.1 Verwendete Datenbanken

Um realistische Daten der manuellen Interaktion mit Maus und Touchscreen in ausreichender Quantität zu erhalten, wurden insgesamt zehn Probanden über einen Zeitraum von vier Wochen mit einer Versuchssoftware ausgestattet. Diese erlaubt ein kontinuierliches Erfassen der Interaktionen im Hintergrund zur laufenden Applikation. Die Software wurde auf den persönlichen Arbeitsplatzrechnern der Versuchspersonen installiert[167]. Die Probanden bedienten ihr System zunächst mit der Maus, im Anschluss dann mittels Touchscreen und waren im Alter von 22 a bis 32 a, zwei davon weiblich. Bei der Bedienung mit der Maus[168] wurde eine Bildschirmauflösung von 1.280 x 1.024 Pixel gewählt. Auf Grund technisch notwendiger Limitationen wurde diese bei der berührsensitiven Steuerung auf 800 x 600 Pixel reduziert.

Die für den Versuch erstellte Software zeichnet jede Maus- oder Touchscreeninteraktion auf. Damit diese auch mit der Emotion in Verbindung gebracht werden kann, wird der Anwender angeregt, seine Interaktion selbst zu bewerten. Hierzu existieren zwei unterschiedliche Formen: der Benutzer kann zunächst direkt auf ein Frusticon klicken. Dieses ist aus Motivationsgründen animiert gestaltet, und reagiert alternierend auf ein Anklicken. Bei Anwahl dieses Icons wird die letzte Interaktion entsprechend gekennzeichnet gespeichert. Zusätzlich fragt die Testsoftware zu zufälligen Zeitpunkten nach, inwiefern ein automatisch erkannter Gefühlszustand richtig ist. Die Emotionen sind dabei - angepasst zur geringen Komplexität des Eingabemediums - folgendermaßen gewählt: Ärger, Irritation, Reflexion und Neutralität als abgrenzender Zustand. Eine genauere Beschreibung dieser affektiven Zustände ist in Kap. 2.3.2 zu finden. Der Benutzer kann die Frage entweder beantworten oder ignorieren, falls er sich in seinem Arbeitsablauf gestört fühlt. Im Falle der Übereinstimmung wird im Anschluss die jeweils letzte Interaktion vor Beantwortung der Frage als annotiertes Datum gespeichert, und das Urteil selbst zu Bewertungszwecken festgehalten. Die Nachfragen sind dabei variabel gestaltet, jedoch eindeutig. Der Hintergrund ist auch hier die Absicht eine höhere Langzeitmotivation zur häufigen Nutzung der Applikation zu erzeugen. Damit

[167] Diese liefen jeweils unter dem Betriebssystem Microsoft® Windows 98®.

[168] Bei den eingesetzten Mäusen handelte es sich um 2-Button optische *Wheel*-Mäuse der Firma Logitech®.

das System eine Einschätzung der aktuellen Emotion zu Beginn der Anwendung leisten kann, wird eingangs eine benutzerunabhängige Erkennung mit einem Datensatz aus 300 Beispielen des Versuchsleiters vollzogen. Diese initialen Daten werden im Anschluss sukzessive durch benutzerspezifische Daten bei konstanter Gesamtzahl ausgetauscht, bis 300 Daten des aktuellen Anwenders in stratifizierter Verteilung vorliegen. Ab diesem Zeitpunkt beginnt die Kollektion der Daten für die nachfolgende Evaluierung. Folgende Abb. 5.4 zeigt die mittlere gemessene Übereinstimmung der Probanden während der beschriebenen Adaptionsphase.

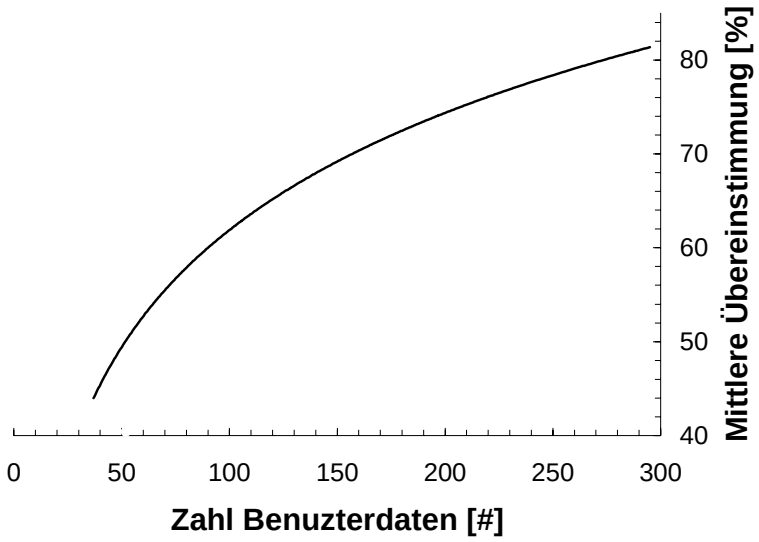


Abb. 5.4: Mittlere Übereinstimmung der Benutzer im Adaptionsprozess bei der Emotionserkennung aus der manuellen Bedienung

Die Übereinstimmung liegt anfangs im Mittel schon über der zugehörigen Ratewahrscheinlichkeit von 25% bei vier Klassen und reicht von 30% bis 68% je nach Proband. Bei vollständig vollzogener Adaption steigt sie auf mindestens 70% für jeden einzelnen Anwender, und über 80% im Mittel über alle Anwender. Dieser Wert liegt auch über dem Wert von 54,6% einer steten Entscheidung zu Gunsten der häufigsten Emotion Neutralität.

Ein Vorteil dieser Vorgehensweise ist, dass der Benutzer sich an den Umgang mit dem System gewöhnt hat, und die Daten vom Anwender in der Annotation selbst bestätigt sind. Diskutabel erscheint hingegen, inwiefern ein Proband zu leicht dazu neigt dem System recht zu geben. Dieser Aspekt wird jedoch durch die im anschließenden Kap. 5.4.2 gezeigte signifikante Trennbarkeit der gesammelten Daten widerlegt. Tab. 5.5 zeigt die Verteilung dieser Daten nach Emotion. Die gewählten Bezeichnungen für die zugehörigen Korpora sind *EM-MAUS*[169] für die Emotionserkennung aus manueller Mausinteraktion und *EM-TS* entsprechend für Daten der Touchscreeninteraktion.

[169] Entsprechend den gewählten Abkürzungen in den vorherigen Betrachtungen aus Kap. 3.12.1 und Kap. 4.7.1 steht EM hier für *Emotionserkennung manuelle Interaktion*.

Anzahl [#]	Neutralität	Irritation	Ärger	Reflexion	Summe
EM-MAUS	1.085	165	529	209	**1.988**
EM-TS	193	210	240	184	**827**

Tab. 5.5: Verteilung in den Datenbanken EM-MAUS und EM-TS

5.4.2 Evaluierung der Bewegungsanalyse

Tab. 5.6 zeigt einen Vergleich von Klassifikatoren auf der gesamten Datenbank EM-MAUS in einer 10-fach SCV. Im Gegensatz zur akustischen Emotionserkennung kann hier durch StackingC ein Gewinn gegenüber den Basisklassifikatoren auch bei Einsatz von SVM erzielt werden. Somit wird hier mit Ensembletechniken die insgesamt höchste Akkuratheit erzielt. Bei Einsatz eines einzigen Basisklassifikators ist die Leistung bei MultiBoosting eines C4.5 höher als die von SVM.

Klassifikator	Akkuratheit [%]
Raten	25,0
Häufigste Klasse	54,6
1NN	86,2
kNN	86,2
MLP	85,8
DT C4.5	85,3
AdaBoosting C4.5	89,4
Bagging C4.5	88,2
MultiBoosting C4.5	89,4
NB	32,9
BN	77,8
SVM 1-vs-1	86,7
StackingC MLR NB 1NN SVM C4.5	88,4
Voting NB 1NN SVM C4.5	87,1
StackingC MLR BN kNN SVM MB C4.5	**89,9**

Tab. 5.6: Erkennungsleistung mit diversen Verfahren, Datenbank EM-MAUS, alle Merkmale, 10-fach SCV

Tab. 5.7 zeigt erzielte Resultate für die Datenbank EM-MAUS in einer gemittelten 10-fach SCV nach Proband sowie eines LOSO-Tests[170].

[170] Die Bezeichnung LOSO wurde hier in Analogie an die Sprachverarbeitung beibehalten. Es handelt sich hierbei zwar nicht um Sprecher, sondern um Anwender. Eine entsprechende, exakte Bezeichnung existiert aber noch nicht.

Akkuratheit [%]	μ
10-fach SCV, personenbezogen	**90,3**
LOSO	**70,5**

Tab. 5.7: Erkennungsleistung Datenbank EM-MAUS , Merkmale Top 49 SVM-SFFS, Klassifikator MultiBoosting C4.5, 10-fach SCV je Nutzer und LOSO

Die Erkennungsleistung erscheint unter Berücksichtigung der Interaktionsform erstaunlich hoch. Es zeigt sich jedoch auch hier (vgl. Kap.3.12.4) eine deutliche Einbuße für eine anwenderunabhängige Erkennung.

Tab. 5.8 zeigt die Konfusionen der Benutzeremotionen. Es lässt sich beobachten, dass Neutralität relativ sicher abgegrenzt werden kann, und Reflexion sich am schlechtesten erkennen lässt.

[%] Wahr Prädiktion	**Ärger**	**Irritation**	**Neutralität**	**Reflexion**
Ärger	**89,9**	2,4	3,6	5,7
Irritation	0,4	**84,3**	0,7	1,0
Neutralität	9,5	13,3	**93,8**	22,0
Reflexion	0,2	0,0	1,9	**71,3**

Tab. 5.8: Konfusion Emotionen, Datenbank EM-MAUS, Merkmale Top 49 SVM-SFFS, Klassifikator Boosting C4.5, 10-fach SCV**,** 84,9% durchschnittliche Leistung

Bei der Erkennung auf dem Touchscreen ergeben sich unter gleichen Bedingungen im Mittel 70,1% Erkennungsleistung für die Datenbank EM-TS. Diese deutlich geringere Akkuratheit ist unter Anderem darauf zurückzuführen, dass Interaktionen am Touchscreen stets solchen mit gedrückter Maustaste entsprechen. Somit entfällt das Bewegen des Mauspfeils zu einem Zielobjekt ebenso wie Positionsänderungen ohne eigentlichen Sinn. In Summe ist die Bedienung mittels Touchscreen also zielgerichteter. Als wichtigste Merkmalsgruppe hat sich hier die Information der Andruckstärke gezeigt, gefolgt von dem Zeitdelta- und schließlich dem Ortdeltaverlauf. Eine Transformation von kartesischen Koordinaten zu Kugelkoordinaten erbrachte bei durchgeführten Versuchen keine Verbesserung der Leistung.

6

Integration und Adaption

„Das Ganze ist mehr als die Summe seiner Teile.“

ARISTOTELES (384 - 322 V. CHR.)

Ziel ist es im Folgenden die Information multipler Quellen synergetisch zu vereinen. Dabei wird sowohl die unterschiedliche Analyse desselben Kanals, als auch die Fusion aus mehreren Informationskanälen betrachtet. Schließlich kann durch eine Adaption an den aktuellen Benutzer eine Verbesserung der Leistung erzielt werden.

6.1 Integration akustischer und linguistischer Information

Als einfachste Variante bietet sich zur Vereinigung von Informationsströmen eine späte semantische Fusion auf Entscheidungsebene, *Late-Semantic-Fusion* genannt, an. In einer strengen Definition dieser Methodik wird während des Fusionsprozesses nur die jeweils erkannte Klasse jeder Instanz, ohne Berücksichtigung der Ausgangssicherheiten für die Sieger- oder konkurrierenden Klassen, berücksichtigt. Durch eine solche späte Entscheidung lässt sich ein hoher Grad an Modularisierbarkeit hinsichtlich Austauschbarkeit, Erweiterbarkeit, Plattformunabhängigkeit und in gewissem Maße auch Unabhängigkeit bezüglich der Darstellungsform der Information erreichen, da nur jeweils die Siegerklasse zum aktuellen Zeitpunkt an eine integrative Schicht weitergegeben werden muss. Zur eigentlichen Fusion bieten sich dann etwa logische UND- und ODER-Verknüpfungen oder mehrheitsbasierte Entscheidungen an [LEE02A]. Abb. 6.1 veranschaulicht dieses Prinzip für die Integration akustischer und linguistischer Information.

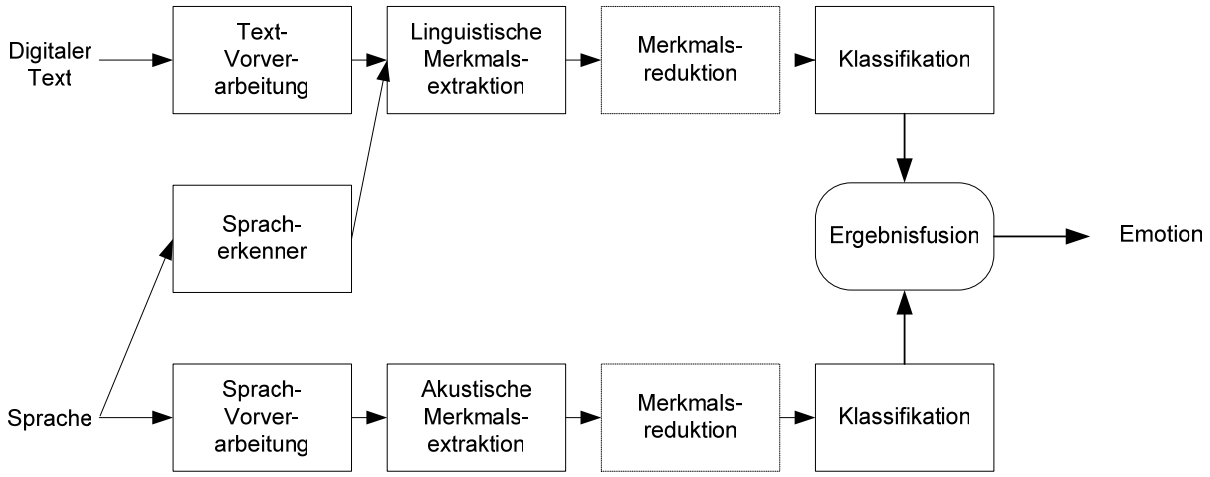


Abb. 6.1: Blockdarstellung semantische Fusion

Um eine hohe Güte zu erzielen, sollte jedoch möglichst alle Information für den finalen Entscheidungsprozess verfügbar sein. Hierin liegt die Hauptschwäche einer späten Fusion. Hinsichtlich optimaler Gesamtleistung ist daher eine Vereinigung auf Merkmalsebene vor der Klassifikation, als *Early-Feature-Fusion* bekannt, vorzuziehen. Letztere birgt jedoch oft Probleme bei der Umsetzbarkeit, da eine Synchronisation und Vereinheitlichung der Merkmalsströme obligatorisch ist. Darüber hinaus wird die Dimensionalität weiter erhöht und eine gesteigerte Suszeptibilität gegenüber Rauschen beobachtet - es kann jedoch eine gemeinsame Merkmalsselektion gewählt werden, mit deren Hilfe die Relevanz der Merkmale gesamtheitlich beurteilt werden kann. Abb. 6.2 zeigt hierzu den Ablauf.

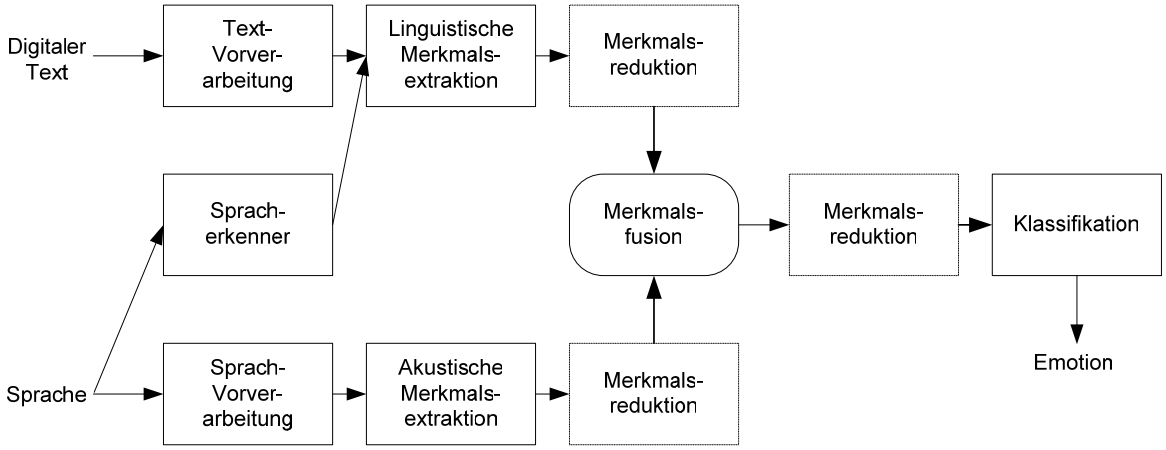


Abb. 6.2: Blockdarstellung der Merkmalsfusion mit Merkmalsreduktion optional unter Ausnutzung merkmalsspezifischer Charakteristika vor, oder gesamtheitlich nach der Fusion der Attribute

Diese Integration von akustischen mit linguistischen Merkmalen ist hier nur bei der Anwendung der vorgeschlagenen Bag-Of-Words Repräsentation von Text möglich (vgl. [SCH05F]). Alternativ können Ausgangssicherheiten je Klasse einer linguistischen Analyse, mittels statistischem Sprachmodell oder grafischer Modellierung, als Attribute in den akustischen Merkmalsvektor aufgenommen werden (vgl. [SCH04A]). Dabei handelt es sich jedoch nicht um eine frühe Vereinigung im strengeren Sinne.

Grundsätzlich besteht bei der linguistischen Analyse schon eine gewisse Distanz zum Signal im Vergleich zu einer akustischen, da bereits eine Verarbeitung zu Text der Merkmale durch die zugrunde liegenden Spracherkennung, hier MFCC, vorausgeht.

Als Kompromisslösung zwischen Durchführbarkeit und Leistungsoptimierung bietet sich auch eine weiche Integration auf Entscheidungsebene, unter Berücksichtigung möglichst aller Ausgangsunsicherheiten sämtlicher Klassen, an. Diese Vorgehensweise wird oft als *Soft-Decision-Fusion* bezeichnet. Vorraussetzung ist die meist nicht triviale Berechenbarkeit der Konfidenzwerte. Als besonderen Vorteil können diese dafür, wie auch bei der Early-Feature-Fusion, in einem geschickten Vereinigungsprozess die Bereitstellung von Ausgangssicherheiten auf Fusionsebene ermöglichen. Der Übergang von der strikten Late-Semantic-Fusion zur idealen Soft-Decision-Fusion gestaltet sich dabei fließend. Als Möglichkeiten einer weichen Integration sind in erster Linie gewichtete Summation oder Multiplikation der Klassenwahrscheinlichkeiten sowie der Einsatz eines geeigneten nachgestellten Klassifikators[171] zu nennen. Die Verwendung eines Lernverfahrens auf Fusionsebene erlaubt das Lernen von Schwächen und Stärken bei der Entscheidung zu Gunsten einzelner Klassen auf Grund einzelner Signalströme. Das Training eines solchen Entscheidungsalgorithmus erfordert dabei ein disjunktes Trainingsset für die jeweiligen Instanzen, was den Gesamtbedarf an Datenmaterial steigert. Der Ablauf erfolgt dann ähnlich wie in Abb. 6.1 dargestellt, mit zusätzlicher Berücksichtigung von Konfidenzwerten bei der Ergebnisfusion.

Auf Grund der genannten Vorteile wird hier eine Fusion auf Merkmalsebene bevorzugt. Es werden somit logTF-Merkmale der BOW-Darstellung in den akustischen Merkmalsvektor aufgenommen. Tab. 6.3 zeigt hierzu erzielte Ergebnisse einer 10-fach SCV mit SVM jeweils mit nur akustischen, linguistischen und fusionierten Merkmalen auf den beiden in Kap. 3.12.1 und 4.7.1 vorgestellten Datenbanken AEC und EAL-F+W.

Akkuratheit [%]	**AEC**	**EAL-F+W**
Akustik, exklusiv	79,1	90,3
Linguistik, exklusiv	79,5	39,6
Akustik + Linguistik	**87,3**	**94,8**

Tab. 6.3: Erkennungsleistung Akustik, Linguistik und Fusion, Klassifikator SVM, 10-fach SCV

Für beide Datenbanken ergibt sich eine signifikante Verbesserung der mittleren Gesamtleistung durch die Integration beider Informationsquellen gegenüber der Verwendung im einzelnen.

Tab. 6.4 zeigt ergänzend, für die Datenbank AEC, die Erkennungsraten nach Emotion aufgeschlüsselt. Zu sehen ist, dass einzig die Bemutterung durch die Fusion schlechter erkannt wurde, was jedoch durch eine einfache regelbasierte Selektion des optimalen Signalstroms gelöst werden kann. Weitere Ergebnisse zu diesem Thema auf einer deutschen Datenbank finden sich in [SCH04A].

[171] Oft als *Post-Classifier* bezeichnet.

Akkuratheit [%]	Ärger	Bemutterung	Betonung	Neutralität
Akustik, exklusiv	82,4	70,3	82,6	66,9
Linguistik, exklusiv	81,6	71,9	74,1	82,0
Akustik + Linguistik	**90,8**	**70,2**	**83,3**	**89,5**

Tab. 6.4: Erkennungsrate nach Emotion, Datenbank AEC, Klassifikation mit SVM, 10-fach SCV

6.2 Multimodale Integration

Multimodalität wird hier im Sinne einer synergetischen Ausnutzung multipler menschlicher Kommunikationskanäle verstanden. Es soll dabei der Umstand ausgenutzt werden, dass der Mensch seine aktuelle Emotion gleichzeitig auf mehreren Kanälen zum Ausdruck bringt. In dieser Arbeit beschränkt sich die Integration auf die manuelle und sprachliche Interaktion. Das Prinzip der multimodalen Modellbildung kann jedoch sinnvoll um weitere Modalitäten wie die Verarbeitung visueller Information ergänzt werden [OVI00B].

Zur multimodalen Integration der Analyse sprachlicher und manueller Interaktion wird hier die in Kap. 6.1 vorgestellte Soft-Decision-Fusion gewählt. Sie umgeht die Probleme die die Synchronisation begleiten und erlaubt eine adäquate modulare Modellierung jeder Modalität hinsichtlich Merkmalsstruktur und optimalem Klassifikator. Dies ist im Besonderen bei der Fusion von Information verschiedener Modalitäten relevant. Auf Grund der genanten Vorteile wird hier ein übergeordneter Klassifikator eingesetzt, der in der Lage ist zu lernen welcher Modalität bezüglich welcher erkannten Emotion besonders zu vertrauen ist, und Ausgangssicherheiten verarbeiten kann. Da nicht zu jeder Zeit auf allen Kanälen Information anliegt, muss das Lernmodell zusätzlich neben unsicherem Wissen auch unvollständiges verarbeiten können. Aus der Reihe in Frage kommender Verfahren, wie C4.5 und BN, wurden hier letztere gewählt, da das mittels BN automatisch gelernte Wissen mathematisch elegant mit Expertenwissen vereint werden kann. Dies erweist sich als notwendig, da Trainingsmaterial zur multimodalen Emotionserkennung schwerer erhältlich ist als solches zur monomodalen.

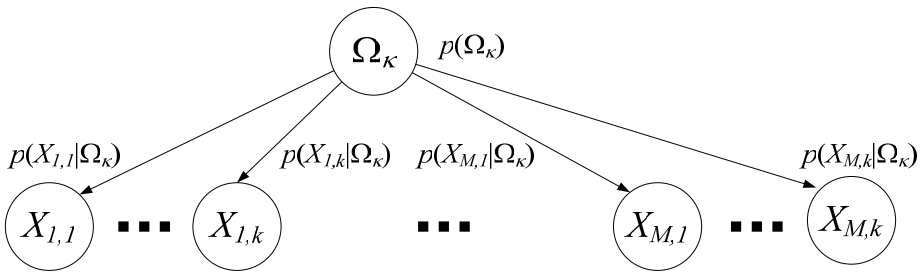


Abb. 6.5: Multimodale Fusion mittels BN

Für die Vereinigung der Information zur Emotionserkennung ist eine singulär verbundene Struktur in zwei Ebenen für das BN ausreichend[172]. Zu jeder Emotion existiert ein eigenes Netz, und es wird

[172] In Kap. 7.2.3 wird dieses Prinzip auf mehrere Ebenen ausgeweitet.

in der Erkennungsphase dasjenige mit höchster Wurzelwahrscheinlichkeit gewählt. Abb. 6.5 zeigt die gewählte Netztopologie. Zu jeder Modalität mit Index m, mit $m = 1,...,M$, existiert ein Kindknoten je Emotion Ω_κ, mit $\kappa = 1,...,k$, und die zugehörigen Zufallsvariablen $X_{m,\kappa}$. Die ebenfalls zugehörige Konfidenz wird jeweils als weiche Evidenz direkt in das Netz eingebracht.

Um eine Distribution der Erkennerinstanzen auf unterschiedliche Rechner zu erlauben, wird auch hier, wie in Kap. 2.5.2, ein einfaches Kommunikationsprotokoll[173] über TCP/IP verwendet.

6.3 Sprechererkennung zur automatischen Adaption

In Kap. 3.12 wurde gezeigt, dass Emotionserkennung aus akustischen Merkmalen mit sprecherspezifischen Modellen eine höhere Erkennungsleistungsfähigkeit als eine sprecherunabhängige aufweist. Dabei erhöht sich die Leistung sowohl wenn der Sprecher mit im Trainingsmaterial enthalten ist, als aber auch dann, wenn sich dieses auf den Sprecher konzentriert, wie schon für die Datenbank DES gezeigt. Dieses Verhalten ist auch in der automatischen Spracherkennung bekannt. Während dort versucht wird akustische Modelle zur Laufzeit an den aktuellen Sprecher zu adaptieren, erweist sich dies hier ungleich schwieriger, wenn nicht unmöglich. Grund hierfür ist der deutlich größere erforderliche Zeitraum zur Sammlung ausreichenden Adaptionsmaterials für alle Klassen. Bestehen jedoch bereits sprecherspezifische Modelle, muss der aktuelle Sprecher bekannt sein, da die Verwendung eines falschen Sprechermodells geringere Präzision in der Klassifikation erbringt, als der Einsatz eines allgemeinen Modells. Findet die Emotionserkennung speziell im offenen Mikrophonbetrieb oder aus einem kontinuierlichen Signalstrom mit Fremdsprechern statt, oder muss die Verarbeitung zeitnah für mehrere Sprecher erfolgen, erfordert dies in der Konsequenz ebenfalls eine automatische Erkennung des Sprechers vor der eigentlichen Zuordnung des Affekts zur zeitrichtigen Bereitstellung adäquater Modelle. Ziel dieses Abschnitts ist es somit, eine Sprecheridentifikation auf Basis akustischer Merkmale umzusetzen. Neben der Hilfe zur Benutzeradaption wird dieser im Einsatz mit offenem Mikrophon auch die Aufgabe der Verifikation des Sprechers zugeteilt. Hierzu wird ein Beispiel aus dem später in Kap. 7.3 beschriebenen Einsatz im Fahrzeug angeführt: Ein verunsichert oder ermüdet klingender Beifahrer könnte akustisch zur Einleitung von Sicherheitsmaßnahmen beitragen, während die Überwachung des Fahrers permanent erfolgt. Die Zuordnung einer Äußerung zu einem Sprecher ist somit unerlässlich, um Fehlern vorzubeugen.

Die Erkennung sollte für den realen Einsatz in Echtzeit und unabhängig vom gesprochenen Inhalt erfolgen, und muss auch unabhängig von der zugrunde liegenden Emotion möglich sein. Dieser Aspekt wird in der allgemeinen Forschung zur Sprechererkennung in der Regel nicht betrachtet [MAR04], da nur neutrale Aufzeichnungen für Test und Training angewandt werden. Die Verzerrungen und Veränderungen der Stimme, die unter ausgeprägt emotionaler Sprechweise auftreten, stellen jedoch eine deutliche Herausforderung an die Robustheit einer Sprechererkennung [KLA00B]. Als besondere Erschwernis erweist sich, dass die emotionalen Äußerungen oft nur kurze Laute von ca. einer Sekunde sind. In der klassischen Sprechererkennung hingegen bewegen sich die

[173] Dieses beinhaltet in erster Linie die Angaben Quelle, Senke, *N* beste Erkennungsergebnisse und Konfidenzbewertungen.

Testsamples im Bereich von 15-30s.

Es werden als Ansatz die 276, in Kap. 3 beschriebenen, akustischen Merkmale für die Identifikation des Sprechers verwendet. Dies bringt den besonderen Vorteil, dass nur eine gemeinschaftliche Merkmalsextraktion für Emotions- und Sprechererkennung erfolgen muss. Da nicht alle Größen hierzu geeignet erscheinen, wird eine Auswahl mittels IGR-FS und alternativ SVM-SFFS getroffen. Zum Vergleich wird auch eine Analyse mit PCA-FS vollzogen.

Für die Evaluation der Sprechererkennung wird hier beispielhaft die öffentliche Wall Street Journal-Datenbank WSJ1 verwendet [PAU92] (vgl. Kap. 3.12.1). In ihr sind von Journalisten diktierte Sätze enthalten. Die Daten entsprechen den Sets SI_ET_H2 mit zehn muttersprachlichen und SI_ET_S3 mit zehn nicht muttersprachlichen Sprechern. In jedem Set ist die Hälfte der Sprecher weiblich. Die Länge der Sprachsamples beläuft sich von 2 s mit zwei Wörtern bis zu 16 s. Die Zahl der Samples pro Sprecher beläuft sich von 61 bis 82. Insgesamt werden 1.430 Äußerungen betrachtet. Tab. 6.6 zeigt hierzu einen Vergleich unterschiedlicher Klassifikatoren mit den erzielten Erkennungsraten.

Klassifikator	Akkuratheit [%]
ND	95,5
C4.5	88,5
NB	97,3
SVM	100,0
Boosting C4.5	96,7

Tab. 6.6: Erkennungsleistung bei der Sprechererkennung, Datenbank WSJ, diverse Klassifikatoren, 3-fach SCV

Tab. 6.7 zeigt die Ergebnisse nach der Merkmalsreduktion.

Dimension *x*	IGR-FS [%]	SVM- SFFS [%]	PCA-FS [%]
276	100,0	100,0	100,0
80	100,0	99,8	100,0
60	99,6	99,4	100,0
40	99,4	98,8	99,4
20	96,9	97,3	98,4

Tab. 6.7: Erkennungsleistung bei Merkmalsreduktion, Datenbank WSJ, SVM, 3-fach SCV

Das hier vorgestellte Verfahren wird in [WAF05A] mit einem alternativen Ansatz sowie konventioneller Sprechererkennung mittels Kurzzeit-MFCC verglichen und zeigt sich dabei als überlegen.

Im Vergleich hierzu wird auch die Datenbank EA-ACT betrachtet: Auf dieser werden in einer 10-fach SCV mit SVM und dem kompletten Merkmalsvektor 95,9% Identifikationsleistung erreicht. Bei Reduktion der Sprecherzahl von 39 auf 20, wie bei dem Testset aus der Datenbank WSJ, erhöht sich diese Leistung auf 99,2%. Reduziert man die Zahl weiter auf zehn oder vier Sprecher, wie bei

den Datenbanken EMO-DB oder DES, ist eine fehlerfreie Zuordnung möglich.

Existiert also ein emotionales Sprechermodell, kann dieses nahezu zuverlässig ausgewählt werden und die Erkennungsrate der Emotion entsprechend gesteigert werden. Für die Datenbank DES kann so beispielsweise in einer 10-fach SCV mit vorhergehender Sprecheradaption, bei vollem Merkmalsset und SVM als Klassifikator, eine mittlere Erkennungsleistung von 69,7% gegenüber 65,9% ohne dieselbe erreicht werden. Noch deutlicher wird dieser Vorteil, wenn neben sprecherspezifischen Daten auch sprecherspezifisch optimierte Merkmalssets verwendet werden. Hier kann die Leistung von 75,4% auf 83,0% gesteigert werden. Bemerkenswert daran ist, das für diesen signifikanten Zugewinn wie beschrieben nur ein Merkmalsextraktionsvorgang erforderlich ist. Lediglich der Klassifikationsaufwand verdoppelt sich.

Eine reine Trennung nach Geschlecht des Sprechers erbringt hier bei Versuchen auf der Datenbank DES im Gegensatz zu [VER04B] jedoch eine vergleichsweise unbedeutende Verbesserung: Die Erkennung des Geschlechts ist dabei mit einer nahezu fehlerlosen Sicherheit von 98,8% bei vollem Merkmalsset und SVM als Klassifikator in einer 10-fach SCV möglich. Bei Integration der Geschlechtsinformation ergeben sich statt 66,2% Akkuratheit 66,7% korrekte Zuordnung unter den gleichen Rahmenbedingungen. Dieser Effekt muss aber nicht direkt mit dem Geschlecht korreliert sein, sondern kann auch in Verbindung mit einer Reduktion der vier Sprecher auf zwei potentielle im Sinne einer sprecherspezifischen Zuordnung gesehen werden. Grundsätzlich ist eine gesteigerte Zahl von Sprechern hilfreich, um einen unbekannten Nutzer besser emotional einschätzen zu können. Die Erkennungsraten können dabei für die affektive Zuordnung einer Frau bei Training ausschließlich mit männlichen Sprachdaten auch höher sein, als bei Verwendung rein weiblicher Beispiele, und vice versa.

7

Anwendung und Transfer

„Von allem, was ausgerechnet wird in der Welt, geschehen zwei Drittel gedankenlos.“

GEORG CHRISTOPH LICHTENBERG (1742-1799)

Innerhalb dieses Kapitels wird die praktische Anwendung der erarbeiteten Methoden anhand von drei Beispielen diskutiert. Darüber hinaus wird ein Transfer in verwandte Anwendungsgebiete aufgezeigt.

7.1 Einsatz in der Spracherkennung

7.1.1 Adaption an emotionale Sprechweise

In der automatischen Spracherkennung soll die erkannte Emotion zur Adaption der akustischen Modelle verwendet werden, um eine robuste Erkennung auch bei emotional gefärbter Stimme sicherstellen zu können, die zur Erschwernis des Problems beitragen kann [HAN97]. Hierzu soll ein Versuch zeigen, inwiefern dies praktikabel erscheint: Die Zahlen von Null bis Neun wurden von einer männlichen Testperson im Alter von 25 a je 50-mal in neutraler und 50-mal in verärgerter Sprechweise in englischer Sprache geäußert. Die Aufzeichnung erfolgte in einem reflexionsarmen Raum, wie in Kap. 2.5.1 beschrieben. Des Weiteren wurden die Aufnahmen etappenweise über einen Zeitraum von vier Wochen durchgeführt und die Reihenfolge der zu spielenden Emotionen wechselte zufällig, um Antizipationseffekte zu reduzieren. Die insgesamt 1.000 Beispiele, im Folgenden als Datenbank *EA-0-9* bezeichnet, werden mit einem phonembasierten Erkenner [STA03] in einem sprecherunabhängigen Szenario evaluiert. Zur Spracherkennung werden als Merkmale 12 MFCC Koeffizienten und die Signalenergie sowie erste und zweite Ableitungen dieser Verläufe verwendet. Zur Erkennung wird ein hybrider Ansatz [GAU94] eingesetzt, der ein sogenanntes *Tied-Posteriors* Modell, bestehend aus einem MLP welches a-posteriori Phonem-Wahrscheinlichkeiten schätzt [ROB94] sowie einem Satz von Monophon-HMM, besitzt. Generell erlaubt die leistungsstarke Kombination aus ANN und HMM [WAF05B] die Vereinigung des diskriminativen Lernansatzes von ANN und eine einfache Kontextintegration mit den Warpingfähigkeiten von

HMM. Mit $P(j|\underline{x})$ als a-posteriori Phonem-Wahrscheinlichkeit des MLP, $c_{i,j}$ als Mixturkoeffizient jedes HMM-Zustands und $P(j)$ als a-priori Wahrscheinlichkeit jeden Phonems j ergibt sich die Phonem-Wahrscheinlichkeit im Tied-Posterior akustischen Modell zu:

$$p(\underline{x}|S_i) \propto \sum_{j=1}^{J} c_{i,j} \frac{P(j|\underline{x})}{P(j)} \tag{7.1}$$

Um zu prüfen, ob eine Adaption der akustischen Modelle eine Verbesserung erbringt, bieten sich zwei Ansatzpunkte [STA05] an: Zunächst kann das Neuronale Netz adaptiert werden, indem entscheidende Gewichte in der Ausgangsschicht des MLP nachtrainiert werden. Diese werden zur Adaption durch Berechnung der Varianz versteckter Neuronen und Suche von Knoten mit hoher selbiger Varianz bestimmt. Als Zweites können im Anschluss die Mixturkoeffizienten $c_{i,j}$ der HMM mittels einer MAP adäquaten Schätzung angepasst werden, wobei β die Lernrate, $\xi_{i,j}$ die Zustands- und $\gamma_{i,j}$ die Mixturokkupationen bezeichnen. Letztere ergeben sich aus den Adaptionsbeispielen, erzielt durch Baum-Welch Nachschätzungsiterationen:

$$c_{i,j,n+1} = \beta \cdot c_{i,j,n} + \frac{e^{\xi_{i,j}}}{e^{\gamma_{i,j}} + \beta} \tag{7.2}$$

Sowohl die Adaption des MLP, als auch der HMM, wird überwacht durchgeführt. Der Erkenner verwendet bei diesem Experiment ein geschlossenes Vokabular der Größe 5k, Monophone und ein Bigramm Sprachmodell. Als Datenbank für das Training wird die Datenbank WSJ [PAU92] benutzt (vgl. Kap. 3.12.1). Tab. 7.1 zeigt die Erkennungsrate ohne Adaption, mit Adaption an ausschließlich neutrale oder ärgerliche und schließlich gemischte Beispiele.

Akkuratheit [%]	**Keine Adaption**	**Adaption Ärger**	**Adaption Neutral**	**Adaption Ä+N**
Neutral	91,4	71,0	99,5	93,0
Ärger	47,2	89,5	46,0	90,5

Tab. 7.1: Adaption an aktuelle Emotion zur Kompensation affektiver Verzerrung in der Spracherkennung, Datenbank EA-0-9

Es zeigt sich, dass die Erkennung bei ärgerlicher Sprache signifikante Einbrüche aufweist. Durch eine jeweils richtige Adaption kann dies zwar kompensiert werden, eine gegenläufige Adaption bringt jedoch Verschlechterungen.

Adaption	**Keine**	**Statisch Neutral**	**Statisch Ärger**	**Statisch N+Ä**	**Dynam. korrekt**	**Dynam. diametral**
Akk. [%]	69,5	72,8	80,3	91,8	**94,5**	58,5

Tab. 7.2: Adaption an aktuelle Emotion zur Kompensation affektiver Verzerrungen in der Spracherkennung, Datenbank EA-0-9

Tab. 7.2 zeigt Ergebnisse für eine statische Adaption und eine dynamische unter Einsatz von sprecherabhängiger Affekterkennung. Zusätzlich wird die Erkennungsleistung bei jeweils exakt gegenläufiger Adaption als Grenzfall einer falschen Anpassung gezeigt. Es ist zu sehen, dass bei einer statischen Adaption die Verwendung von Beispielen beider Emotionen im Mittel die beste Wahl ist und eine deutliche Verbesserung gegenüber mangelnder Anpassung erbringt. Eine diametrale Adaption führt insgesamt zu einer Verschlechterung, wohingegen eine Anpassung an den jeweils aktuellen Sprecheraffekt unter Einsatz einer automatischen Emotionserkennung in jedem Fall die beste Wahl ist, und eine absolute Verbesserung von 2,7% sowie eine relative von 2,9% aufweist.

Bei einem weiteren Test über Einzelziffern hinaus, wird die in Kap. 3.12.1 vorgestellte Datenbank EA-WSJ verwendet, um dieses Verhalten in ganzen Phrasen bei mehreren Sprechern zu verifizieren. Hier ergibt sich bei Verwendung von 40 disjunkten Sätzen zur Adaption an den Sprecher und die jeweilig erkannte Emotion eine signifikante absolute Verbesserung von 1,1% sowie eine relative von 2,6% für die WER im Vergleich zu statischer Anpassung. Das weitere Verhalten entspricht exakt dem für die Datenbank EA-09 geschilderten.

Insgesamt kann man somit feststellen, dass emotional verzerrte Sprache sich als erwartungsgemäß problematisch für eine ASR-Einheit erweist, die, wie allgemein üblich, nur auf neutralen Daten trainiert wurde. Eine statische Adaption bringt hier schon deutliche Zugewinne, das Optimum ist jedoch eine dynamisch korrekte Anpassung mittels vorhergehender Erkennung des Sprecheraffekts.

7.1.2 Sprachverstehen

Mit Hilfe der in Kap. 4.6 vorgestellten grafischen Modellierung können über emotionale Sätze hinaus auch natürlichsprachliche Nutzeranweisungen dargestellt und erkannt werden. Anstelle von Emotionen sind dann einzelne Benutzerintentionen in den Wurzelknoten repräsentiert. Jede Intention erhält hierzu ihr eigenes Netz, und es wird analog zu Kap. 4.6 diejenige Benutzerabsicht maximaler Wurzelwahrscheinlichkeit gewählt. Durch diesen Ansatz erhält man auch hier Ausgangssicherheiten beruhend auf Eingangssicherheiten, wodurch das Bereitstellen von gewichteten Listen wahrscheinlicher Intentionen möglich wird. Die notwendige Vorraussetzung Parameter eines Benutzerwunsches zu speichern und auszugeben ist bereits vorgestellt worden, jedoch bisher nur zur Ausgabe der Ausprägung einer Emotion genutzt worden.

In den folgenden Kap. 7.2 und Kap. 7.3 wird so realisiertes Sprachverstehen näher, anhand zweier konkreter Beispielszenarien, demonstriert.

7.2 Transfer in die Musikverarbeitung

Besonders die in Kap. 3 zur Erkennung von Emotion aus dem akustischen Sprachsignal beschriebenen Merkmale sowie die Verfahren zur Selektion und Klassifikation lassen sich teilweise direkt zur Erkennung und automatischen Verarbeitung von Musik im Bereich des *Music Information Retrieval*, kurz *MIR,* erfolgreich nutzen. Ziel dieser jungen Disziplin ist es, die Suche von digitaler Musik in großen Archiven mittels innovativen Ansätzen intuitiv, effizient und effektiv zu gestalten. In den folgenden Abschnitten wird gezeigt, wie die gebildeten akustischen Funktionale

zur Unterscheidung von Signaltypen und ihre dynamischen Basisverläufe zur Erkennung von Melodien angewandt werden können. Im Anschluss wird ein Gesamtsystem zur multimodalen Suche nach Musiktiteln präsentiert, welches die in den bisherigen Abschnitten vorgestellte Entwicklung von Erkennungstechnologie einerseits transferiert und andererseits vereint. Zur Einführung wird in den folgenden Abschnitten kurz der Hintergrund im Rahmen des MIR erläutert.

7.2.1 Diskrimination von Sprache Gesang und Musik

Die automatische Unterscheidung des Signaltyps nach Sprache, monophonem Gesang und polyphoner Musik erschließt eine Reihe von Anwendungsgebieten. Zu diesen gehören unter Anderen der Einsatz in Hörgeräten, bei denen, je nach Signaltyp, verschiedene akustische Parameter zum besseren Sprachverständnis oder optimierten Hörerlebnis beitragen sollen, oder automatische Segmentierung von Audiodatenströmen, wie zum Beispiel Radiosendungen, aus denen etwa nur Gesprächsteile oder nur Musik herausgeschnitten werden sollen. Von besonderem Interesse ist hier jedoch die Unterteilung von Signalanteilen beim sogenannten offenen Mikrophonbetrieb eines Sprachdialogsystems (vgl. Kap. 4.3.2). Diese Form der Interaktion erweist sich als besonders natürlich, da zu jeder Zeit gesprochen werden kann. Bei der im später folgenden Kap. 7.2.3 vorgestellten Suche von Musik mittels Sprache und Vorsingen, erfordert dies eine automatische Unterteilung in gesprochene und gesungene Abschnitte zur weiteren Verarbeitung. Sobald ein Musikstück ausgewählt ist, angespielt wird und vom Mikrophon erfasst wird, soll diese polyphone Musik zur Filterung auch abgegrenzt werden. Wie später in Kap. 7.3.1 beschrieben, wird dieses Prinzip im Fahrzeug ebenfalls verwendet, um bei der Sprachbedienung Signalanteile reiner Musik eines Audiogeräts zu eliminieren. Darüber hinaus ist es denkbar, innerhalb von Audiostücken gesungene Abschnitte ohne starke Hintergrundanteile gezielt zu suchen, um das Finden von Audiodateien durch ein kurzes Vorsingen zu ermöglichen (vgl. Kap. 7.2.2).

Während zur Diskrimination von polyphoner Musik und Sprache eine Reihe von Forschungsarbeiten bestehen [SEI97], [PAA05], ist der diffizilere Fall der Unterscheidung von Gesang und Sprache, ins Besondere von der gleichen Person sowie in kurzen Abschnitten noch wenig betrachtet [CHO01], [GER02]. In direkter Anwendung der in Kap. 3 vorgestellten akustischen Funktionale, Selektionsverfahren und Klassifikationsalgorithmen lässt sich ein System entwickeln, mit dem man die gestellten Anforderungen umsetzen kann. Dabei wird im Besonderen einer Echtzeitfähigkeit, durch starke Reduktion der insgesamt 276 Funktionale, Rechnung getragen, um den Signalstrom schritthaltend verarbeiten zu können.

Zunächst muss jedoch auch hier als Ausgangsbasis eine adäquate Lernmenge $\mathcal{L}$ zum Lernen und Testen von geeigneten Methoden des maschinellen Lernens geschaffen werden. Hierzu existieren derzeit keine öffentlichen Standarddatenbanken, weswegen ein Korpus erstellt wurde, der zur Vereinfachung als *MIR-SMS*[174] abgekürzt wird. In diesem Datensatz finden sich in annähernd gleicher Distribution 1.000 kurze Sprachclips *(MIR-TALK)*, 1.114 Clips monophonen Gesangs *(MIR-SING)* und 1.000 Clips polyphoner Musik *(MIR-MTV)*. Bei den Sprachaufzeichnungen

[174] Die Datenbanken zur Musikverarbeitung werden entsprechend der eingeführten Bezeichnung mit MIR und einem Zusatz abgekürzt. *SMS* steht dabei für Sprache, Musik und Singen.

handelt es sich um Interaktionsdaten mit den in Kap. 7.2.3 und Kap. 7.3.1 vorgestellten Sprachinterfaces. Grundlage für die Gesangs- und Musikdatenbanken sind alle 200 Lieder der *MTV*[175] *Europe Top 10* der Jahre 1981-2000. Diese wurden in einem Versuch in einem RAR von 11 Personen nachgesungen, wobei die Stelle und die Länge frei wählbar waren [SCH05D]. Die gewählten Liedtitel sind dabei einerseits weit bekannt, und andererseits durch diese Auswahl klar definiert. Polyphone Musikclips wurden erhalten, indem aus den jeweiligen Originalen an fünf festgelegten relativen Positionen Clips ausgeschnitten wurden. Tab. 7.3 gibt einen Überblick zu den näheren Eckdaten der einzelnen Teilmengen von MIR-SMS.

Beschreibung	**MIR-TALK**	**MIR-SING**	**MIR-MTV**
Domäne	Sprache	Monophoner Gesang	Polyphone Musik
Inhalt	Frei gewählt	MTV Europe Top 10 1981-2000	MTV Europe Top 10 1981-2000, 5 fixe Intervalle
Sprecher / Sänger	45, 3 weiblich 21-30a, Ø 25,0 a	11, 2 weiblich 20-28 a, Ø 24,0 a	-
Samplezahl	1.000	1.114	1.000
Samplelänge	1,0 s - 10,0s Ø 4,3 s	2,0 s - 21,0 s Ø 9,5 s	10,0 s
Verteilung	gleich	gleich	gleich
Sampling / Kodierung	44,1 kHz, 16 Bit, PCM	44,1 kHz, 16 Bit, PCM	MP3, 128 kBit/s
Mikrophon	Kondensator Yoga EM240	Kondensator Yoga EM240	-
Umgebung	Büro, Fahrzeug	RAR	-
Öffentlich	Nein	Nein	Ja

Tab. 7.3: Übersicht über die verwendeten akustischen Datenbanken zur Musikverarbeitung

Einen Aufschluss zu den erzielten Erkennungsleistungen gibt Tab. 7.4.

Akkuratheit [%]	**Sprache**	**Gesang**	**Musik**	**Gesamt**
SVM	99,5	99,2	99,9	**99,7**

Tab. 7.4: Diskrimination von Sprache, Gesang und polyphoner Musik, Datenbank MIR SMS, Top 15 SVM-SFFS Merkmale, Klassifikator SVM, 10-fach SCV

Diese Erkennungsraten werden noch bei einer Reduktion mit SVM-SFFS auf 15 Attribute erzielt. Tab. 7.5 zeigt diese 15 relevantesten Merkmale sowie die Angabe ihrer IGR. Es ist deutlich zu erkennen, dass IGR-FS ein anderes Ranking ergibt. Die Tatsache, dass auch hier Merkmale mit geringer Einzelrelevanz im finalen Set enthalten sind, zeigt den Verbundoptimierungscharakter von

[175] Kommerzieller Musikfernsehsender.

SVM-SFFS. Bei gleicher Dimensionalität werden mit IGR-FS nur noch 97,5% und mit PCA-FS 99,2% Akkuratheit erreicht.

Rang	IGR	Merkmal
1	0,4941	MFCC2StdDev
2	0,5959	MFCC5Min
3	0,9137	HNRMean
4	0,5203	MFCC2Mean
5	0,5958	MFCC11De2Mean
6	0,9389	PoSpe650
7	0,3148	SpecPoCenStdDev
8	0,3041	PitchStdDev
9	0,5432	MFCC15De2Mean
10	0,0663	MFCC12Mean
11	0,3509	F7ReDiF0
12	0,6116	MFCC3Mean
13	0,6619	MFCC1Max
14	0,2978	PitchDe1Mean
15	0,3849	PitchRePoDistStdDev

Tab. 7.5: Top 15 Merkmale Reihenfolge nach SVM-SFFS, Datenbank MIR-SMS, Angabe IGR

7.2.2 Melodieerkennung

Eine Reihe von Aspekten spricht für die automatische Erkennung von Melodien, die als Suchanfrage an ein Gerät gerichtet sind, anstelle von konventioneller textbasierter Suche nach Titel, Interpret, etc.: Oft liegt unvollständige, fehlende oder falsche Information beim Benutzer oder systemseitig vor, falls digitale Musikstücke fehlerhafte ID3-Tags[176] besitzen. Ein großes Problem bereitet auch der mehrsprachige Charakter von Musiktiteln und Interpreten sowohl für den Nutzer beim Schreiben oder Sprechen, als auch für ein Erkennungssystem auf Grund der sich ergebenden Zeichen- und Phonemvielfalt. Die Melodie eines Stückes ist diesem im Gegensatz hierzu stets immanent, und dem Suchenden meist gegenwärtig. Dies gestaltet das Finden von Musik auf Basis von Vorsummen[177], -singen, oder -pfeifen sehr effizient und intuitiv. Darüber hinaus ist auch eine Suche mittels dem Audiostück selbst von Interesse, um etwa ein im Radio gehörtes Lied, oder ähnliche Stücke wie Liveaufnahmen oder alterierte Versionen zu finden.

Mehrere Probleme erschweren jedoch den wohl interessantesten Fall eines direkten Vergleichs gesungener Passagen mit polyphonen Aufnahmen als Suchanfrage [PIK01], [REI02]: Das originale Audiosignal ist in der Regel von Hintergrundmusik wie Begleitung oder Perkussion überlagert, welche gefiltert werden muss. Ein Benutzer wird ferner meist eine konkrete, besonders einprägsame

[176] Dateianhang für digitale Musikstücke, der die Informationen Titel, Interpret, Veröffentlichungsjahr, Album, Genre und Platz für einen Kommentar bietet. In der ursprünglichen Version sind die Felder Jahr auf 4 Zeichen, Genre auf 1 Byte und alle weiteren auf 30 Zeichen beschränkt, was zusammen 128 Byte entspricht. Um diese Limitationen aufzuweichen, wurde der Standard in der Version ID3v2 in einen Header vor dem Musikstück in offener Struktur umgewandelt.

[177] Als *Query-By-Humming (QBH)* bezeichnet.

Stelle vorsingen, welche innerhalb des gesamten Stücks gefunden werden muss. Die Melodie wird dabei in der Regel, auf Grund eingeschränkten Singvermögens und fehlender Referenz, rhythmisch und tonal verzerrt sowie in einer anderen Tonart wiedergegeben sein [SON02]. Zusätzlich kann es sein, dass der Suchende das Thema variiert oder nicht mehr vollständig in Erinnerung hat. Auch kommt es vor, dass Nutzer eventuell nicht singen wollen, vor allem nicht in der Öffentlichkeit. Manche Genres wie avantgardistischer Jazz machen ein Vorsummen darüber hinaus nahezu unmöglich. Schließlich existiert meist eine Vielzahl an Versionen und Interpretationen eines musikalischen Themas, wodurch eine melodiebasierte Suche alleine nicht immer ausreichend ist. Auf alle Fälle erlaubt sie eine Einschränkung des Hypothesenraums auf eine Zahl, die mit Hilfe einfacher Selektionsparadigmen zu bewältigen ist.

Prinzipiell können bei der Merkmalsextraktion Verläufe wie in Kap. 3.2 beschrieben angewandt werden. Da die dynamische Entwicklung einer Melodie eine besonders relevante Kerngröße bei dieser Aufgabe darstellt, werden hier keine Funktionale gebildet, sondern direkt Verfahren der Klassifikation multivariater Zeitreihen angewandt (vgl. Kap. 3.11). Beim Vergleich monophoner Suchanfragen zu polyphonen Originalaufzeichnungen sollen darüber hinaus Signalanteile neben der Kernmelodie reduziert werden und mit Schwerpunkt der musikalischen Natur Rechnung getragen werden.

Zunächst kann, soweit vorhanden, stereophone Information genutzt werden, da das zentrale musikalische Thema im Regelfall im stereophonen Spektrum mittig angeordnet ist. Grund hierfür ist die so verstärkte Wahrnehmbarkeit unabhängig von der räumlichen Position des Zuhörers. Mittig positioniert findet sich sonst meist nur der Bass, weil sich diese niederfrequenten Anteile vom menschlichen Gehör schwer orten lassen und oft, wie in kontemporärer Popmusik, der Wunsch nach starker Basswiedergabe besteht. Diese wird zusätzlich durch die Verwendung beider stereophoner Lautsprecher betont. Für den betrachteten Zweck kann der hier störende Bassanteil jedoch durch einfache Hochpassfilterung eliminiert werden. Andere Begleitinstrumente und eventuelle Schlagzeug- und Perkussionsanteile werden hingegen bevorzugt seitlich arrangiert, um das räumliche Spektrum aufzuweiten. Entgegen einigen Systemen, wie zum Beispiel für Karaoke verwendet, bei denen der Mittelanteil durch Kanalsubtraktion gelöscht wird, soll hier gerade dieser Anteil extrahiert werden. Dies gestaltet sich weitaus anspruchsvoller, und ist nur in Mehrkanalaufzeichnungen verlustfrei möglich. Daher wird folgende, schnell berechenbare Näherung für stereophone Aufnahmen vorgeschlagen (vgl. [SCH03C]). Die Bezeichnungen $s_{mon}[n]$, $s_l[n]$ und $s_r[n]$ stehen dabei für das erhaltene einkanalige Signal sowie den ursprünglich linken und rechten stereophonen Kanal zum Abtastzeitpunkt n. Bei λ_d handelt es sich um einen korrigierenden Dämpfungsfaktor, der hier zu $\lambda_d = 0,5$ gewählt wurde. Dieser ist erforderlich, weil sich durch die Kanaladdition Spitzen im Verzerrungsbereich ergeben können, die durch den gewählten Wert auch bei maximalem Signalwert in beiden Kanälen vermieden werden.

$$s_{mon}[n] = \lambda_d \cdot \left(\operatorname{sign}\left(s_r[n] + s_l[n]\right)\right) \cdot \max\left(\left(\left|s_r[n] + s_l[n]\right| - \left|s_r[n] - s_l[n]\right|\right), 0\right) \tag{7.3}$$

Der Anteil $s_r[n] + s_l[n]$ ist dabei der konventionellen Bildung des monophonen Signals äquivalent, und entspricht ohne Dämpfung bereits einer Anhebung des Mittenanteils um 6 dB. Der Term

$s_r[n]-s_l[n]$ hingegen ist die beschriebene Subtraktion zur Mitteneliminierung. Die Signumfunktion $\text{sign}\left(s_r[n]+s_l[n]\right)$ als Vorfaktor restauriert die originale Phase, was durch die Betragsbildung erforderlich wird. Sollte kein Mittenanteil vorhanden sein, wird der Signalwert zu Null gesetzt, um ein Übersprechen der Seitenanteile zu vermeiden. Für eine akustische Wiedergabe eignet sich das so verstärkte Mittensignal zwar nicht, für einen Vergleich auf Signalebene kann jedoch ein nutzbarer Vorteil verifiziert werden (vgl. [SCH03C]).

Als Merkmal wird zunächst der Verlauf der Grundfrequenz analog zu Kap. 3.4.3 betrachtet. Dieser bietet sich nur für einen Vergleich zwischen gesungenen Abschnitten an, da er bei einem polyphonen Signal auf Grund der starken Überlagerungen nicht sinnvoll extrahiert werden kann[178]. Eine Quantisierung in diatonischen Halbtonschritten bietet sich vor dem Hintergrund der Signalcharakteristik dabei als optimale Glättung des Konturverlaufs an. Nachteilig hieran ist die Notwendigkeit eines Bezugspunktes, wie hier der Kammerton a'=440 Hz[179] und die sich ergebende Rasterung. Letztere erscheint grob genug, um den Grenzfall eines Gesangsstückes, welches sich fortlaufend an der Intervallgrenze bewegt, vernachlässigen zu können. Des Weiteren kann eine Quantisierung in Halbtonschritten, bei nicht wohltemperierter Stimmung oder sogenannten Viertelton- oder *Smearbends* wie sie in der Bluesmusik vorkommen, nicht fein genug erscheinen. Sie hat sich aber hier als ausreichend und zweckmäßig erwiesen, um die Zahl der Notenbänder für die Klassifikation zu beschränken.

Um rhythmische Expression zu erfassen, wird auch die Intensität durch die Signalenergie und ihre höheren Ableitungen, wie in Kap. 3.4.2 beschrieben, extrahiert.

Für polyphone Signale wird eine spektrale Repräsentation als Alternative gewählt. Neben MFCC Koeffizienten (vgl. Kap. 3.5.4) oder LPC Koeffizienten (vgl. Kap. 3.5.1) bietet sich vor dem Hintergrund der musikalischen Charakteristik vor allem die Bestimmung dominanter Partialtöne in einem an die menschliche Lautheitswahrnehmung angepassten FFT-Spektrum (vgl. Kap. 3.5.3) an [NAG02].

Zunächst wird berechnet, wie stark ein Frequenzband der Mittenfrequenz f und der Breite B aus seiner unmittelbaren Umgebung von $2\cdot\Delta$ Nachbarbändern hervortritt: ausgehend von der spektralen Bandenergie $E_t[f]$ bei der Mittenfrequenz f im Signalrahmen mit Index t wird die Dominanz der spektralen Bandenergie eines Partialtons $E_{p,t}[f]$, *Emphase* genannt, bestimmt:

$$E_{p,t}[f]=\sum_{j=-\Delta}^{\Delta}\left(E_t[f]-E_t[f-j\cdot B]\right) \tag{7.4}$$

---

[178] Eine Trennung der Signalquellen etwa mittels *Independent Component Analysis* kann eine Grundfrequenzextraktion prinzipiell begünstigen, doch die sich ergebende Robustheit reicht auch hier meist nicht aus.

[179] Dies entspricht der ISO 16 aus dem Jahr 1939. Tatsächlich schwankt der Kammerton als Stimmton für Orchester bis zu zwei Ganztöne zwischen 392 Hz, im Frankreich des 18. Jahrhunderts verwendet - entsprechend der hier getroffenen Stimmung gleich einem g' - bis zu 490 Hz - hier entsprechend einem h'. Generell ist eine steigende Tendenz der Frequenz beobachtbar. Für die vorrangig betrachtete Populärmusik stellen die ausgewählten 440 Hz jedoch einen sinnvollen Bezugspunkt dar.

Auf dieser Basis wird die harmonische Ausgeprägtheit $H_{p,t}[f]$, anhand der Dominanz n_p nachfolgender zugehöriger Partialtöne, bis zur maximalen Frequenz $n_p \cdot f$ bestimmt:

$$H_{p,t}[f] = \frac{1}{n_p} \cdot \sum_{j=1}^{n_p} E_{p,t}[j \cdot f] \tag{7.5}$$

Insgesamt ergeben sich somit bei Wahl der diatonischen Quantisierung innerhalb des menschlichen Gesangsbereichs 47 Koeffizienten, die die harmonische Ausgeprägtheit einzelner Semitöne auf Basis ihrer Dominanz gegenüber Nachbarbändern repräsentieren.

Da eine frei gesungene Suchanfrage im Normalfall nicht in der Tonart des Originalstücks gesungen wird, ist bei Verwendung der Grundfrequenz entweder nur die Tondifferenz zu betrachten, oder eine Normierung zu treffen. In letzterem Fall hat sich die Durchschnittstonhöhe in stimmhaften Bereichen gegenüber der Anfangs- oder Minimaltonhöhe als vorteilhaft erwiesen. Für die harmonische Summe erfolgt eine Anpassung durch eine Verschiebung in Halbtonschritten über die Bänder, bis sich mittels DTW der geringste Abstand ergibt. Generell werden nur Differenzwerte erster und zweiter Ordnung verwendet, um invariante Begleitanteile zu eliminieren.

Zur Klassifikation eignen sich die in Kap. 3.11 vorgestellten dynamischen Verfahren HMM oder DTW. Für das Trainieren dieser Lernalgorithmen stehen entweder nur wenige gesungene Referenzen, oder als einziges Datum das originale Lied zur Verfügung. Dies ist zunächst nicht ausreichend, um ein statistisches Modell im Sinne eines HMM zu erstellen. Im Gegensatz hierzu ist bei Verwendung von DTW bereits mit sehr wenigen Daten ein Vergleich möglich, weswegen dieser in [SON02] zur Verwendung empfohlen wird.

Auf Grund ihrer im Allgemeinen höheren Klassifikationsleistung wird hier ein Verfahren vorgestellt, um dennoch HMM nutzen zu können. Grundgedanke dabei ist, dass ein Musikstück meist mehrere, sich wiederholende Teile besitzt. Es ist davon auszugehen und hat sich bei der Kollektion der Datenbank MIR-SING bewahrheitet, dass bei Suchanfragen in der Regel solche typische, sich oft wiederholenden Stellen, etwa der Refrain, verwendet werden. Gelingt es also, ein polyphones Musikstück automatisch in wiederkehrende Abschnitte wie Strophe oder Refrain[180] zu unterteilen, kann je ein Links-Rechts-HMM pro Abschnitt mit mehreren Beispielen trainiert werden. Hierzu erfolgt zunächst eine Vorsegmentierung nach Stellen starker Änderung mittels der Merkmale Energie, 12 MFCC, Roll-Off-Punkt und spektraler Zentroid sowie Euklidschem Distanzmaß. Im Anschluss werden Abschnitte mit der Mindestlänge von 5 s und der maximalen Länge von 12 s zwischen Punkten der Vorsegmentierung gesucht[181]. In drei Schritten erfolgt dann die Suche nach dominanten Repetitionen mittels DTW (vgl. [SCH05B]):

---

[180] In [SOO04] werden diese Abschnitte technisch als Eigen-Texturen bezeichnet.

[181] Diese Werte basieren auf den für Populärmusik typischen Tempogrenzen von 60 bpm bis 180 bpm. Ausgehend von einem Viervierteltakt und einer Dauer von vier bis acht Takten pro charakteristischem Segment ergeben sich hieraus die zeitlichen Grenzen 5,3 s bis 16 s. Es hat sich jedoch bei der Aufnahme der Datenbank MIR-SING gezeigt, dass Nutzer dazu neigen kürzere Clips zu singen. Insgesamt ergibt sich daher für die Datenbank MIR-MTV das Optimum im angegebenen Bereich zwischen 5 s und 12 s.

- Suche der 20 Kandidaten mit minimaler kumulativer Distanz beim Vergleich aller Abschnitte untereinander als potentielle Repetitionen. Dieser Wert hat sich als Optimum für die in Kap. 7.2.1 beschriebenen Musiktitel der MTV Europe Top 10 der Jahre 1981-2000 gezeigt. Die jeweiligen Musikstücke sind hierzu im MP3 Format mit einer Rate von 128 kBit/s kodiert in der Datenbank MIR-MP3 gesammelt.

- Suche der Abschnitte mit maximaler kumulativer Distanz innerhalb der 20 gefundenen Segmente als verschiedene Elemente. Die Anzahl der Elemente wird dabei dynamisch anhand eines Schwellwerts bezüglich der aufgetretenen Abstände bestimmt.

- Suche der Abschnitte mit minimaler kumulativer Distanz zu den gefundenen Elementen innerhalb aller Abschnitte.

Es wird also erst nach Elementen gesucht und ihre Zahl festgelegt, um dann zugehörige Repetitionen zu diesen aufzufinden. Für die Titel einer Datenbank ergeben sich somit unterschiedliche Zahlen von HMM. Da bei vorgesungenen Suchclips in der Regel nur einmalig der Refrain auftritt, können beim Vergleich zwischen gesungenen Beispielen nur dann sinnvoll HMM angewendet werden, wenn mehrere Beispiele existieren. Bei der Anwendung innerhalb eines Interfaces zur Musiksuche, wie in Kap. 7.2.3 vorgestellt, wird dieser Aspekt aber eher als uninteressant betrachtet: Möchte ein Anwender das Lied seiner Wahl durch Singen aufrufen, sollte er es vorab nur ein einziges mal dem System vorsingen müssen. Daher wird für diesen Fall DTW zum Matching gewählt, und in der Evaluation nur ein paarweiser Vergleich betrachtet.

Vor der Evaluierung der Erkennungsleistung soll auch hier ein Versuch die menschliche Zuordnungsleistung zeigen. Acht Probanden, zwei davon weiblich, im mittleren Alter von 23,0 a, haben nach Anhören von 50 der gesungenen Clips der Datenbank MIR-SING diese zu den dazu passenden 50 Titeln im Original zugeordnet. Dabei konnten sie jederzeit die Aufnahme des gesungenen Stücks wiederholen lassen und jedes der 50 Originalstücke beliebig oft anspielen. Letztere waren den Probanden durchwegs bekannt. Die mittlere Zuordnungsleistung betrug dabei 53,6% ± 6,0%. Die maximale Zuordnungsleistung eines Probanden betrug 86,3%.

Im Zuge einer Merkmalsselektion auf Basis einer Vollsuche nach Merkmalsgruppen mit DTW hat sich für das Matching gesungener Suchclips, im paarweisen Vergleich auf der Datenbank MIR-SING, die Kombination aus Grundfrequenz, Energie und 12 MFCC als Optimum ergeben. Für die Zuordnung von gesungenen Clips der Datenbank MIR-SING zu polyphonem Audio der Datenbank MIR-MP3 haben sich die 47 Koeffizienten der harmonischen Ausprägung von Semitönen dann als optimal erwiesen, wenn die Emphase über $2 \cdot \Delta = 8$ benachbarte Bänder betrachtet wird, und $n_p = 3$ nachfolgende Harmonische analysiert werden.

Abb. 7.6 zeigt die so jeweils maximal erzielten mittleren Erkennungsleistungen einer 4-fach Kreuzvalidierung. Diese sind nach Auftreten innerhalb der Top N Treffern in einer geordneten Ergebnisliste angeführt. Erwartungsgemäß zeigt sich ein Vergleich von monophonen Gesangsclips der gleichen Person als die einfachste Aufgabe. Für einen personenunabhängigen Vergleich sinkt die Erkennungsrate, was auch daran liegt, das nicht gewährleistet ist, dass zwei Personen den selben Ausschnitt eines Liedes singen. Beim Vergleich einer Suchanfrage innerhalb der Datenbank MIR-

MP3 zeigt sich in der 4-fach Kreuzvalidierung deutlich die Schwierigkeit dieser Aufgabe. Erst innerhalb der ersten fünf Titel respektive ergibt sich ein Ergebnis über 90% korrekter Zuordnung.

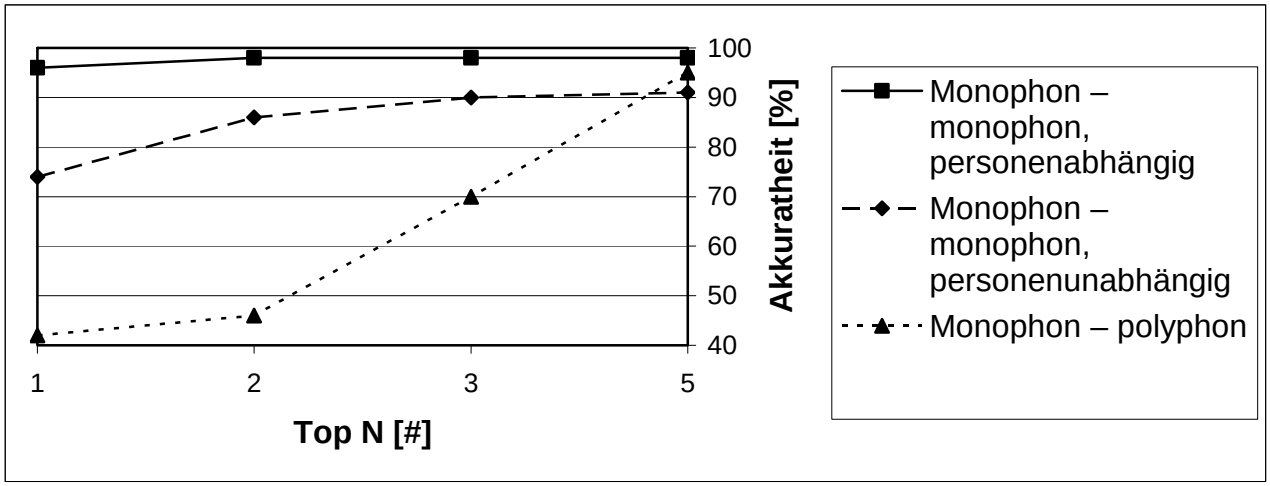


Abb. 7.6: Erkennungsleistung für Melodieerkennung bei verschiedenen Audiotypen, Datenbanken MIR-SING und MIR-MP3, Klassifikator DTW und HMM, paarweiser Vergleich, 4-fach Kreuzvalidierung

Abb. 7.7 zeigt den Verlauf der Erkennungsleistung bei variabler Suchcliplänge. Prinzipiell steigt diese mit zunehmender Länge an. Bei der Aufnahme der Datenbank MIR-SING hat sich jedoch gezeigt, dass die Versuchspersonen ab einer gewissen Dauer die Suchmelodie oft abbrechen, weswegen eine generelle Grenze von 5 s gewählt wurde.

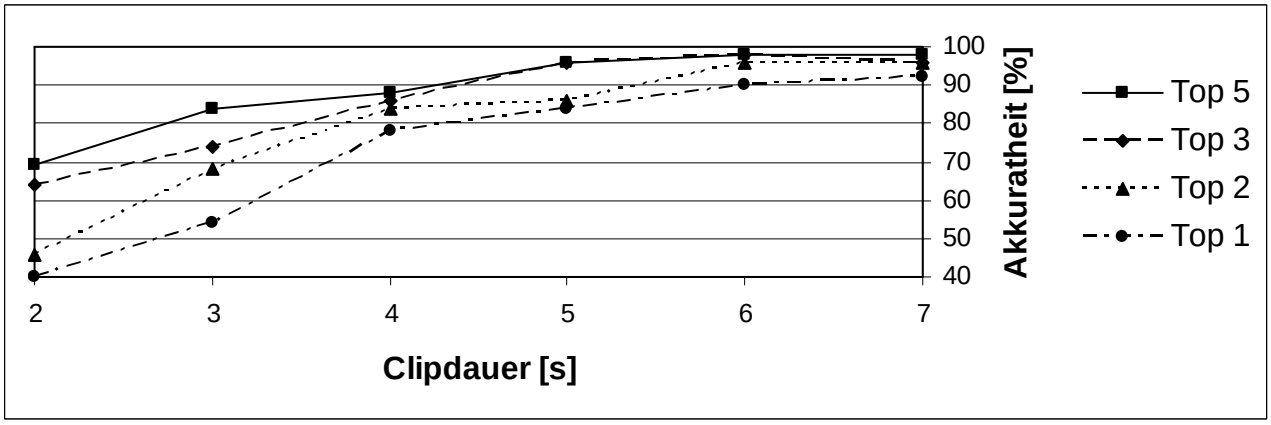


Abb. 7.7: Abhängigkeit der Erkennungsleistung von der Länge der gesungenen Clips, Datenbank MIR-SING, Klassifikator DTW, paarweiser Vergleich, 4-fach Kreuzvalidierung

7.2.3 Demonstrator multimodale Musiksuche

Als erste Demonstrationsplattform für den Transfer und die Anwendung der gezeigten Methoden wird ein realisiertes System zur multimodalen Suche nach Musiktiteln im MP3-Format vorgestellt. Das MIR-Interface erlaubt die Suche von Musiktiteln sowohl schlüsselwort-, als auch inhaltsbasiert per Sprache, Handschrift, Tastatur, Maus und Vorspielen oder Singen eines Liedes in kombinierter

Form (vgl. [SCH03D], [SCH04E]). Die Visualisierung und Kommunikationsstruktur basiert auf einer im Rahmen dieser Forschungsarbeiten entwickelten Rapid-Prototyping Umgebung für multimodale Benutzerschnittstellen (vgl. [SCH02A] und Kap. A.4). Abb. 7.8 zeigt zunächst eine Aufnahme der Bildschirmdarstellung des Interface. Neben der visuellen Darstellung und der akustischen Musikwiedergabe erfolgt auch eine Sprachausgabe für einfache Systemrückfragen. Die Ausgabe beruht auf vorab aufgezeichneten Sprachsamples und der Dialogablauf wird über Zustandsautomaten gesteuert.

Als Schlüsselwörter werden die ID3-Tags Titel, Interpret, Veröffentlichungsjahr, Album und Genre verwendet. Da in vielen MP3-Dateien diese Informationen fehlen oder fehlerhaft sind, werden als weitere Suchwörter auch der Dateiname und das Verzeichnis, in dem die Datei abgelegt ist, verwendet. Über Maus, oder durch Druck bei einem berührsensitiven Bildschirm, kann zunächst eine konventionelle Navigation erfolgen.

Abb. 7.8: Grafische Benutzeroberfläche multimodales MIR-Interface

Ferner erfolgt, wie in Kap. 4.2 gezeigt, eine Verarbeitung von geschriebener Eingabe über Tastatur oder Schrift per Maus oder Hand auf einem Touchscreen, wie in Abb. 7.8 unten ersichtlich. Soft-String-Matching erweist sich hier in besonderem Maße als nützlich, da Nutzer oft nicht den exakten Suchbegriff wissen, oder mehrere Schreibweisen existieren. Das selbst bei hoher Levenshtein-Distanz noch ein Auffinden in einer großen Datenbank möglich ist, zeigen Tab. 4.7 und Tab. 4.8 in Kap. 4.2. Mittels der Schrifterkennung sind auch Bediengesten umgesetzt. Zu diesen gehören beispielsweise Löschen des letzten Zeichens, Suchstart, Listenscrollen, „*Ja*“ und „*Nein*“ durch Wischbewegungen oder einen Haken.

Zur inhaltsbasierten Suche wird die Melodie des Stücks verwendet. Diese kann entweder vom Anwender einmalig vorgesummt werden, was ein robustes Auffinden ermöglicht, oder komfortabler automatisch aus einer MP3-Datei extrahiert werden (siehe Kap. 7.2.2). Der beste Treffer wird angespielt, und die fünf besten Treffer zusätzlich zur erweiterten Anwahl bei Fehlern im Display

angezeigt.

Eine Bedienung mit Sprache kann durch Steuerungs-, Informations- und Navigationsanweisungen in natürlicher Form erfolgen. Hierzu wird das in Kap. 7.1.2 gezeigte Prinzip zur natürlichsprachlichen Verarbeitung eingesetzt. Einzelne Suchattribute werden zur Laufzeit aus den ID3-Tags der aktuell dargestellten Spielliste ausgelesen und phonemisiert. Dies ist dank des in Kap. 4.3.1 vorgestellten, auf String-Matching basierenden Ansatzes zur Spracherkennung durchführbar[182]. Der Benutzer kann so jederzeit anhand eines gewünschten Attributs direkt per Sprache suchen, insofern dieses in der Datenbank enthalten ist. Das Intentionsmodell zur natürlichsprachlichen Interpretation umfasst 26 Intentionen. Zu diesen wurden in einem Usability-Experiment von insgesamt 27 Testpersonen, 10 davon weiblich, im mittleren Alter von 32,4 a ± 14,1 a (Minimum 19 a, Maximum 70 a) 514 Einzelphrasen gesammelt (Minimum pro Intention 12 Phrasen, Maximum 32). Nach Verschriftung und Training der BN, wie in Kap. 7.1.2 gezeigt, erfolgte zunächst ein Reklassifikationstest zur Plausibilität des semantischen Modells. Die Rate korrekter Zuordnung beträgt dabei 95,9%. Durch Erweiterung des Ansatzes um einen endlichen Zustandsautomat, kurz FSA[183], der sicherstellt dass nur aktuell sinnvolle Aktionen gewählt werden, kann die Leistung auf 97,1% erhöht werden. In einer fünffachen SCV ergibt sich die reale Erkennungsleistung durchschnittlich zu 77,9% und 80,6% mit Zustandsautomat. Im Mittel umfasst eine der 26 Intentionen 2,2 Lexeme (Minimum Eins, Maximum Drei) und eines der 57 Lexeme 5,9 Wörter (Minimum drei, Maximum elf). Im Gegensatz hierzu besitzen die gesammelten Äußerungen selbst im Mittel 4,7 Wörter (Minimum 3,3, Maximum 7,9). Das Vokabular ist jeweils offen, um dynamisch Attribute einbinden zu können. Die beschriebene natürlichsprachliche Steuerung inklusive Sprachausgabe und Dialogsteuerung konnte auch erfolgreich als Plug-In für die frei erhältliche Abspielsoftware digitaler Musikdaten *Winamp* [NUL05] umgesetzt werden.

Um dem Benutzer größtmögliche Flexibilität zu gewährleisten, ist ein offener Mikrophonbetrieb gewählt. Das System reagiert auf die Nennung eines frei wählbaren Schlüsselwortes, welches innerhalb einer Äußerung enthalten sein muss. Zusätzlich wird die in Kap. 7.2.1 gezeigte Signaltypdiskrimination eingesetzt. Der Anwender kann so jederzeit vorsingen oder sprechen, ohne dem System explizit den Beginn eines Sprach- oder Gesangsabschnitts mitteilen zu müssen. Auch kombinierte Suchanfragen wie *„Bitte spiele* [singt] *von* [Interpret]*“* werden dadurch ermöglicht. Die Erkennung polyphoner Musik anstelle von Sprache oder Gesang ist erforderlich, da das System sonst auf die über das Mikrofon erfasste polyphone Musik reagieren könnte. Im Falle einer Entscheidung zu Gunsten von Sprache wird zusätzlich die Emotion des Nutzers aus dieser erkannt und die Stimme des Nutzers verifiziert (siehe Kap. 6.3), um andere Sprachanteile zu filtern.

Abb. 7.9 gibt einen Überblick zur Systemarchitektur. Das Audiostück selbst, die Schlüsselwörter und die Melodie sowie aufgezeichnete Daten des Kontexts bilden das digitale Musikarchiv. Die multimodale Fusion ist dabei, wie in Kap. 6.2 gezeigt, mittels BN realisiert. Dabei sind die Klassen hier die Musiktitel in der Datenbank. Zusätzlich zu den Ergebnissen einzelner Erkennerinstanzen wird auch Kontext modelliert. Die Hörgewohnheiten des Anwenders werden gelernt, um den

[182] Bei der Phonemisierung unbekannter Attribute erfolgt keine Erkennung der Sprache.

[183] Kurzform von *Finite-State-Automat*.

Hypothesenraum bei künftigen Suchen stärker einschränken zu können. Hierzu wird die Häufigkeit, die Jahreszeit, die Tageszeit (früh, tagsüber, abends, nachts) und die Emotion (Ärger, Freude, Neutralität und Trauer) zu jedem gespielten Titel protokolliert[184]. Ein Lied wird so mit größerer Wahrscheinlichkeit ausgewählt, wenn es zur aktuellen Situation passt. Darüber hinaus kann hierdurch auch eine automatische Suggestion umgesetzt werden. Ein Nutzer kann so etwa dass System um einen Vorschlag bitten, der zu seiner aktuellen Stimmung und seinen Hörgewohnheiten passt. Sollte dieser nicht zusagen, kann das System dies ebenfalls aus der Emotion des Nutzers erkennen, und einen alternativen Vorschlag unterbreiten.

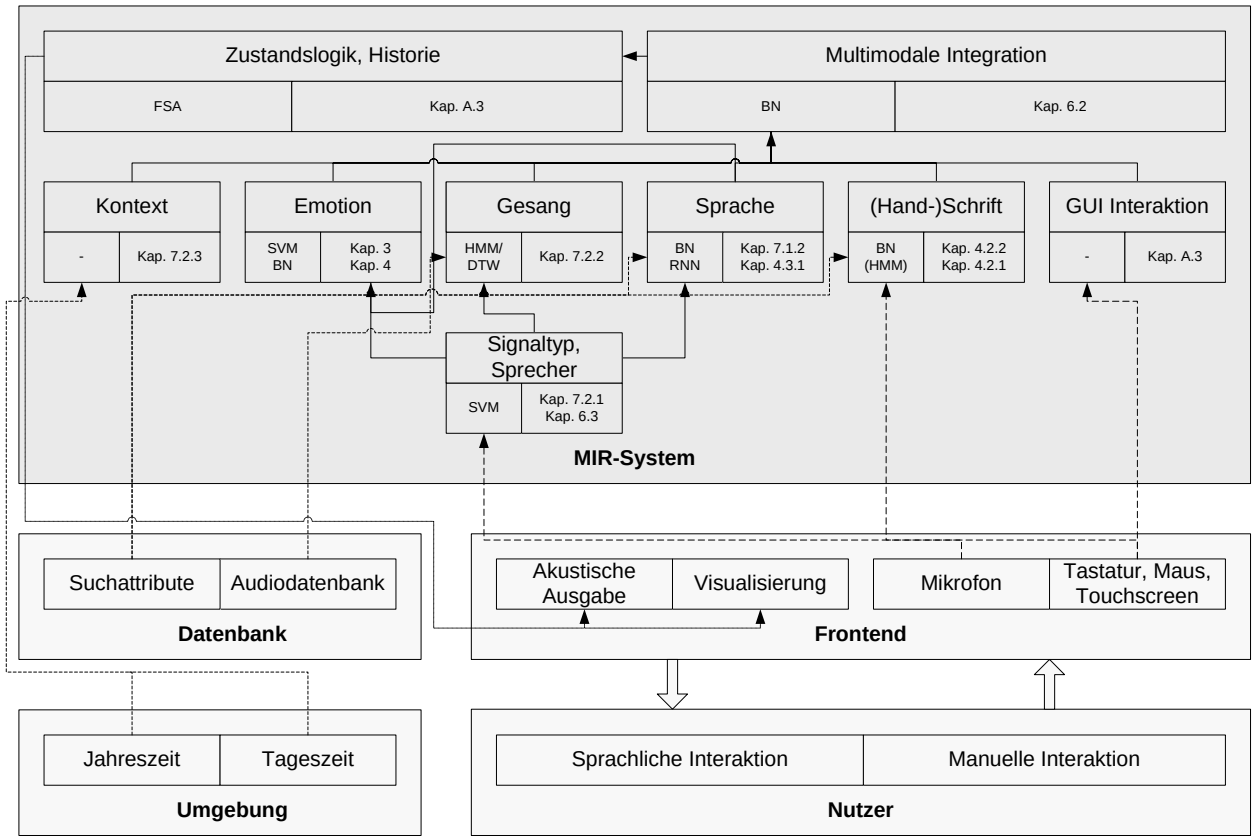


Abb. 7.9: Architekturüberblick multimodales MIR-Interface mit Angabe von Klassifikationsmethode und Kapitelverweisen

7.3 Einsatz im Fahrzeug

Dieser Abschnitt zeigt, wie die entwickelten Verfahren erfolgreich für die Bedienung von Infotainmentdiensten in Fahrzeugen eingesetzt werden können. Dabei spielen sowohl die Erkennung der Fahreremotion, als auch die bereits vorgestellten abgeleiteten Methoden des Erkennens und Verstehens natürlichsprachlicher und handschriftlicher Eingabe, die Suche vorgesungener Melodien und Benutzeridentifikation eine zentrale Rolle.

[184] Die Festlegung auf diese Kontextdaten basiert auf Auswertung einer Nutzerbefragung mit 33 Teilnehmern aus der Hauptzielgruppe, 9 davon weiblich, im durchschnittlichen Alter von 26,0 a (Minimum 21 a, Maximum 35 a). Diese Studie beschäftigte sich mit Hörgewohnheiten im Sinne kontextueller Einflüsse.

Der hierfür entstandene Demonstrator basiert auf dem dreijährigen Kooperationsprojekt *FERMUS*[185]. Ziel dieses Projekts war die Realisierung einer robusten und intuitiven Bedienung eines Infotainmentsystems im Fahrzeug durch massiven Einsatz multimodaler Technologien zur automatischen Erkennung von Benutzerinteraktionen.

7.3.1 Demonstrator Infotainmentbedienung im Fahrzeug

Die in Kap. 7.2.3 vorgestellten Methoden kommen hier in ähnlicher Weise zum Einsatz. Mittels eines Touchscreens und einer manuellen Bedieneinheit mit zwei Drehdrückstellern zur Listennavigation und Lautstärkenregelung sowie acht hintergrundbeleuchteten Drucktastern in Übereinstimmung zur Bildschirmdarstellung kann eine manuelle Bedienung erfolgen. Die Einheiten sind in der Mittelkonsole eines Fahrzeugs angebracht. Es können so durch Direktanwahl per Hand und Menünavigation diverse Informations- und Kommunikationsfunktionalitäten erreicht werden. Der simulierte Umfang an Systemfunktionalität setzt sich dabei aus den Audiodiensten *MP3-Spieler, CD-Spieler* und *Radio* sowie den Kommunikationsdiensten *Telefon, Internet via WAP* und *Maildienste via SMS* zusammen. Abb. 7.10 zeigt hierzu die grafische Benutzeroberfläche, die mit Hilfe von Rapid-Prototyping (siehe Kap. A.4) erstellt wurde sowie die manuelle Bedieneinheit im Konzeptbild.

Abb. 7.10: Grafische Benutzeroberfläche zur Bedienung von Infotainmentdiensten im Fahrzeug (links) und zugehörige Bedienkonsole (rechts) als Konzeptbild

Neben dieser eher traditionell ausgerichteten Bedienung können auch natürlichsprachliche Anweisungen in einem offenen Mikrofonbetrieb an das System gerichtet werden. Es wird dabei verifiziert, ob der Sprecher der Fahrer ist, das System angesprochen wurde und die aktuelle Anweisung vor dem Kontext des Systemzustands sinnvoll erscheint, um Fehlaktivierungen zu vermeiden. Insgesamt werden 60 Basisintentionen über eine Matrix mit 4.440 Einträgen auf 74

[185] Abkürzung für *FEhlerRobuste MUltimodale Sprachdialoge*. Kooperation des Lehrstuhls für Mensch-Maschine-Kommunikation der Technischen Universität München als Auftragnehmer mit der BMW AG, der DaimlerChrysler AG und der Siemens VDO AG.

Grundfunktionen des MMI abgebildet. Das Vokabular umfasst dabei 646 Einträge, mit Aussprachevarianten zur Behandlung von Dialekt. Zur Erstellung des Vokabulars und um Trainingsmaterial zu sammeln, wurden sowohl ein WOO-Test, als auch Versuche mit einem realen Erkenner und insgesamt 10 Probanden durchgeführt (siehe Kap. 7.3.2).

Des Weiteren kann auch eine Anwahl von Musiktiteln durch Vorsingen dieser, wie in Kap. 7.2.2 geschildert, durchgeführt werden. Wie auch bei der Spracheingabe, erlaubt dies eine Suche ohne Blickverlust auf den Verkehr und unter Gewährleistung der Handfreiheit zur Fahrzeugsteuerung. Bei großer Datenbankgröße, wie MP3 von Festplatte, kann der Nutzer Titel einmalig vorsingen, um sie später wieder durch Singen abzurufen. Bei geringeren Datenbankgrößen wie dem Inhalt mehrerer Audio CDs im CD-Wechsler kann dieser Schritt entfallen und ein Vergleich direkt zum originalen Audiotitel erfolgen. Dabei muss eine Merkmalsextraktion bezüglich der CDs vorab zugelassen werden. Die Suche selbst kann dann beliebig oft in Echtzeit erfolgen.

Auf dem in der Mittelkonsole angebrachten Touchscreen kann neben einer Direktwahl auf dem angezeigten GUI mit den Fingern geschrieben werden, um etwa SMS oder Telefonnummern einzugeben. Darüber hinaus können auch hier Berührgesten zur Steuerung vollzogen werden. Hierzu wird die in Kap. 4.2 vorgestellte Zeichen und Strichzugerkennung sowie das Soft-String-Matching mit grafischer Modellierung verwendet, um Toleranz gegenüber fehlerhafter Eingabe zu gewährleisten.

Es ist somit prinzipiell möglich über jede Modalität auf die volle Systemfunktionalität zuzugreifen. Der Anwender kann sich also wie beim multimodalen Interface zur Musiksuche aus Kap. 7.2.3 zu jeder Zeit seine präferierte Modalität wählen und diese parallel und sequentiell kombinieren. Die Integration erfolgt hier wie in Kap. 6.2 geschildert, wobei anstelle von Emotionen mit BN Intentionen modelliert werden.

7.3.2 Anwendung der Emotionserkennung

Ziel der Emotionserkennung ist es hier, im Falle eines verärgerten oder irritierten Fahrers Fehlerauflösungsstrategien einzuleiten oder Hilfe zu leisten. Freude oder Verärgerung nach einer Systemreaktion sollen ferner, wie in Kap. 1.1.1 genannt, zur Adaption weiterer Lernalgorithmen genutzt werden. Die Erkennung erfolgt hier aus der akustischen und linguistischen Analyse, wobei schritthaltend der Fahrer anhand der Stimme verifiziert wird (siehe Kap. 6.3). Auf Grund starker Vibrationen im Fahrzeug während einer realen Fahrt wird hier auf eine affektbezogene Analyse der manuellen Bedienung verzichtet.

Um eine Aussage über die Zuverlässigkeit der vorgestellten Konzepte unter Anwendung realer spontaner Daten treffen zu können, und gleichzeitig zu analysieren wie häufig emotionale Benutzerreaktionen auftreten, wurde eine Studie im Fahrsimulator (siehe 2.5.1 und Kap. A.5) durchgeführt. 10 Versuchpersonen im mittleren Alter von 23,4 a (Minimum 22 a, Maximum 26 a), zwei davon weiblich, nahmen an dem Versuch teil. In zwei gleichlangen Versuchsteilen bedienten diese das in Kap. 7.3.1 vorgestellte Interface über eine reale ASR-Einheit sowie über einen mittels kooperativem WOO simulierten Spracherkenner durch natürliche Sprache in freien Sätzen. Der Versuchsablauf und die Bedienung erfolgten ausschließlich in deutscher Sprache und die

Versuchspersonen sprachen Deutsch jeweils als Muttersprache. Als Versuchsziel wurde den Probanden die Evaluierung zweier Spracherkenner genannt. Die Versuchspersonen wurden in zwei Gruppen eingeteilt, die unterschiedlich mit simuliertem oder realem Erkenner begannen. Die Aufklärung über den Versuchshintergrund erfolgte im Anschluss.

Als primäre Aufgabe wurde den Personen die Führung des Fahrzeuges mit der sekundären Aufgabe der fahrzeugorientierten Kontrolle aufgegeben [GEI85]. Ein unangenehmer Signalton wurde eingespielt, sobald die Fahrspur verlassen wurde. Als tertiäre Aufgabe wurden den Versuchspersonen zu erfüllende Bedienziele automatisch gesteuert über einen fahrzeugexternen Lautsprecher genannt. Die Aufgaben[186] konnten in 45 teilweise komplexen Teilschritten erfüllt werden, und wurden im zweiten Durchlauf in anderer Reihenfolge und mit variierten Parametern wiederholt, so dass jeweils ein sinnvoller Handlungsbogen gewährleistet war. Die generelle Ablaufsteuerung sowie die Simulation der Systemfunktionalität im WOO Teil des Versuchs erfolgten unter Verwendung des in Kap. 2.5 vorgestellten Systems zur semiautomatischen Steuerung von Versuchen. Insbesondere durch Spracherkennungsfehler bedingt kam es bei den Probanden teilweise zu Verärgerungen und Irritationen, es traten jedoch auch erfreute Phrasen auf. Im Display des MMI, wie in Abb. 7.10 ersichtlich, waren zusätzlich drei Knöpfe dargestellt: Hilfebedarf, Verärgerung und Lob für das Interface. Diese wurden von den Probanden regelmäßig genutzt und halfen bei der Annotation der zeitgleich erfolgten Sprachaufzeichnung. Die Versuchsdauer bewegte sich je Teilnehmer zwischen 70 min und 90 min.

2.022 Äußerungen wurden insgesamt während des Versuchs gesammelt. Diese wurden von drei nicht am Versuch teilnehmenden Annotatoren, davon zwei männlich, im Alter von 23 a, 24 a und 30 a, innerhalb des geschlossenen Emotionssets Ärger, Freude, Irritation und Neutralität verschriftet. Für die Auswertung der Erkennungsleistung wurden nur Beispiele mit übereinstimmender Annotation verwendet. Tab. 7.11 zeigt die Verteilung, die sich hieraus ergibt. Diese vermittelt gleichzeitig ein Bild über das Auftreten emotional gefärbter Äußerungen. Ausgehend von der Gesamtzahl, im Vergleich zu der Zahl übereinstimmend emotional gewerteter Äußerungen, ergibt sich ein Verhältnis von unter 20% nicht neutraler Phrasen. Die entsprechende Datenbank wird als EA-CAR bezeichnet. In Tab. 3.44, Kap. 3.12.1 sind weitere Details zu dieser ersichtlich.

Ärger	Freude	Neutralität	Konfusion	Summe
225	25	390	135	**775**

Tab. 7.11: Übersicht Verteilung nach Emotionen Datenbank EA-CAR

Bei der Evaluierung zur automatischen Erkennung auf dieser spontanen Datenbank aus einem

[186] Diese waren im Einzelnen: konkreten Radiosender einstellen, Lautstärke merklich erhöhen, Lautstärke wieder reduzieren, CD-Modus explorieren, von einer bestimmten CD ein bestimmtes Lied abspielen, aus der allgemeinen Liste der MP3-Sammlung ein beliebiges Lied abspielen, Lied anhalten, auf Visitenkarte im Fahrzeug bereit liegende Servicenummer wählen, letzte eingegangenen SMS anzeigen lassen, Verkehrsmeldungen ausgeben lassen sowie Gespräch mit konkretem Telefonbucheintrag einleiten.

angestrebten Einsatzszenario ergeben sich 75,1% Erkennungsleistung bei Wahl der Top 92 akustischen Merkmale mittels SVM SFFS mit SVM als Klassifikator, in einer 10-fach SCV. Dieses Ergebnis entspricht in der Größenordnung den auf der Datenbank AEC erzielten Leistungen, die jeweils unterhalb derer für DES, EMO-DB sowie EA-WSJ liegen (siehe Kap. 3.12.4). Dies beruht auf der Tatsache, dass es sich einerseits um spontane Daten unter realen akustischen Bedingungen handelt. Andererseits variiert, im Gegensatz zu den genannten anderen Korpora, bei AEC und EA-CAR auch der Inhalt der Äußerungen. Dies bringt eine weitere Herausforderung mit sich, da keine Beispiele des gleichen gesprochenen Inhalts im Trainingsmaterial zu einem Testbeispiel in der jeweiligen Emotion vorhanden sind. Insgesamt kann somit unter realen Bedingungen Emotion bereits erkannt werden, jedoch maximal im Bereich menschlichen Urteilsvermögens. In einer Probandenbefragung im Anschluss an den Versuch zur Datensammlung gaben 83,3% der Personen an, dass sie sich ein emotionales System im Einsatz sehr gut vorstellen zu können.

8

Diskussion und Ausblick

„Wissenschaft kann die letzten Rätsel der Natur nicht lösen. Sie kann es deswegen nicht, weil wir selbst ein Teil der Natur und damit auch ein Teil des Rätsels sind, das wir lösen wollen.“

MAX PLANCK (1858 - 1947)

In dieser Arbeit wurden neuartige Verfahren zur robusten Erkennung der Benutzeremotion aus gesprochener und geschriebener Sprache sowie Daten der manuellen Interaktion gezeigt.

Zunächst wurden Einsatzgebiete und Modalitäten hierzu vorgestellt und der Begriff Emotion abgegrenzt sowie Modelle zur technischen Repräsentation derselben vor dem Hintergrund einer Applikation diskutiert. Zielsetzung war, Emotionserkennung im Sinne eines realen Einsatzes robust zu gestalten. Dabei wurde auf die Optimierung jedes wesentlichen Teilaspektes erfolgreichen maschinellen Lernens geachtet: Kollektion repräsentativer Datenbanken, Findung geeigneter Merkmale in adäquater Repräsentationsform sowie Bereitstellung leistungsstarker Ansätze zur Klassifikation.

Zur Behandlung von Korpora wurden Verfahren zur Akquisition und Annotation emotionaler Daten vorgestellt, welche Ausgangspunkt für die angewandten Verfahren der automatischen Mustererkennung sind. Weiterhin wurden in allen Teilbereichen diverse Lernschätze erstellt. Im Sinne eines Vergleichs zu existierenden Arbeiten wurden ferner aktuelle öffentliche Datensammlungen vorgestellt.

Im Rahmen der akustischen Emotionserkennung konnte eine Reihe neuer Merkmale erfolgreich eingesetzt werden, die bisher nicht betrachtet wurden. Es wurde gezeigt, dass eine statische Modellierung einer dynamischen vorzuziehen ist. Hierdurch wird eine weitere Abstraktion erzielt, um die Emotion unabhängig vom gesprochenen Inhalt einer Äußerung erkennen zu können. Darüber hinaus wurde ein Gesamtkonzept zur Verbindung von systematischer Merkmalsgenerierung mit Verfahren der Zeitreihenanalyse, darauf aufsetzender sequentieller gleitender Suche und genetischer Generierung im Verbund mit dem Zielklassifikator vorgestellt. Hierzu wurde unter anderem die letztlich gewählte Gleitende Suche einer Reduktion auf Basis einer

Hauptachsentransformation sowie einer Suche nach Informationsgehalt komparativ gegenübergestellt. Die Stärke des vollständigen Ansatzes zur selbstständigen Optimierung des Merkmalsraumes wurde auf drei öffentlichen Datenbanken unter Erzielung signifikanter Erkennungsleistungssteigerungen verifiziert. Die erzielte kompakte Darstellung bringt neben der Steigerung der Effektivität auch eine Erhöhung der Effizienz im Sinne einer Echtzeitfähigkeit. Insgesamt zeigt sich bei der Wahl geeigneter Größen eine starke Abhängigkeit von der Datenbank. Die finalen Merkmalssets wurden hierzu vorgestellt.

Im Zuge günstiger Klassifikation wurde eine Vielzahl unterschiedlicher Verfahren auf Ihre Eignung hin untersucht: die modellbildenden statistischen Verfahren Naive-Bayes, Bayessche Netze, Hidden-Markov-Modelle und k-Nächste-Nachbarn, instanzbasierte Verfahren sowie die weiteren Verfahren maschinellen Lernens, Künstliche Neuronale Netze und Support-Vektor-Maschinen. Durch Ensembletechniken konnten die Ergebnisse mehrerer Basisklassifikatoren verbessert werden. Als insgesamt stärkster Klassifikator haben sich bei der akustischen Analyse Support-Vektor-Maschinen gezeigt. Insbesondere mit der in dieser Arbeit vorgestellten Entscheidungsbaumbildung anhand der Konfusionsmatrix konnte ein weiterer Zugewinn gegenüber konventionellen Mehrklassenstrategien erzielt werden.

Zur Evaluation wurden diverse Strategien vorgestellt. Hierbei hat sich vor Allem gezeigt, dass eine sprecherunabhängige Erkennung nicht weit hinter sprecherabhängige zurückfallen muss. Im Ergebnis konnte die Leistung durch den gewählten Lösungsansatz auf allen betrachteten öffentlichen Datenbanken gegenüber früherer vergleichbaren Arbeiten auf diesen signifikant gesteigert werden. Ferner konnte eine Zuordnungsleistung in der Größenordnung menschlicher Entscheider für gespielte Daten unter idealisierten akustischen Bedingungen erzielt werden. Die hohe Robustheit des erarbeiteten Ansatzes hierzu zeigt sich vor allem bei Verwendung spontaner Daten unter Feldaufnahmebedingungen.

Zur inhaltlichen Analyse emotionaler Äußerungen wurde die Vektorraumrepräsentation von Text aus dem Bereich der Dokumentanalyse erfolgreich portiert und Phrasensuche mittels grafischer Modellierung als neuer Ansatz vorgestellt. Im direkten Vergleich zum weit verbreiteten N-Gramm Ansatz konnte auch hier jeweils eine Leistungssteigerung erzielt werden. Grundlage dieser Analyse ist eine fehlertolerante Erfassung von Text aus Schrift und Sprache. Hierzu wurde ein String-Matching-Verfahren auf Basis von Bayesschen Netzen präsentiert und mit konventioneller Levenshtein-Distanz verglichen. Es zeigte sich eine klare Überlegenheit im Fall von exklusiver Löschung von Symbolen. Dieser Umstand wurde bei der Kombination des neuen Ansatzes mit einem Phoneme erkennenden Rekurrenten Neuronalen Netz zur Spracherkennung erfolgreich genutzt. Ohne Sprachmodell zeigte sich hier eine Überlegenheit gegenüber konventioneller Methoden bei einem Test auf der Wall Street Journal-Datenbank.

Für die Erkennung von Emotion aus manueller Bedienung über eine Computermaus oder einen berührsensitiven Bildschirm wurde die in der akustischen Analyse bewährte Methodik erfolgreich übernommen. Geeignete Basisverläufe wurden vorgestellt, und es konnte demonstriert werden, dass auch unter Verzicht auf zusätzliche Sensoren eine Erkennung hier möglich ist. Dabei wurde eine Überlegenheit der Analyse aus Mausinteraktionsdaten gegenüber solchen aus der Bedienung von Touchscreens festgestellt.

Der Diskussion dieser singulären Ansatzpunkte folgte die Vorstellung von Strategien zur synergetischen Integration der Analyse einzelner Aspekte. Verfahren der Multistream- und multimodalen Fusion auf Merkmals- sowie semantischer Ebene wurden im Einsatz analysiert. Hierbei hat sich ein klarer Zugewinn für die Fusion linguistischer und akustischer Information gezeigt. Im Hinblick auf ein reales akustisches Umfeld und multiple Nutzer wurde im Weiteren eine Adaption an den aktuellen Anwender zur Steigerung der Robustheit vorgestellt.

Abschließend wurden drei Gebiete zur Demonstration des praxistauglichen Einsatzes und Transfers ausgewählt: robuste Spracherkennung unter emotionalen Verzerrungen, multimodale Musiksuche und Einsatz der Emotionserkennung und multimodaler Bedienung im Fahrzeug.

Im Bereich der Spracherkennung konnte festgestellt werden, dass emotional verzerrte Sprache die Erkennung erwartungskonform negativ beeinflusst. Eine statische Adaption an emotionale Sprechweise kann dies bereits kompensieren. Ein optimales Ergebnis zeigt sich jedoch nur bei einer dynamischen Adaption, die aber eine Erkennung des aktuellen Affekts des Sprechers voraussetzt, um die akustischen Modelle entsprechend anzupassen. Weiterhin wurde gezeigt, wie die vorgestellte grafische Modellierung zur linguistischen Emotionserkennung zur Interpretation natürlicher Sprache verwendet werden kann.

Die statischen Merkmale aus der akustischen Analyse konnten in den Bereich der musikalischen Signaltypdiskrimination zur Trennung von Sprache, monophonem Gesang und polyphoner Musik übertragen werden. Ebenfalls erfolgreich wurde auch die dynamische akustische Modellierung zur Erkennung von Melodien übertragen. Im Anschluss wurde eine Plattform zur multimodalen Suche von Musiktiteln unter Transfer der gezeigten Methodik realisiert. Erkannte Emotion kommt dabei zur Nutzermodellierung und zur Suggestion von Musik nach Gefühlslage des Hörers zum Einsatz.

Als letztes Anwendungsbeispiel wurde ein ebenfalls multimodales Interface zur Bedienung von Infotainmentdiensten im Fahrzeug unter Nutzung vorgestellter Technologie realisiert. Der Affekt des Fahrers wird hier zur Einleitung von Fehlerauflösungsstrategien und Hilfeleistung verwendet. In diesem Bereich wurde ferner eine Datenbank spontaner Emotion aus dem Feld aufgezeichnet. Im finalen Ergebnis konnte auch hier eine generelle Erkennungsleistung im Bereich menschlicher Güte festgestellt werden.

Zusammengefasst wurden somit die Ziele der Arbeit, automatische Emotionserkennung aus sprachlicher und manueller Interaktion robuster zu gestalten, Praxistauglichkeit zu demonstrieren und Transfermöglichkeiten aufzuzeigen, in jedem geplanten Teilbereich erreicht.

Abschließend wird hier noch auf künftige Forschungsaktivitäten hingewiesen, die sich aus dieser Arbeit ergeben. Besonderer Bedarf herrscht an erweiterten Datenbanken, vor allem spontaner Emotionen, mit besonderem Bezug auf die linguistische Analyse, aber auch die manuelle Interaktion. Ferner ist künftig Datenmaterial unter gestörten akustischen Verhältnissen zu betrachten. Darüber hinaus können weitere Sprecherzustände wie Schmerzeinfluss, Müdigkeit, oder Alkoholeinfluss in die Betrachtung mit aufgenommen werden. Da sich emotionale Zustände durchaus auch überlagern können, wie beispielsweise positive oder negative Überraschung, kann eine Klassifikation betrachtet werden, die dies zulässt. Hierzu sind jedoch geeignete Datenbanken

die Vorraussetzung. Neben einer weiteren Steigerung der Personenunabhängigkeit, etwa durch Kategorisierung nach affektiven Typen und automatischer Zuordnung, ist für den künftigen Einsatz auch eine interkulturelle Analyse erforderlich. Hier sind nach [SCE00] Erfolge bei Verwendung eines einheitlichen Modells möglich.

Speziell auf akustischer Ebene kann eine evaluative Abgrenzung zwischen dem hier verfolgten Ansatz einer phrasenweisen Betrachtung und einer auf Phonem- oder Silbenbasis erfolgen. Dabei ist ferner die Kombination dynamischer und statischer Modellierung denkbar.

Bei der semantischen Analyse ist anzunehmen, das sich über die reine Wortwahl hinaus auch auf höherer Ebene ein Ansatzpunkt zur Erkennung bietet. Dies legt ein im Rahmen der Arbeit durchgeführter WOO-Versuch zur gezielten Provokation von Ärger durch simulierte Systemfehler eines sprachbedienten Infotainmentinterfaces im Fahrzeug (siehe Kap. 7.3.1) nahe[187]. Mit steigender Verärgerung der Versuchspersonen sank die Verbosität und die Vokabulargröße, während auffällig viele Wiederholungen auftraten. Als ergänzende Merkmale auf dieser Ebene bieten sich daher beispielsweise die Phrasenlänge in Wörtern, der Grad der Verbosität aus dem Verhältnis der Wortzahl zu der redundanter Termini, der aktive Wortschatz anhand der Vielfalt innerhalb von Lexemen, oder die Zahl unsinniger Wiederholungen oder Widersprüche an. Diese Merkmale setzen jedoch teilweise ein Verständnis auf höherer Ebene oder eine Betrachtung über mehrere Äußerungen voraus. Auch eine akustische Erkennung von Interjektionen wie lachen, räuspern, oder husten kann hier die semantische Analyse stützen. Erfolgt diese auf Basis von geschriebenem Text, kann als Pendant zur akustischen Analyse eine Betrachtung des Drucksatzes wie Fettschrift, oder Kursivschrift einen weiteren Beitrag leisten.

Bezüglich maschinellen Lernens zur Emotionserkennung kann sich der Effekt von Multitask- und lokal gewichtetem Lernen [ATK97] als hilfreich erweisen.

In der multimodalen Fusion ist ein Zugewinn an Erkennungssicherheit vor allem durch Integration physiologischer und mimischer Information [ARS05B], [MÜL05B] sowie weiterer Modalitäten wie das Fahrverhalten im automobilen Bereich oder Einbringung von Kontextwissen oder eine Betrachtung über mehrere Interaktionen zu erwarten. Letzteres kann etwa durch Erweiterung des aktuellen Merkmalsvektors um einen oder mehrere vorhergehende Merkmalsvektoren, Zeitdeltawerte oder Klassifikationsergebnisse geschehen. Hierzu sind jedoch verschriftete Daten aus dem Verlauf heraus erforderlich.

Auf Basis der aufgezeigten erzielbaren Erkennungsgüte und zu erwartender Leistungen des genannten weiteren Forschungsbedarfes kann automatische Emotionserkennung als feste Komponente künftiger Mensch-Maschine-Kommunikation angenommen werden und in vielen praktischen Gebieten erfolgreich eingesetzt werden.

---

[187] Die Versuchspersonen hatten während des Versuchs in einem Fahrzeugsimulator (siehe Kap. 2.5.1) als primäre Aufgabe eine Fahrsimulation zu steuern (siehe Kap. A.5). Durch eine zentrale Ablaufsteuerung (siehe Kap. 2.5.2) wurden den Probanden darüber hinaus Aufgaben zur Bedienung gestellt. Ebenfalls zentral gesteuert wurden gezielt Fehler in der Systemreaktion integriert. 17 Studenten, davon 3 weiblich, im mittleren Alter von 26,0 a ohne größere Standardabweichung, nahmen an der Studie teil (siehe [SCH01], [ALT02A], [ALT02B]).

A

Anhang

A.1 BNF Versuchsablaufsteuerung

<S>	::= <CMD_SEQ>
<CMD_SEQ>	::= <CMD_LINE> | <CMD_LINE> <CMD_SEQ>
<CMD_LINE>	::= <TASK_DEF> | <TU_CMD> | <TS_CMD> | <GUI_CMD> | <WAIT_CMD> | <JUMP_CMD> | <REACT_CMD> | <WIZMSG_CMD>
<TASK_DEF>	::= task <TASK_ID> <TASK_DESC>
<TASK_ID>	::= <chars>
<TASK_DESC>	::= <chars>
<TU_CMD>	::= sim <tu_command> | tu <TU_ID> <tu_command>
<TS_CMD>	::= app <ts_command>| ts <TS_ID> <ts_command>
<GUI_CMD>	::= gui <gui_command> | gui <GUI_ID> <gui_command>
<WAIT_CMD>	::= wait <digits>
<WIZMSG_CMD>	::= wizmsg <chars>
<REACT_CMD>	::= <WIZWAIT_SCR> | <WIZWAIT_CMD>
<WIZWAIT_SCR>	::= wizwait_script | wizwait_script <rcscript> | …
<WIZWAIT_CMD>	::= wizwait_cmd | wizwait_cmd <INTERN_CMD> | …
<INTERN_CMD>	::= <WAIT_CMD> | <TU_CMD> | <TS_CMD> | <GUI_CMD> | <WIZMSG_CMD> | <JUMP_CMD> | <SCRIPT_CMD>
<JUMP_CMD>	::= skip_task | repeat_task
<SCRIPT_CMD>	::= script <rcscript>
<WIZMSG_CMD>	::= wizmsg <chars>
<digits>	= numeric charset
<chars>	= alphanumeric charset
<rcscript>	= run-chart filename
<tu_command>	= cfg-command of a task unit
<ts_command>	= cfg-command of test-system
<gui_command>	= cfg-command of test-wizard GUI

Tab. A.1: BNF der Beschreibungssprache der semiautomatischen Versuchsablaufsteuerung

A.2 Ergebnisse Merkmalsselektion

Rang	IGR	Merkmal	Rang	IGR	Merkmal
1	0,4722	MFCC3Mean	51	0,2297	B3Min
2	0,1529	MFCC7Mean	52	0,1354	MFCC13Min
3	0,2933	MFCC14Mean	53	0,0000	MFCC12De1Max
4	0,2614	MFCC13StdDev	54	0,3891	PitchRange
5	0,3182	MFCC2Mean	55	0,1046	MFCC14De2Max
6	0,2788	MFCC6Mean	56	0,0000	MFCC4Max
7	0,2186	SpecFluxMax	57	0,0907	MFCC12Min
8	0,2185	SpecPoCenDe1Mean	58	0,2261	HNRStdDev
9	0,2119	PitchDe1Max	59	0,2219	MFCC12Mean
10	0,0851	B2Mean	60	0,2787	F6ReDiF0
11	0,3732	PitchStdDev	61	0,0000	MFCC1De2Max
12	0,1065	SpecPoCenMax	62	0,2514	MFCC4Mean
13	0,1670	SpecPoCenDe1StdDev	63	0,1698	MFCC3De1Max
14	0,2292	F1Mean	64	0,0000	SpecPoCenDe1Max
15	0,2224	SpecFluxStdDev	65	0,0925	F1De1Mean
16	0,3635	PitchDe1Mean	66	0,0000	F5Mean
17	0,1621	MFCC5De2Mean	67	0,0000	F2Min
18	0,1921	MFCC1De2Mean	68	0,1097	B1Min
19	0,1319	MFCC9Mean	69	0,2942	F1Min
20	0,2700	MFCC15StdDev	70	0,3098	IntensPosMax
21	0,2233	HNRMean	71	0,2232	MFCC4De2Mean
22	0,2112	SpecFluxMean	72	0,1238	MFCC10Mean
23	0,2669	MFCC6StdDev	73	0,0000	MFCC4De1Max
24	0,1950	RollOffStdDev	74	0,0000	MFCC8De2Mean
25	0,0000	F7Min	75	0,1960	MFCC14StdDev
26	0,2642	MFCC13Mean	76	0,1445	MFCC2StdDev
27	0,1199	F2Mean	77	0,1712	IntensRisetimeMean
28	0,0000	RollOffDe1Max	78	0,0954	SpecPoCenStdDev
29	0,1385	MFCC9De1Mean	79	0,0000	F4De1Max
30	0,1557	MFCC12StdDev	80	0,1499	MFCC10De2Mean
31	0,1596	RateVoiSo	81	0,0000	MFCC12Max
32	0,0000	MFCC10StdDev	82	0,4264	SpecPoCenMean
33	0,1358	RollOffDe1StdDev	83	0,1486	F4Mean
34	0,1595	MFCC11De2Mean	84	0,0000	F5De1Mean
35	0,2318	RollOffMean	85	0,1772	MFCC7De2Mean
36	0,3089	F7ReDiF0	86	0,3024	PitchRePoDistMean
37	0,2529	PitchRePoDistStdDev	87	0,2246	PitchPosMax
38	0,1930	RollOffDe1Mean	88	0,1406	MFCC11De1Mean
39	0,1829	MFCC9De2Mean	89	0,0000	F5StdDev
40	0,3221	MFCC5Mean	90	0,1722	MFCC6De2Mean
41	0,2548	IntensMean	91	0,2702	HNRMax
42	0,1491	B5Mean	92	0,1138	MFCC1Max
43	0,2159	IntensRePoDistMean	93	0,0000	MFCC6De1Max
44	0,1737	RollOffMax	94	0,3183	F4ReDiF0
45	0,3133	MFCC3De1Mean	95	0,0931	MFCC11Min
46	0,3121	MFCC3Max	96	0,0000	DurationSilMed
47	0,2595	B3Mean	97	0,2969	PitchMin
48	0,0889	ElongMean	98	0,0000	F6Mean
49	0,2099	MFCC1De1Mean	99	0,2681	F1ReDiF0
50	0,1781	MFCC8Mean	100	0,0000	MFCC8De1Max

Tab. A.2: Top 100 Merkmale Reihenfolge nach SVM-SFFS, Datenbank EMO-DB, Angabe IGR

Rang	IGR	Merkmal	Rang	IGR	Merkmal
1	0,1314	MFCC3De2Mean	51	0,0588	MFCC1De1Mean
2	0,4176	SpecFluxMax	52	0,0000	SpecPoCenDe1Mean
3	0,3666	ElongMean	53	0,0000	F5StdDev
4	0,1493	MFCC13Mean	54	0,0835	PitchRePoDistMean
5	0,3179	MFCC3Mean	55	0,0703	MFCC7De1Mean
6	0,6124	SpecFluxStdDev	56	0,0000	F7De1Mean
7	0,4288	ElongMed	57	0,0000	SpecPoCenMax
8	0,1014	MFCC4StdDev	58	0,0412	DurationSilMed
9	0,2164	PitchStdDev	59	0,1713	IntensMean
10	0,0000	MFCC6Mean	60	0,0000	SpecPoCenStdDev
11	0,6021	SpecFluxMean	61	0,0000	MFCC11StdDev
12	0,0000	MFCC8Mean	62	0,0000	MFCC4Min
13	0,1328	MFCC7Mean	63	0,0000	B2Range
14	0,0000	F4De1Mean	64	0,0000	SpecPoCenMean
15	0,0000	MFCC14Mean	65	0,1681	MFCC1Max
16	0,0000	RollOffDe1Max	66	0,0000	F2ReDiF0
17	0,0000	MFCC11De2Mean	67	0,0000	HNRStdDev
18	0,0000	RollOffDe1Mean	68	0,0000	B2StdDev
19	0,0751	F1ReDiF0	69	0,0000	B4De1Max
20	0,0000	B5De1Mean	70	0,0722	MFCC7Max
21	0,0000	RollOffDe1StdDev	71	0,1385	HNRMax
22	0,0000	MFCC9De1Mean	72	0,0000	MFCC7Min
23	0,0000	RollOffMean	73	0,1118	MFCC4De2Mean
24	0,0000	MFCC15Mean	74	0,0813	MFCC7De2Mean
25	0,0000	MFCC11De1Mean	75	0,0000	HNRMean
26	0,0804	PitchDe1Max	76	0,0000	MFCC10Min
27	0,0000	MFCC11Mean	77	0,0000	B3De1Mean
28	0,0524	DurationSilMean	78	0,0000	B6StdDev
29	0,0000	SpecPoCenDe1StdDev	79	0,0000	MFCC15De2Max
30	0,0000	RollOffMax	80	0,0000	F4StdDev
31	0,0900	PitchRePoDistStdDev	81	0,0000	IntensPosMax
32	0,0000	MFCC1Min	82	0,0000	MFCC7De1Max
33	0,1436	MFCC3StdDev	83	0,0000	MFCC14De2Max
34	0,0000	RollOffStdDev	84	0,0000	B6Mean
35	0,1694	PitchMean	85	0,0000	F4Range
36	0,0000	MFCC12Min	86	0,0000	MFCC5Min
37	0,0969	MFCC5Mean	87	0,0000	MFCC13De2Max
38	0,1003	MFCC9Mean	88	0,0000	MFCC10Mean
39	0,0000	DurationVoiSoMean	89	0,0000	B3Mean
40	0,0000	MFCC9De2Mean	90	0,0000	IntensFalltimeMed
41	0,1253	MFCC3De1Mean	91	0,0966	MFCC4Mean
42	0,0000	ZeroCrossingRate	92	0,0000	F5De1Mean
43	0,0000	MFCC10De2Mean	93	0,0000	MFCC12De2Max
44	0,0000	SpecPoCenDe1Max	94	0,0000	MFCC1De1Max
45	0,0000	MFCC12Mean	95	0,0000	IntensFalltimeMean
46	0,0000	F2StdDev	96	0,0000	MFCC12Max
47	0,0000	MFCC4De1Max	97	0,0000	MFCC9De2Max
48	0,0000	F7Mean	98	0,0000	MFCC11De2Max
49	0,0000	F3De1Mean	99	0,1378	PitchDe1Mean
50	0,0000	B1De1Mean	100	0,0000	MFCC15De1Max

Tab. A.3: Top 100 Merkmale Reihenfolge nach SVM-SFFS, Datenbank DES, Angabe IGR

Rang	IGR	Merkmal	Rang	IGR	Merkmal
1	0,1025	F3StdDev	55	0,0830	IntensMax
2	0,1681	MFCC2Mean	56	0,0464	F6De1Max
3	0,1255	ZeroCrossingRate	57	0,0478	B3De1Max
4	0,1379	MFCC1Max	58	0,0504	F6StdDev
5	0,1089	IntensMean	59	0,1090	MFCC5Max
6	0,0624	MFCC6Mean	60	0,0471	B1Min
7	0,1129	MFCC1De1Mean	61	0,0755	MFCC2Max
8	0,0761	RollOffStdDev	62	0,0364	F6Mean
9	0,0384	B7De1Mean	63	0,0700	RollOffMax
10	0,1510	ElongMean	64	0,0440	MFCC15Mean
11	0,0866	MFCC10Mean	65	0,0569	MFCC14StdDev
12	0,0691	F7De1Mean	66	0,0843	DurationVoiSoMean
13	0,0468	IntensRisetimeMean	67	0,0861	MFCC3Min
14	0,0950	MFCC4Mean	68	0,1013	F1ReDiF0
15	0,1509	MFCC5Mean	69	0,0400	SpecPoCenDe1StdDev
16	0,0329	MFCC6StdDev	70	0,0465	MFCC14Max
17	0,0977	IntensStdDev	71	0,0552	DurationSilMed
18	0,1011	F2ReDiF0	72	0,1345	F6Max
19	0,0633	MFCC9De1Mean	73	0,0895	RollOffMean
20	0,0629	MFCC3Max	74	0,0000	B5De1Mean
21	0,0314	F7StdDev	75	0,0220	F4StdDev
22	0,0867	MFCC12Mean	76	0,0000	MFCC15De1Max
23	0,0506	MFCC7Mean	77	0,0322	MFCC13Max
24	0,0285	MFCC9Max	78	0,0886	MFCC1Mean
25	0,0364	F7Min	79	0,0818	PitchDe1Max
26	0,0960	MFCC5De2Max	80	0,0000	MFCC13De2Max
27	0,0424	DurationSilMean	81	0,0000	MFCC6De1Max
28	0,1385	ElongMed	82	0,0754	MFCC9De2Mean
29	0,0782	SpecPoCenStdDev	83	0,0472	MFCC14De1Mean
30	0,1168	PitchMean	84	0,0319	SpecPoCenDe1Max
31	0,0542	B3Min	85	0,0324	F1De1Max
32	0,0783	F1StdDev	86	0,0393	MFCC11StdDev
33	0,0303	MFCC7De1Mean	87	0,0000	MFCC7StdDev
34	0,0174	PoSpe250	88	0,0526	MFCC15De2Mean
35	0,0332	MFCC12Max	89	0,0393	F4Range
36	0,1051	MFCC14Mean	90	0,0500	MFCC15StdDev
37	0,0574	MFCC8Mean	91	0,0260	F4Mean
38	0,0877	MFCC3StdDev	92	0,0635	B7Min
39	0,0000	MFCC11Mean	93	0,0629	MFCC1StdDev
40	0,0688	B4Min	94	0,0410	F5De1Max
41	0,0000	MFCC7Max	95	0,0174	PitchRePoDistMean
42	0,0000	MFCC4De2Max	96	0,0654	MFCC10De1Mean
43	0,0693	F1Max	97	0,0529	F2De1Mean
44	0,0495	F1De1Mean	98	0,0000	MFCC10Max
45	0,1612	SpecFluxMean	99	0,0791	IntensRePoDistStdDev
46	0,0530	MFCC4StdDev	100	0,0600	MFCC13StdDev
47	0,0290	MFCC4Max	101	0,0000	MFCC7Min
48	0,0672	IntensRePoDistMean	102	0,0563	MFCC10StdDev
49	0,0338	MFCC3De1Mean	103	0,0378	B3De1Mean
50	0,0319	MFCC4De1Max	104	0,0787	MFCC5StdDev
51	0,0000	B2Mean	105	0,0444	MFCC9Mean
52	0,0498	F4De1Mean	106	0,0216	MFCC14De2Max
53	0,1548	SpecFluxStdDev	107	0,0690	F5Max
54	0,0606	RateVoiSo	108	0,0722	B1Range

Tab. A.4: Top 108 Merkmale Reihenfolge nach SVM-SFFS, Datenbank AEC, Angabe IGR

A.3 Auszüge gesprochener Inhalt

A.3.1 Zuordnungstest linguistische Emotionserkennung

Phrase	Ärger	Ekel	Furcht	Freude	Neutralität	Trauer	Überraschung
a big spider crawls over your face	7,1	**64,3**	25,7	2,9	0,0	0,0	0,0
a car you could ride around in and not cause a stink	16,7	0,0	0,0	25,0	**50,0**	3,3	5,0
a cold coke would be perfect now	0,0	0,0	0,0	**85,7**	14,3	0,0	0,0
a little louder I think one of them may have missed it	30,9	9,1	9,1	0,0	**32,7**	9,1	9,1
a recording - a damned automatic recording	**100,0**	0,0	0,0	0,0	0,0	0,0	0,0
a toast to Bernie	0,0	0,0	0,0	**50,7**	42,7	6,7	0,0
but do you know they will not even let us have it	25,0	0,0	**28,6**	0,0	14,3	30,7	1,4
but I can't have any is that what you mean	30,8	0,0	0,0	7,7	4,6	**49,2**	7,7
but I feel silly telling you	29,3	6,7	20,0	0,0	6,7	**37,3**	0,0
but I used to get my aggression out through those cars all the time	**80,0**	0,0	0,0	6,7	6,7	0,0	6,7
but it just seems that I cant get up	0,0	0,0	26,7	0,0	16,0	**50,7**	6,7
carry on my good man	13,3	0,0	0,0	**44,0**	22,7	13,3	6,7
cause I wanna watch the Knicks on television	**37,3**	6,7	0,0	9,3	**37,3**	6,7	2,7
do I drive uh no I gotta problem with driving	28,6	7,1	**35,7**	0,0	0,0	7,1	21,4
do we have to talk about this kind of crap at the dinner table	**84,0**	0,0	0,0	0,0	6,7	6,7	2,7
do you even know my name screw boy	**46,7**	6,7	4,0	6,7	16,0	6,7	13,3
do you have no concept of the social code	**56,7**	6,7	0,0	0,0	6,7	1,3	28,7
do you listen to this crap	**40,0**	20,0	0,0	0,0	6,7	0,0	33,3
do you realize how paranoid you are	**32,7**	13,3	16,0	6,7	6,7	14,0	10,7
even then I knew they were just jerks	**37,1**	21,4	7,1	4,3	16,4	13,6	0,0
even when you got famous you still distrusted the world	**37,9**	7,1	7,1	0,0	15,7	28,6	3,6
every time I turn around something else happens	16,7	0,0	31,9	0,0	7,2	**32,6**	11,6
he doesn't even do much for me around the house	36,7	0,0	0,0	0,0	15,3	**48,0**	0,0
he doesn't even know who the hell I am	**72,9**	0,0	0,0	0,0	0,0	21,4	5,7
he grasped both my hands	0,0	10,8	15,4	23,1	15,4	7,7	**27,7**
how did it get so big	0,0	2,3	3,8	0,0	12,3	0,0	**81,5**
how disgusting	7,1	**92,9**	0,0	0,0	0,0	0,0	0,0
how do you know his name	0,0	0,0	0,0	0,0	0,0	0,0	**100,0**
how far are you going to	8,6	14,3	18,6	0,0	**32,1**	7,1	19,3
how I don't want you to live with me	**50,7**	14,3	1,4	0,0	7,1	23,6	2,9
how long have you been staying here	0,0	10,0	0,0	0,0	**50,0**	0,0	40,0
I am sorry to have bothered you	3,6	0,0	2,9	0,0	27,9	**65,7**	0,0
I am told that you dance wonderfully well	0,0	0,0	0,0	**82,9**	2,9	0,0	14,3
I love to write sentences	0,0	0,0	0,0	**78,6**	21,4	0,0	0,0
I may throw up	4,0	**40,0**	10,0	10,0	20,0	16,0	0,0
I mean how could you not play anymore that's so strange to me	15,4	0,0	7,7	7,7	7,7	12,3	**49,2**
I mean I can't believe this	21,8	0,0	0,0	0,0	9,1	0,0	**69,1**
let me put it to you this way	18,5	0,0	0,0	0,0	**73,8**	7,7	0,0
let me tell you the man is something else	14,6	0,0	10,0	0,0	**63,1**	12,3	0,0
lets be the best team in the championship	0,0	0,0	0,0	**76,9**	15,4	0,0	7,7
my god I mean you know how I am about insects	16,9	**39,2**	28,5	0,0	7,7	0,0	7,7
my grandfather died of hunger in the old days	4,6	0,0	0,0	7,7	7,7	**80,0**	0,0
oh please stop	19,2	7,7	**46,2**	1,5	2,3	7,7	15,4
oh right right sorry always been a bit of an ass	**38,3**	1,7	12,5	1,7	34,2	11,7	0,0
playing piano all day and then jumping into cold water	0,0	0,0	0,0	**58,5**	15,4	15,4	10,8
but don't you think its right though that you should see him at least once	10,7	7,1	7,1	0,0	16,4	**51,4**	7,1
cause I'm about as tired of your mouth as I am working this stinking hole	**68,6**	14,3	0,0	0,0	0,0	17,1	0,0
every time my heart gets broken it gets splashed across the newspapers as entertainment	16,7	8,7	13,3	0,0	0,0	**61,3**	0,0
I am talking with her now a little but I am very cautious of what I say to her	0,0	7,1	21,4	14,3	**35,7**	11,4	10,0
my mind tends to jump around a little and have some trouble between fantasy and reality	4,6	0,0	4,6	11,5	**35,4**	28,5	15,4

Tab. A.5: Klassifikation von 50 Phrasen der Datenbank EL-FILM hinsichtlich ihrer Emotion durch 15 Probanden. Mehrfachnennungen wurden gewichtet gewertet. Die mittlere maximale Übereinstimmung unter den Probanden beträgt 55,7%.

A.3.2 Trainingssätze EA-WSJ

Neither Bank America nor Mr. Schwab would comment.
The acquisitions are financed through public offerings of the partnerships.
But Mr. Mc Gillicuddy cautioned that too much is being made of the trade deficit.
The plan was announced October nineteenth the day of the stock market crash.
It seems that few people have anything good to say about the recent budget compromise.
The strong yen encouraged more Japanese to travel abroad.
San Diego currently pays sixty five cents quarterly while S. C. E. Pays sixty two cents.
Rates on short term treasury bills fell slightly.
Still the subject is sensitive with corporate America.
This newspaper contacted dozens of companies and leasing brokers.
Most companies took the advice of their tax attorneys and declined comment.
The difference between these figures is the overstatement of the U. S. Trade deficit cited by Canada.
One out of five employees in its remaining businesses have been let go.
The company said Marshall and Ilsley proposed the merger of the two bank holding companies.
As more Japanese auto makers open assembly plants in the U. S.
The trend is exposing U. S. Based suppliers to unprecedented competition.
Energy prices fell zero dot two percent after a zero dot five percent rise in may.
Apparel prices which rose briskly in the spring fell zero dot three percent after being unchanged in may.
This is largely because trees naturally bear less fruit the year after a big crop.
It depends on what happens with the city of Oklahoma City he said.
Sales rose zero two percent in July.
The current boom started sixteen months ago.
They should invest in Mexico he said.
Mexico is safe.
No clearances have been received.
Mr. Murphy declined to comment.

Tab. A.6: Neutral und ärgerlich eingesprochene Testsätze aus der Datenbank EA-WSJ

A.3.3 Sätze der EMO-DB

Der Lappen liegt auf dem Eisschrank.
Das will sie am Mittwoch abgeben.
Heute Abend könnte ich es ihm sagen.
Das schwarze Stück Papier befindet sich da oben neben dem Holzstück.
In sieben Stunden wird es soweit sein.
Was sind denn das für Tüten, die da unter dem Tisch stehen?
Sie haben es gerade hoch getragen und jetzt gehen sie wieder runter.
An den Wochenenden bin ich jetzt immer nach Hause gefahren und habe Agnes besucht.
Ich will das eben wegbringen und dann mit Karl was trinken gehen.
Die wird auf dem Platz sein, wo wir sie immer hinlegen.

Tab. A.7: Gesprochener Inhalt der akustischen Emotionsdatenbank EMO-DB

A.4 MMI Prototyping

Die im Rahmen der Arbeit vorgestellten Benutzerschnittstellen (siehe Kap. 7.2.3 und Kap. 7.3.1) sind mit einer hierfür erstellten generischen Schnittstellenlösung umgesetzt worden (siehe [SCH02A]). Während eine einheitliche Architektur und Kommunikationsstandards für unterschiedliche Schnittstellentypen festgelegt sind, können konkrete Design- und Funktionsvorgaben mittels einer definierten Skriptsprache spezifiziert werden. In ihr sind Verknüpfungen von Elementen zu Aktionen, Ausgaben und Statusmeldungen sowie Bindungen von dynamischen Inhalten vorgesehen. Es können so hierarchische Menüebenen in Baum- oder Netzform und Abkürzungen umgesetzt werden.

Im Besonderen ist ein Integrationskonzept zur Verarbeitung multimodaler Benutzereingaben realisiert, welches die Anbindung diverser Einzelerkennungsmodule mittels BN vorsieht (siehe Kap. 6.2). Die gesamte Kommunikation erfolgt dabei wahlweise über TCP/IP Internetverbindungen, um eine Distribution auf mehrere Plattformen zu ermöglichen, oder über Pipekommunikation für eine schnelle Umsetzung auf einem einzelnen Rechner. Die Kommunikation mit Erkennungsmodulen verläuft prinzipiell bidirektional. Es besteht ein einfaches Kommunikationsprotokoll in Form einer formalen Grammatik, welches im Wesentlichen die Angaben Ziel, Quelle und Nachricht enthält (vgl. Kap. 2.5.2 und Kap. 6.2). Der Nachrichtenkern enthält seitens der Instanzen von Erkennern *n* beste Hypothesen mit Bewertungen über die Konfidenz oder Auskünfte über Störungen. Durch dieses Prinzip ist jederzeit auch eine Fernsteuerung zur Simulation von Systemfunktionalität im Rahmen von WOO Untersuchungen realisierbar.

Die Ausgabe konzentriert sich auf visuelle zweidimensionale und akustische Informationswiedergabe. Hier sind Schablonen vorgegeben, die eine schnelle prototypische Lösung eines Interfaces erlauben. Mögliche Einsatzgebiete sind unter Anderen Infotainmentapplikationen

auf Palmtops, Bordcomputern, elektronischen Infokiosken oder Desktops im Officebereich.

Benutzerintentionen werden manuell durch als Verknüpfung dargestellte Objekte erreicht, die im Fließtext oder als Piktogramm in eine Displayseite eingebunden werden können. Diese Aktionsseiten sind ebenfalls lose verbunden. In den Seiten kann mittels Verknüpfungswahl sowie vorwärts- und rückwärts navigiert werden. Dabei können beliebig viele, zunächst voneinander getrennte Hauptzweige realisiert werden, in die durch einen Direktsprung auf oberste oder zuletzt gewählte Ebenen gewechselt werden kann. Optional können zusätzliche Statuszeilen integriert werden. Über diese kann zum Beispiel die aktuelle Seite oder Systemaktion angezeigt werden. Sie eignen sich insbesondere für die Präsentation dynamischer Inhalte. Als weitere Basisfunktionen stehen eine Hilfe-, eine Undo- und eine Stopp-Funktionalität jederzeit bereit, und erweisen sich gerade beim Einsatz unsicherer Erkennertechnologien als unerlässlich. Hierzu ist die Definition von adäquaten Hilfen und gegenläufigen Aktionen erforderlich.

Ein zu entwickelndes prototypisches System lernt selbstständig die auftretenden Applikationszustände kennen, und speichert in ihnen beobachtete Aktionen und Übergänge ab. In einer ersten Phase können vom Entwickler einmalig möglichst alle vorhandenen Zustände vorgeführt werden. Dies erlaubt später die Priorisierung von als sinnvoll gelernten Aktionen unter Auswertung multimodaler Konfidenzbewertungen. Alternativ kann das System auch untrainiert im realen Einsatz lernen. Ferner werden präferierte Aktionen und die Wahl der Modalitäten aufgezeichnet, um schnelle Zugriffsmöglichkeiten und erwartungsgesteuerte Hypothesenraumbeschränkungen zu ermöglichen und Langzeitstudien zum Gebrauch automatisch protokollieren zu lassen.

Abb. A.8 zeigt einen Überblick zum Konzept sowie eine erstellte Internet-Browser Applikation. Weitere hier vorgestellte Beispiele finden sich in Kap. 7.2.3 und Kap. 7.3.1.

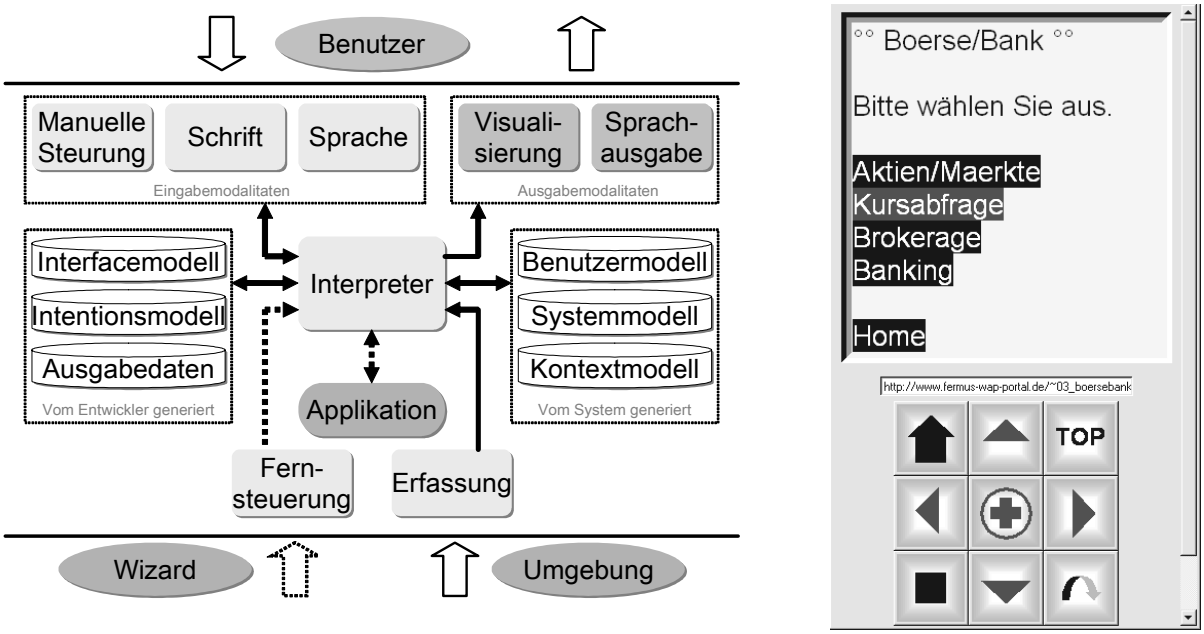


Abb. A.8: Schematischer Architekturüberblick des generischen Benutzerschnittstellenkonzepts (links) und beispielhafter mit dem System erzeugter multimodaler WAP-Browser (rechts)

A.5 Fahraufgabe

Die in den Versuchen in Fahrzeugumgebung verwendete Fahraufgabe (siehe Kap. 7.3) wird, wie in Kap. 2.5.1 beschrieben, an eine senkrechte Leinwand vor dem Fahrzeug projiziert. Abb. 4.9 zeigt hierzu eine Momentaufnahme der Simulation und des zugehörigen Streckeneditors. Die Fahraufgabe stellt keinen Anspruch an eine möglicht realistische Simulation. Vielmehr geht es um eine exakt steuerbare Regelaufgabe, die auf Grund der gewählten perspektivischen Darstellung auch Vorausschau erlaubt.

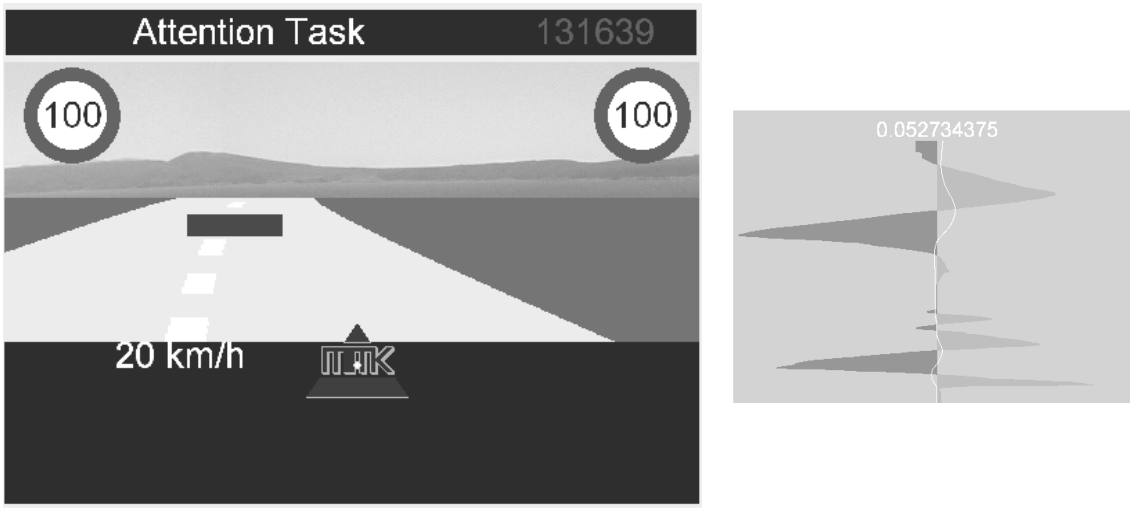


Abb. A.9: Exemplarische Momentaufnahme der Fahraufgabe (links) und des Streckeneditors (rechts)

Die Grundaufgabe besteht für den Probanden in der Spurführung, möglichst entlang der imaginären Ideallinie. Wird die Spur verlassen, wird ein unangenehmer, sich in Frequenz und Intensität kontinuierlich steigender Signalton eingespielt. Vor einem Versuch kann eine Messung zum Fahrkönnen der Versuchspersonen erfolgen. Dieses Fahrvermögen wird in fünf Stufen anhand der kumulativen Abweichung von der Ideallinie bewertet [McG02]. Dabei wird das Schwierigkeitsniveau während einer festen Messzeit variiert. Im Anschluss kann dann während des Hauptversuchs sichergestellt werden, dass keine Überforderung von Probanden besteht. Eine Reihe von Parametern, in Tab. A.10 aufgeführt, erlaubt generell eine Gestaltung nach variabler Anforderung an den Fahrer.

Gruppe	**Einzelparameter**
Fahrzeug	Lenkempfindlichkeit, Höchstgeschwindigkeit, Beschleunigung, Bremsverhalten, Fahrzeugträgheit
Sichtfeld	Sichtbreite, Sichthöhe, Sichtweite
Tracking	Fahrzeugbreite, Straßenbreite, Spurenzahl, Objektverhalten
Visualisierung	Perspektivische Verzerrung, Erscheinungsbild der Straßenmarkierungen und Verkehrszeichen, Jahreszeit, Beleuchtung

Tab. A.10: Parameter der Fahraufgabe

Ferner kann die Fahrt über die Dauer eines Versuchs durch alternierende Umgebung oder Lichtverhältnisse interessanter gestaltet werden. Um Schreck- oder Stresssituationen zu provozieren, können Hindernisse auf der Fahrbahn integriert werden und Signaltöne oder Lichtblitze eingespielt werden. Des Weiteren können Geschwindigkeitsbegrenzungen platziert werden, an die sich die Versuchsperson halten soll.

Die Fahraufgabe lässt sich zentral von der semiautomatischen Versuchsablaufsteuerung aus Kap. 2.5.2 steuern. Es ist somit möglich, spezielle Fahrsituationen gezielt mit Bedienaufgaben zu vereinen. Weiterhin erfolgt die Auswertung über die Fahrleistung synchron mit der restlichen Versuchsprotokollierung. Die Kommunikation verläuft dabei entsprechend über TCP/IP. In Tab. A.11 sind die Kommandos zur Ansteuerung mit zugehörigen Parametern aufgelistet.

Gruppe	**Kommando**	**Beschreibung**
Ablauf	rcfreeze, rcrelease creset do_the_stats terminate	Pausieren, Pause aufheben Zähler rücksetzen Statistikausgabe Baseline-Erhebung Beenden und Protokollausgabe
Akustik	noise <file> enginevol <volume> enginemute engineunmute WizSpeech <file>	Einspielen von Störgeräuschen Motorlautstärke setzen Motorgeräusch aus Motorgeräusch an Einspielen von Ansagen
Effekte	sign <limit> object <x1, x2, y, color> flash <time>	Verkehrszeichen zeigen Objekt auf Fahrbahn Aufblitzen für angegebene Zeit
Fahrt	control <steermode> speed <actualspeed> trackload <file>	Lenkmodus setzen Geschwindigkeit sperren Laden eines Straßenabschnitts
Info	persid <text> show_infotxt <text>	Versuchsperson kennzeichnen Versuchsabschnitt kennzeichnen
Sichtfeld	day, night, light season <seasons>	Tageslicht, Nacht, Lichtmodus Jahreszeit setzen
Parameterdefinition		
actualspeed ::= 0 ... maxspeed color ::= white | yellow | ... | black file ::= text height ::= 1 | 2 | ... | 100 limit ::= 10 | 20 | ... | 130 relpos ::= 0 | 0.1 | 0.2 | ... | 1 seasons ::= winter | summer size ::= xs | s | m | l | xl		steermode ::= nowheel | nogas | all text ::= (a | b | ... | Z) | {text} time ::= 0 | 1 | ... | 10000 volume ::= 0...5000 x1 ::= relpos x2 ::= relpos y ::= size | height

Tab. A.11: Anweisungen zur Ansteuerung der Fahraufgabe in BNF

Die Fahrsimulation wurde auch in einer Reihe weiterer Studien erfolgreich eingesetzt (unter Anderem [ALT02A], [ALT02B]).

Index

Verwendete Abkürzungen

AEC	Aibo Emotion Corpus
AFSM	Adaptive Floating Search Method
AKF	Autokorrelationsfunktion
AMDF	Average Magnitude Difference Function
API	Application Program Interface
ASCII	American Standard Code for Information Interchange
ASR	Automatic Speech Recognition
BN	Bayessches Netz
BNF	Backus Naur Form
BOW	Bag-Of-Words
BPM	Beats-Per-Minute
C4.5	Erweitertes Lernverfahren für Entscheidungsbäume
CPT	Conditional-Probability-Table
DAG	Directed-Acyclic-Graph
DES	Danish Emotional Speech Database
DF	Document Frequency
DCT	Diskrete Cosinus-Transformation
DFT	Diskrete Fourier-Transformation
DT	Entscheidungsbaum (Decision Tree)
DTW	Dynamic-Time-Warping
EA-X	Emotionsdatenbank *X* zur akustischen Analyse
EAL-X	Emotionsdatenbank *X* zur akustischen und linguistischen Analyse
EDR	Electrodermal Response
EL-X	Emotionsdatenbank *X* für linguistische Analyse
EM	Expectation Maximization
EM-X	Emotionsdatenbank *X* mit Daten manueller Bedienung
FFT	Fast-Fourier-Transformation
FSA	Finite-State-Automat
GA	Genetischer Algorithmus
GMM	Gaussian-Mixture-Modell
GSA	General Somatic Activity
GSR	Galvanic Skin Response
GUI	Graphical User Interface
HMM	Hidden-Markov-Modell
HNR	Harmonics-To-Noise Ratio
ID3	Ursprünglicher Entscheidungsbaumalgorithmus
IDF	Inverse Document Frequency
IEEE	Institute of Electrical and Electronics Engineers, Inc.
IG	Information-Gain
IGR	Information-Gain-Ratio
IGR-FS	Information-Gain-Ratio Feature-Selection
IP	Internet Protocol
IRC	Inter-Relate-Chat
JPD	Joint-Probability-Distribution
ISO	International Organization for Standardization
KI	Künstliche Intelligenz
1NN	1-Nächster-Nachbar
kNN	k-Nächste-Nachbarn
ANN	Artificial Neural Network
LBG-VQ	Linde-Buzo-Gray Vektorquantisierer
LD	Levenshtein-Distanz
LDA	Lineare Diskriminanz Analyse
LLD	Low-Level-Descriptor

LOO	Leave-One-Out
LOSO	Leave-One-Speaker-Out
LPC	Linear Prediction Coding
MAP	Maximum-A-Posteriori
MFCC	Mel Frequency Cepstral Coefficient
MIR	Music Information Retrieval
MIR-X	Musikdatenbank *X*
MLE	Maximum-Likelihood-Estimation
MLP	Multi-Layer-Perceptron
MLR	Multiple Lineare Regression
ML-SVM	Multi-Layer SVM
MMI	Mensch-Maschine-Interface
MP3	ISO MPEG 1 Audio-Layer-3
MPEG	Motion Picture Expert Group
MSE	Mean Square Error
MTV	Music Television
OOV	Out-Of-Vocabulary
PC	Principal-Component
PCA	Principal-Component-Analysis
PCA-FS	Principal-Component-Analysis Feature-Selection
PCM	Pulse Code Modulation
PDA	Pitch Detection Algorithmus
PDA	Personal Digital Assistant
PTT	Push-To-Talk
QBH	Query-By-Humming
RAR	Reflexionsarmer Raum
RBF	Radiale Basisfunktion
RNN	Rekurrente Neuronale Netze
SBS	Sequential-Backward-Search
SCV	Stratified Cross Validation
SFBS	Sequential-Floating-Backward-Search
SFFS	Sequential-Floating-Forward-Search
SFS	Sequential-Forward-Search
SFSM	Sequential-Floating-Search-Method
SMA	Simple Moving Average
SMI	Self-Mutual-Information
SMS	Short Message Service
SNR	Signal-To-Noise Ratio
SVM	Support Vector Machine
SVM-SFFS	Sequential-Floating-Forward-Search mit SVM-Wrapper
TCP	Transfer Control Protocol (auch Transmission Control Protocol)
TF	Term-Frequency
TFT	Thin Film Transistor
VQ	Vektorquantisierung
VRML	Virtual Reality Markup Language
WAP	Wireless Access Protocol (auch Wireless Application Protocol)
WSJ	Wall Street Journal
XML	Extendible Markup Language
WDF	Wahrscheinlichkeitsdichtefunktion
WER	Word Error Rate
WOO	Wizard-Of-Oz
ZCR	Zero-Crossing-Rate

Nomenklatur

Die verwendete Nomenklatur ist in Anlehnung an die international jeweils gebräuchlichste Notation gewählt. Auf Grund der Vielzahl diverser behandelter Verfahren kommt es dabei teilweise zur mehrfachen Besetzung des gleichen Symbols. Dieser Umstand wurde dann zu Gunsten eines vereinfachten Vergleichs mit der Literatur in Kauf genommen.

Allgemein

$\underline{x}$	Vektor
x_i	Element i des Vektors $\underline{x}$ mit $i = 1,...,N$
$\underline{A}$	Matrix
$a_{i,j}$	Element i, j der $N \times M$ Matrix $\underline{A}$ mit $\underline{A} = \{a_{i,j}\}$ und $1 \le i \le N,\ 1 \le j \le M$
x	Univariate Werte- oder Zeitreihe mit $x = (x_t)$ und $t = 0,...,T-1$
$\underline{x}$	Multivariate Werte- oder Zeitreihe mit $\underline{x} = (\underline{x}_t)$ und $t = 0,...,T-1$
x_t	Element mit diskretem Zeitindex t der Zeitreihe x und $t = 0,...,T-1$
$\mathbb{R}^n$	Reeller Raum der Dimension n
$\mathbb{N}$	Menge der natürlichen Zahlen
j	Imaginäre Einheit
$\underline{e}_n$	Eigenvektor n
$\underline{E}$	Matrix der Eigenvektoren mit $E = \{\underline{e}_n\}^T$ und $n = 1,...,N$
$\underline{C}_x$	Kovarianzmatrix bezüglich der Vektoren $\underline{x}_l$ mit $l = 1,...,L$
λ_n	Eigenwert n
p	Polynomordnung
f	Funktional
F	Funktionsraum
Φ	Abbildungsvorschrift
$\underline{m}_x$	Mittelpunktvektor der Vektoren $\underline{x}$
$\mu(x)$	Mittelwert der Werte x
$\sigma(x)$	Standardabweichung der Werte x
$c(x)$	Zentroid der Werte x
ε	Fehler
E	Quadratische Fehlersumme
$E(\underline{x})$	Erwartungswert von $\underline{x}$
$g(x, \mu, \sigma)$	Gaußverteilung
L	Likelihood
O	Kardinalität
α	Signifikanzniveau
H_0	Nullhypothese
a	Aktivität
v	Valenz
$\underline{e}$	Emotionaler Raumpunkt

Signalverarbeitung

f	Frequenz
f_s	Samplingfrequenz
$f_{\max}$	Obere Grenzfrequenz
$F[f]$	Überlagerungswert bei dB(A)-Anpassung zur Frequenz f
T	Periodendauer eines periodischen Signals
B	Spektrale Bandbreite
B_f	Filterbreite
$H(f)$	Allgemeine Übertragungsfunktion
H_{LPC}	Übertragungsfunktion eines LPC-Filters
H_{SMA}	Übertragungsfunktion eines SMA-Filters
H_{pre}	Übertragungsfunktion eines Präemphasefilters
α_{pre}	Präemphasefaktor
E_t	Energie im Rahmen mit Zeitindex t
$E_t(f)$	Spektrale Bandenergie bei der Mittenfrequenz f im Rahmen mit Zeitindex t
T_0	Periodendauer eines harmonischen Signals
ZCR_t	Nulldurchgangsrate im Rahmen mit Zeitindex t
$w[n]$	Allgemeine Fensterfunktion
$w_{Han}[n]$	Hanning-Funktion
$w_{Ham}[n]$	Hamming-Funktion
$\overline{x}$	Geglättetes Signal
$\tilde{x}$	Normiertes Signal
$\hat{x}$	Schätzwert
N	Gesamtzahl der Abtastwerte je Rahmen
T_w	Zeitliche Fensterlänge
$s(t)$	Zeitsignal
$s[n]$	Zeitsignal mit diskretisierter Zeitachse
$s_t[n]$	Signalwert zum Abtastwert n im Rahmen mit Zeitindex t
$S_t[f]$	Fourier-Transformierte zu $s_t[n]$
$\underline{S}_t$	Spektraler Vektor mit $\underline{S}_t = \{S_t[f]\}^T$ und $f = 0,..., f_{\max}$
$S_t[u]$	Cosinus-Transformierte zu $s_t[n]$
n	Diskrete Zeitvariable innerhalb eines Rahmens
t	Diskreter Zeitindex auf Rahmenebene
u	Bereich der DCT
$\mathrm{DFT}\{x\}$	Diskrete Fourier-Transformation von x
$\mathrm{DFT}^{-1}\{x\}$	Inverse Diskrete Fourier-Transformation von x
$\mathrm{DCT}\{x\}$	Diskrete Cosinus-Transformation von x

Informationstheorie

$H(\mathcal{L})$	Entropie bezüglich der Menge $\mathcal{L}$
$IG(\mathcal{L},x)$	Information-Gain bezüglich der Menge $\mathcal{L}$ und dem Attribut x
$IGR(\mathcal{L},x)$	Information-Gain-Ratio bezüglich der Menge $\mathcal{L}$ und dem Attribut x
$\mathrm{smi}(\Omega_\kappa, w_i)$	Self-Mutual-Information von w_i bezüglich Ω_κ
$\mathrm{sal}(w_i)$	Salience des Wortes w_i

Allgemeine Mustererkennung

x	Nichtkategoriales Attribut
$\underline{x}$	Merkmalsvektor
$f(\underline{x})$	Muster
U	Umwelt
Ω	Menge der Klassen
Ω_κ	Musterklasse mit $\Omega_\kappa \in \Omega$ und $\kappa = 1,...,K$
Ω_0	Rückweisungsklasse
κ	Klassenindex
κ_e	Index der erkannten Klasse
K	Gesamtzahl der Klassen mit $K = \lvert\Omega\rvert$
$\mathcal{L}$	Menge der Lernvektoren
$\underline{x}_l$	Lernvektor $\underline{x}_l \in \mathcal{L}$ mit Index l und $l = 1,...,L$
$\underline{x}_{l,\kappa}$	Lernvektor $x_l \in \mathcal{L}$ mit Index l und $l = 1,...,L$ der Klasse κ
y_l	Kategoriales Attribut zum Lernvektor mit Index l
$\mathcal{L}_\kappa$	Menge der Lernvektoren der Klasse κ
$\mathcal{L}_j^x$	Teilmenge j der nach Werten des Attributs x partitionierten Lernbeispiele
L	Gesamtzahl der Lernbeispiele mit $L = \lvert\mathcal{L}\rvert$
$d(x)$	Entscheidungsfunktion
$d(\underline{x},\underline{x}_l)$	Distanzfunktion zwischen den Vektoren $\underline{x}$ und $\underline{x}_l$
d_r	Minkowski-Metrik
$d_{\cos}$	Cosinus-Distanzfunktion
$\mathcal{M}$	Menge der Lernmodelle
$\mathcal{M}_m$	Lernmodell m mit $\mathcal{M}_m \in \mathcal{M}$ und $m = 1,...,M$
j	Partitionskoeffizient der *j*-fachen Kreuzvalidierung
$\alpha_{m,\kappa}$	Regressionskoeffizient α bezüglich Lernmodell m und Klasse κ
$MLR(\underline{x})$	Multiple Lineare Regressionsfunktion in Abhängigkeit von $\underline{x}$
$P_{m,\kappa}(\underline{x})$	Konfidenz $P(\underline{x})$ zur Beobachtung $\underline{x}$ bezüglich Lernmodell m und Klasse κ

Support-Vektor-Maschinen

$H(\underline{w}, b)$	Hyperebene mit Normalenvektor $\underline{w}$ und Bias b
K^{Φ}	Kernel-Funktion
$\underline{x}^{sv}$	Stützvektor
$r(\underline{x})$	Punktabstand von $\underline{x}$ zur Hyperebene H
$\mu_{\mathcal{L}}$	Trennbreite μ bezüglich der Datenmenge $\mathcal{L}$
ξ	Schlupfvariable
G	Fehlergewichtungsfaktor
k	Verstärkungsfaktor des Sigmoid-Kernels
Θ	Offset des Sigmoid-Kernels

Neuronale Netze

$w_{i,j}$	Gewicht i, j im MLP
$\underline{W}$	Gewichtsmatrix eines MLP
$T(u)$	Transferfunktion
α	Steigungsparameter der Sigmoid-Funktion
$F(\underline{x}, \underline{W})$	Gütefunktion für das Lernen in Abhängigkeit von $\underline{x}$ und $\underline{W}$
β	Schrittweite beim Gradientenabstieg

Bayessche Netze

X_n	Zustandsvariable des Knotens mit Index n und $n = 1,...,N$
$E_{n,i}$	Zustandsvariable eines Elternknotens von Knoten n mit Index i und $i = 1,...,I$
$K_{n,j}$	Zustandsvariable eines Kindknotens von Knoten n mit Index j und $j = 1,...,J$
$\pi(X_n)$	π-Nachricht des Knotens n
$\lambda(X_n)$	λ-Nachricht des Knotens n
$BEL(X_n)$	Wahrscheinlichkeit des Knotens n
α	Korrekturfaktor zur Normalisierung der Wahrscheinlichkeiten
c	Konfidenzmaß

Dynamische Klassifikation

$\underline{D}$	Distanzmatrix beim DTW mit $\underline{D} = \{d(t,z)\}$ und $t = 0,...,T-1$, $z = (0,...,Z-1)$
λ	HMM / GMM
S_i	HMM-Zustand i
N	Gesamtzahl der Zustände
q_t	Zum Zeitpunkt t aktueller Zustand
$a_{i,j}$	Übergangswahrscheinlichkeit von Zustand S_i in Zustand S_j
$\underline{A}$	Zustandsübergangsmatrix mit $\underline{A} = \{a_{i,j}\}$ und $1 \le i, j \le N$
$b_i(\underline{x}_t)$	Beobachtungswahrscheinlichkeit, dass in Zustand i die Beobachtung $\underline{x}_t$ erfolgt
$\alpha_t(i)$	Partielle Vorwärtswahrscheinlichkeit zum Zeitindex t im Zustand i

$\beta_t(i)$	Partielle Rückwärtswahrscheinlichkeit zum Zeitindex t im Zustand i
C	Codebuchgröße
$\psi_t(j)$	Pfad des Viterbi-Algorithmus
$c_{i,m}$	Mixturgewicht m mit $m = 1,...,M$ im Zustand i
$\xi_{i,j}$	Zustandsokkupation im Zustand i für das Symbol j
$\gamma_{i,j}$	Mixturokkupation im Zustand i für das Symbol j

Sprachverarbeitung

E_k	Kurzzeitenergie
E_l	Langzeitenergie
$F_{0,t}$	Sprachgrundfrequenz im Rahmen mit Zeitindex t
FX_t	Formant X im Rahmen mit Zeitindex t
BX_t	Bandbreite des Formanten X im Rahmen mit Zeitindex t
HNR_t	Harmonische Ausgeprägtheit im Rahmen mit Zeitindex t
$MFCCX_t$	MFCC Koeffizient X im Rahmen mit Zeitindex t
$f_{rop,t}$	Roll-Off-Punkt Frequenz im Rahmen mit Zeitindex t
$S_{flux,t}$	Spektraler Fluss im Rahmen mit Zeitindex t
$Mel[f]$	Mel-Skala in Abhängigkeit von der Frequenz f
w_j	Wort j
$W_{\mathcal{L}}$	Gesamtzahl aller Wörter in $\mathcal{L}$
L_w	Wortlänge in Buchstaben
$\boldsymbol{\mathcal{S}}$	Phrase mit den Wörtern w_j und $j = 1,...,S$
S	Phrasenlänge mit $S = \lvert\boldsymbol{\mathcal{S}}\rvert$
w_i	Wort i des Vokabulars
$\boldsymbol{\mathcal{V}}$	Vokabular mit den Wörtern w_i und $i = 1,...,V$
V	Vokabulargröße mit $V = \lvert\boldsymbol{\mathcal{V}}\rvert$
$TF(w_i, \boldsymbol{\mathcal{S}})$	Häufigkeit des Wortes w_i innerhalb $\boldsymbol{\mathcal{S}}$
$DF(w_i)$	Zahl der Phrasen, in denen das Wort w_i enthalten ist
$IDF(w_i)$	Rate der Phrasen, in denen das Wort w_i enthalten ist, bezogen auf $\mathcal{L}$
λ	Lidstonekoeffizient
P_{MAP}	MAP-Wahrscheinlichkeit
P_{MLE}	MLE-Wahrscheinlichkeit
P_λ	Wahrscheinlichkeit nach Interpolation mit Lidstonekoeffizient λ
$\mathrm{smi}(\Omega_\kappa, w_j)$	SMI des Wortes w_j bezüglich der Klasse Ω_κ
s_n	Zeichensequenzgruppe mit Index n und $n = 1,...,N$
$z_{n,m}$	Zeichensequenz der Sequenzgruppe n mit Index m und $m = 1,...,M$
$D_n(w_i, w_j)$	Abstand in Zeichensequenzen zwischen w_i und w_j

$x_{TF,i}$	Vektorkomponente i bei der Vektorraumrepräsentation mit TF
$x_{TFIDF,i}$	Vektorkomponente i bei der Vektorraumrepräsentation mit TFIDF
$x_{\log TF,i}$	Vektorkomponente i bei der Vektorraumrepräsentation mit logTF
$x_{\log TFIDF,i}$	Vektorkomponente i bei der Vektorraumrepräsentation mit logTFIDF
o_{TF}	Offset-Konstante bei der Berechnung von logTF

Bewegungsanalyse

$\underline{k}$	Bildschirmpunkt mit $k = (k_1, k_2)^T$
$\underline{k}_s$	Startpunkt einer Bewegung
$\underline{k}_e$	Endpunkt einer Bewegung
α	Winkel zwischen Ideallinie einer Bewegung und Abszisse
χ_t	Ortdeltaverlauf mit Zeitindex t
τ_t	Zeitdeltaverlauf mit Zeitindex t
χ_{ges}	Gesamtabweichung einer Bewegung zur Ideallinie
τ_{ges}	Gesamtzeit einer Bewegung
l	Länge der Ideallinie einer Bewegung
z_t	Andruckstärke mit Zeitindex t

Musikverarbeitung

$s_{mon}[n]$	Signalwert des monophonen Audiosignals zum Abtastzeitpunkt n
$s_l[n]$	Signalwert des linken stereophonen Kanals zum Abtastzeitpunkt n
$s_r[n]$	Signalwert des rechten stereophonen Kanals zum Abtastzeitpunkt n
λ_d	Dämpfungsfaktor zur Spitzenkompensation bei Kanaladdition
Δ	Halbe Zahl betrachteter Nachbarbänder bei Partialtonemphasenbestimmung
$E_{p,t}[f]$	Emphase des Partialtons bei der Frequenz f im Rahmen mit Index t
n_p	Partialtonzahl
$H_{p,t}[f]$	Harmonische Ausprägung des Partialtons bei der Frequenz f im Rahmen t

Bibliographie

[ABE00] ABELIN, A.; ALLWOOD, J.: *Cross linguistic interpretation of emotional prosody,* Tagungsband ISCA Workshop on Speech and Emotion: A conceptual framework for research, S. 25-28, Belfast, Irland, 2000.

[ACE99] ACERO, A.: *Formant Analysis and Synthesis using Hidden Markov Models*, Tagungsband Eurospeech ’99, Budapest, Ungarn, 1999.

[ADE03] ADELHARDT, J.; SHI, R. P.; FRANK, C.; ZEIßLER, V.; BATLINER, A.; NÖTH, E.; NIEMANN, H.: *Multimodal User State Recognition in a Modern Dialogue System*, Tagungsband 26th German Conference on Artificial Intelligence, KI '03, S. 591-605, 2003.

[ALT01] ALTHOFF, F.; MCGLAUN, G.; SCHULLER, B.; MORGUET, P.; LANG, M.: *Using Multimodal Interaction to Navigate in Arbitrary Virtual VRML Worlds,* Tagungsband Workshop on Perceptive User Interfaces (PUI 2001), ACM Digital Library, Orlando, Florida, USA, November 2001.

[ALT02A] ALTHOFF, F.; GEISS, K.; MCGLAUN, G.; SCHULLER, B.; LANG, M.: *Experimental Evaluation of User Errors at the Skill-Based Level in an Automotive Environment,* Tagungsband International Conference on Human Factors in Computing Systems CHI 02, ACM SIGCHI, S. 782-783, Minneapolis, USA, April 2002.

[ALT02B] ALTHOFF, F.; MCGLAUN, G.; SCHULLER, B.; LANG, M.; RIGOLL, G.: *Evaluating Misinterpretations during Human-Machine Communication in Automotive Environments,* Tagungsband SCI 2002, 6th World Multiconference on Systemics, Cybernetics and Informatics, Proceedings Vol. VII, "Information Systems Development II", IIIS, S. 13-17, Orlando, Florida, USA, 2002.

[ANG02] ANG, J.; DHILLON, R.; KRUPSKI, A.; SHRIBERG, E.; STOLCKE, A.: *Prosody-Based Automatic Detection of Annoyance and Frustration in Human-Computer Dialog*, Tagungsband ICSLP'02, Denver, Colorado, USA, September, 2002.

[ARK99] ARK, W.; DRYER, D. C.; LU, D.J.: *The Emotion Mouse,* Tagungsband 8th HCI International '99, Ergonomics and User Interfaces, Vol. I, S. 818-823, München, Deutschland, August 1999.

[ARS05A] ARSIC, D.; WALLHOFF, F.; SCHULLER, B.; RIGOLL, G.: *Video Based Online Behavior Detection Using Probabilistic Multi-Stream Fusion*, Tagungsband ICIP 2005, International Conference on Image Processing, IEEE, Genua, Italien, September 2005.

[ARS05B] ARSIC, D.; WALLHOFF, F.; SCHULLER, B.; RIGOLL, G: *Vision-Based Online Multi-Stream Behavior Detection Applying Bayesian Networks*, Tagungsband ICME 2005, 6th International Conference on Multimedia and Expo, IEEE, Amsterdam, Niederlande, Juli 2005.

[ARS05C] ARSIC, D.; WALLHOFF, F.; SCHULLER, B.; RIGOLL, G: *Bayesian Networks Based Multi-Stream Fusion for Automated Online Video Surveillance*, Tagungsband EUROCON 2005, IEEE, Belgrad, Jugoslavien, 2005.

[ARU01] ARUNACHALAM; S., GOULD, D.; ANDERSON, E.; BYRD, D.; NARAYANAN, S. S.: *Politeness and frustration language in childmachine interactions,* Tagungsband Eurospeech, Aalborg, Denmark, 2001.

[ATK97] ATKESON, G.; MOORE, A.; SCHAAL, S.: *Locally Weighted Learning*, in Artificial Learning Review, Vol. 11, Nr. 1-5, S. 11-73, 1997.

[AVE80] AVERILL, J. R.: *A constructivist view of emotion*, in Emotion: Theory, research and experience, Vol. 1, Hrsg.: Plutchik, R. und Kellerman, H., S. 305-339, New York: Academic Press, 1980.

[BAK75] BAKER, J.K.: *The Dragon System – An Overview*, IEEE Trans. ASSP, Bd. ASSP-23 (1975), Vol. 1, S.24-29, 1975.

[BAL96] BALBO, S.: EMA: *Automatic Analysis Mechanism for the Ergonomic Evaluation of user interfaces*, in CSIRO - Division of Information Technology – Sydney, Technical Report 96/44, 1996.

[BAL99] BALL, G.; BREESE, J.: *Modeling the emotional state of computer users*, Tagungsband Workshop on Attitude, Personality and Emotions in User-Adapted Interaction, UM '99, 1999.

[BAT00A] BATLINER, A.; HUBER, R.; NIEMANN, H.; NÖTH, E.; SPILKER, J.; FISCHER, K.: *The Recognition of Emotion*, in Verbmobil: Foundations of Speech-to-Speech Translation, Berlin, Springer-Verlag, 2000.

[BAT00B] BATLINER, A.; FISCHER, K.; HUBER, R.; SPILKER, J.; NÖTH, E..: *Desperately seeking emotions: Actors, wizards, and human beings*, Tagungsband ISCA Workshop on Speech and Emotion, 2000.

[BAT04] BATLINER, A.; HACKER, C.; STEIDL, S.; NÖTH, E.; D'ARCY, S; RUSSELL, M.; WONG, M.: "*You stupid tin box" – Children interacting with the AIBO robot: A cross-linguisitc emotional speech corpus,* Tagungsband LREC 2004, Lissabon, Portugal, 2004.

[BAT05] BATLINER, A.; SEIDL, S.; HACKER, C.; NÖTH, E.; NIEMANN, H.: *Tales of Tuning – Prototyping for Automatic Classification of Emotional User States*, Tagungsband Interspeech 2005, ISCA, S. 489-492, Lissabon, Portugal, 2005.

[BEZ84] VAN BEZOOIJEN, R.: *The Characteristics and recognizability of vocal expression of emotions,* ISBN 311013277X , Drodrecht, Foris Publications, Niederlande, 1984.

[BOE93] BOERSMA, P.: *Accurate Short-Term Analysis of the Fundamental Frequency and the Harmonics-To-Noise Ratio of a Sampled Sound*, in Proceedings of the Institute of Phonetic Sciences, Vol. 17, S. 97-110, Amsterdam, 1993.

[BRA00] BREAZEAL, C.; ARYANANDA, L.: *Recognition of Affective Communicative Intent in Robot-Directed Speech*, Tagungsband 1st International Conference on Humanoid Robots (Humanoids 2000), Cambridge, MA, 2000.

[BRE96] BREIMAN, L.: *Bagging Predictors*, Machine Learning, Vol. 24(2), S. 123-140, 1996.

[BRS98] BREESE, J.; BALL, G.: *Modeling Emotional State and Personality for Conversational Agents,* Technischer Bericht MS-TR-98-41, Microsoft, März 1998.

[BUR00] BURKHARDT, F.: *Simulation emotionaler Sprachweise mit Sprachsyntheseverfahren*, Dissertation, Technische Universität Berlin, 2000.

[CAM02] CAMPBELL, W. N.: *The recording of emotional speech: JST/CREST database research,* Tagungsband LREC 2002, 2002.

[CAR79] CARLSON, R.; GRANSTRÖM, G.; KLATT, D. H.: *Vowel Perception: The Relative Perceptual Salience of Selected Acoustic Manipulations,* in Speech Transmission Lab – Q.Prog.Stat.Report (STL-QPSR), Vol. 2-3, S. 73-83, Royal Institute of Technology, Stockholm, 1979.

[CHA94] CHARWAT, H. J.: *Lexikon der Mensch-Maschine-Kommunikation,* 2. Auflage, Oldenburg Verlag, ISBN 3486226185, S. 291-292, München, 1994.

[CHO01] CHOU, W.; GU, L.: *Robust Singing Detection in Speech/Music Discriminator Design,* Tagungsband ICASSP 2001, International Conference on Acoustics, Speech, and Signal Processing, IEEE, 2001.

[CHR00] CRISTIANINI, N.; SHAWE-TAYLOR, J.: *An Introduction to Support Vector Machines and other kernel-based learning methods*, Cambridge University Press, 2000.

[CHU04] CHUANG, Z.; WU, C.: *Emotion Recognition using Acoustic Features and Textual Content*, Tagungsband ICME 2004, 5th International Conference on Multimedia and Expo, IEEE, Taipei, Taiwan, 2004.

[CLE95] CLEARY, J. G.; TRIGG, L. E.: *K*: An Instance-based Learner Using an Entropic Distance Measure*, Tagungsband 12th International Conference on Machine Learning, S. 108-114, 1995.

[COH03] COHEN, W. W.; RAVIKUMAR, P.; FIENBERG, S. E.: *A Comparison of String Metrics for name-matching tasks*, Tagungsband IJCAI-2003: Workshop on Information Integration on the Web (IIWeb-03), 2003.

[COR96] CORNELIUS, R.: *The Science of Emotion*, Upper Saddle River, New Jersey, Prentice-Hall, 1996.

[COR00] CORNELIUS, R.: *Theoretical approaches to emotion,* Tagungsband ISCA Workshop on Speech and Emotion, Belfast, Irland, 2000.

[COV91] COVER, T. M.; THOMAS, J. A.: *Elements of Information Theory*, John Wiley & Sons, New York, NY, 1991.

[COW95] COWIE, R.; DOUGLAS-COWIE, E.: *Speakers and hearers are people: Reflections on speech deterioration as a consequence of acquired deafness*, in Profound Deafness and Speech Communication, Hrsg.: Spens, K-E. und Plant, G., S. 510-527, London: Whurr, 1995.

[COW99] COWIE, R.; DOUGLAS-COWIE, E.; APOLLONI, B.; TAYLOR, J.; ROMANO, A.; FELLENZ, W.: *What a neural net needs to know about emotion words*, in Computational Intelligence and Applications, Hrsg.: Mastorakis, N., S. 109-114. World Scientific & Engineering Society Press, 1999.

[COW01] COWIE, R.; DOUGLAS-COWIE, E.; TSAPATSOULIS, N.; VOTSIS, G.; KOLLIAS, S.; FELLENZ, W.; TAYLOR, J. G.: *Emotion recognition in human-computer interaction,* IEEE Signal Processing Magazine, Vol. 18, Nr. 1, S. 32-80, Januar 2001.

[DAR72] DARWIN, C.: *The Expression of the emotions in man and animals,* Chicago, University of Chicago Press, 1872 / 1965.

[DAW74] DAWSON, J. L.: *Suffix removal for word conflation*, Bulletin of the Association for Literary & Linguistic Computing, Vol. 2(3), S. 33-46, 1974.

[DAY05] DAWY, Z.; HAGENAUER, J.; HANUS, P.; MUELLER, J. C.: *Mutual Information Based Distance Measures for Classification and Content Recognition with Applications to Genetics,* Tagungsband IEEE International Conference on Communications (ICC 2005), Seoul, Südkorea, Mai 2005.

[DEL96] DELLAERT, F.; POLZIN, T.; WAIBEL, A.: *Recognizing Emotions in Speech*, Tagungsband ICSLP 96, Vol. 3, S. 1970-1973, Philadelphia, PA, USA, 1996.

[DEV03] DEVILLERS, L.; LAMEL, L.: *Emotion Detection in Task-Oriented Dialogs*, Tagungsband ICME 2003, 4th International Conference on Multimedia and Expo, IEEE, Multimedia Human-Machine Interface and Interaction I, Vol. III, S. 549-552, Baltimore, MD, USA, 2003.

[DOU00] DOUGLAS-COWIE, E.; COWIE, R.; SCHRÖDER, M.: *A New Emotion Database: Considerations, Sources and Scope*, Tagungsband ISCA Workshop on Speech and Emotion: A conceptual Framework for Research, S. 39-44, Belfast, Irland, 2000.

[DOU03] DOUGLAS-COWIE, E.; CAMPBELL, N.; COWIE, R.; ROACH, P.: *Emotional Speech: Towards a new generation of databases,* in Speech Communication, Vol. 40, S. 33-60, 2003.

[DUM98] DUMAIS, S.; PLATT, J.; HACKERMAN, D.; SAHAMI, M.: *Inductive learning algorithms and representations for text categorization*, Tagungsband ACM-CIKM'98, November, 1998.

[EAG04] EAGLE, N: *Can Serendipity be Planned?*, in MIT Sloan Journal, Vol. 46, Nr. 1, S. 9-14, 2004.

[EKM72] EKMAN, P.: *Universals and cultural differences in facial expressions of emotion*, in Nebraska Symposium on Motivation 1971, Hrsg.: Cole, J., Vol. 19, S. 207-283, Lincoln, NE: University of Nebraska Press, 1972.

[ELL92] ELLIOTT, C.: *The Affective Reasoner: A Process Model of Emotions in a Multi-agent System*, Dissertation, Northwestern University, The Institute for the Learning Sciences, Technischer Bericht Nr. 32, Mai 1992.

[EMO05] BURKHARDT, F.; PAESCHKE, A.; ROLFES, M.; SENDLMEIER, W.; WEISS, B.: *A Database of German Emotional Speech,* Tagungsband Interspeech 2005, S. 1517-1520, Lissabon, Portugal, 2005.

[ENG96] ENGBERG, I. S.; HANSEN, A. V.: *Documentation of the Danish Emotional Speech Database DES,* Aalborg, Dänemark, September 1996.

[FEL84] FELLBAUM, K.: *Sprachverarbeitung und Sprachübertragung.* Springer-Verlag, Berlin, Heidelberg, New York, Tokyo, 1984.

[FER00] FERNANDEZ, R.; PICARD, R.: *Modeling Drivers' Speech under Stress*, Tagungsband ISCA Workshop ITRW on Speech and Emotion, New Castle, UK, 2000.

[FIN03] FINN, A.; KUSHMERICK, N.: *Learning to classify documents according to genre,* Tagungsband IJCAI-03 Workshop on Computational Approaches to Style Analysis and Synthesis, 2003.

[FRE96] FREUND, Y.; SCHAPIRE, R. E.: *Experiments with a New Boosting Algorithm*, Tagungsband International Conference on Machine Learning, S. 148-156, 1996.

[FUK90] FUKUNAGA, K.: *Introduction to Statistical Pattern Recognition*, Academic Press, S. 400-407, 1990.

[FUL77] FULLER, G.: *BIOFEEDBACK Methods and Procedures in Clinical Practice*, ISBN 0686251385, 1977.

[GAU94] GAUVAIN, J.-L.; LEE, C.-H.: *Maximum A-Posteriori Estimation for Multivariate Gaussian Observations of Markov Chains,* in IEEE Transactions on Speech and Audio Processing, Vol. 2, S. 291-298, 1994.

[GEI85] GEISER, G.: *Mensch-Maschine-Kommunikation im Kraftfahrzeug*, ATZ, 87, Vol. 2, S.74-77, 1985.

[GER02] GERHARD, D.: *Pitch-based acoustic feature analysis for the discrimination of speech and monophonic singing*, in Journal of the Canadian Acoustical Association, Vol. 30:3, S. 152-153, 2002.

[GIL87] GILES, H.; MULAC, A.; BRADAC, J. J.; JOHNSON, P.: *Speech accommodation theory: The first decade and beyond,* in Communication Yearbook, Hrsg.: McLaughlin, M. L., Vol. 10, S. 13-48, Newbury Park, CA: Sage, 1987.

[GOC00] GOECKS, J.; SHAVLIK, J.: *Learning Users' Interests by Unobtrusively Observing Their Normal Behavior,* Tagungsband IUI 2000, S. 129-132, 2000.

[GOE00] GOERTZEL, B.; SILVERMAN, K.; HARTLEY, C.; BUGAJ, S.; ROSS, M.: *The Baby Webmind Project*, Tagungsband AISB 2000, The Annual Conference of The Society for the Study of Artificial Intelligence and the Simulation of Behaviour, 2000.

[GOL89] GOLDBERG, D. E.: *Genetic Algorithms in Search, Optimization & Machine Learning*, Addison-Wesley Publishing Company, Inc., 1989.

[GOU83] GOULD, J. D.; CONTI, J.; HOVANYECZ, T.: *Composing letters with a simulated listening typewriter,* in Communications of the ACM Vol.26, S. 295-308, April 1983.

[GUO02] GUO, H.; HSU, W.: *A Survey of Algorithms for Real-Time Bayesian Network Inference*, Tagungsband AAAI02, 2002.

[HAM05] HAMMAL, Z.; BOZKURT, B.; COUVREUR, L.; UNAY, D.; CAPLIER, A.; DUTOIT, T.: *Passive versus Active : Vocal Classification System,* Tagungsband 13th European Signal Processing Conference EUSPICO 2005, Türkei, 2005.

[HAN97] HANSEN, J.H.L.; BOU-GHAZALE, S.: *Getting Started with SUSAS: A Speech Under Simulated and Actual Stress Database*, Tagungsband EUROSPEECH 97, Vol. 4, S. 1743-1746, Rhodos, Griechenland, September 1997.

[HAO03] HAO, H.; LIU, C., SAKO, H.: *Comparison of Genetic Algorithm and Sequential Search Methods for Classifier Subset Selection,* Tagungsband 7th International Conference on Document Analysis and Recognition, ICDAR 2003, Edinburgh, Schottland, S. 765-769, 2003.

[HAR86] HARRÉ, R.: *The social construction of emotions,* Oxford: Basil Blackwell, 1986.

[HAS98] HASTIE, T.; TIBSHIRANI, R.: *Classification by pairwise coupling*, in The Annals of Statistics, Vol. 26, Nr. 2, S. 451-471, 1998.

[HES05] HESS, T.; SCHULLER, D.: *Business Process Reengineering als nachhaltiger Trend? Eine Analyse der Praxis in deutschen Großunternehmen nach einer Dekade*, in ZFBF Schmalenbachs Zeitschrift für betriebswirtschaftliche Forschung, Jahrgang 57, S. 355-373, Juni 2005.

[HIR05] HIRSCHBERG, J.; BENUS, S.; BRENIER, J. M.; ENOS, F.; FRIEDMANN, S.; GILMAN, S.; GIRAND, C.; GRACIARENA, M.; KATHOL, A.; MICHAELIS, L.; PELLOM, B.; SHRIBERG, E.; STOLCKE, A.: *Distinguishing Deceptive from Non-Deceptive Speech,* Tagungsband Interspeech 2005, ISCA, S. 1833-1836, Lissabon, Portugal, 2005.

[HOC02] HOCHREITER, S; MOZER, M. C.; OBERMAYER, K.: *Coulomb Classifiers: Generalizing Support Vector Machines via an Analogy to Electrostatic Systems*, in Advances in Neural Information Processing Systems, Vol. 15, MIT Press, 2002.

[HOL03] HOLZMAN, L. E.; POTTENGER, W. M.: *Classification of Emotions in Internet Chat: An Application of Machine Learning Using Speech Phonemes*, in Lehigh CSE 2003 Technical Reports, LU-CSE-03-002, 2003.

[ISO98] ISO 9241 PART 11: *Ergonomic requirements for office work with visual display terminals- Guidance on usability*, International Organization for Standardization, Genf, Schweiz, 1998.

[ISO98A] ISO 13407: *Benutzerorientierte Gestaltung interaktiver Systeme*, Normentwurf, Ausgabe 1998-2, Beuth-Verlag, Berlin, 1998.

[ISO98B] ISO N205: *Suitability of TICS for use while driving*, Work Item 17 287, International Organization for Standardization, Genf, 1998.

[JAM84] JAMES, W.: *What is an emotion?,* in Mind, Vol. 19, S. 188-205, 1884.

[JEN96] JENSEN, V. F.: *An introduction to Bayesian Networks*, UCL-Press, Springer-Verlag New York, Inc., ISBN 0387915028, 1996.

[JOA97A] JOACHIMS, T.: *Text Categorization with Support Vector Machines: Learning with Many Relevant Features*, Technischer Bericht LS-8 Report 23, Dortmund, 1997.

[JOA97B] JOACHIMS, T.: *A probabilistic analysis of the Rocchio Algorithm with TFIDF for text categorization*, Tagungsband 14th International Conference on Machine Learning, 1997.

[JOH94] JOHN, G. H.; KOHAVI, R.; PFLEGER, K.: *Irrelevant features and the subset selection problem*, Tagungsband ICML-94, 11th International Conference on Machine Learning, Morgan Kaufmann Publishers, S.121-129, San Francisco, CA, USA, 1994.

[JOH95] JOHN, G. H.; LANGLEY, P.: *Estimating Continuous Distributions in Bayesian Classifiers.* Tagungsband 11th Conference on Uncertainty in Artificial Intelligence, S. 338-345, Morgan Kaufmann, San Mateo, 1995.

[JOL86] JOLLIFE, I.T.: *Principal Component Analysis*, Springer-Verlag, New York, 1986.

[JSN93] JOHANNSEN, G.: *Mensch-Maschine-Systeme*, Springer-Verlag, Berlin Heidelberg New York, ISBN 3-540-56152-8, S. 3-6, 1993.

[KAP04] KAPOOR, A.; PICARD, R.; IVANOV, Y.: *Probabilistic Combination of Multiple Modalities to Detect Interest,* Tagungsband ICPR 2004, International Conference on Pattern Recognition, 2004.

[KIE00] KIENAST, M.; SENDLMEIER, W. F.: *Acoustical analysis of spectral and temporal changes in emotional speech*, Tagungsband ISCA Workshop on Speech and Emotion, Belfast, Irland, 2000.

[KLA00A] KLASMEYER, G.; SENDLMEIER, W. F.: *Voice and Emotional States,* in Voice Quality Measurement, Hrsg.: Kent, R. D. und Ball, M. J., S. 339-357, Singular Publishing Group, 2000.

[KLA00B] KLASMEYER, G.; JOHNSTONE, T.; BÄNZIGER, T.; SAPPOK, C.; SCHERER, K. R.: *Emotional Voice Variability in Speaker Verification*, Tagungsband ISCA Workshop on Speech and Emotion: A conceptual framework for research, Belfast, Irland, 2000.

[KRO93] KROVETZ, R.: *Viewing morphology as an inference process*, Tagungsband 16th ACM SIGIR Conference, S. 191-202, 1993.

[KÜS04] KÜSTNER, D.; TATO, R.; KEMP, T.; MEFFERT, B.: *Towards Real Life Applications in Emotion Recognition: Comparing Different Databases, Feature Sets, and Reinforcement Methods for Recognizing Emotions from Speech*, Tagungsband ADS 2004, S. 25-35, 2004.

[KWO03] KWON, O.-W.; CHAN, K.; HAO, J.; LEE, T. W.: *Emotion Recognition by Speech Signals*, Tagungsband Eurospeech 2003, 2003.

[LAN94] LANG, M.: *Aspects of Human-Machine-Communication,* in Progress and Prospects of Speech Research and Technology, Tagungsband CRIM/FORWISS-Workshop, infix-Verlag, Sankt Augustin, S. 1-8, 1994.

[LAN95] LANG, M; MORGUET, P.: *Skript zum Praktikum Mensch-Maschine-Kommunikation*, Lehrskript, Lehrstuhl für Mensch-Maschine-Kommunikation, Technische Universität München, 1995.

[LAN01] LANG, M.: *1. Kommunikationsdienste und Netze, 2. Informationsdarstellung*, in Handbuch der Ergonomie, Band 2, Teil A-7, Kommunikation und Information, Hrsg.: Bundesamt für Wehrtechnik und Beschaffung, Koblenz, Carl Hanser Verlag, München, 2001.

[LAN02] LANG, M.: *Usability Engineering,* Themenheft der Zeitschrift "it+ti", Informationstechnik und Technische Informatik, Oldenbourg Wissenschaftsverlag, München, Vol. 1/2002, S. 3-4, 2002.

[LAU96] LAURITZEN, S. L.: *Graphical Models*, Oxford University Press Inc, ISBN 0-19-852219-3, New York, 1996.

[LAV80] LAVER, J.: *Principles of Phonetics,* Cambridge University Press, 1994.

[LEE02A] LEE, C. M.; PIERACCINI, R.: *Combining acoustic and language information for emotion recognition*, Tagungsband ICSLP 2002, Denver, CO, USA, 2002.

[LEE02B] LEE, C. M.; NARAYANAN, R.; PIERACCINI, R.: *Classifying Emotion in Human-Machine Spoken Dialogs*, Tagungsband ICME 2002, 3rd International Conference on Multimedia and Expo, Vol. 1, S. 737-740, Lausanne, Schweiz, 2002.

[LEE04] LEE, C.M., NARAYANAN, S.: *Towards Detecting Emotions in Spoken Dialogs*, IEEE Transactions on Speech and Audio Processing, Vol. 13, Nr. 2, März 2004.

[LEV66] LEVENSHTEIN, V.: *Binary codes capable of correcting insertions and reversals*, in Soviet Physics Doklady, Vol. 10, S. 707-710, 1966.

[LID20] LIDSTONE, G.: *Note on the general case of the Bayes-Laplace formula for inductive or a posteriori probabilities*, in Transactions of the Faculty of Actuaries, Vol. 8, S. 182-192, 1920.

[LIE04] LIE, W.-N.; SU, C.-K.: *Content-based Retrieval of MP3 Songs Based on Query by Singing,* Tagungsband ICASSP 2004, International Conference on Acoustics, Speech, and Signal Processing, IEEE, Vol. V, S. 929-932, Montreal, Kanada, 2004.

[LIT03A] LITMAN, D.; FORBES RILEY, K. M.; SILLIMAN, S.: *Towards Emotion Prediction in Spoken Tutoring Dialogues*, Tagungsband Human Language Technology Conference / North American Chapter of the Association for Computational Linguistics (HLT/NAACL 2003), 2003.

[LIT03B] LITMAN, D.; FORBES, K.: *Recognizing Emotions from Student Speech in Tutoring Dialogues*, Tagungsband ASRU 2003, 2003.

[LIU03] LIU, H.; LIEBERMANN, H.; SELKER, T.: *A Model of Textual Affect Sensing using Real-World Knowledge,* Tagungsband 7th International Conference on Intelligent User Interfaces (IUI 2003), S. 125-132, 2003.

[LNG95] LANG, P.: *The emotion probe: Studies of motivation and attention*, in American Psychologist, Vol. 50(5), S.372-385, 1995.

[LOV68] LOVINS, J. B.: *Development of a stemming algorithm*, in Mechanical Translation and Computational Linguistics, Vol. 11, S. 22-31, 1968.

[MAR04] MARTIN, A.; MILLER, D.; PRZYBOCKI, M.; CAMPBELL, J.; NAKASONE, H.: *Conversational telephone speech corpus collection for the NIST speaker recognition evaluation 2004*, Tagungsband 4th International Conference on Language Resources and Evaluation, LREC 2004, ELRA, Lissabon, Portugal, 26-28 May, 2004.

[MCC43] MCCULLOCH, W; PITTS, W.: *A logical calculus of the ideas immanent in nervous activity*, in Bulletin of Mathematical Biophysics, Vol. 9, S. 115-133, 1943.

[MCI00] MCGILLOWAY, S.; COWIE, R.; DOUGLAS-COWIE, E.; GIELEN, C.; WESTERDIJK, M.; STROEVE, S.: *Approaching automatic recognition of emotion from voice: a rough benchmark*, Tagungsband ISCA Workshop on Speech and Emotion, S. 207-212, 2000.

[MCG02] MCGLAUN, G.; ALTHOFF, F.; SCHULLER, B.; LANG, M.: *A new technique for adjusting distraction moments in multi-tasking non-field usability tests,* Tagungsband International Conference on Human Factors in Computing Systems CHI 02, ACM SIGCHI, S. 666-667, Minneapolis, USA, März 2002.

[MEH68] MEHRABIAN, A.: *Communication without words,* in Psychology Today, Vol. 2(9), S. 52-55, 1968.

[MET05A] METACRITIC.COM, http://www.metacritic.com, Stand 26.04.2005.

[MET05B] METACRITIC.COM: *About Metascores – How we calculate our scores,* http://www.metacritic.com/about/scoring.shtml, Stand 26.04.2005.

[MOR97] MORIYAMA, T.; OZAWA, S.: *A measurement of emotional content in speech and its application to cyber commerce*, Tagungsband IEICE Intelligence and Information Symposium, S. 1-5, Dezember 1997.

[MOU96] MOUSSET, E.; AINSWORTH, W. A.; FONOLLOSA, J. A. R.: *A comparison of several recent methods of fundamental frequency and voicing decision estimation*, Tagungsband ICSLP '96, S. 1273-1276, Philadelphia, USA, 1996.

[MOZ98] MOZZICONACCI, S. J. L.: *Speech variability and emotion: production and perception*, Dissertation, Eindhoven, Niederlande, 1998.

[MUE01] MUELLER, F.; LOCKERD, A.: *Cheese: Tracking Mouse Movement Activity on Websites, a Tool for User Modeling*, in Computers that Recognise and Respond to User Emotion: Theoretical and Practical Implications, MIT Media Lab Tech Report 538, Hrsg.: Picard, R.W. und Klein, J., 2001.

[MÜL04] MÜLLER, R.; SCHULLER, B.; RIGOLL, G.: *Enhanced Robustness in Speech Emotion Recognition Combining Acoustic and Semantic Analyses*, Tagungsband Workshop "From Signals To Signs of Emotion and Vice Versa", EU-IST FP6 Network of Excellence HUMAINE, Hrsg.: Inst. of Communication and Computer Systems of the National Techn. University of Athens. S. 4-5, Santorini, Griechenland, September 2004.

[MÜL05A] MÜLLER, R.; SCHULLER, B.; RIGOLL, G.: *Belief Networks in Natural Language Processing for Improved Speech Emotion Recognition,* Tagungsband 2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms, MLMI 2005, Edinburgh, Schottland, Juli 2005.

[MÜL05B] MÜLLER, R; SCHREIBER, S.; SCHULLER, B.; RIGOLL, G.: *A System Structure for Multimodal Emotion Recognition in Meeting Environments*, Tagungsband 2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms, MLMI 2005, Edinburgh, Schottland, Juli 2005.

[NAG02] NAGANO, H.; KASHINO, K.; MURASE, H.: *Fast Music Retrieval using polyphonic binary feature vectors*, Tagungsband ICME 2002, 3rd International Conference on Multimedia and Expo, IEEE, Lausanne, Switzerland, 2002.

[NAS03] NASOZ, F.; LISETTI C.L.; ALVAREZ, K.; FINKELSTEIN, N.: *Emotion Recognition from Physiological Signals for User Modeling of Affect*, Tagungsband 9th International Conference on User Modeling, Johnstown, USA, Juni 2003.

[NAV01] NAVARRO, G.: *A Guided Tour to Approximate String Matching*, in ACM Computing Surveys, Vol. 33(1), S. 31–88, 2001.

[NEM04] NEMESYSCO VOICE ANALYSIS TECHNIQUES, http://www.nemesysco.com, Stand 05.05.2005.

[NIE02] NIESCHULZ, R.; SCHULLER, B., GEIGER, M.; NEUSS, R.: *Aspekte effizienten Usability Engineerings,* in Themenheft der Zeitschrift "it+ti", Schwerpunktthema "Usability Engineering", Oldenbourg Wissenschaftsverlag, München, Vol. 1/2002, S. 23-30, 2002.

[NIE03] NIEMANN, H.: *Klassifikation von Mustern*, 2. überarbeitete Auflage, 2003.

[NIL93] NIELSEN, J.: *Usability Engineering,* Boston Academic Press, ISBN 0-12-518405-0, San Diego, CA, USA, 1993.

[NOG01] NOGUEIRAS, A.; MORENO, A.; BONAFONTE, A.; MARIÑO, J.: *Speech Emotion Recognition Using Hidden Markov Models*, Tagungsband Eurospeech 2001, Poster Proceedings, S. 2679-2682, Skandinavien, 2001.

[NUL05] NULLSOFT, http://www.winamp.com, Stand 18.10.2005.

[ORT88] ORTONY, A; CLORE, G.L.; COLLINS, A.: *The cognitive structure of emotions*, Cambridge University Press, ISBN 0-521-38664-0, UK, 1988.

[ORT90] ORTONY, A.; TURNER, T. J.: *What's basic about basic emotions?,* in Psychological Review, Vol. 97, S. 315-331, 1990.

[OVI00A] OVIATT, S.; COHEN, P.: *Multimodal Interfaces that process what comes naturally*, in Communications of the ACM, Vol. 43, Nr. 3, S. 45-53, März 2003.

[OVI00B] OVIATT, S.: *Multimodal interface research: A science without borders*, Tagungsband ICSLP 2000, 6th International Conference on Spoken Language Processing, Beijing, China, Oktober 2000.

[PAA05] PANAGIOTAKIS, C.; TZIRITAS, G.: *A speech/music discriminator based on RMS and zero-crossings*, in IEEE Transactions on Multimedia, Vol. 7, Nr. 1, Februar 2005.

[PAD96] PADGHAM, L.; TAYLOR, G.: *A system for modelling agents having emotion and personality*, Tagungsband PRICAI Workshop on Intelligent Agent Systems, S. 59-71, 1996.

[PAI90] PAICE, C. D.: *Another Stemmer*, in SIGIR Forum, Vol. 24(3), S. 56-61, 1990.

[PAN03] PANTIC, M; ROTHKRANTZ, L..: *Toward an Affect-Sensitive Multimodal Human-Computer Interaction*, in Proccedings of the IEEE, Vol. 91, S. 1370-1390, September 2003.

[PAU92] PAUL, D. B.; BAKER, J. M.: *The Design for the Wall Street Journal-based CSR Corpus*, Tagungsband DARPA Speech and Natural Language Workshop, S. 357-362, Morgan Kaufmann, Pacific Grove, Kalifornien, USA, 1992.

[PEA88] PEARL, J.: *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann Publishers, Inc., 1988.

[PER00] PEREIRA, C.: *Dimensions of emotional meaning in speech,* Tagungsband ISCA Workshop on Speech and Emotion: A conceptual framework for research, S. 25-28, Belfast, Irland, 2000.

[PEK02] PETKOVIC, M.; MIHAJLOVIC, V.; JONKER, W.; DJORDJEVIC-KAJAN, S.: *Multi-Modal Extraction of Highlights from TV Formula 1 Programs*, Tagungsband ICME 2002, 3rd International Conference on Multimedia and Expo, IEEE, Lausanne, Schweiz, August 2002.

[PET99] PETRUSHIN, V.: *Emotion in Speech: Recognition and Application to Call Centers*, Tagungsband Conference on Artificial Neural Networks in Engineering ANNIE '99, 1999.

[PIC97] PICARD, R. W.; HEALEY, J.: *Affective wearables*, in Personal Technologies, Vol. 1, S. 231-240, 1997.

[PIC98] PICARD, R. W.: *Affective Computing*, zweite Auflage, ISBN 0-262-16170-2, MIT Press, Cambridge, London, England, 1998.

[PIC00] PICARD, R. W.: *Towards computers that recognize and respond to user emotion,* in IBM Systems Journal, Vol. 39, NOS 3&4, S. 705-719, 2000.

[PIK01] PICKENS, J.: *A Survey of Feature Selection Techniques for Music Information Retrieval,* Technischer Bericht, Center for Intelligent Information Retrieval, University of Massachusetts, Amherst, MA, 2001.

[PIW03] PIWEK, P.: *A flexible pragmatics-driven language generator for animated agents*, Tagungsband EACL03, 2003.

[PLU94] PLUTCHIK, R.: *The Psychology and Biology of Emotions*, HarperCollins College, New York, 1994.

[POL00] POLZIN, T. S.; WAIBEL, A.: *Emotion-sensitive human-computer interfaces*, Tagungsband ISCA Workshop, Speech and Emotion, 2000.

[POR80] PORTER, M.: *An algorithm for suffix stripping*, in Program, Vol. 14(3), S. 130-137, 1980.

[PUD94] PUDIL, P; NOVOVIČOVÁ, J.; KITTLER, J.: *Floating search methods in feature selection*, in Pattern Recognition Letters, Vol. 15/11, S. 1119–1125, November 1994.

[QUA98] QUASTHOFF, U.: *Tools for Automatic Lexicon Maintenance: Acquisition, Error Correction, and the Generation of Missing Values*, Tagungsband First International Conference on Language Resources & Evaluation, ELRA 1998, S. 853-856, 1998.

[QUA03] QUASTHOFF, U.: *Projekt Deutscher Wortschatz*, Abteilung Automatische Sprachverarbeitung, Institut für Informatik, Universität Leipzig, http://wortschatz.informatik.uni-leipzig.de, 2003.

[QUI87] QUINLAN, J. R.: *Simplifying Decision Trees*, in International Journal of Man-Machine Studies, Vol. 27, S. 221-234, 1987.

[QUI93] QUINLAN, J. R.: *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, San Mateo, CA, 1993.

[QUI96] QUINLAN, J. R.: *Bagging, Boosting and C4.5*, Tagungsband 14th National Conference on AI, 1996.

[RAB89] RABINER, L. R.: *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*, in Proceedings of the IEEE, Vol. 77, Nr. 2, Februar 1989.

[RAB93] RABINER, L. R.; JUANG, B. H.: *Fundamentals of Speech Recognition,* Prentice-Hall, Englewood Cliffs, New Jersey, 1993.

[RAO03] RAOUZAIOU, A; KARPOUZIS, K.; KOLLIAS, S.: *Emotion Representation for Online Gaming*, Tagungsband ICME 2003, 4th International Conference on Multimedia and Expo, IEEE, Juli 2003.

[REE96] REEVES, B.; NASS, C.: *The Media Equation*, Cambridge University Press, ISBN 1575860538, 1996.

[REI02] REISS, J.; SANDLER, M.: *Benchmarking Music Information Retrieval Systems*, Tagungsband JCDL Workshop on the Creation of Standardized Test Collections, Tasks, and Metrics for Music Information Retrieval (MIR) and Music Digital Library (MDL) Evaluation, Portland, Oregon, Juli 2002.

[RIG99] RIGOLL, G.: *Neuronale Netze*, ISBN 3816909752, Expert-Verlag, März 1999.

[RIG04] RIGOLL, G.: *Pattern Recognition,* Lehrskript, Lehrstuhl für Mensch-Maschine-Kommunikation, Technische Universität München, April 2004.

[RIG05] RIGOLL, G.; MÜLLER, R.; SCHULLER, B.: *Speech Emotion Recognition Exploiting Acoustic and Linguistic Information Sources,* Tagungsband SPECOM 2005, 10th International Conference Speech and Computer, Patras, Griechenland, University of Patras, Vol. 1, S. 61-67, Oktober 2005.

[ROB94] ROBINSON, A. J: *An Application of Recurrent Nets to Phone Probability Estimation*, in IEEE Transactions on Neural Networks, Vol. 5, Issue 2, S. 298-305, März 1994.

[ROS74] ROSS, M.; SHAFFER, H.; COHEN, A.; FREUDBERG, R.; MANLEY, H.: *Average magnitude difference function pitch extractor*, in IEEE Transactions on Acoustics, Speech and Signal Processing, ASSP-22, Vol. 5, S. 353-362, Oktober 1974.

[RUE00] RÜGER, S. M.; GAUCH, S. E.: *Feature Reduction for Document Clustering and Classification*, Technischer Bericht, ISSN 1469-4166, UK, 2000.

[RUM87] RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J..: *Learning internal representations by error propagation,* in Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1, S. 318-362, MIT, 1987.

[RUS88] RUSKE, G.: *Automatische Spracherkennung: Methoden der Klassifikation und Merkmalsextraktion*, Oldenbourg Wissenschaftsverlag, ISBN 3-486-20877-2, München,1988.

[RUS03] RUSKE, G.: *Mustererkennung in der Sprachverarbeitung*, Lehrskript, Lehrstuhl für Mensch-Maschine-Kommunikation, Technische Universität München, 2003.

[SAL88] SALTON, G.; BUCKLEY, C.: *Term Weighting Approaches in Automatic Text Retrieval*, in Information Processing and Management, Vol. 24, Nr. 5, S. 513-523, 1988.

[SAK94] SCHALKOFF, R. J.: *Artificial Neural Networks*, McGraw-Hill, 1994.

[SCA03] SCHAUTEN, D.; KIENDL, H.; MEYER, J.; MACHE, D. H.: *Rekonstruktionsbasierte Selektion relevanter Einflussgrößen*, in Computational Intelligence (Sonderforschungsbereich 531 „Design und Management komplexer technischer Prozesse & Systeme mit Methoden der Computational Intelligence“), ISSN 1433-3325, 2003.

[SCE82] SCHERER, K. R.: *Akustische Parameter der Vokalen Kommunikation*, in Vokale Kommunikation, Hrsg.: Scherer, K. R., Beltzverlag, S. 122-137, 1982.

[SCE00] SCHERER, K., R.: *A cross-cultural investigation of emotion inferences from voice and speech: Implications for speech technology*, Tagungsband ICSLP, Beijing, China, 2000.

[SCH01A] SCHULLER, B.; ALTHOFF, F.; MCGLAUN, G.; GEISS, K.; LANG, M.: *FERMUS Bericht,* Technischer Bericht, Technische Universität München, Juli 2001.

[SCH01B] SCHULLER, B.; ALTHOFF, F.; MCGLAUN, G.; LANG, M.: *Navigation in virtual worlds via natural speech,* Tagungsband 9th International Conference on Human-Computer Interaction (HCI International 2001), Hrsg.: Lawrence Erlbaum Ass., New Jersey, Poster Sessions: Abridged Proceedings, S. 19-21, New Orleans, Louisiana, USA, August 2001.

[SCH02A] SCHULLER, B.; LANG, M.: *Integratives Konzept zur prototypischen Implementierung multimodaler Benutzerschnittstellen - Integrative rapid-prototyping for multimodal user interfaces,* Tagungsband VDI/VDE - GMA Fachtagung USEWARE 2002, Düsseldorf: VDI-Verlag, Hrsg.: VDI, VDI-Berichte, Vol. 1678 "USEWARE 2002 Mensch-Maschine-Kommunikation/Design", S. 279-284, Darmstadt, Deutschland, Juni 2002.

[SCH02B] SCHULLER, B.; LANG, M.; RIGOLL, G.: *Automatic Emotion Recognition by the Speech Signal,* Tagungsband SCI 2002, 6th World Multiconference on Systemics, Cybernetics and Informatics, Vol. IX, "Image, Acoustic, Speech and Signal Processing II", IIIS, S. 367-372, Orlando, Florida, USA, Juli 2002.

[SCH02C] SCHULLER, B.; LANG, M.; RIGOLL, G.: *Multimodal Emotion Recognition in Audiovisual Communication,* Tagungsband ICME 2002, 3rd International Conference on Multimedia and Expo, IEEE, Lausanne, Schweiz, August 2002.

[SCH02D] SCHULLER, B.: *Towards intuitive speech interaction by the integration of emotional aspects,* Tagungsband IEEE International Conference on Systems, Man and Cybernetics, SMC 2002, "Bridging the Digital Divide", Vol. 6, Yasmine Hammamet, Tunesien, Oktober 2002.

[SCH02E] SCHULLER, B.; ALTHOFF, F.; MCGLAUN, G.; LANG, M.; RIGOLL, G.: *Towards Automation of Usability Studies,* Tagungsband IEEE International Conference on Systems, Man and Cybernetics, SMC 2002, "Bridging the Digital Divide", Hrsg.: A. El Kamel, K. Mellouli, P. Borne, Vol. 4, TP1N6, Yasmine Hammamet, Tunesien, Oktober 2002.

[SCH03A] SCHULLER, B.; RIGOLL, G., LANG, M.: *Hidden Markov Model-Based Speech Emotion Recognition.* Tagungsband ICASSP 2003, International Conference on Acoustics, Speech, and Signal Processing, IEEE, Vol. II, S. 1-4, Hong Kong, China, April 2003.

[SCH03B] SCHULLER, B.; RIGOLL, G., LANG, M.: *Hidden Markov Model-Based Speech Emotion Recognition,* Tagungsband ICME 2003, 4th International Conference on Multimedia and Expo, IEEE, Speech Coding, Analysis and Synthesis, Vol. II, S. 401-404, Baltimore, MD, USA, Juli 2003.

[SCH03C] SCHULLER, B.; RIGOLL, G., LANG, M.: *HMM-Based Music Retrieval Using Stereophonic Feature Information and Framelength Adaptation,* Tagungsband ICME 2003, 4th International Conference on Multimedia and Expo, IEEE, Speech and Audio Processing III, Vol. II, S. 713-716, Baltimore, MD, USA, Juli 2003.

[SCH03D] SCHULLER, B.; ZOBL, M.; RIGOLL, G., LANG, M.: *A Hybrid Music Retrieval System using Belief Networks to Integrate Queries and Contextual Knowledge,* Tagungsband ICME 2003, 4th International Conference on Multimedia and Expo, IEEE, Multimedia Human-Machine Interface and Interaction I, Vol. I, S. 57-60, Baltimore, MD, USA, Juli 2003.

[SCH03E] SCHULLER, B.; RIGOLL, G.; LANG, M.: *Sprachliche Emotionserkennung im Fahrzeug,* DGLR Bericht 2003-04, Fachausschusssitzung Anthropotechnik, Entscheidungsunterstützung für die Fahrzeug- und Prozessführung, Universität der Bundeswehr München, Hrsg.: Grandt, M., S. 227-240, Neubiberg, Deutschland, Oktober 2003.

[SCH04A] SCHULLER, B.; RIGOLL, G.; LANG, M.: *Speech Emotion Recognition Combining Acoustic Features and Linguistic Information in a Hybrid Support Vector Machine - Belief Network Architecture,* Tagungsband ICASSP 2004, International Conference on Acoustics, Speech, and Signal Processing, IEEE, Vol. 1, S. 577-580, Montreal, Quebec, Kanada, Mai 2004.

[SCH04B] SCHULLER, B.; MÜLLER, R.; RIGOLL, G.; LANG, M.: *Applying Bayesian Belief Networks in Approximate String Matching for Robust Keyword-based Retrieval,* Tagungsband ICME 2004, 5th International Conference on Multimedia and Expo, IEEE, Taipei, Taiwan, Juni 2004.

[SCH04C] SCHULLER, B.; RIGOLL, G.; LANG, M.: *Emotion Recognition in the Manual Interaction with Graphical User Interfaces,* Tagungsband ICME 2004, 5th International Conference on Multimedia and Expo, IEEE, Taipei, Taiwan, Juni 2004.

[SCH04D] SCHULLER, B.; RIGOLL, G.; LANG, M.: *Discrimination of Speech and Monophonic Singing in Continuous Audio Streams Applying Multi-Layer Support Vector Machines,* Tagungsband ICME 2004, 5th International Conference on Multimedia and Expo, IEEE, Taipei, Taiwan, Juni 2004.

[SCH04E] SCHULLER, B.; RIGOLL, G.; LANG, M.: *Multimodal Music Retrieval for Large Databases*, Tagungsband ICME 2004, 5th International Conference on Multimedia and Expo, IEEE, Taipei, Taiwan, Juni 2004.

[SCH05A] SCHULLER, B.; LANG, M.; RIGOLL, G.: *Robust Acoustic Speech Emotion Recognition by Ensembles of Classifiers*, Tagungsband DAGA'05, 31. Deutsche Jahrestagung für Akustik, DEGA, Strukturierte Sitzung “Automatische Spracherkennung in gestörter Umgebung“, S. 329-330, München, Deutschland, März 2005.

[SCH05B] SCHULLER, B.; RIGOLL, G.; LANG, M.: *Matching Monophonic Audio Clips to Polyphonic Recordings*, Tagungsband DAGA'05, 31. Deutsche Jahrestagung für Akustik, DEGA, Strukturierte Sitzung "Music Processing", S. 299-300, München, Deutschland, März 2005.

[SCH05C] SCHULLER, B.; JIMENEZ VILLAR, R.; RIGOLL, G.; LANG, M.: *Meta-Classifiers in Acoustic and Linguistic Feature Fusion-Based Affect Recognition*, Tagungsband ICASSP 2005, International Conference on Acoustics, Speech, and Signal Processing, IEEE, Philadelphia, PA, USA, März 2005.

[SCH05D] SCHULLER, B.; BRÜNING SCHMITT, B. J.; ARSIC, D.; REITER, S.; LANG, M.; RIGOLL, G.: *Feature Selection and Stacking for Robust Discrimination of Speech, Monophonic Singing, and Polyphonic Music*, Tagungsband ICME 2005, 6th International Conference on Multimedia and Expo, IEEE, Amsterdam, Niederlande, Juli 2005.

[SCH05E] SCHULLER, B.; REITER, S.; MÜLLER, R.; AL-HAMES, M.; LANG, M.; RIGOLL, G.: *Speaker Independent Speech Emotion Recognition by Ensemble Classification*, Tagungsband ICME 2005, 6th International Conference on Multimedia and Expo, IEEE, Amsterdam, Niederlande, Juli 2005.

[SCH05F] SCHULLER, B.; MÜLLER, R.; LANG, M.; RIGOLL, G.: *Speaker Independent Emotion Recognition by Early Fusion of Acoustic and Linguistic Features within Ensembles.* Tagungsband Interspeech 2005, Special Session "Emotional Speech Analysis and Synthesis: Towards a Multimodal Approach", ISCA, S. 805-809, Lissabon, Portugal, 2005.

[SCH05G] SCHULLER, B.; ARSIC, D.; WALLHOFF, F.; LANG, M.; RIGOLL, G.: *Bioanalog Acoustic Emotion Recognition by Genetic Feature Generation Based on Low-Level-Descriptors,* Tagungsband Eurocon 2005, IEEE, Belgrad, Jugoslavien, 2005.

[SCR00] SCHRÖDER, M.: *Experimental study of affect bursts*, in SpeechEmotion 2000, S. 132-137, 2000.

[SCT95] SCHUKAT-TALAMAZZINI, E. G.: *Automatische Spracherkennung – Statistische Verfahren der Musteranalyse*, Vieweg Verlag, 1995.

[SCU01] SCHULTZ, T.; WAIBEL, A.; BETT, M.; METZE, F.; PAN, Y.; RIES, K.; SCHAAF, T.; SOLTAU, H.; WESTPHAL, M.; YU, H.; ZECHNER, K.: *The ISL Meeting Room System,* Tagungsband Workshop on Hands-Free Speech Communication (HSC-2001), Kyoto Japan, April 2001.

[SEE03] SEEWALD, A.: *Towards understanding stacking – Studies of a general ensemble learning scheme*, Dissertation, TU Wien, 2003.

[SEI97] SCHEIRER, E.; SLANEY, M.: *Construction And Evaluation Of A Robust Multifeature Speech/Music Discriminator*, Tagungsband ICASSP 97, International Conference on Acoustics, Speech, and Signal Processing, IEEE, S. 1331-1334, München, Deutschland, 1997.

[SHA63] SHANNON, C.; WEAVER, W.: *The Mathematical Theory of Communication*, University of Illinois Press, ISBN 0252725468, 1963.

[SHV92] SHAVER, P. R.; WU, S.; SCHWARTZ, J. C.: *Cross-cultural similarities and differences in emotion and its representation: A prototype approach*, in Emotion, Hrsg.: Clark, M. S., S. 175-212, Newbury Park: Sage, 1992.

[SIE95] SIEGLE, G.: *The Balanced Affective Word List Project*, http://www.sci.sdsu.edu/cal/wordlist, Stand 1995.

[SMI99] SMITH, S. W.: *The Scientist and Engineer's Guide to Digital Signal Processing*, 2. Auflage, California Technical Publishing, San Diego, CA, 1999, S. 277-284, 1999.

[SOM99] SOMOL, P.; PUDIL, P; NOVOVIČOVÁ, J.; PACLÍK, P.: *Adaptive Floating search methods in feature selection*, in Pattern Recognition Letters, Vol. 20/11-13, Special Issue on Pattern Recognition in Practice VI, S. 1157–1163, November 1999.

[SON02] SONG, J.; BAE, S.; YOON, K.: *Query by humming: Matching humming query to polyphonic audio*, Tagungsband ICME 2002, 3rd International Conference on Multimedia and Expo, IEEE, Lausanne, Schweiz, 2002.

[SOO04] SOOD, S.; KRISHNAMURTHY, A.: *Extraction of characteristic music textures (Eigen-Textures) via graph spectra and Eigen-clusters*, Tagungsband ICASSP 2004, International Conference on Acoustics, Speech, and Signal Processing, IEEE, Vol. IV, S. 229-232, Montreal, Kanada, Mai 2004.

[STA03] STADERMANN, J.; RIGOLL, G.: *Comparing NN Paradigms in Hybrid NN/HMM Speech Recognition using Tied Posteriors*, Tagungsband IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2003), U.S. Virgin Islands, 2003.

[STA05] STADERMANN, J.; RIGOLL, G.: *Two-Stage Speaker Adaptation of Hybrid Tied-Posterior Acoustic Models*, Tagungsband ICASSP 2005, International Conference on Acoustics, Speech, and Signal Processing, IEEE, Philadelphia, USA, 2005.

[STE04] STEIDL, S.; HACKER, C.; RUFF, C.; BATLINER, A.; NÖTH, E.; HAAS, J.: *Looking at the Last Two Turns, I'd Say This Dialogue is Doomed – Measuring Dialogue Success,* Tagungsband TSD, Text, Speech and Dialog, S. 629-636, Tschechei, 2004.

[STI00] STIBBARD, R.: *Automated Extraction of ToBI Annotation Data from the Reeding/Leeds Emotional Speech Corpus,* Tagungsband ISCA Workshop on Speech and Emotion: A conceptual framework for research, S. 60-65, Belfast, Irland, 2000.

[TAT02] TATO, R.; SANTOSA, R.; KOMPE, R.; PARDO, J.: *Emotional Space Improves Emotion Recognition*, Tagungsband ICSLP 2002, International Conference on Spoken Language Processing, 2002.

[TIM00] TIMPE, K.-P.; JÜRGENSOHN, T.; KOHLREP, H.: *Mensch-Maschine-Systemtechnik - Konzepte, Modellierung, Gestaltung, Evaluation*, Symposion Publishing, Düsseldorf, 2000.

[TIN99] TING, K. M.; WITTEN, I. H.: *Issues in Stacked Generalization*, in Journal of Artificial Intelligence Research, Vol. 10, S. 271-289, 1999.

[TOI02] TOIVANEN, J.; SEPPÄNEN, T.: *Prosody-based search features in information retrieval*, in Fonetik 2002, Rapport TMH-QPSR Vol. 44, S. 105-108, Schweden, 2002.

[VAL84] VALINAT, L. G.: *A theory of the learnable*, in Communications of the ACM, Vol. 27, Issue 11, ACM Press, S. 1134-1142, 1984.

[VAN99] VAN SON, R. J. J. H.; POLS, L. W. C.: *An Acoustic Description of Consonant Reduction,* Speech Communication, Vol. 28, S. 125-140, 1999.

[VAP95] VAPNIK, V.; CORTES, C.: *Support vector networks*, Machine Learning, Vol. 20, S.273-297, November 1995.

[VER03] VERVERIDIS, D.; KOTROPOULOS, C.: *A State of the Art Review on Emotional Speech Databases*, Tagungsband 1st Richmedia Conference, S. 109-119, Lausanne, Schweiz, Oktober 2003.

[VER04A] VERVERIDIS, D.; KOTROPOULOS, C., PITAS, I.: *Automatic Emotional Speech Classification*, Tagungsband ICASSP 2004, International Conference on Acoustics, Speech, and Signal Processing, IEEE, S. 593-596, Montreal, Kanada, 2004.

[VER04B] VERVERIDIS, D.; KOTROPOULOS, C.: *Automatic Speech Classification to Five Emotional States Based on Gender Information*, Tagungsband XII. European Signal Processing Conference EUSIPCO 2004, Wien, Österreich, 2004.

[VOG05] VOGT, T.; ANDRE, E.: *Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition,* Tagungsband ICME 2005, 6th International Conference on Multimedia and Expo, IEEE, Amsterdam, Holland, 2005.

[WAF05A] WALLHOFF, F.; SCHULLER, B.; RIGOLL, G.: *Speaker Identification - Comparing Linear Regression Based Adaptation and Acoustic High-Level Features*, Tagungsband DAGA'05, 31. Deutsche Jahrestagung für Akustik, DEGA, München, Deutschland, S. 221-222, März 2005.

[WAF05B] WALLHOFF, F.; ARSIC, D.; SCHULLER, B.; STADERMANN, J.; STÖRMER, A.; RIGOLL, G: *Hybrid Profile Recognition on the MUGSHOT Database*, Tagungsband EUROCON 2005, IEEE, Belgrad, Jugoslavien, 2005.

[WAN00] WANG, C.; SENEFF, S.: *Robust Pitch Tracking for Prosodic Modeling in Telephonic Speech.* Tagungsband ICASSP 2000, International Conference on Acoustics, Speech, and Signal Processing, IEEE, S. 1143-1146, Istanbul, Türkei, 2000.

[WEB00] WEBB, G. I.: *MultiBoosting: A Technique for Combining Boosting and Wagging*, in Machine Learning, Vol. 40, S. 159-198, Kluwer Academic Publishers, Boston, 2000.

[WIT00] WITTEN, I. H.; FRANK, E.: *Data Mining: Practical machine learning tools with Java implementations*, Morgan Kaufmann, San Francisco, S. 133 ff., 2000.

[WOL92] WOLPERT, D. H.: *Stacked generalization*, in Neural Networks, Volume 5, S. 241-259, Pergamon Press, 1992.

[WUT02] WU, T.; KHAN, F. M.; FISHER, T. A.; SHULER, L. A.; POTTENGER, W. M.: *Posting Act Tagging Using Transformation-Based Learning*, Tagungsband Workshop on Foundations of Data Mining and Discovery, IEEE International Conference on Data Mining (ICDM'02), Dezember 2002.

[XUC03] XU, C.; MADDAGE, N.; SHAO, X.; CAO, F.; TIAN, Q.: *Musical Genre Classification Using Support Vector Machines*, Tagungsband ICASSP 2003, International Conference on Acoustics, Speech, and Signal Processing, IEEE, Vol. V, S. 429-432, Hong Kong, China, 2003.

[YAC03] YACOUB, S.; SIMSKE, S.; LIN, X.; BURNS, J.: *Recognition of Emotions in Interactive Voice Response Systems*, Tagungsband Eurospeech 2003, 8th European Conference on Speech Communication and Technology, Genf, Schweiz, 2003.

[YAG97] YANG, Y.; PEDERSEN, J.: *A comparative study on feature selection in text categorization*, Tagungsband International Conference on Machine Learning (ICML), 1997.

[YAN01] YANG, L.; CAMPBELL, N.: *Linking Form to Meaning: The Expression and Recognition of Emotions Through Prosody*, in SSW4 Proceedings, 4th ISCA Tutorial and Research Workshop on Speech Synthesis, 2001.

[YIN96] YING, G. S.; JAMIESON L. H.; MICHELL C. D.: *A Probabilistic Approach to Pitch Detection*, Tagungsband ICSLP '96, Vol. 2, S. 1201-1204, Philadelphia, PA, USA, 1996.

[YOU02] YOUNG, S.; EVERMANN, G.; KERSHAW, D.; MOORE, G.; ODELL, J.; OLLASON, D.; POVEY, D.; VALTCHEV, V.; WOODLAND, P.: *The HTK-Book 3.2*, Cambridge University, Cambridge, England, 2002.

[YUF01] YU, F.; CHANG, E.; XU, Y.; SHUM, H.: *Emotion Detection from Speech to Enrich Multimedia Content*, IEEE Pacific Rim Conference on Multimedia, S. 550-557, 2001.

[ZHA05] ZHANG, L.; JACK, L. B.; NANDI, A. K.: *Extending Genetic Programming for Multi-Class Classification by Combining k-Nearest Neighbor*, Tagungsband ICASSP 2005, International Conference on Acoustics, Speech, and Signal Processing, IEEE, Vol. V, S. 349-352, Philadelphia, USA, 2005.

[ZHE02] ZHE, X.; BOUCOUVALAS, A. C.: *Text-to-Emotion Engine for Real Time Internet Communication*, in International Symposium on CSNDSP 2002, Staffordshire University, S. 164-168, Juli 2002.

[ZOB03] ZOBL, M.; GEIGER, M.; SCHULLER, B.; RIGOLL, G., LANG, M.: *A Realtime System for Hand-Gesture Controlled Operation of In-Car Devices,* Tagungsband ICME 2003, 4th International Conference on Multimedia and Expo, IEEE, Multimedia Human Machine Interface and Interaction I, Vol. III, S. 541-544, Baltimore, MD, USA, 2003.

[ZON97] ZONGKER, D.; JAIN, A.: *Algorithms for feature selection: An evaluation*, Tagungsband International Conference on Pattern Recognition, ICPR 96, S. 18-22, 1996.

[ZWI90] ZWICKER, E.; FASTL, H.: *Psychoacoustics. Facts and Models*, in Series in Information Sciences, Vol. 22, Springer-Verlag, Berlin, 1990.