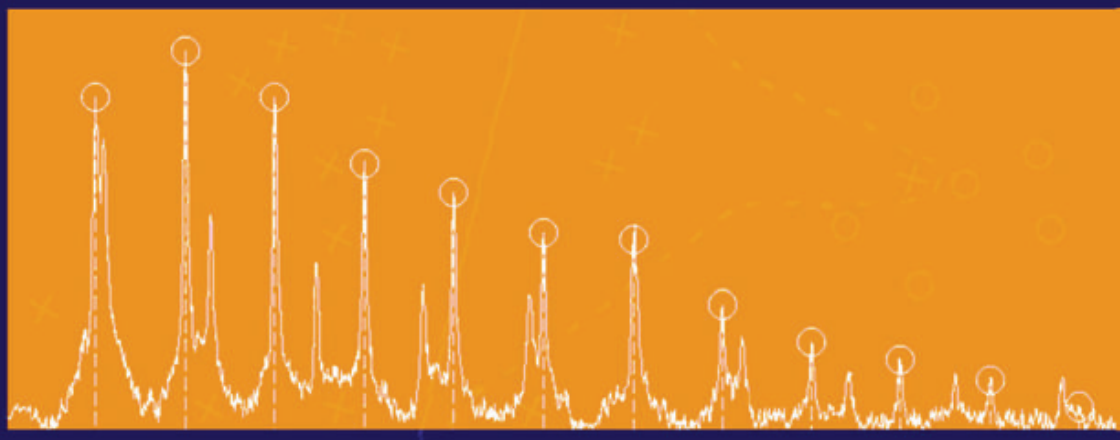
Gunnar Eisenberg

Identifikation und Klassifikation von Musikinstrumentenklängen in monophoner und polyphoner Musik



Cuvillier Verlag

Identifikation und Klassifikation von Musikinstrumentenklängen in monophoner und polyphoner Musik

von Diplom-Ingenieur
Gunnar Eisenberg
aus Berlin

von der Fakultät IV – Elektrotechnik und Informatik
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Ingenieurswissenschaften
– Dr.-Ing. –

genehmigte Dissertation

Berlin 2008

D 83

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.ddb.de abrufbar.
1. Aufl. - Göttingen : Cuvillier, 2008
Zugl.: (TU) Berlin, Univ., Diss., 2008

978-3-86727-825-6

Promotionsausschuss:

Vorsitzender: Prof. Dr. Manfred Opper
1. Gutachter: Prof. Dr. Thomas Sikora
2. Gutachter: Prof. Dr. Stefan Weinzierl

Tag der wissenschaftlichen Aussprache: 7. November 2008

Nonnenstieg 8, 37075 Göttingen
Telefon: 0551-54724-0
Telefax: 0551-54724-21
www.cuvillier.de

1. Auflage, 2008
Gedruckt auf säurefreiem Papier

978-3-86727-825-6

INHALTSVERZEICHNIS

Inhaltsverzeichnis ..... **v**

Abbildungsverzeichnis ..... **xi**

Tabellenverzeichnis ..... **xvii**

Abkürzungsverzeichnis ..... **xix**

1 Einleitung **1**

1.1 Inhaltsbezogene Datensuche 1

1.2 Musikinstrumentenerkennung 4

1.3 Studiotechnologie 5

1.4 Kapitelübersicht 6

2 Musikterminologie **9**

2.1 Klangereignisse 9

2.1.1 Anschlag 10

2.1.2 Hüllkurve 10

2.1.3 Einordnung 13

2.2 Hörereignisse 15

2.2.1 Lautheit 15

2.2.2 Tondauer 16

2.2.3 Tonhöhe 16

2.2.4 Klangfarbe 16

2.3 Musikparameter 17

2.3.1 Zählzeiten ... 18
2.3.2 Tempo ... 18
2.3.3 Dynamik ... 19
2.3.4 Rhythmus ... 20
2.3.5 Intervalle ... 20
2.3.6 Harmonie ... 22
2.3.7 Melodie ... 22
2.3.8 Klang ... 23
2.4 Tonarten ... 23
2.4.1 Stammtonreihe ... 23
2.4.2 Gleichstufige Chromatik ... 25
2.4.3 Gängige Tonarten ... 27
2.4.4 Cent ... 28
3 Musikinstrumente ... 29
3.1 Systematiken ... 29
3.1.1 Genre ... 30
3.1.2 Spielart ... 30
3.1.3 Hornbostel-Sachs ... 31
3.1.4 Klangcharakteristik ... 32
3.2 Instrumentenfamilien ... 34
3.2.1 Flöteninstrumente ... 34
3.2.2 Rohrblattinstrumente ... 36
3.2.3 Trompeteninstrumente ... 39
3.2.4 Harmonikainstrumente ... 42
3.2.5 Streichinstrumente ... 45
3.2.6 Zupfinstrumente ... 47
3.2.7 Schlaginstrumente ... 50
4 Merkmalsextraktion ... 53
4.1 Merkmale ... 53
4.1.1 Aufbau ... 53

4.1.2 Anforderungen 54
4.2 Extraktionsszenarien 54
4.2.1 Monophone Musiksignale 55
4.2.2 Eingeschränkt Polyphone Musiksignale 55
4.2.3 Komplexe Polyphone Musiksignale 55
4.3 Zeitbereichsmerkmale 56
4.3.1 Nulldurchgangsrate 56
4.3.2 Effektivwert 57
4.3.3 Hüllkurve 58
4.3.4 Hüllkurvenparameter 58
4.3.5 Lineare Prädiktion 59
4.4 Spektrale Merkmale 60
4.4.1 Spectral Centroid 61
4.4.2 Spectral Spread 62
4.4.3 Spectral Skewness 63
4.4.4 Spectral Kurtosis 64
4.4.5 Spectral Flatness 66
4.4.6 Spectral Crest Factor 67
4.4.7 Spectral Rolloff 67
4.4.8 Spectral Flux 68
4.4.9 Audio Spectrum Envelope 69
4.4.10 Mel-scale Frequency Cepstrum Coefficients 70
4.5 Harmonische Merkmale 73
4.5.1 Monophone Grundfrequenzerkennung 73
4.5.2 Polyphone Grundfrequenzerkennung 74
4.5.3 Harmonic Peak Spectrum 76
4.5.4 Harmonizität 78
4.5.5 Rauschartigkeit 79
4.5.6 Harmonic Spectral Deviation 80
4.5.7 Harmonic Inner-Ratio 81

4.6 Dimensionsreduktion ... 82
4.6.1 Fluch der Dimensionalität ... 83
4.6.2 Matrixfaktorisierung ... 84
4.6.3 Hauptkomponentenanalyse ... 85
4.6.4 Singulärwertzerlegung ... 89
4.6.5 Analyse der unabhängigen Komponenten ... 93
4.6.6 Nicht-negative Matrixfaktorisierung ... 96

5 Klassifikation ... 99

5.1 Klassifikationsverfahren ... 100
5.1.1 Syntaktische Verfahren ... 100
5.1.2 Statistische Verfahren ... 101
5.1.3 Neuronale Verfahren ... 101
5.2 Lernverfahren ... 102
5.2.1 Überwachtes Lernen ... 102
5.2.2 Unüberwachtes Lernen ... 102
5.2.3 Kontinuierliches Lernen ... 103
5.2.4 Überanpassung ... 103
5.3 Distanzbasierte Modelle ... 105
5.3.1 Hierarchische Clusterbildung ... 105
5.3.2 Vektorquantisierung ... 107
5.3.3 K-Nearest Neighbour ... 111
5.4 Wahrscheinlichkeitsmodelle ... 112
5.4.1 Bayes'sche Klassifikation ... 112
5.4.2 Gaussian Mixture Models ... 113
5.4.3 Hidden Markov Models ... 120
5.5 Neuronale Netze ... 124
5.5.1 Biologische Neuronale Netze ... 125
5.5.2 Künstliche Neuronale Netze ... 126
5.5.3 Vorwärtsgerichtete Netze ... 128
5.5.4 Wettbewerbsnetze ... 130

5.5.5 Hopfield-Netze ... 131

6 Implementierung ... 133

6.1 Bisherige Ansätze ... 134
6.1.1 Monophone Klassifikation ... 134
6.1.2 Eingeschränkt Polyphone Klassifikation ... 137
6.1.3 Komplexe Polyphone Klassifikation ... 139
6.2 Monophones Experimentiersystem ... 140
6.2.1 Aufbau ... 141
6.2.2 Merkmalsextraktion ... 143
6.2.3 Merkmalsaufbereitung ... 143
6.2.4 Training ... 144
6.2.5 Klassifikation ... 144
6.3 Monophones Echtzeitsystem ... 145
6.3.1 Aufbau ... 146
6.3.2 Pluginbetrieb ... 146
6.3.3 Training ... 147
6.3.4 Segmentierung ... 148
6.3.5 Spektrogramm ... 150
6.3.6 Normierung ... 151
6.3.7 Faktorisierung ... 151
6.3.8 Hidden Markov Model ... 153
6.3.9 MPEG-7 Modell ... 154
6.3.10 Klassifikation ... 154
6.3.11 Basisprojektion ... 155
6.3.12 Klassenzuordnung ... 156
6.4 Polyphones System ... 157
6.4.1 Aufbau ... 157
6.4.2 Merkmalsextraktion ... 157
6.4.3 Resynthesemodell ... 160
6.4.4 Training ... 161

6.4.5 Klassifikation 162

7 Auswertung 165

7.1 Testumgebung 165

7.1.1 Trainingsmaterial 165

7.1.2 Evaluierung 166

7.2 Monophone Klassifikation 167

7.2.1 Standardparameter 167

7.2.2 Spektrogrammparameter 168

7.2.3 ICA-Parameter 170

7.2.4 HMM-Parameter 171

7.3 Polyphone Klassifikation 173

7.3.1 Trainingsparameter 173

7.3.2 Einzelklänge 174

7.3.3 Musikstücke 175

7.3.4 Zufallsmusik 177

7.3.5 Zweiklänge 178

7.3.6 Dreiklänge 179

8 Zusammenfassung 181

8.1 Musikinstrumentenerkennung 181

8.2 Anwendung 182

9 Literaturverzeichnis 185

ABBILDUNGSVERZEICHNIS

Bild 2.1: Anschläge (schwarz) eines Musiksignals (grau) 11

Bild 2.2: Zeitsignal eines gestrichenen Kontrabasses (grau) mit der dazugehörigen Hüllkurve (schwarz) 12

Bild 2.3: Idealisierte ADSR-Hüllkurve mit exponentiellem Ein- und Ausschwingverhalten 13

Bild 2.4: Idealisiertes Kurzzeitspektrum eines Tons mit 3960 Hz (gestrichelte Linie) und eines Klangs mit der Grundfrequenz 440 Hz (durchgehende Linie) 14

Bild 2.5: Kurzzeitspektrum eines Geräuschs 14

Bild 2.6: Noten (links) sowie Pausen (rechts) der Notenwerte ganz, halb, viertel, achtel, sechzehntel, zweiunddreißigstel 17

Bild 2.7: Hauptzählzeiten (schwarz) eines Musiksignals (grau) zusammen mit unterschiedlich gewichtigen Nebenzählzeiten (gestrichelt) 18

Bild 2.8: Zuordnung der Notennamen der Stammtonreihe zu den Linien bzw. Zwischenräumen, hervorgehoben ist der Kammerton **a** 24

Bild 2.9: Notation und Bezeichnung der Töne der gleichstufigen chromatischen Tonleiter 26

Bild 3.1: Verschiedene Flöteninstrumente: Blockflöte und Querflöte [Brin98] 35

Bild 3.2: Klangerzeugung in einem Flöteninstrument an einer Schneidkante [Brin98] 35

Bild 3.3: Kurzzeitspektrum eines Flötenklanges mit einer Grundfrequenz von 130,81 Hz (SPN: C3) 36

Bild 3.4: Zeitsignal eines Flötenklanges mit einer Grundfrequenz von 130,81 Hz (SPN: C3) 36

Bild 3.5: Verschiedene Rohrblattinstrumente: Klarinette, Saxophon, Oboe und Fagott [Brin98] 37

Bild 3.6: Klangerzeugung in einem Rohrblattinstrument mit Doppelrohrblatt [Brin98] 37

Bild 3.7: Kurzzeitspektrum eines Oboenklanges mit einer Grundfrequenz von 130,81 Hz (SPN: C3) 38

Bild 3.8: Zeitsignal eines Oboenklanges mit einer Grundfrequenz von 130,81 Hz (SPN: C3) 38

Bild 3.9: Verschiedene Trompeteninstrumente: Posaune, Horn und Trompete [Brin98] 40

Bild 3.10: Klangerzeugung in einem Trompeteninstrument mit Kesselmundstück [Brin98] 40

Bild 3.11: Kurzzeitspektrum eines Trompetenklanges mit einer Grundfrequenz von 130,81 Hz (SPN: C3) 41

Bild 3.12: Zeitsignal eines Trompetenklanges mit einer Grundfrequenz von 130,81 Hz (SPN: C3) 41

Bild 3.13: Klangerzeugung in einem Harmonikainstrument mit durchschlagender Zunge [Bär 03] 43

Bild 3.14: Kurzzeitspektrum eines Harmoniumklanges mit einer Grundfrequenz von 130,81 Hz (SPN: C3) 44

Bild 3.15: Zeitsignal eines Harmoniumklanges mit einer Grundfrequenz von 130,81 Hz (SPN: C3) 44

Bild 3.16: Verschiedene Streichinstrumente: Geige, Bratsche, Cello und Kontrabass [Brin98] 46

Bild 3.17: Kurzzeitspektrum eines Geigenklanges mit einer Grundfrequenz von 130,81 Hz (SPN: C3) 47

Bild 3.18: Zeitsignal eines Geigenklanges mit einer Grundfrequenz von 130,81 Hz (SPN: C3) 47

Bild 3.19: Verschiedene Zupfinstrumente: Laute und Mandoline [Brin98] 48

Bild 3.20: Kurzzeitspektrum eines Gitarrenklanges mit einer Grundfrequenz von 130,81 Hz (SPN: C3) 49

Bild 3.21: Zeitsignal eines Gitarrenklanges mit einer Grundfrequenz von 130,81 Hz (SPN: C3) 49

Bild 3.22: Verschiedene Schlaginstrumente: Pauken, Schnarrtrommel, Reibe und Rasseln [Brin98] 51

Bild 4.1: Ermittlung der Hüllkurvenparameter: Effektivwert des Signals (schwarz), Schwellenwert bei 1 (horizontal gepunktet), Phasengrenzen (vertikal gepunktet) 59

Bild 4.2: Drei Filterbänder einer MFCC-Filterbank, obere Beschriftung: linearer Bereich, untere Beschriftung: logarithmischer Bereich 72

Bild 4.3: Schablone für die Grundfrequenz f_0 = 261,63 Hz (SPN: C4) 75

Bild 4.4: Betragsfrequenzgang eines Klanggemischs (schwarz) mit dem daraus abgeleiteten Sinusmodell (gestrichelt) 77

Bild 4.5: Polyphone Grundfrequenzermittlung für ein Sinusmodell (gestrichelt) mit einer Schablone (schwarz) der Grundfrequenz f_0 = 261,63 Hz (SPN: C4) 77

Bild 4.6: Harmonic Peak Spectrum (gestrichelt) eines Klanges mit dem Betragsfrequenzgang des dazugehörigen Klanggemischs (schwarz) 78

Bild 4.7: Anwendung der PCA auf zweidimensionale Merkmalsvektoren, links: Originalmatrix **A**, rechts: Transformationsergebnis **B**, Standardabweichung mit ihren Hauptachsen (Ellipsen) 86

Bild 4.8: Anwendung der SVD auf zweidimensionale Merkmalsvektoren, links: Originalmatrix **A**, rechts: Transformationsergebnis **B**, Standardabweichung mit ihren Hauptachsen (Ellipsen) 90

Bild 4.9: Anwendung der ICA auf nicht normalverteilte, zweidimensionale Merkmalsvektoren, links: Originalmatrix **A**, rechts: Transformationsergebnis **B**, Achsen der unabhängigen Komponenten (gestrichelt) 94

Bild 4.10: Anwendung der PCA auf nicht normalverteilte, zweidimensionale Merkmalsvektoren, links: Originalmatrix **A**, rechts: Transformationsergebnis **B**, Standardabweichung mit ihren Hauptachsen (Ellipsen) 95

Bild 5.1: Drei Klassenwolken (ω_1, ω_2, ω_3) im zweidimensionalen Merkmalsraum mit den dazugehörigen Merkmalsausprägungen 101

Bild 5.2: Zwei Klassenwolken im zweidimensionalen Merkmalsraum (ω_1: Kreise, ω_2: Punkte) mit überangepasster Klassengrenze (gestrichelt) und generalisierender Klassengrenze (durchgezogen) 104

Bild 5.3: Mittlere Klassifikationsrate P abhängig vom Trainingszyklus n, Klassifikationsrate der Trainingsmatrix (gestrichelt), Klassifikationsrate der Validierungsmatrix (durchgezogen) 105

Bild 5.4: Hierarchische Clusterbildung eines skalaren Merkmals a_1 in 19 Iterationsschritten n 106

Bild 5.5: Vektorquantisierung im zweidimensionalen Merkmalsraum, die Codebuchvektoren sind als Punkte eingetragen, die ebenfalls eingetragenen Clustergrenzen bilden so genannte Voronoizellen 108

Bild 5.6: Klassifikation eines neuen Merkmalsvektors (Stern) mittels KNN. Für $K = 1$ wird der neue Merkmalsvektor der Klasse ω_1 zugeordnet (Kreise), für $K = 7$ hingegen der Klasse ω_2 (Punkte). 111

Bild 5.7: Verteilungsdichtemodellierung im zweidimensionalen Merkmalsraum durch ein GMM mit drei Komponenten, Standardabweichung der einzelnen Komponenten mit ihren Hauptachsen (Ellipsen) 114

Bild 5.8: Markov-Kette mit drei Zuständen 121

Bild 5.9: Biologisches Neuron eines Wirbeltieres [Cofe08] 126

Bild 5.10: Schematischer Aufbau eines künstlichen Neurons 127

Bild 5.11: Verschiedene Schwellenfunktionen mit dem Schwellenwert $c = 5$, links einfache Sprungfunktion, rechts Sigmoidfunktion 128

Bild 5.12: Vorwärtsgerichtetes Netz mit drei Schichten 129

Bild 5.13: Hopfield-Netz mit drei Ein- und Ausgängen 132

Bild 6.1: Schematischer Aufbau des monophonen Experimentiersystems im Trainingsbetrieb 142

Bild 6.2: Schematischer Aufbau des monophonen Experimentiersystems im Klassifikationsbetrieb 142

Bild 6.3: Benutzeroberfläche des monophonen Echtzeitsystems als VST-Plugin im Betrieb mit dem Programm Cubase SX 2 der Firma Steinberg, links: Parametereingabe, rechts: Ergebnisliste 147

Bild 6.4: Schematischer Aufbau des monophonen Echtzeitsystems im Trainingsbetrieb 148

Bild 6.5: Audiosignal mit vier Klangereignissen, Effektivwert (schwarz), adaptive Schwelle (gestrichelt) und Segmentgrenzen (strichpunktiert) 149

Bild 6.6: HMM mit fünf Zuständen in einer Bakis-Topologie 154

Bild 6.7: Schematischer Aufbau des monophonen Echtzeitsystems im Klassifikationsbetrieb 155

Bild 6.8: Schematischer Aufbau der Merkmalsextraktion des polyphonen Systems 158

Bild 6.9: Resynthese eines Klangereignisses in einer Klangmischung (schwarz) aus den ersten 15 Harmonischen (gestrichelt). 160

Bild 6.10: Schematischer Aufbau des polyphonen Systems im Trainingsbetrieb 161

Bild 6.11: Schematischer Aufbau des polyphonen Systems im Klassifikationsbetrieb 162

Bild 7.1: Mittlere Klassifikationsraten der mit dem monophonen Echtzeitsystem durchgeführten Tests in Prozent, Random Guess: 14,28 % (durchgezogene Linie) 168

Bild 7.2: Mittlere Klassifikationsraten der mit dem polyphonen System durchgeführten Tests in Prozent, Random Guess: 14,28 % (durchgezogene Linie) 174

TABELLENVERZEICHNIS

Tabelle 2.1: Qualitative Tempiangaben [Hemp01, Diet02] 19

Tabelle 2.2: Musikalische Intervalle [Hemp01, Diet02] 21

Tabelle 2.3: Bezeichnung der Oktavlagen sowie der dazugehörigen Frequenzbereiche in der konventionellen musikalischen Notation und der Scientific Pitch Notation (SPN) 25

Tabelle 2.4: Tonintervalle der gleichstufigen Chromatik und ihre mathematischen Verhältnisse 27

Tabelle 2.5: Gängige Tonarten und die in ihnen enthaltenen Halb- und Ganztonschritte ausgehend vom Grundton 28

Tabelle 7.1: Test 1, Klassifikationsraten des monophonen Echtzeitsystems in Prozent bei Standardparametern, mittlere Klassifikationsrate: 84,84 %, Random Guess: 14,28 % 169

Tabelle 7.2: Test 2, Klassifikationsraten des monophonen Echtzeitsystems in Prozent bei einer Fensterbreite von 40 ms, mittlere Klassifikationsrate: 74,34 %, Random Guess: 14,28 % 170

Tabelle 7.3: Test 4, Klassifikationsraten des monophonen Echtzeitsystems in Prozent bei 15 ICA-Komponenten, mittlere Klassifikationsrate: 82,28 %, Random Guess: 14,28 % 171

Tabelle 7.4: Test 6, Klassifikationsraten des monophonen Echtzeitsystems in Prozent bei 8 HMM-Zuständen, mittlere Klassifikationsrate: 83,76 %, Random Guess: 14,28 % 172

Tabelle 7.5: Test 8, Klassifikationsraten des polyphonen Systems in Prozent für monophone Einzelklänge, mittlere Klassifikationsrate: 74,57 %, Random Guess: 14,28 % ... 175

Tabelle 7.6: Test 10, Klassifikationsraten des polyphonen Systems in Prozent für komplexe Musiksignale, mittlere Klassifikationsrate: 55,47 %, Random Guess: 14,28 % ... 176

Tabelle 7.7: Test 12, Klassifikationsraten des polyphonen Systems in Prozent für komplexe Zufallsmusik, mittlere Klassifikationsrate: 56,99 %, Random Guess: 14,28 % ... 177

Tabelle 7.8: Test 14, Klassifikationsraten des polyphonen Systems in Prozent für Zweiklänge, mittlere Klassifikationsrate: 67,60 %, Random Guess: 14,28 % 179

Tabelle 7.9: Test 15, Klassifikationsraten des polyphonen Systems in Prozent für Dreiklänge, mittlere Klassifikationsrate: 52,96 %, Random Guess: 14,28 % 180

ABKÜRZUNGSVERZEICHNIS

ADSR – Attack Decay Sustain Release

API – Application Programming Interface

ASE – Audio Spectrum Envelope

BPM – Beats per Minute

DCT – Diskrete Cosinustransformation

DFT – Diskrete Fouriertransformation

DIN – Deutsches Institut für Normung

EM – Expectation Maximisation

FEAPI – Feature Extracion API

FFT – Fast Fourier Transform

GB – Gigabyte

GM – General MIDI

GMM – Gaussian Mixture Model

HMM – Hidden Markov Model

ICA – Independent Component Analysis

KNN – K-Nearest-Neighbour

LBG – Linde Buzo Gray

LDS – Leistungsdichtespektrum

MAP – Maximum A-Posteriori

MFCC – Mel-scale Frequency Cepstrum Coefficients

MIDI – Musical Instrument Digital Interface

MM – Metronom Mälzel

MPEG – Moving Picture Experts Group

MTS – MIDI Tuning Standard

NMF – Non-negative Matrix Factorization

PCA – Principal Component Analysis

PCM – Puls Code Modulation

QBH – Query by Humming

QBT – Query by Tapping

SOM – Self Organizing Maps

SPN – Scientific Pitch Notation

STFT – Short Term Fourier Transformation

SVD – Singular Value Decomposition

SVM – Support Vector Machine

TB – Terabyte

VST – Virtual Studio Technology

WWW – World Wide Web

XML – Extensible Markup Language

1 EINLEITUNG

Durch die rasante Entwicklung im Bereich der Netzwerk- und Speichertechnologien sind heutzutage große Datenarchive und weit verzweigte Netze zum Alltag geworden. Allen voran steht hier das World Wide Web (WWW) mit seiner äußerst heterogenen Netzwerk- und Speicherstruktur. Das WWW ermöglicht es einer großen Anzahl von unterschiedlichen Nutzern, gleichzeitig Daten abzurufen sowie dem Netzwerk hinzuzufügen. Diese dynamische Struktur macht es unmöglich, ein statisches Verzeichnis der Inhalte zu erstellen, da dieses bereits zum Zeitpunkt des Erscheinens überholt wäre. Die Katalogisierung von Webinhalten wird aktuell durch schlagwortbasierte Suchmaschinen wie Google oder Yahoo vorgenommen, die permanent den Textinhalt des WWW durchsuchen und in Datenbanken indizierte Verweise speichern [Goog08, Yaho08]. Nutzer des WWW können nach bestimmten Inhalten suchen, indem sie diese Datenbanken durchsuchen und den gespeicherten Verweisen zum eigentlichen Inhalt folgen. Es besteht jedoch keine Garantie, dass der Inhalt zum Zeitpunkt des Nutzerzugriffs noch vorhanden ist.

1.1 INHALTSBEZOGENE DATENSUCHE

Anfang des Jahres 2006 wurden mindestens 10 Milliarden Webseiten mit einem Umfang von ca. 200 TB in Suchmaschinen gelistet, wobei der Gesamtdatenumfang des WWW aufgrund der heterogenen Netzwerkstruktur und der volatilen Inhalte schwer zu bestimmen ist. Der nach außen hin sichtbare Teil des Netzes, das so genannte Surface Web, hat einen mindestens 20 mal größeren Datenbestand als die größte Biblio-

thek der Welt, die Library of Congress. Der Umfang des WWW wächst seit der Gründung im Jahre 1989 exponentiell und verdoppelt sich ungefähr alle 24 Monate [ShPr01, Tant06].

Bei den Inhalten des WWW handelt es sich heutzutage neben reinen Textdaten auch um große Sammlungen multimedialer Inhalte wie Musik-, Video- und Bilddaten. Diese sind entweder in Webseiten eingebettet oder in eigenständigen, dem WWW angegliederten Archiven verfügbar. Um multimediale Daten über klassische, schlagwortbasierte Suchmaschinen finden zu können, ist es erforderlich, dass zusätzlich zu den eigentlichen Inhalten Metainformationen verfügbar sind, die den Inhalt in Textform beschreiben und von Suchmaschinen abgefragt werden können. Bei in Webseiten eingebetteten Daten müssen zusätzlich Schlagwörter oder beschreibender Text vorhanden sein, bei Archiven müssen Schlagwörter oder beschreibender Text separat gespeichert werden. Inhaltsbeschreibende Metadaten sind im Zusammenhang mit Musikstücken beispielsweise der Interpret, der Titel, das Genre, das Tempo, die Tonart, die Notendarstellung oder, dem Thema dieser Arbeit entsprechend, Informationen über die in dem Musikstück enthaltenen Instrumente.

Im einfachsten Fall können Metadaten ohne spezielle Formatierung als Text oder als Aufzählung von Schlagwörtern gespeichert werden, bei größeren Datenbeständen bietet sich eine Auszeichnungssprache wie die Extensible Markup Language (XML) oder eine Datenbanklösung an, was die Möglichkeiten einer gezielten Suche vergrößert. Im Jahre 2002 wurde in diesem Zusammenhang der Standard MPEG-7 [ISO15938] veröffentlicht, der unter anderem eine einheitliche Annotation von multimedialen Daten mit Metadaten spezifiziert und so eine erhöhte Vergleichbarkeit und erweiterte Suchfunktionen ermöglicht.

Mit textbasierten Metadaten ist es möglich, die hoch entwickelte Technologie der vorhandenen, schlagwortbasierten Suchmaschinen direkt weiter zu verwenden, inklusive der vorhandenen Fehlerkorrektur bei Tippfehlern und der Möglichkeit von Alternativvorschlägen bei wenigen gefundenen Suchergebnissen. Typische Anfragen für Musikstücke könnten beispielsweise lauten: „langsames Tempo, Michael Jackson, Gitarre“ oder „Klavier, männlicher Sänger, Pop“.

Metadaten müssen heutzutage größtenteils noch manuell annotiert werden, oftmals sind deshalb, gerade wegen der anwenderoffenen Struktur des WWW, Metadaten nicht oder nur unzureichend vorhanden. Somit ist eine große Menge an multimedialen Daten im WWW gespeichert, deren Inhalte von außen jedoch nicht zu erkennen sind und einem suchenden Benutzer verborgen bleiben. Der Umfang dieses nicht durch textbasierte Suchmaschinen erfassten, so genannten Deep Web wird etwa 50 bis 500 mal größer als das gelistete Surface Web geschätzt [ShPr01, Harr06, Tant06].

Damit die inhaltsbeschreibenden Metadaten nicht zu jeder Datei manuell erstellt werden müssen, konzentriert sich die Forschung derzeit auf die Entwicklung von Algorithmen, mit denen diese inhaltsbeschreibenden Metadaten automatisch aus den zu beschreibenden Dateien extrahiert werden können. Durch die automatische Extraktion von Metadaten wird es Suchmaschinen ermöglicht, multimediale Inhalte auf semantischer Ebene zu durchsuchen, ohne vollständig auf eine vorhandene korrekte Annotation des Inhalts angewiesen zu sein. Weiterhin werden durch die automatische Extraktion von Metadaten inhaltliche Vergleiche von multimedialen Inhalten möglich. Somit kann der Nutzer einer Suchmaschine neben schlagwortbasierten Anfragen auch Ähnlichkeitsabfragen formulieren. In Query by Humming Systemen (QBH, Anfrage durch Summen) kann ein Nutzer beispielsweise als Anfrage die Melodie eines gesuchten Musikstücks summen [BEW+04a, Batk06], Query by Tapping Systeme (QBT, Anfrage durch Klopfen) erlauben die Formulierung einer Suchanfrage über das Klopfen der Rhythmik des Stücks und sind auch auf Endgeräten ohne Mikrofon implementierbar [EiBS04a, EiBS04b]. Der Nutzer könnte als weitere Ähnlichkeitsabfrage auch ein Musikstück angeben, das eine große klangliche Ähnlichkeit mit einem von ihm gesuchten Musikstück aufweist [MaSS02, TzEC02, WeCr03].

Für Ähnlichkeitsabfragen ist neben der Qualität des verwendeten Algorithmus zur Extraktion der Metadaten der Detailgrad der Beschreibung von entscheidender Bedeutung. Bei einer zu detaillierten Beschreibung wird es schwierig, Ähnlichkeiten bei geringen Abweichungen noch zu erkennen, ist die Beschreibung hingegen zu ungenau, wird eine zu große Ähnlichkeit bei unterschiedlichen Inhalten erzeugt. Für die Verarbeitungsdauer einer Ähnlichkeitsabfrage ist es von entscheidender Bedeu-

tung, wie schnell und effizient sich die einzelnen Metadatensätze vergleichen lassen.

1.2 MUSIKINSTRUMENTENERKENNUNG

Im Zusammenhang mit der automatischen Extraktion von Metadaten aus Musikstücken wurden im Rahmen dieser Arbeit verschiedene Algorithmen umfangreich untersucht, die automatisch die in Musikstücken auftretenden Musikinstrumentenklänge und Geräusche identifizieren und klassifizieren. Diese Aufgabe lässt sich in die zwei Teilgebiete der Merkmalsextraktion und der Klassifikation unterteilen. Algorithmen der Merkmalsextraktion gewinnen aus Musikstücken oder einzelnen Musikinstrumentenklängen Informationen, die bestimmte Aspekte des analysierten Signals widerspiegeln. Die Klassifikation beruht auf Algorithmen der Mustererkennung und ordnet bestimmte Merkmalsmuster bekannten Musikinstrumentenklassen zu. Somit wird eine Entscheidung über die zu hörenden Musikinstrumente und Geräusche gefällt.

Im Rahmen dieser Arbeit wurden weiterhin drei Systeme zur automatischen Identifikation und Klassifikation von Musikinstrumentenklängen entwickelt, mit denen sich unterschiedliche theoretische und technische Aspekte erforschen und demonstrieren lassen. Unterschiedliche Musikinstrumentenklänge und Geräusche werden von den Systemen ihrem Ursprung entsprechend klassifiziert, wobei Klänge von gleichen Instrumentenarten jeweils in einer Klasse zusammengefasst werden. Die Systeme sind lernfähig und weisen ein Ordnungsverhalten auf, das dem eines musikalisch geschulten Menschen entspricht.

Das erste System ist ein monophones Experimentiersystem, das einzelne monophone Klänge oder Geräusche klassifizieren kann. Es ist modular aufgebaut, so dass sich die einzelnen Prozessstufen sehr einfach austauschen lassen. Somit lassen sich unterschiedlichste Algorithmen der Merkmalsextraktion, Merkmalsaufbereitung und Klassifikation sehr effizient sowohl einzeln als auch im Zusammenspiel testen. Dieses System bildet die Basis für die beiden anderen Systeme, die aufgrund der gewonnenen Erkenntnisse in unterschiedliche technische Richtungen weiterentwickelt und optimiert wurden.

Das zweite System ist ein monophones Echtzeitsystem, das ebenfalls einzelne aufeinander folgende monophone Klänge oder Geräusche klassifizieren kann, allerdings arbeitet es als echtzeitfähiger Prozess. Das System stellt eine Studie für den industriellen Gebrauch dar und wurde als Plugin für die im Bereich der Studiotechnik weit verbreitete VST-Schnittstelle [Stei08] implementiert. Es verwendet Algorithmen, die ein für den Echtzeitbetrieb nötiges, gutes Verhältnis zwischen effizienter Berechenbarkeit und guten Klassifikationsergebnissen haben.

Das dritte System ist ein polyphones Klassifikationssystem, das in der Lage ist, harmonische Musikinstrumentenklänge in komplexer polyphoner Musik zu identifizieren und zu klassifizieren. Dieses System benutzt Algorithmen, die sehr robust gegenüber Störeinflüssen durch überlappende oder gleichzeitig auftretende Klänge oder Geräusche sind. Das entwickelte System verwendet die Merkmale der Musikinstrumentenklänge nicht nur zur Analyse und Identifikation bzw. Klassifikation, sondern ist auch in der Lage, die Klänge durch ein auf dem Prinzip der Additiven Resynthese basierendes harmonisches Modell zu resynthetisieren.

1.3 STUDIOTECHNOLOGIE

Die entwickelten Systeme leisten neben den genannten Anwendungsmöglichkeiten im Bereich der semantischen Suche und automatischen Metadatenerzeugung zusätzlich einen wesentlichen Beitrag im Rahmen der Studiotechnologie. So können die extrahierten Informationen über die in einem Musikstück erkannten Musikinstrumente beispielsweise dazu genutzt werden, die Melodie des Stücks von einem ähnlich klingenden Instrument doppeln zu lassen. Die Auswahl eines geeigneten Instruments könnte hierbei nach gewissen Vorgaben automatisch durchgeführt werden. Datenbanken mit Synthesizerklängen, wie sie in Musikstudios vielfältig verwendet werden, könnten automatisch nach ihren klanglichen Eigenschaften in einer Datenbank angeordnet werden. Diese Datenbank könnte durch die Benutzung bestimmter klanglicher Eigenschaften und Ähnlichkeiten durchsucht werden.

Da sich im Bereich der musikalischen Metadatengewinnung viele Verfahren gegenseitig beeinflussen und eng miteinander verflochten sind, können die von den implementierten Systemen ermittelten Informatio-

nen vielfältige andere Aufgaben erleichtern. So können beispielsweise für Verfahren der Quellentrennung sowie der automatischen Transkription Signalmodelle verwendet werden, die auf die enthaltenen Musikinstrumente angepasst sind. Weiterhin ergeben sich Vereinfachungen für die in QBH- oder QBT-Systemen durchzuführenden Transkriptionen von komplexen Musikstücken in monophone Melodien oder Rhythmen [EiBS04a, EiBS04b, BEW+04b, Batk06].

1.4 KAPITELÜBERSICHT

Aus den für die vorliegende Arbeit betrachteten Aspekten und durchgeführten Forschungen ergibt sich thematisch die folgende Gliederung:

- **Musikterminologie**: Neben der Definition von Hör- und Klangereignissen und ihrer Parameter werden in Kapitel 2 alle für diese Arbeit relevanten musikalischen Phänomene sowie die Parameter der Musik erläutert. Weiterhin werden Tonarten und ihre Bedeutung im Zusammenhang mit der Musikinstrumentenidentifikation und Klassifikation beschrieben.
- **Musikinstrumente**: Die physikalischen Eigenschaften von Musikinstrumenten sowie die daraus resultierenden Klangeigenschaften werden in Kapitel 3 untersucht. Hierbei werden insbesondere charakteristische Eigenschaften der Dynamik und spektrale Eigenschaften betrachtet, die für die Identifikation von Musikinstrumentenklängen verwendet werden können.
- **Merkmalsextraktion**: Verfahren zur Extraktion relevanter Merkmale aus Musikstücken und den in ihnen vorkommenden Musikinstrumentenklängen werden in Kapitel 4 analysiert. Die vorgestellten Merkmale werden hinsichtlich ihrer Verwendbarkeit für die Beschreibung von Musikinstrumentenklängen und hinsichtlich ihrer effizienten Berechenbarkeit untersucht.
- **Klassifikation**: In Kapitel 5 werden verschiedene Verfahren der Mustererkennung vorgestellt und bezüglich ihrer Verwendbarkeit im Zusammenhang mit der Identifikation und Klassifikation von Musikinstrumentenklängen analysiert.

- **Implementierung**: Die drei im Rahmen dieser Arbeit implementierten Klassifikationssysteme werden ausführlich in Kapitel 6 beschrieben.
- **Auswertung**: In Kapitel 7 werden die verschiedenen mit den implementierten Systemen durchgeführten Testreihen sowie ihre Ergebnisse dargestellt.

Die **Zusammenfassung** in Kapitel 8 gibt die wesentlichen Ergebnisse der Arbeit wieder und gibt einen Ausblick auf mögliche Anwendungsgebiete.

2 MUSIKTERMINOLOGIE

Die in dieser Arbeit vorgestellten Algorithmen und Systeme verarbeiten Musiksignale und die darin enthaltenen Klangereignisse. Hierfür werden verschiedene physikalische und perzeptive Parameter modelliert und ausgewertet. Für das weitere Verständnis sind deshalb exakte Definitionen von physikalischen sowie perzeptiven Musikphänomenen nötig. In diesem Kapitel werden deshalb die für die Musikinstrumentenidentifikation und Klassifikation relevanten Parameter zusammen mit weiteren wichtigen musiktheoretischen Grundlagen beschreiben.

2.1 KLANGEREIGNISSE

Musiksignale setzen sich aus vielen einzelnen, teilweise überlagerten Klangereignissen zusammen. Als Klangereignisse bezeichnet man kurze akustische Signale, die den unterschiedlichen, in dem entsprechenden Musikstück verwendeten Musikinstrumenten entspringen. Klangereignisse haben als Schallsignale eine Reihe von physikalischen Parametern, die direkt messbare Kenngrößen wie den Schalldruck darstellen [DIN1320, Wein08]. Weiterhin haben sie einen klar definierten Anfang sowie einen zeitabhängigen Amplitudenverlauf, der von der Hüllkurve beschrieben wird. Aufgrund ihrer spektralen Eigenschaften lassen sie sich in verschiedene Kategorien unterteilen.

Die Auswahl der verwendeten Musikinstrumente und die zeitliche Verankerung der aus ihnen entspringenden Klangereignisse ergeben sich direkt aus der dem Musikstück zugrunde liegenden Komposition. Die Freiheit, die ein Komponist hierbei hat, spielt im Zusammenhang mit

dem Thema dieser Arbeit eine große Rolle. So können selbstverständlich nur Musikinstrumentenklänge identifiziert werden, wenn sie von bekannten Instrumenten stammen. Werden synthetische Klänge verwendet, kann von automatischen Identifikationssystemen allenfalls eine Ähnlichkeit mit realen Instrumenten festgestellt werden. Hierbei unterscheiden sich automatische Identifikationssysteme jedoch nicht von menschlichen Zuhörern.

Die für diese Arbeit relevanten Eigenschaften einzelner Klangereignisse sowie die Unterscheidung in die Kategorien Ton, Klang und Geräusch werden in den folgenden Abschnitten beschrieben [Hemp01, Mich05].

2.1.1 ANSCHLAG

Der Begriff Anschlag kommt ursprünglich aus der musikalischen Spielpraxis der Saiten- und Schlaginstrumente. Er bezeichnet den Zeitpunkt, zu dem die Saiten bzw. Felle der Instrumente angeschlagen und somit zum Klingen gebracht werden. Verallgemeinernd wird der Begriff jedoch auch für andere Instrumente und synthetische Klangerzeuger verwendet und bezeichnet den Beginn eines Klangereignisses [EiHu73, Diet02]. Bild 2.1 zeigt hierzu ein Beispiel, in dem neben dem Signalverlauf die Anschlagszeitpunkte der daraus resultierenden Klangereignisse eingetragen sind. Das Auffinden der Anschlagszeitpunkte spielt eine große Rolle bei der automatischen Metadatengewinnung aus Musiksignalen und stellt selbst ein weit reichendes Forschungsgebiet dar [Diso06].

2.1.2 HÜLLKURVE

Der zeitliche Verlauf der Signalamplitude wird von der Hüllkurve (auch Einhüllende) beschrieben. Die Hüllkurve ist das Signal, das entsteht, wenn alle relevanten Spitzen des Zeitsignals miteinander verbunden werden. Das Konzept der Hüllkurve erlaubt es, ein Klangereignis $x(n)$ als das Ergebnis einer Amplitudenmodulation zwischen dem Hüllkurvensignal $w(n)$ als Modulator und dem normierten Klangereignis $x\_(n)$ als Träger aufzufassen:

$$x(n) = w(n) \cdot x\_(n). \tag{2.1}$$

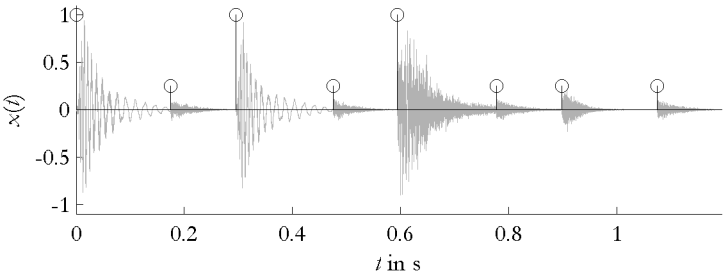


Bild 2.1: Anschläge (schwarz) eines Musiksignals (grau)

Die Grundeigenschaften des Hüllkurvensignals sind, dass seine Amplitudenänderungen mit geringer Frequenz im Infraschall-Bereich stattfinden und dass es keine hörbaren Sprünge oder Knicke enthält.

Ein Hüllkurvensignal ist zeitbegrenzt und hat die Form einer beliebigen Fensterfunktion, d.h. es steigt auf einen Maximalwert und fällt dann wieder auf Null. Somit ergibt sich aus der Länge der Hüllkurve direkt die Länge des zugrunde liegenden Klangereignisses.

Bild 2.2 zeigt ein Beispiel für ein Zeitsignal eines gestrichenen Kontrabasses und seine Hüllkurve. Die charakteristische Fensterform entsteht durch den bei allen Musikinstrumenten physikalisch ähnlichen Prozess der Klangerzeugung (vgl. Kapitel 3). Hierbei wird ein Medium zum Schwingen angeregt, so z.B. die Saiten einer Gitarre oder eines Klaviers, das Fell einer Trommel oder die Luftsäule einer Flöte oder Orgel [FlRo91, Veit96, Wink98]. Ein Hüllkurvensignal lässt sich in die folgenden vier Bereiche einteilen [Anwa00, Hoen01]:

- Die **Anstiegsphase** (engl. Attack) ist der erste Teil des Einschwingvorgangs des Instrumentenklangs, in der das tonerzeugende, schwingende Medium angeregt wird. Die Anregung kann impulsartig erfolgen, wie beispielsweise bei gezupften Saiten und geschlagenen Trommelfellen, oder kontinuierlich, wie bei gestrichenen Saiten oder geblasenen Instrumenten. Je träger das angeregte Medium ist, desto ausgeprägter ist die Anstiegsphase. So ist sie z.B. bei einer gestrichenen Geige sehr ausgeprägt, bei einer Trompete weniger stark ausgeprägt, bei einer Trommel kaum

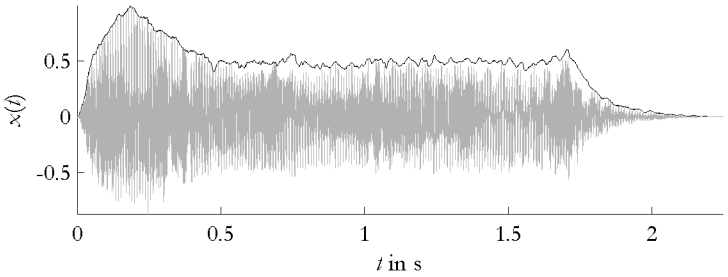


Bild 2.2: Zeitsignal eines gestrichenen Kontrabasses (grau) mit der dazugehörigen Hüllkurve (schwarz)

noch wahrnehmbar. Die Trägheit des schwingenden Mediums führt weiterhin dazu, dass die zu Beginn der Anregung aufgenommene Energie verzögert abgegeben wird und zu einem Überschwingen des Mediums und zur Ausprägung eines Maximums der Lautstärke führt. Die Anstiegsphase ist beendet, wenn die überschüssige Energie wieder abgegeben wurde und das Maximum der Hüllkurve erreicht ist.

- In der **Abklingphase** (engl. Decay), die den zweiten Teil des Einschwingvorgangs darstellt, fallen die Lautstärke und mit ihr die Hüllkurve auf das Niveau, das dem eingeschwungenen Zustand entspricht. Die Abklingphase tritt nur auf, wenn die Schwingungsanregung über einen gewissen Zeitraum anhält.
- In der **Haltephase** (engl. Sustain) ist das System im eingeschwungenen Zustand, und die zugeführte Leistung wird gleichmäßig und vollständig in Schallleistung umgewandelt. Fehlt diese konstante Anregung, entfallen sowohl die Haltephase als auch die Abklingphase, und nach der Anstiegsphase kommt direkt die Ausschwingphase. Der Klang einer Orgel hat beispielsweise eine sehr ausgeprägte Haltephase, wohingegen eine gezupfte Saite keine Haltephase hat.
- Die **Ausschwingphase** (engl. Release) tritt ein, wenn die Energiezufuhr beendet ist und das Medium aufgrund der fehlenden weiteren Anregung ausschwingt. Je stärker die Dämpfung des Mediums ist, desto schneller erreicht die Hüllkurve wieder den

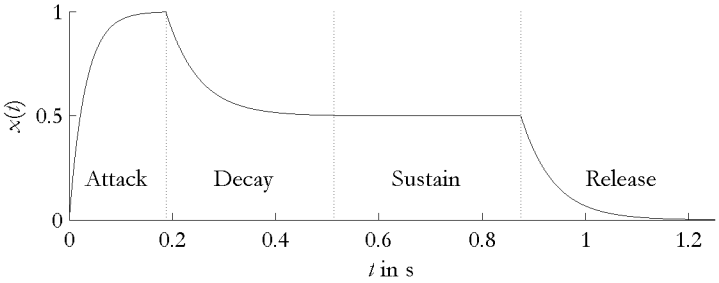


Bild 2.3: Idealisierte ADSR-Hüllkurve mit exponentiellem Ein- und Ausschwingverhalten

Wert Null. Oftmals folgt die Ausschwingphase einer exponentiell abklingenden Funktion. Eine ausgeprägte Ausschwingphase findet sich bei Instrumenten mit schwingenden Saiten, Instrumente mit schwingenden Luftsäulen hingegen haben eine kaum wahrnehmbare Ausschwingphase.

Die unterschiedlichen Phasen der Hüllkurve können durch ein Hüllkurvenmodell approximiert werden, bei dem die verschiedenen Bereiche durch Exponentialkurven angenähert werden. Aufgrund der Anfangsbuchstaben der englischen Bezeichnungen wird dieses Hüllkurvenmodell vor allem im Bereich der Synthesizertechnologie als ADSR-Hüllkurve bezeichnet (Bild 2.3).

Für die Erzeugung von synthetischen Klangereignissen können selbstverständlich beliebig komplexe Hüllkurven verwendet werden, jedoch lassen sich auch komplexe Hüllkurven durch eine Kombination von mehreren Hüllkurven mit dem beschriebenen ADSR-Verlauf zerlegen.

2.1.3 EINORDNUNG

Klangereignisse lassen sich unterscheiden in die drei Kategorien Ton, Klang und Geräusch (Bild 2.4 und Bild 2.5). Die Einordnung erfolgt hierbei über den spektralen Charakter, wobei eine hinreichend große spektrale Konstanz vorausgesetzt wird. Dies bedeutet, dass sich die Kurzzeitspektren der Signale innerhalb eines Zeitraums von ca. 20 ms nicht oder nur sehr wenig ändern [Hanu94, Zieg00].

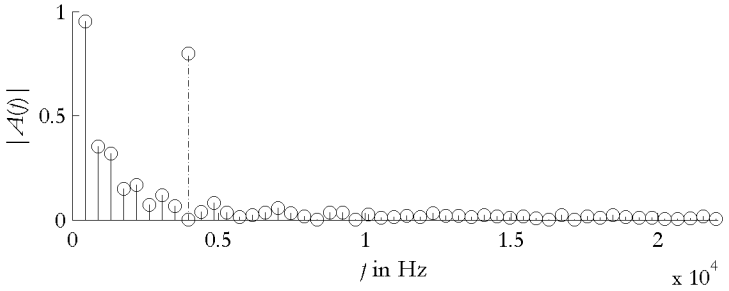


Bild 2.4: Idealisiertes Kurzzeitspektrum eines Tons mit 3960 Hz (gestrichelte Linie) und eines Klangs mit der Grundfrequenz 440 Hz (durchgehende Linie)

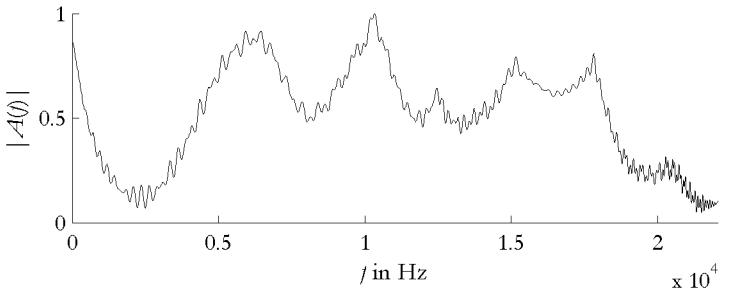


Bild 2.5: Kurzzeitspektrum eines Geräuschs

- **Töne** haben eine sehr große Ähnlichkeit mit Sinussignalen, d.h. ihre Spektren haben nur eine stark ausgeprägte, schmale Spitze bei der Grundfrequenz f_0 des Tons (Bild 2.4).
- **Klänge** haben Spektren mit mehreren stark ausgeprägten, schmalen Spitzen, hauptsächlich bei Vielfachen der Grundfrequenz f_0. Die Vielfache der Grundfrequenz werden Harmonische oder Obertöne genannt und aufsteigend nummeriert. Die Grundfrequenz stellt hierbei die erste Harmonische und den nullten Oberton dar. Anstelle der Bezeichnung nullter Oberton verwendet man in der Regel die Bezeichnung Grundton (Bild 2.4).

- **Geräusche** sind Signale, deren Spektren disharmonisch sind, d.h. sie haben keine ausgeprägte Grundfrequenz und keine ausgeprägten Obertöne (Bild 2.5).

Weist das Klangereignis keine hinreichende spektrale Konstanz auf, so kann die Einordnung für Teile des Signals vorgenommen werden. So hat z.B. das von einem Saxophon erzeugte akustische Signal eine stark geräuschhafte Einschwingphase, die dann in einen Klang übergeht.

2.2 HÖREREIGNISSE

Als Hörereignisse bezeichnet man den beim Anhören eines Klangereignisses hervorgerufenen auditiven Wahrnehmungsgegenstand. Hörereignisse haben eine Reihe von perzeptiven Parametern, die sich im Gegensatz zu den physikalischen Parametern der zugrundeliegenden Klangereignisse oftmals nicht direkt, sondern nur über Modelle beschreiben lassen.

Die wichtigsten perzeptiven Parameter einzelner Hörereignisse sind die Lautheit, die Tondauer, die Tonhöhe und die Klangfarbe, die im Rahmen der Komposition weitestgehend festgelegt werden und in den folgenden Abschnitten beschrieben sind [DIN1320, Hemp01, Mich05, Wein08].

Teilweise werden noch weitere Eigenschaften des einzelnen Klangereignisses wie Artikulation, Akzente und Verzierungen als Parameter bezeichnet, sie spielen jedoch eine untergeordnete Rolle und lassen sich aus den genannten Hauptparametern ableiten.

2.2.1 LAUTHEIT

Die Lautheit gibt die subjektive Wahrnehmung der Lautstärke eines Hörereignisses auf einer Skala leise-laut an. Sie unterscheidet sich somit gegenüber dem physikalisch messbaren Schalldruck eines Klangereignisses dadurch, dass die starke Frequenzabhängigkeit des Gehörs mit einbezogen wird. Die Lautheit wird in den Einheiten Sone oder Phon angegeben. In Kompositionen wird die Lautheit einzelner Hörereignisse über ein Skala von *pp* – pianissimo, sehr leise, bis *ff* – fortissimo, sehr laut, notiert.

Über genormte Bewertungsfilter lässt sich die Lautheit unter bestimmten, festgelegten Rahmenbedingungen aus dem Schalldruck des korrespondierenden Klangereignisses schätzen, indem frequenzabhängige Bewertungsfilter verwendet werden. Am häufigsten wird in diesem Zusammenhang die A-Kurve verwendet, was in der Einheit des Schalldruckpegels db(A) gekennzeichnet wird [DIN61672].

2.2.2 TONDAUER

Die Tondauer eines Hörereignisses beschreibt die Dauer zwischen dem Anschlag des zugrundeliegenden Klangereignisses und dem Moment, in dem es nicht mehr wahrgenommen werden kann. Die Tondauer ist stark davon abhängig, in welchem musikalischen Umfeld das Hörereignis zu hören ist, da es von anderen Ereignissen zeitlich oder spektral maskiert werden kann. Die Tondauer kann in Sekunden angegeben werden, in der westlichen Notation werden jedoch Tondauern als normierte Verhältnisse einer Grunddauer mittels Notenwerten dargestellt (Bild 2.6).

2.2.3 TONHÖHE

Die Tonhöhe beschreibt für Töne und Klänge die subjektive Hörempfindung, die auf einer Skala tief-hoch skaliert wird. Häufig wird als Tonhöhe direkt die Frequenz des Grundtons des zugrundeliegenden Klangereignisses wahrgenommen, teilweise dominieren jedoch auch Obertöne das Frequenzspektrum, so dass diese die Tonhöhenempfindung entscheiden. Dies kommt beispielsweise bei Blasinstrumenten vor, bei denen der Grundton teilweise kaum wahrnehmbar ist.

Je nach Notation werden den Tonhöhen feste Namen zugeordnet, in der westlichen Musik richten sich die Notennamen am Kammerton a^1 aus, der 440 Hz entspricht. Teilweise werden auch Geräuschen Tonhöhen zugeordnet, die sich durch dominante Bereiche im Frequenzspektrum ergeben.

2.2.4 KLANGFARBE

Die Klangfarbe eines Hörereignisses hängt von der Verteilung und Lage der Obertöne sowie dem Gemisch aus Grundton, Obertönen und Rauschanteilen des zugrundeliegenden Klangereignisses ab. Weiterhin

Bild 2.6: Noten (links) sowie Pausen (rechts) der Notenwerte ganz, halb, viertel, achtel, sechzehntel, zweiunddreißigstel

beeinflusst der zeitliche Verlauf des Spektrums und der Lautheit die Klangfarbe. Bei Musikinstrumenten gibt die Bauform durch bestimmte Resonanzen einen charakteristischen Frequenzbereich vor, in dem die Obertöne unabhängig von der Tonhöhe relativ stark oder schwach ausgeprägt sind. Diese Bereiche werden Formanten genannt und spielen bei der automatischen Musikinstrumentenerkennung eine wesentliche Rolle (vgl. auch die Abschnitte 4.4 und 4.5).

2.3 MUSIKPARAMETER

Die Parameter der Musik ergeben sich direkt aus den jeweiligen Parametern der in dem Musiksignal enthaltenen Klangereignisse und den hieraus resultierenden Hörereignissen. Sie gliedern sich in die Bereiche Rhythmik, Melodik und Klang [Step68, Zieg00, Hemp01, Mich05].

Musikalische Parameter der Rhythmik sind die Zählzeiten, das Tempo, die Dynamik und der Rhythmus. Sie beschreiben die zeitliche Abfolge im musikalischen Kontext. Hierbei bilden die Zeitabstände, die Dauer sowie die jeweilige Lautheit der vorkommenden Hörereignisse Muster.

Musikalische Parameter der Melodik sind die Harmonie und die Melodie. Sie beschreiben die Schichtung und Abfolge von Tönen und Klängen mit unterschiedlichen Tonhöhen im musikalischen Kontext. Die Tonhöhen der Töne und Klänge stehen hierbei in bestimmten musikalischen Intervallen zueinander.

Der Klang als Parameter der Musik ergibt sich direkt als Ergebnis aus den Klangfarben der auftretenden Hörereignisse.

Die genannten Parameter der Musik stellen die wichtigsten Parameter dar und werden in den folgenden Abschnitten genauer beleuchtet. Teilweise werden in der Literatur neben den aufgeführten Musikparametern noch weitere Parameter wie Agogik oder Phrasierung genannt, sie lassen

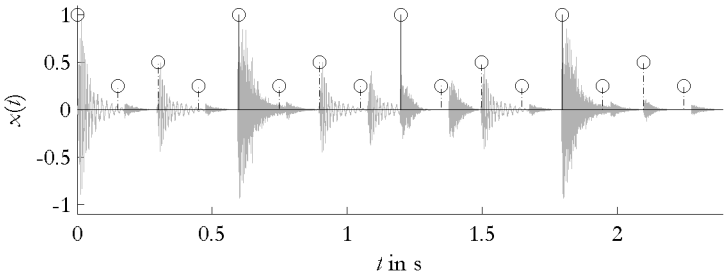


Bild 2.7: Hauptzählzeiten (schwarz) eines Musiksignals (grau) zusammen mit unterschiedlich gewichtigen Nebenzählzeiten (gestrichelt)

sich jedoch aus den aufgeführten Parametern herleiten und sind für diese Arbeit nicht relevant.

2.3.1 ZÄHLZEITEN

In der Regel läuft Musik vor einem zeitlichen Grundraster ab, das aus Zeitpunkten mit einem kontinuierlichem Abstand besteht. Diese Zeitpunkte werden Zählzeiten genannt und bilden einen kontinuierlichen Puls mit einer konstanten Frequenz. Die Zählzeiten empfinden auch musikalische Laien nach relativ kurzer Hördauer als Zeitpunkte, zu denen sie klatschen oder mit dem Fuß stampfen können, ohne den musikalischen Kontext zu zerstören. Die Zählzeiten eines Musikstücks, an denen sich auch Tänze orientieren, stellen quasi sein musikalisches Rückgrat dar und ergeben sich aus dem Musikstück selbst, indem die meisten betonten Klangereignisse des Stücks mit den Zählzeiten zusammenfallen [Hemp01] (Bild 2.7). Als Synonyme für den Begriff Zählzeiten werden auch die Begriffe Puls und Metrum verwendet [Thie73, Diet02].

2.3.2 TEMPO

Das Tempo legt fest, wie schnell, d.h. mit welcher Frequenz, die Schläge der Zählzeiten gezählt werden und wird in einer der gleichwertigen Einheiten Metronom Mälzel (MM) oder Beats per Minute (BPM, Schläge pro Minute) angegeben. Im deutschsprachigen Raum wird MM bevorzugt, im englischsprachigen Raum hingegen BPM. Der Zahlenwert der Tempoangabe ist jedoch bei MM und BPM gleich.

Tabelle 2.1: Qualitative Tempiangaben [Hemp01, Diet02]

Bezeichnung	MM	Übersetzung / Eigenschaften
Largo	40 – 60	breit, ruhig
Larghetto	60 – 66	etwas flüssiger als largo
Lento	60 – 66	langsam, schleppend
Grave	66 – 76	schwer, ernst, bedächtig
Adagio	66 – 76	gemächlich
Andante	76 – 108	gehend, mäßig bewegt
Andantino	84 – 108	etwas schneller als andante
Moderato	108 – 120	mäßig
Allegretto	108 – 120	etwas langsamer als allegro
Allegro	120 – 168	lustig, heiter, munter, schnell
Vivace	120 – 168	lebhaft
Presto	168 – 208	schnell, sehr schnell
Prestissimo	>208	äußerst schnell

Historische Kompositionen geben das Tempo nicht in MM an, sondern gehen für die Festlegung eines Basistempos vom menschlichen Herzschlag bzw. von einem gehenden Menschen aus, dessen Schritte mit den Zählzeiten zusammenfallen [Thie73] (Tabelle 2.1). Dieses Basistempo entspricht ungefähr einem Tempowert von 75 MM. Tempovariationen werden hierbei mit italienischen Bezeichnungen für Laufgeschwindigkeiten definiert und lassen einen gewissen interpretatorischen Spielraum für die Musiker.

2.3.3 DYNAMIK

Der Unterschied zwischen lauten, betonten und leisen, unbetonten Klangereignissen bildet die Dynamik eines Musikstücks. Die Dynamik ist somit ein Muster von gleich bleibenden oder sich in der Zeit entwickelnden Lautstärken von einzelnen Klangereignissen und kann zum Spannungsauf- oder -abbau beitragen, indem die Gesamtlautstärke anschwillt (crescendo) oder abschwillt (decrescendo). Bild 2.7 beispielsweise weist eine relativ große Dynamik auf, da der Lautstärkeunterschied zwischen den lauten und leisen Schlägen verhältnismäßig groß ist. Ein typischer, durch die Dynamik geprägter Tanz ist der Walzer, der aus drei zeitlich gleichwertigen Zählzeiten besteht, von denen die erste betont ist [Zieg00, Hemp01, Thie73].

2.3.4 RHYTHMUS

Im Gegensatz zur Dynamik ergibt sich beim Rhythmus nicht aus den unterschiedlichen Lautstärken, sondern aus den Zeitabständen und Längen der einzelnen Klangereignisse ein Muster. Die Zeitabstände stehen dabei in der Regel in ganzzahligen Verhältnissen zueinander (1:2, 1:3, 1:4, 2:3, 3:4 usw.). Rhythmen zeichnen sich dadurch aus, dass sie relativ einprägsam sind und sich beim Zuhörer nach einer gewissen Hördauer ein Gefühl der Vorhersagbarkeit einzelner Klangereignisse einstellt [Step68, Zieg00]. Ein typischer Rhythmus ist der Anfang des Kinderliedes „Hänschen Klein", das mit kleinen Variationen aus dem wiederholten Muster kurz-kurz-lang besteht.

2.3.5 INTERVALLE

Das Verhältnis der Tonhöhen zweier Töne oder Klänge zueinander bezeichnet man als Intervall. Einige Intervalle, vor allem die, die aus einfachen ganzzahligen Verhältnissen gebildet werden können (1:2, 2:3, 3:4, 4:5, 5:6), werden als wohlklingend empfunden. Sie tragen den Namen konsonante Intervalle. Dieses Empfinden beruht auf der Tatsache, dass die Obertöne in allen Klängen, die aus natürlichen Schwingungsvorgängen wie schwingenden Saiten oder Luftsäulen entstehen, ganzzahlige Vielfache des Grundtons sind und somit in ganzzahligen Verhältnissen zueinander stehen [Mich05] (vgl. auch Kapitel 3). Intervalle mit komplizierten oder nicht ganzzahligen Tonhöhenverhältnissen werden als dissonant bezeichnet.

Den Standardintervallen werden in der westlichen Musik Namen zugeordnet, die aus den lateinischen Ordnungszahlen abgeleitet sind. Diese Namen werden auch beibehalten, wenn das ganzzahlige Verhältnis nicht mehr genau stimmt, sondern zugunsten einer übergeordneten Stimmungstemperatur leicht verändert ist (vgl. auch Abschnitt 2.4.2). Tabelle 2.2 zeigt die Standardintervalle und die ihnen entsprechenden mathematischen Verhältnisse [Zieg00, Hemp01, Diet02].

Tabelle 2.2: Musikalische Intervalle [Hemp01, Diet02]

Intervall	Verhältnis	Eigenschaft
Prime	1:1	sehr konsonant
kleine Sekunde	16:15	dissonant
große Sekunde	9:8	dissonant
kleine Terz	6:5	konsonant
große Terz	5:4	konsonant
Quarte	4:3	sehr konsonant
Tritonus	7:5	sehr dissonant
Quinte	3:2	sehr konsonant
kleine Sexte	8:5	konsonant/dissonant
große Sexte	5:3	konsonant/dissonant
kleine Septime	7:4	dissonant
große Septime	15:8	dissonant
Oktave	2:1	sehr konsonant

Die harmonischen Intervalle ergeben sich direkt aus den Verhältnissen zweier benachbarter Obertöne eines Klangs: Prime, Oktave, Quinte, Quarte, gr. Terz, kl. Terz. Die größte harmonische Ähnlichkeit haben hierbei zwei Töne oder Klänge, deren Intervall eine Prime ist. Ähnlich harmonisch werden auch Töne oder Klänge empfunden, deren Intervall eine Oktave ist. Diese harmonische Ähnlichkeit führt weiterhin dazu, dass im Rahmen der Notation von Musik Töne und Klänge, deren Tonhöhenabstand eine oder mehrere Oktaven ist, den gleichen Basisnamen erhalten (vgl. Abschnitt 2.4.1).

Das Transponieren eines Tons oder Klangs um ein bestimmtes Intervall bedeutet im musikalischen Zusammenhang, dass die Grundfrequenz mit dem Verhältnis des Intervalls multipliziert wird. Ist der Abstand zweier Töne mit mehreren Intervallen angegeben, ist aus mathematischer Sicht die Multiplikation der Intervallverhältnisse gemeint. So bedeutet beispielsweise die Aussage, Ton x mit der Grundfrequenz f_x ist zwei Quarten höher als Ton y mit der Grundfrequenz f_y, das folgende Verhältnis, wobei q das Intervall der Quarte angibt:

$$\frac{f_x}{f_y} = q^2 = \frac{16}{9}. \tag{2.2}$$

2.3.6 HARMONIE

Die Harmonie beschreibt den Zusammenklang verschiedener Töne oder Klänge mit unterschiedlichen oder gleichen Tonhöhen. Hierbei werden bestimmte Gruppen von Klängen oder Tönen als Akkorde derart zusammengespielt, dass sie vom Zuhörer als eigenständige Hörereignisse wahrgenommen werden. Abhängig davon, wie viele Obertöne der einzelnen Klänge zusammenfallen, werden die Akkorde als harmonisch (konsonant) oder disharmonisch (dissonant) wahrgenommen.

Die klassische Akkordlehre geht von Dreiklängen aus, bei denen drei im Abstand von Terzen geschichtete Klänge gleichzeitig erklingen. Hierbei spricht man von einem Dur-Akkord, wenn die erste Terz groß und die zweite klein ist. Ist umgekehrt die erste Terz klein und die zweite groß, so wird die Bezeichnung Moll-Akkord verwendet. In der heutigen Musik gibt es eine Reihe von weiteren gängigen Akkorden, die auch Vier-, Fünf- oder Mehrklänge enthalten können. Der genaue Aufbau derartiger Akkorde geht allerdings über den Rahmen dieser Arbeit hinaus [Zieg00, Hemp01, Diet02].

Für die automatische Musikinstrumentenerkennung spielen Harmonien eine große Rolle, weil sie die Entscheidung, ob eine Gruppe von Obertönen aus einem eigenständigen Klangereignis oder aus einem Akkord entsprungen ist, stark erschwert.

2.3.7 MELODIE

Eine Melodie ist im Gegensatz zur Harmonie eine Folge von Tönen oder Klängen, die eine künstlerisch geformte, in sich geschlossene Kombination aus Tonhöhe und Tondauer bildet. Melodien bilden eine Einheit und haben in der Regel eine in sich schlüssige, einprägsame Abfolge.

Die wesentliche Eigenschaft einer Melodie ist, dass sie vom Menschen nachgesungen oder gesummt werden kann. Hieraus folgen direkt die weiteren Eigenschaften von Melodien [Zieg00]:

- Melodien sind in der Regel monophon, die Töne oder Klänge erklingen also nacheinander, was eine klare Aussage über die Tonhöhe zu jedem Zeitpunkt erlaubt.

- Nicht die absoluten Tonhöhen der Melodie sind relevant, sondern nur die Abfolge der musikalischen Intervalle. Dies erlaubt eine einfache Transponierbarkeit der absoluten Tonhöhen in den individuell singbaren Frequenzbereich eines Sängers.
- Weiterhin ist das Tempo einer Melodie für die Wahrnehmung der übergeordneten Gestalt nicht relevant. Das Tempo kann in weiten Bereichen frei gewählt werden.

2.3.8 KLANG

Der Klang als Parameter der Musik ergibt sich aus den Klangfarben der auftretenden Hörereignisse. Hierbei können bewusst Frequenzbereiche betont oder auch ausgespart werden, beispielsweise um Platz für eine Gesangsstimme zu schaffen. Es können auch interessante klangliche Effekte und Spannungen entstehen, wenn bewusst Klangereignisse mit gegenläufigen Klangfarben verwendet werden. Dies ist besonders in elektronischer Musik häufig der Fall.

2.4 TONARTEN

Eine Tonart bildet eine Menge von Tonhöhen, die zu einer Gruppe zusammengefasst werden und die Basis für Melodien bilden. Da die Tonhöhen einer Tonart in der Regel das Intervall der Oktave stufenförmig in kleinere Intervalle unterteilen, spricht man auch von Tonleitern. Für die Bildung einer Tonart gibt es unterschiedliche Bildungsgesetze, von denen in den folgenden Abschnitten die wichtigsten zusammen mit der dazugehörigen Notation im Fünf-Linien-System beschrieben sind. Im Zusammenhang mit der Musikinstrumentenidentifikation und Klassifikation spielen Tonarten eine wichtige Rolle, da sie a-priori Informationen über die potenziellen Tonhöhen der analysierten Klangereignisse liefern.

2.4.1 STAMMTONREIHE

Die Stammtonreihe unterteilt eine Oktave in sieben Töne und lässt sich direkt ohne weitere Vorzeichen wie # oder *b* (vgl. Abschnitt 2.4.2) im Fünf-Linien-System darstellen. Die Notennamen entsprechen den ersten sieben Buchstaben des Alphabets, wobei das b historisch bedingt durch

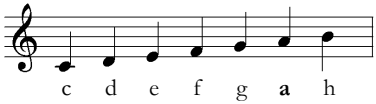


Bild 2.8: Zuordnung der Notennamen der Stammtonreihe zu den Linien bzw. Zwischenräumen, hervorgehoben ist der Kammerton **a**

das h ersetzt wird und bei der Benennung mit der Note c begonnen wird. Somit ergeben sich die Notennamen der Stammtonreihe: c, d, e, f, g, a, h.

Die Notennamen werden ausgehend von den Stammtönen für höhere oder tiefere Noten zyklisch fortgesetzt. Hierdurch haben zwei Töne, deren Tonhöhe sich nur um eine oder mehrere Oktaven unterscheidet, identische Notennamen. Die Noten der Stammtonreihe entsprechen den weißen Tasten eines Klaviers.

Bild 2.8 zeigt die Zuordnung der Notennamen zu den Linien bzw. Zwischenräumen. Die Verankerung einer bestimmten Note mit einer bestimmten Linie wird durch den am Anfang der Zeile stehenden Notenschlüssel festgelegt. Am gängigsten ist der in Bild 2.8 gezeigte Violinschlüssel, der die Note g mit der zweiten Linie von unten verankert.

Die durch die Notennamen repräsentierten Tonhöhen lassen sich über das Verfahren der Quintenschichtung gewinnen. Hierbei wird der Grundton, im Fall der Stammtonreihe das c, in seiner Tonhöhe festgelegt. Ausgehend von diesem Grundton werden weitere Töne gewonnen, deren Tonhöhen sich jeweils um eine Quinte, d.h. um das Verhältnis 2:3, unterscheiden. Es werden eine Quinte unterhalb des Grundtones und fünf Quinten oberhalb des Grundtones gebildet. Somit ergeben sich die Noten f, c, g, d, a, e, h, oder, zurücktransponiert in die Oktave, die Stammtonreihe c, d, e, f, g, a, h.

Die musikalischen Intervalle zwischen den Tonhöhen der Stammtonreihe sind ausschließlich kleine und große Sekunden. Die großen Sekunden werden als Schritte oder genauer als Ganztonschritte bezeichnet. Da eine große Sekunde ungefähr dem Verhältnis von zwei kleinen Sekunden entspricht, werden die kleinen Sekunden als Halbtonschritte bezeichnet. So ergeben sich, ausgehend vom c, die folgenden Ganzton- und Halbtonschritte: 1, 1, ½, 1, 1, 1, ½.

Tabelle 2.3: Bezeichnung der Oktavlagen sowie der dazugehörigen Frequenzbereiche in der konventionellen musikalischen Notation und der Scientific Pitch Notation (SPN)

Oktavname	Frequenz-umfang in Hz	konventionelle Notation	SPN
Subkontra-Oktave	16,35 - 30,87	C_2 - H_2	C0 - H0
Kontra-Oktave	32,70 - 61,74	C_1 - H_1	C1 - H1
große Oktave	65,41 - 123,47	C - H	C2 - H2
kleine Oktave	130,81 - 246,94	c - h	C3 - H3
eingestrichene Oktave	261,63 - 493,88	c^1 - h^1	C4 - H4
zweigestrichene Oktave	523,25 - 987,77	c^2 - h^2	C5 - H5
dreigestrichene Oktave	1046,50 - 1975,53	c^3 - h^3	C6 - H6
viergestrichene Oktave	2093,00 - 3951,07	c^4 - h^4	C7 - H7

Für die Bezeichnung der absoluten Tonhöhen der Stammtonreihe in unterschiedlichen Oktavlagen gibt es unterschiedliche Systeme (Tabelle 2.3). Die konventionelle musikalische Notation unterscheidet die Tonhöhen durch Groß- und Kleinschreibung der Notennamen sowie zusätzliche Indices und ist aus dem musikalischen Gebrauch entstanden.

Die Scientific Pitch Notation (SPN) wird hingegen häufig im technischen Zusammenhang benutzt. Hierbei wird der eigentliche Notenname als Großbuchstabe notiert, gefolgt von der Oktavlage als ganze Zahl. Die Oktavlagen sind so nummeriert, dass sie mit C0 für 16,35 Hz ungefähr bei der unteren Grenze des menschlichen Hörbereichs beginnen.

Die Verankerung der Notennamen zur absoluten Tonhöhe erfolgt über den Kammerton a, der 1939 im Rahmen der zweiten internationalen Stimmtonkonferenz in London auf 440 Hz bei 20 °C festgelegt wurde [Mich05]. Der Kammerton wird in der konventionellen Notation als a^1 (eingestrichenes a) und in der SPN als A4 notiert.

2.4.2 GLEICHSTUFIGE CHROMATIK

Die gleichstufige chromatische Tonleiter oder kurz gleichstufige Chromatik bildet eine Erweiterung der Stammtonreihe, indem jeder Ganztonschritt der Stammtonreihe in zwei Halbtonschritte aufgespalten wird. Somit wird die Oktave in zwölf Töne aufgeteilt, deren Abstand jeweils Halbtonschritte sind. Werden die Töne der Stammtonreihe, wie im letzten Abschnitt beschrieben, durch Quintenschichtung gebildet, so haben

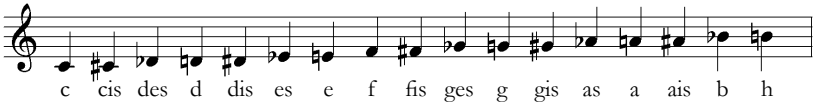


Bild 2.9: Notation und Bezeichnung der Töne der gleichstufigen chromatischen Tonleiter

sie alle ein ganzzahliges Verhältnis zum Grundton. Dies spiegelt die Verhältnisse der Obertöne zur Grundfrequenz einer schwingenden Saite oder einer schwingenden Luftsäule wider. Aufgrund dieser direkten Verknüpfung der physikalischen Vorgänge mit den für die Tonart ausgewählten Tönen wird die entstehende Tonreihe auch Naturtonreihe genannt. Sie hat allerdings den Nachteil, dass bei der Transponierung der Töne in eine Oktave die Verhältnisse der einzelnen Töne zueinander keine reinen kleinen oder großen Sekunden mehr sind. Dieses Manko wird durch die Einführung der gleichstufigen chromatischen Tonleiter ausgeglichen. Hier wird das Intervall der kleinen Sekunde zwischen zwei Tönen exakt auf $2^{(1/12)}$ festgelegt, was zur Folge hat, dass zwölf Halbtonabstände ganz genau eine Oktave ergeben. Somit lässt sich das Intervall *I* zwischen zwei Tönen über die folgende Formel berechnen, wobei *n* den Halbtonabstand der beiden Töne angibt:

$$I(n) = 2^{(n/12)}. \qquad (2.3)$$

Die Notennamen der gleichstufigen chromatischen Tonleiter werden größtenteils aus der Stammtonreihe übernommen, wobei die neu hinzugekommenen Töne als Erweiterungen der Stammtonreihe notiert werden (Bild 2.9). Hierbei gibt es für jeden hinzugekommenen Ton zwei unterschiedliche Notationsformen mit dazugehörigen Bezeichnungen. Ist der Ton einen Halbton höher als der entsprechende Ton der Stammtonreihe, so wird er mit einem # notiert, und an den Namen des Stammtons wird die Erweiterung -is angehangen. Ist der Ton einen Halbton tiefer als der entsprechende Ton der Stammtonreihe, so wird er mit einem *b* notiert, und an den Namen des Stammtons wird die Erweiterung -es angehangen. Eine Ausnahme bildet der Stammton h, der zu b wird. Welche der beiden Notationsformen gewählt wird, ergibt sich aus dem musikalischen Kontext. Die Noten der gleichstufigen chromatischen Tonleiter entsprechen den weißen und schwarzen Tasten eines Klaviers.

Tabelle 2.4: Tonintervalle der gleichstufigen Chromatik und ihre mathematischen Verhältnisse

Halbtonabstand n	Intervall	Verhältnis $I(n)$
0	Prime	1,000
1	kleine Sekunde	1,059
2	große Sekunde	1,122
3	kleine Terz	1,189
4	große Terz	1,260
5	Quarte	1,335
6	Tritonus	1,414
7	Quinte	1,498
8	kleine Sexte	1,587
9	große Sexte	1,682
10	kleine Septime	1,782
11	große Septime	1,888
12	Oktave	2,000

Für die Bezeichnung der Intervalle werden auch weiterhin die Namen der reinen musikalischen Intervalle benutzt, die sich aus den ganzzahligen Verhältnissen der Obertonreihe ergeben (Tabelle 2.4).

2.4.3 GÄNGIGE TONARTEN

Aus der Tonmenge der gleichstufigen chromatischen Tonleiter lassen sich viele Tonarten der westlichen Musik direkt ableiten. Für die Benennung der Tonarten wird der Grundton des Bildungsgesetzes dem Namen vorangestellt. Die wichtigsten Tonarten sind Heptatoniken, bei denen sieben Töne der chromatischen Tonleiter ausgewählt werden. Je nach Auswahl spricht man von Dur-Tonleitern, Moll-Tonleitern, Blues-Tonleitern und anderen. Auch die Stammtonreihe bildet eine Heptatonik und wird c-Dur-Tonleiter genannt. Tabelle 2.5 zeigt verschiedene Tonarten und die in ihnen enthaltenen Halb- und Ganztonschritte ausgehend vom Grundton. Für die Notation im Fünf-Linien-System werden am Anfang der Zeile hinter dem Notenschlüssel Tonartzeichen gesetzt, die als Vorzeichen für alle auf dieser Stufe folgenden Noten gelten. Somit können die Noten selbst weitestgehend vorzeichenlos notiert werden.

Neben den Tonarten, die als Untermenge der gleichförmigen chromatischen Tonleiter gebildet werden können, gibt es auch Tonarten, deren Intervalle sich nicht aus kleinen Sekunden zusammensetzen lassen und

Tabelle 2.5: Gängige Tonarten und die in ihnen enthaltenen Halb- und Ganztonschritte ausgehend vom Grundton

Tonart	Ganztonschritte
Pentatonik	1, 1, $^3/_2$, 1, $^3/_2$
Ganztonleiter	sechs Ganztonschritte
Dur-Tonleiter	1, 1, $^1/_2$, 1, 1, 1, $^1/_2$
Moll-Tonleiter	1, $^1/_2$, 1, 1, $^1/_2$, 1, 1
Blues-Tonleiter	$^3/_2$, $^1/_2$, $^1/_2$, $^1/_2$, $^1/_2$, 1
Chromatik	zwölf Halbtonschritte

deren Tonhöhen somit nicht mit den Tonhöhen der chromatischen Tonleiter zusammenfallen. Dies gilt beispielsweise für historische Tonarten oder Tonarten der modernen Musik und streng genommen auch für die Naturtonreihe.

2.4.4 CENT

Um noch feinere Intervalle als den Halbton darstellen zu können, wurde von Ellis das Cent-System eingeführt [Elli85]. Hierbei wird zusätzlich zur gleichstufigen chromatischen Tonleiter ein Halbton in 100 Cent aufgeteilt. Die Einteilung in Cent ist so gewählt, dass sich hiermit das kleinste vom Menschen noch wahrnehmbare Intervall zwischen zwei Sinustönen ausdrücken lässt. Dieses liegt beim direkten Vergleich zweier Töne mit Frequenzen um 1000 Hz bei etwa zwei bis drei Cent [Stai75, Grei05]. Das mathematische Intervall *I*, das durch einen Cent-Wert *c* angegeben wird, lässt sich aus der folgenden Formel berechnen:

$$I(c) = 2^{(c/1200)}. \tag{2.4}$$

Die große Ähnlichkeit der Formel für die Intervalle der gleichförmigen chromatischen Intervalle (2.3) und der Formel für die Cent-Intervalle (2.4) zeigt die enge Verknüpfung der beiden Systeme. Da die Definition von Cent rein mathematisch ist, lassen sich beispielsweise für technische Anwendungen auch problemlos Bruchteile von Cent oder negative Werte verwenden [DIN13320].

3 MUSIKINSTRUMENTE

Musikinstrumente sind Gegenstände, mit denen Klangereignisse zum Musizieren erzeugt werden können. Den weltweit größten Bekanntheitsgrad und die größte Verbreitung haben die klassischen Instrumente des Symphonieorchesters [Meye04, Wein08]. Weiterhin gibt es eine Vielzahl von Instrumenten aus unterschiedlichen Kulturkreisen oder historischen Epochen, auf die jedoch wegen ihres speziellen Charakters in dieser Arbeit nicht eingegangen wird. In diesem Kapitel werden zuerst verschiedene Systematiken zur Klassifikation von Musikinstrumenten vorgestellt, im zweiten Teil werden, der in dieser Arbeit verwendeten Systematik folgend, die charakteristischen Eigenschaften der verschiedenen Instrumentenfamilien beschrieben. Hierbei sind neben den allgemeinen Eigenschaften der einzelnen Instrumentenfamilien die für eine automatische Identifikation bzw. Klassifikation besonders relevanten Merkmale hervorgehoben.

3.1 SYSTEMATIKEN

Die Unterteilung der Musikinstrumente in verschiedene Gattungen nennt man Systematik. In den gängigen Systematiken sind die zur Klassifikation herangezogenen Hauptmerkmale das Genre, in dem die Instrumente primär auftauchen, die Spielart, das verwendete Baumaterial und das primär schwingende Bauteil sowie die daraus resultierenden Klangeigenschaften. Weiterhin ist es im Zusammenhang mit historischen oder ethnologischen Betrachtungen üblich, Systematiken zu verwenden, die

Musikinstrumente nach Epochen und Kulturkreisen klassifizieren [Vo-Sa14, Sach65, Kueh75, Mart99, Bär 03, BuLe04,].

3.1.1 GENRE

Bei der Klassifikation der Musikinstrumente nach dem Genre werden die für jedes Genre typischerweise benötigten Instrumente in einer Klasse zusammengefasst. In dieser Systematik fallen einige Musikinstrumente in mehrere Klassen, so dass eine eindeutige Zuordnung nicht mehr möglich ist. Gängige Unterteilungen nach dem Genre sind:

- die Instrumente des klassischen **Symphonieorchesters**, bestehend aus Geigen, Bratschen, Celli, Kontrabässen, Querflöten, Oboen, Klarinetten, Fagotten, Waldhörnern, Trompeten, Posaunen, Tuben und Schlaginstrumenten,
- das **Streichquartett**, bestehend aus zwei Geigen, einer Bratsche und einem Cello,
- die Instrumente einer **Rock-/Popband** unter anderem bestehend aus einem Schlagzeug, Gitarren und Bassgitarren.

Obwohl die Klassifikation nach dem Genre für die Unterteilung der Musikinstrumente Nachteile aufweist, wird sie häufig verwendet. Die Nachteile für die automatische Klassifikation im Sinne dieser Arbeit sind, dass die Identifikation des Musikinstrumentes eng an die Identifikation des Genres geknüpft ist. Die Genreklassifikation selbst ist jedoch nicht trivial und stellt ein weitläufiges Forschungsgebiet dar [BuLe03, BuLe04].

3.1.2 SPIELART

Die reine Unterteilung nach der Spielart wird oftmals im Musikunterricht verwendet und teilt die Instrumente nur danach ein, über welche Techniken der Musiker Klänge erzeugt. Die fünf Hauptgruppen der Systematik sind [Bär 03]:

- **Blasinstrumente**, bei denen der Musiker Luft in das Instrument bläst und evtl. mit den Händen die Tonhöhe durch Griffe festlegt, z.B. Flöte, Oboe, Fagott, Klarinette, Saxophon, Horn, Trompete, Posaune und Mundharmonika,

- **Streichinstrumente**, bei denen der Musiker mit der einen Hand Saiten mit einem Bogen anstreicht und mit der anderen Hand die Saiten durch Griffe verkürzt, um so die Tonhöhe festzulegen, z.B. Geige, Bratsche, Cello und Kontrabass,
- **Zupfinstrumente**, bei denen der Musiker mit der einen Hand Saiten zupft und mit der anderen Hand die Saiten durch Griffe verkürzt, um so die Tonhöhe festzulegen, z.B. Gitarre, Laute und Mandoline,
- **Schlaginstrumente**, die vom Musiker mit den Händen oder Schlägeln zur Klangerzeugung angeschlagen werden, z.B. Trommeln, Pauken, Xylophone, Marimbas, Glocken, Becken und Klanghölzer,
- **Tasteninstrumente**, bei denen der Musiker Tasten drückt, um verschiedene Klänge zu erzeugen, z.B. Harmonium, Klavier, Cembalo und Synthesizer-Keyboard.

Obwohl eine Unterteilung nach der Spielart für den Menschen sehr intuitiv und einfach erfolgen kann, ist sie für die automatische Identifikation und Klassifikation anhand des Klanges unbrauchbar, da die Spielart den Klangereignissen in der Regel nicht angehört werden kann.

3.1.3 HORNBOSTEL-SACHS

Die ursprünglich im Jahre 1914 veröffentlichte Systematik von Hornbostel und Sachs klassifiziert Musikinstrumente nach dem primär schwingenden, Klang gebenden Bauteil [VoSa14]. Die Systematik ist sehr umfangreich und sieht bis zu sechs Gliederungsebenen vor. Sie bildet aufgrund ihres Umfangs und ihres Detailgrades bis heute in einer leicht modifizierten Version den Standard für musikwissenschaftliche Betrachtungen. Die fünf Hauptgruppen der Systematik sind:

1. **Idiophone** (Selbstklinger): z.B. Xylophon, Marimba, Glocke, Becken, Klanghölzer, Rassel und Reibe,
2. **Membranophone** (Fellklinger): z.B. Trommel und Pauke,

3. **Chordophone** (Saiteninstrumente): z.B. Geige, Bratsche, Cello, Kontrabass, Gitarre, Laute, Mandoline, Harfe, Klavier und Cembalo,
4. **Aerophone** (Blasinstrumente): z.B. Oboe, Fagott, Klarinette, Saxophon, Horn, Trompete, Posaune, Tuba, Flöte und Kirchenorgel,
5. **Elektrophone** (Elektromechanische/Elektronische Instrumente): z.B. E-Gitarre, Trautonium, Theremin und Synthesizer.

Die Hornbostel-Sachs-Systematik erlaubt neben der übergeordneten Unterscheidung nach dem primär schwingenden Bauteil eine Unterteilung in Untergruppen anhand der Spielart. Jeder Klassifikationsebene ist eine feste Nummer zugeordnet, so dass alle Instrumente durch eine feste Gliederungsnummer eindeutig identifiziert werden können. So bedeutet beispielsweise die Gliederungsnummer 423.233 eine chromatisch in der Tonhöhe veränderbare Trompete mit Ventilen, deren Röhre überwiegend zylindrisch verläuft (vgl. auch Abschnitt 3.2.3):

4 – Aerophon,
42 – Blasinstrument,
423 – Trompete,
423.2 – Chromatische Trompete,
423.23 – Ventiltrompete,
423.233 – zylindrische Röhre.

3.1.4 KLANGCHARAKTERISTIK

Das Schema von Hornbostel und Sachs stellt eine gute Systematik für musiktheoretische Betrachtungen dar, für die automatische Identifikation und Klassifikation von Musikinstrumenten anhand ihres Klanges kann es aber in seiner direkten Form nicht verwendet werden. So gibt es beispielsweise Instrumente, von denen sich gewisse Töne sehr ähneln, obwohl sie in dem baumartigen Schema von Hornbostel-Sachs in völlig unterschiedliche Zweige der Systematik eingeordnet werden. Beispiele hierfür sind tiefe Basstöne eines Klaviers und gezupfte Klänge eines Kontrabasses oder gestrichene Töne einer Bratsche und die Klänge eines Fagotts. Diese Klänge sind als Einzelklänge für musikalisch geschulte Menschen zwar noch unterscheidbar, treten sie jedoch in einer Klangmi-

schung ohne den dazugehörigen Kontext einer Melodie auf, sind sie faktisch nicht mehr zu unterscheiden. Somit kann die Hornbostel-Sachs-Systematik nicht direkt für die automatische Identifikation und Klassifikation von Musikinstrumenten anhand ihres Klanges verwendet werden.

Für die in der vorliegenden Arbeit entwickelte automatische Musikinstrumentenidentifikation und Klassifikation wurde deshalb eine auf dem Schema von Hornbostel und Sachs aufbauende Systematik verwendet, die den Aspekt der Spielweise und die daraus resultierenden Eigenschaften der erzeugten Klangereignisse stärker berücksichtigt. Eine ähnliche Systematik wurde bereits von Martin vorgeschlagen [Mart99], für die vorliegende Arbeit wurde sie jedoch zusätzlich verfeinert. Die sieben Hauptgruppen der verwendeten Systematik sind:

- **Flöteninstrumente**: z.B. Blockflöte, Querflöte, Panflöte und Kirchenorgel,
- **Rohrblattinstrumente**: z.B. Oboe, Fagott, Klarinette und Saxophon,
- **Trompeteninstrumente**: z.B. Horn, Trompete, Posaune und Tuba,
- **Harmonikainstrumente**: z.B. Akkordeon, Mundharmonika und Harmonium,
- **Streichinstrumente**: z.B. Geige, Bratsche, Cello und Kontrabass,
- **Zupfinstrumente**: z.B. Gitarre, Laute, Mandoline, Harfe, Klavier und Cembalo
- **Schlaginstrumente**: z.B. Trommel, Pauke, Xylophon, Marimba, Glocke, Becken, Klanghölzer, Rassel und Reibe.

Durch die verwendete Systematik lassen sich die Instrumente bereits auf der höchsten Gliederungsebene anhand des zeitlichen Lautstärkeverlaufs in zwei Kategorien unterteilen, in Instrumente mit konstantem Lautstärkeverlauf (Flöteninstrumente, Rohrblattinstrumente, Trompeteninstrumente, Harmonikainstrumente, Streichinstrumente) und Instrumente mit exponentiell abklingendem Lautstärkeverlauf (Zupfinstrumente und Schlaginstrumente), die sich klanglich erheblich unterscheiden.

3.2 INSTRUMENTENFAMILIEN

Aufgrund der Bauart und Spielweise der verschiedenen Musikinstrumente ergeben sich charakteristische Eigenschaften der produzierten Klänge und Geräusche. Dies bedeutet, dass die Menge aller mit einem bestimmten Musikinstrument erzeugbaren Klänge begrenzt ist. In der Regel lassen sich die charakteristischen Klangeigenschaften eines Musikinstrumentes sehr eng umschreiben, was die automatische Identifikation und Klassifikation von Musikinstrumenten, wie sie in dieser Arbeit behandelt wird, erst ermöglicht [FlRo91, Wink98].

Die Instrumente der einzelnen Instrumentenfamilien und ihre Klangeigenschaften werden in den folgenden Abschnitten detailliert beschrieben.

3.2.1 FLÖTENINSTRUMENTE

Flöteninstrumente bestehen aus einer oder mehreren Klangröhren, die in der Regel aus Holz oder dünnwandigem Messing gefertigt sind. An der einen Seite der Klangröhre ist ein Schnabelmundstück oder ein Schneidkantenmundstück angebracht. Über die Röhre verteilt befinden sich verschieden große Bohrungen. Die wichtigsten Flöteninstrumente sind die Blockflöte und die Querflöte (Bild 3.1) sowie die Panflöte und die Kirchenorgel.

Für die Klangerzeugung wird ein Luftstrom auf eine scharfe Kante gelenkt und von dieser zerschnitten. Hierbei bilden sich Luftwirbel innerhalb des Klangkörpers, die je nach Resonanzfrequenz der Luftsäule Klänge mit unterschiedlichen Tonhöhen erzeugen (Bild 3.2). Die Resonanzfrequenz der Luftsäule und damit die Tonhöhe wird durch das Verschließen einzelner Löcher über bestimmte Griffmuster festgelegt. Hierzu werden die Löcher entweder direkt mit dem Finger oder aber über eine Klappenmechanik verschlossen [Hemp01].

Bei Schnabelmundstücken (Blockflöte) wird der Ton innerhalb des Instruments auf die Schneidkante gelenkt, bei Schneidkantenmundstücken bläst der Musiker direkt auf die Schneidkante und kann hierdurch den Ton zusätzlich beeinflussen (Querflöte, Panflöte).

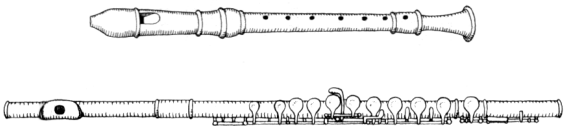


Bild 3.1: Verschiedene Flöteninstrumente: Blockflöte und Querflöte [Brin98]

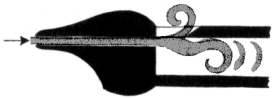


Bild 3.2: Klangerzeugung in einem Flöteninstrument an einer Schneidkante [Brin98]

Durch die Bohrungen werden die Resonanzeigenschaften der Klangröhre beeinflusst, indem die Lage bestimmter Wellenbäuche durch die Löcher erzwungen wird. Durch diese komplexe Veränderung der Resonanzeigenschaften werden einige Obertöne unterdrückt bzw. stark abgeschwächt, was zum charakteristischen Klang von Flöteninstrumenten beiträgt.

Die charakteristischen klanglichen Eigenschaften von Flöteninstrumenten sind eine dunkle, milde Klangfarbe mit einer relativ ausgeprägten Formantstruktur, bei der gewisse Frequenzbereiche und die in ihnen enthaltenen Harmonischen sehr dominant sind. Dies kann auch dazu führen, dass einzelne Obertöne lauter als der Grundton sind. Die an der Schneidkante auftretenden chaotischen Luftwirbel tragen weiterhin zu einem hohen Rauschanteil, der vor allem im hohen Frequenzbereich auftritt, bei [Brin98] (Bild 3.3).

Die Klänge von Flöteninstrumenten haben, wie die Klänge aller geblasenen Instrumente, eine ca. 50 bis 100 ms dauernde Einschwingphase und gehen dann in der Haltephase in einen konstanten Verlauf über. Die Ausschwingphase ist ebenfalls sehr kurz, und der Zusammenbruch der schwingenden Luftsäule, und damit des Klanges, erfolgt nach dem Ausbleiben der Luft als Schwingungsanregung innerhalb von ca. 50 ms. Somit ergibt sich eine blockartige Hüllkurve, die auch als Orgelhüllkurve

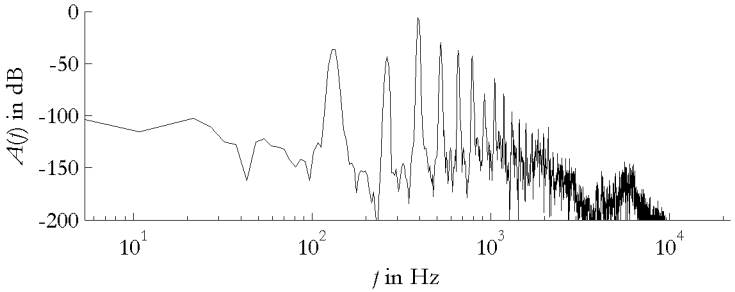


Bild 3.3: Kurzzeitspektrum eines Flötenklanges mit einer Grundfrequenz von 130,81 Hz (SPN: C3)

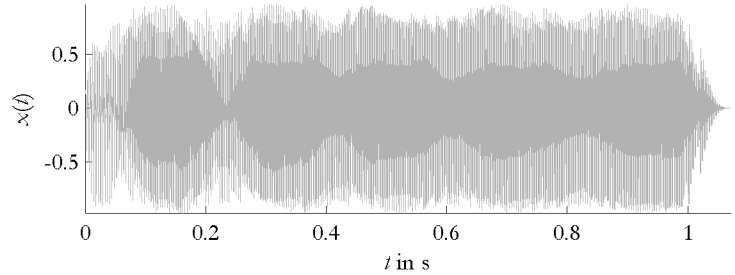


Bild 3.4: Zeitsignal eines Flötenklanges mit einer Grundfrequenz von 130,81 Hz (SPN: C3)

bezeichnet wird, deren Länge direkt vom Spieler bestimmt wird. Während der kurzen Einschwingphase kann der Klang je nach Spielweise sehr rauschartig sein. In der Haltephase haben Flöteninstrumente eine relativ statische Klangfarbe, die lediglich in der Lautstärke variiert [Brin98] (Bild 3.4).

3.2.2 ROHRBLATTINSTRUMENTE

Rohrblattinstrumente bestehen aus einer Klangröhre aus Holz oder dünnwandigem Messing, die in ihrem Aufbau den Klangröhren von Flöteninstrumenten recht ähnlich ist. Als Mundstück tragen Rohrblattinstrumente entweder ein Doppelrohrblatt oder ein Einzelrohrblatt. Die Rohrblätter der Mundstücke sind aus flachen, dünnen Holzscheiben

Bild 3.5: Verschiedene Rohrblattinstrumente: Klarinette, Saxophon, Oboe und Fagott [Brin98]

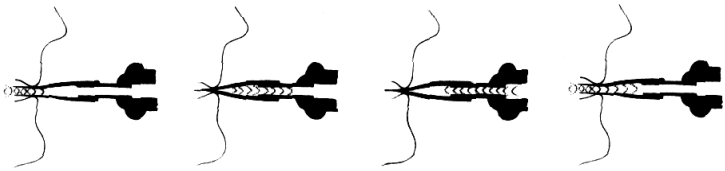


Bild 3.6: Klangerzeugung in einem Rohrblattinstrument mit Doppelrohrblatt [Brin98]

gefertigt. Die wichtigsten Rohrblattinstrumente sind die Klarinette, das Saxophon, die Oboe und das Fagott (Bild 3.5).

Für die Klangerzeugung werden bei Mundstücken mit Doppelrohrblättern die Rohrblätter mit den Lippen umschlossen und durch die Lippenspannung leicht zusammengedrückt (Bild 3.6), so dass sie im Atemstrom anfangen zu schwingen und der Luftstrom periodisch unterbrochen wird. Bei Mundstücken mit Einzelrohrblatt wird das einzelne Rohrblatt gegen eine feststehende Wand gedrückt, das Prinzip der Klangerzeugung ist jedoch das gleiche wie bei Doppelrohrblättern [Step68].

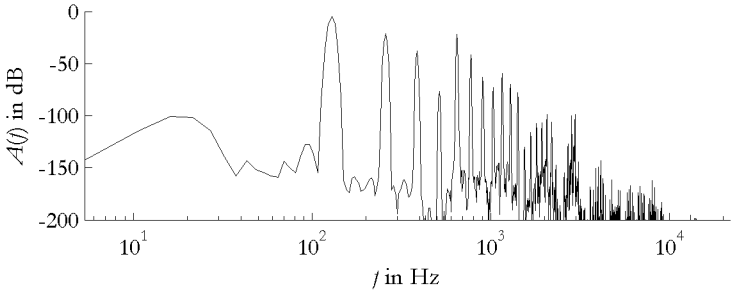


Bild 3.7: Kurzzeitspektrum eines Oboenklanges mit einer Grundfrequenz von 130,81 Hz (SPN: C3)

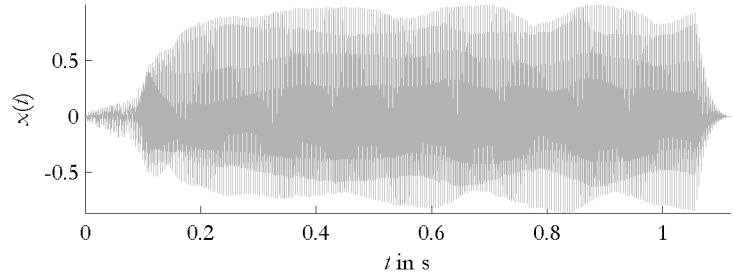


Bild 3.8: Zeitsignal eines Oboenklanges mit einer Grundfrequenz von 130,81 Hz (SPN: C3)

Die Grundfrequenz der erzeugten Schwingung ist, ähnlich wie bei Flöteninstrumenten, von der Resonanzfrequenz der Klangröhre abhängig. Die Resonanzfrequenz wird durch das Verschließen einzelner Löcher über bestimmte Griffmuster festgelegt. Hierzu werden die Löcher entweder direkt mit dem Finger oder aber über eine Klappenmechanik verschlossen.

Von ihren klanglichen Eigenschaften sind Rohrblattinstrumente eine Mischung aus Flöteninstrumenten und Trompeteninstrumenten (vgl. Abschnitt 3.2.3). Die Schwingungsanregung hat Ähnlichkeit mit Trompeteninstrumenten, wobei das abruptere, impulsartige Öffnen und Schließen des Rohrblatts sehr obertonreiche Schwingungen hervorruft. Die Bauweise der Klangröhre aus relativ dickwandigem Holz und die

Erzwingung einer bestimmten Resonanzfrequenz durch das Öffnen und Schließen der Bohrungen hingegen erzeugen ähnliche Formanteigenschaften im Obertonspektrum wie bei Flöteninstrumenten. Da es keine Schnittkante wie bei Flöteninstrumenten gibt, treten keine rauschartigen Nebengeräusche auf [Zieg00] (Bild 3.7).

Bei Klängen von Rohrblattinstrumenten sind die Einschwingphasen stärker ausgeprägt als bei anderen Blasinstrumenten. Es ergibt sich ein rauschartiger Verlauf von bis zu 300 ms Länge, der durch das Strömen der Luft in der Klangröhre entsteht. Besonders bei Saxophonen ist dieses Verhalten sehr charakteristisch. Die eigentliche tonale Klangausprägung hat eine für Blasinstrumente typische blockartige Hüllkurve mit einer Ausschwingphase von ca. 50 ms [Zieg00] (Bild 3.8).

Flöteninstrumente und Rohrblattinstrumente werden oft unter dem Begriff Holzblasinstrumente zusammengefasst, der aber fälschlicherweise auf das verwendete Material hinweist [Brin98].

3.2.3 TROMPETENINSTRUMENTE

Trompeteninstrumente bestehen aus einer Tonröhre, meist aus dünnwandigem Messing, an deren oberer Seite ein Kessel- oder Trichtermundstück angebracht ist. Die Tonröhre weitet sich zum unteren Ende hin auf und endet in einem Schalltrichter. Um das Instrument kompakt zu halten, ist die Tonröhre in der Regel aufgewickelt. Die wichtigsten Trompeteninstrumente sind die Posaune, das Horn und die Trompete (Bild 3.9) sowie die Tuba.

Ein Trompeteninstrument wird gespielt, indem die in der Tonröhre befindliche Luftsäule zu einer stehenden Welle angeregt wird. Hierzu wird die Atemluft des Bläsers durch seine Lippen periodisch unterbrochen, was als Resonanzreaktion in der Tonröhre eine stehende Welle ausprägt. Die Tonhöhe ergibt sich durch die Spannung der Lippen des Bläsers, die somit auch die primär schwingenden Elemente eines Trompeteninstruments darstellen [Hemp01]. Physikalisch gesehen haben Trompeteninstrumente den Aufbau einer Polsterpfeife (Bild 3.10).

Die Luftsäule in der Tonröhre kann prinzipiell nur in ihren Grundmoden angeregt werden, d.h. für die erzeugten Klänge muss die Rohrlänge L ein Vielfaches n der halben Wellenlänge λ der Tonhöhe sein:

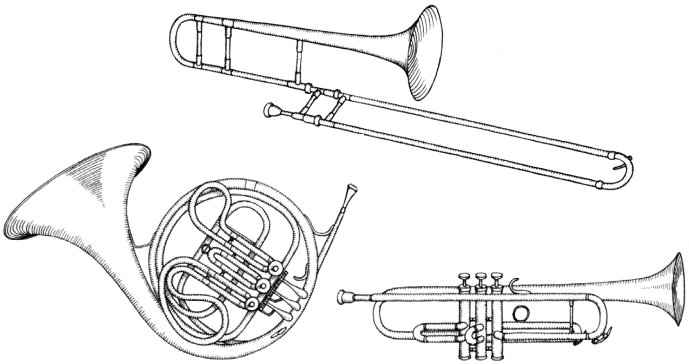


Bild 3.9: Verschiedene Trompeteninstrumente: Posaune, Horn und Trompete [Brin98]

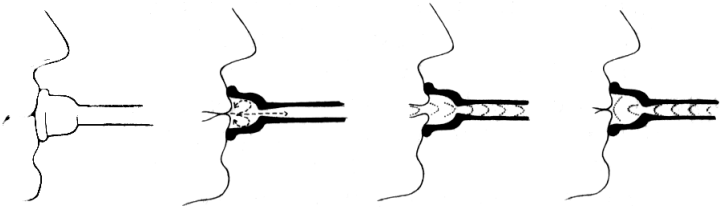


Bild 3.10: Klangerzeugung in einem Trompeteninstrument mit Kesselmundstück [Brin98]

$$\lambda = \frac{2L}{n}. \tag{3.1}$$

Aufgrund der festen Länge ergibt sich aus dieser Einschränkung für die erzeugbaren Tonhöhen die Naturtonreihe. Um trotzdem alle Töne der chromatischen Tonleiter spielen zu können, haben Trompeten und Hörner einen Ventilmechanismus, über den die Länge der Tonröhre verändert werden kann, um so einen anderen Grundton zu erzeugen, dessen Naturtonreihe die fehlenden Töne enthält. Bei der Posaune kann die Rohrlänge, und somit die Tonröhre, direkt vom Spieler über einen Zugriegel eingestellt werden, so dass die Posaune das einzige Trompeteninstrument ist, das echte Glissandi spielen kann.

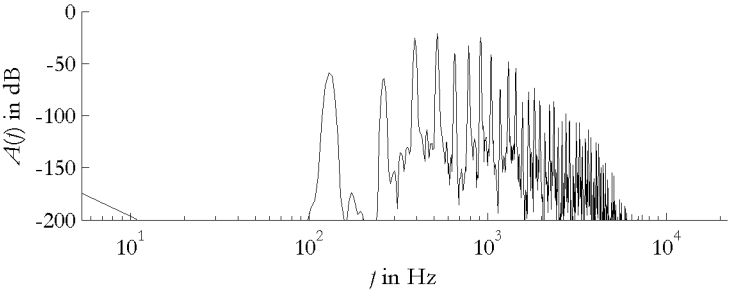


Bild 3.11: Kurzzeitspektrum eines Trompetenklanges mit einer Grundfrequenz von 130,81 Hz (SPN: C3)

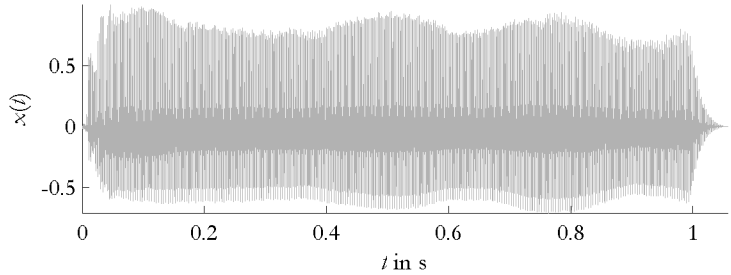


Bild 3.12: Zeitsignal eines Trompetenklanges mit einer Grundfrequenz von 130,81 Hz (SPN: C3)

Die charakteristischen klanglichen Eigenschaften von Trompeteninstrumenten sind eine sehr helle Klangfarbe und eine ausgewogene Obertonausprägung, die an eine Sägezahnschwingung erinnert. Da die Anregung der Luftsäule durch den Bläser impulsartig ist, schwingen neben dem eigentlichen, in der Tonröhre angeregten Ton immer auch die anderen Moden der Luftsäule mit, die im Klang in einem ausgewogenen Verhältnis enthalten sind. Somit entstehen sowohl gerade als auch ungerade Harmonische. Durch das dünnwandige Außenmaterial der Tonröhre bilden sich relativ starke Resonanzen für eine große Anzahl von Obertönen aus, was die helle Klangfarbe ausmacht [Hemp01] (Bild 3.11).

Auch die Klänge von Trompeteninstrumenten haben eine blockartige Hüllkurve mit einer Einschwingphase von ca. 10 - 20 ms. Während der

Einschwingphase ist die Anregung durch die Lippen relativ impulsartig und unregelmäßig, was zu geräuschhaften, teilweise perkussiv anmutenden Klangkomponenten führt. Nach dem Ausbleiben der Luft als Schwingungsanregung bricht die schwingende Luftsäule, und damit der Klang, innerhalb einer kurzen Ausschwingphase von ca. 50 ms zusammen [Hemp01] (Bild 3.12).

Für die Klasse der Trompeteninstrumente wird oftmals auch synonym der Begriff Blechblasinstrumente verwendet, der aber fälschlicherweise auf das verwendete Material hinweist [Brin98]. Entscheidend für die Klassifikation als Trompeteninstrument ist die Erzeugung der schwingenden Luftsäule durch die Lippenspannung des Bläsers. So werden z.B. Alphörner auch als Trompeteninstrumente bzw. Blechblasinstrumente bezeichnet, obwohl sie aus Holz gefertigt sind. Alphörner sind einige der wenigen Trompeteninstrumente, deren Tonröhre nicht gewickelt ist [Bär 03].

3.2.4 HARMONIKAINSTRUMENTE

Bei Harmonikainstrumenten sorgt eine dünne, schwingende Zunge für die Klangerzeugung. Die Zunge ist in der Regel aus Metall gefertigt und ist mittig über einem Luftkanal angebracht, der von der Zunge im Ruhezustand größtenteils verschlossen wird. Wird ein Ton geblasen, biegt sich die Zunge durch den entstehenden Luftdruck und gibt den Luftkanal kurzzeitig frei. Die Luft kann durch die entstehende Öffnung entweichen, und der Luftdruck auf die Zunge lässt nach, die aufgrund der Elastizität des Materials zurückschnellt und die Öffnung erneut verschließt. Durch das abwechselnde Öffnen und Schließen des Luftkanals wird die hindurchströmende Luft in Schwingung versetzt [Hemp01] (Bild 3.13).

Da die Metallzunge frei schwingen kann, wird sie auch als durchschlagende Zunge bezeichnet. Die Grundfrequenz der erzeugten Schwingung hängt von den Schwingungseigenschaften der Metallzunge ab. Um verschiedene Tonhöhen spielen zu können, muss für jede Tonhöhe eine Durchschlagzunge mit entsprechendem Luftkanal vorhanden sein, der entweder direkt oder über eine Klappenmechanik mit Luft versorgt wird.

Im Gegensatz zu anderen Blasinstrumenten ist eine der primären Schwingungsquelle nachgeschaltete Resonanzröhre nicht unbedingt er-

Bild 3.13: Klangerzeugung in einem Harmonikainstrument mit durchschlagender Zunge [Bär 03]

forderlich. Die durch die Durchschlagzunge angestoßene Schwingung braucht vielmehr einen festen Resonanzkörper, um sich entfalten zu können. Deshalb sind die Durchschlagzungen direkt auf einem in der Regel hölzernen Resonanzkörper befestigt. Durch diese Eigenschaft unterscheiden sich Harmonikainstrumente von anderen Blasinstrumenten und nehmen eine Sonderstellung ein. Die wichtigsten Harmonikainstrumente sind das Akkordeon, die Mundharmonika und das Harmonium.

Die Durchschlagzungen erzeugen eine Grundschwingung, die aufgrund der Materialträgheit und Steifheit der Zungen die Form einer harmonisch verzerrten Sinusschwingung hat. Neben den Materialeigenschaften hängen die Verzerrungen weiterhin von der Stärke des Luftstroms ab, was dazu führt, dass die erzeugte Grundschwingung mit stärker werdendem Luftstrom nicht nur lauter, sondern auch rechteckförmiger wird. Somit haben Klänge von Harmonikainstrumenten prinzipiell viele Harmonische als ungerade Vielfache der Grundfrequenz. Durch Materialungenauigkeiten schwingen die Zungen nicht völlig symmetrisch, so dass auch gerade Vielfache der Grundfrequenz entstehen (Bild 3.14).

Der durch die Durchschlagzungen erzeugte Klang ist in seiner Klangfarbe relativ statisch und somit aus musikalischer Sicht monoton. Um den Gesamtklang eines Harmonikainstrumentes interessanter zu machen, kann der Spieler über die eingebauten Mechaniken relativ einfach mehrere Grundklänge gleichzeitig als Akkorde anregen. Der Klang wird weiterhin stark von der Beschaffenheit des Resonanzkörpers beeinflusst. Durch den gemeinsamen Resonanzkörper für alle Klänge ergeben sich Formanteigenschaften, die besonders den Mittenbereich verstärken und die Höhen abschwächen. So haben Harmonikainstrumente zwar einen hellen Klang, der aber in den Höhen je nach Bauart des Resonanzkörpers abgemildert ist. So klingt eine Mundharmonika deutlich heller als

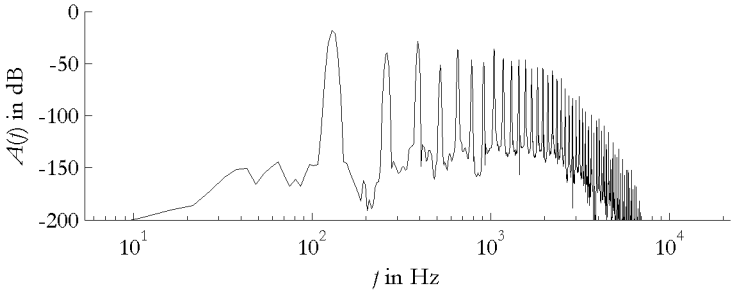


Bild 3.14: Kurzzeitspektrum eines Harmoniumklanges mit einer Grundfrequenz von 130,81 Hz (SPN: C3)

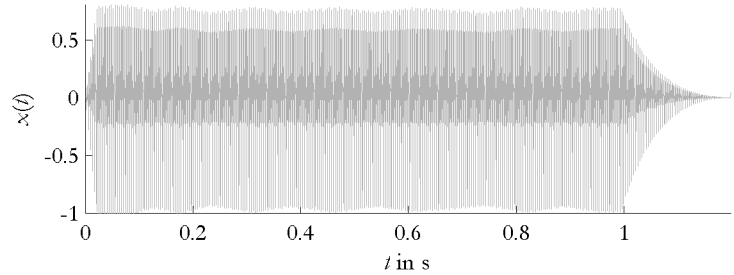


Bild 3.15: Zeitsignal eines Harmoniumklanges mit einer Grundfrequenz von 130,81 Hz (SPN: C3)

ein Akkordeon, was wiederum heller klingt als ein Harmonium [Hemp01].

Die zeitliche Entwicklung des Klanges hängt stark von dem restlichen Aufbau des Instruments und von der Tatsache ab, ob eine Mechanik zum Ansteuern der verschiedenen Luftkanäle verwendet wird, oder ob die Kanäle direkt angeblasen werden (Bild 3.15). Bei der Verwendung einer Mechanik ist oftmals während der ca. 30 ms dauernden Einschwingphase ein leichtes, perkussiv anmutendes Überschwingen der Durchschlagzungen vernehmbar, das unharmonisch wirkt, bevor der Klang einen eingeschwungenen Zustand einnimmt. Bei einem großen Resonanzkörper (z.B. Harmonium) ist die Ausschwingphase relativ deutlich wahrnehmbar und beträgt bis zu 300 ms [Hemp01].

3.2.5 STREICHINSTRUMENTE

Streichinstrumente bestehen aus einem hohlen, hölzernen Resonanzkörper, dem Korpus, von dem aus eine hölzerne Brücke, der Hals, abgeht. Zwischen dem Ende des Halses und dem Korpus sind Saiten gespannt, die zur primären Klangerzeugung dienen. Die Saiten sind über einen dünnen Holzsteg gespannt, der die Schwingung der Saiten an den Korpus weiterleitet. Die wichtigsten Streichinstrumente sind die Geige, die Bratsche, das Cello und der Kontrabass (Bild 3.16).

Zur Klangerzeugung werden die Saiten über einen klassischerweise mit Pferdehaar bespannten Bogen angestrichen und so in Schwingung versetzt. Die Tonhöhen der erzeugten Klänge werden über die Längen der gestrichenen Saiten festgelegt. Hierzu drückt der Musiker die Saiten mit den Fingern auf das Griffbrett des Halses. Durch die Resonanz des Korpus wird die Schwingung einerseits laut genug, um sich in mehrstimmiger Musik durchsetzen zu können, zum anderen wird hierdurch die Klangfarbe maßgeblich beeinflusst.

Beim Anstreichen der Saite bleibt diese durch die raue Oberflächenstruktur der Bogenhaare kurzzeitig am Bogen haften und wird der Streichbewegung folgend wenige Millimeter ausgelenkt. Sobald die durch die Auslenkung auftretende Spannung größer als die Haftung wird, löst sich die Saite und gleitet zurück. Durch diesen ständig wiederkehrenden Vorgang wird der Saite eine sägezahnartige Schwingung aufgeprägt, die sich in der Saite ausbreitet und an deren Enden reflektiert wird. Weiterhin schwingt die Saite in verschiedenen Moden mit [FlRo91, Wink98].

Durch die Schwingungserzeugung haben die primär erzeugten Schwingungen der Saiten ein sehr obertonreiches Klangspektrum. Besonders durch die konstante Überlagerung der Saitenschwingung mit der Anregung durch den Bogen ergibt sich hier ein rauschartiger Klanganteil. Allerdings entfaltet sich der eigentliche Klang über den in Resonanz mitschwingenden Korpus, der als Bandpassfilter auftritt und dem Klang eine charakteristische Formantstruktur aufprägt. Diese bleibt aufgrund des gemeinsamen Resonanzkörpers für alle Saiten und Tonhöhen bestehen und hat bei hochwertigen Instrumenten eine Kerbe im Frequenzspektrum bei ca. 4000 Hz. Diese Kerbe trennt die tiefen Harmonischen

Bild 3.16: Verschiedene Streichinstrumente: Geige, Bratsche, Cello und Kontrabass [Brin98]

von den hochfrequenten Anteilen und verleiht dem Klang Brillanz und Lebendigkeit [Brin98] (Bild 3.17).

Streicherklänge haben eine gut wahrnehmbare Einschwingphase von ca. 100 bis 300 ms, in der der Klang einen erhöhten Rauschanteil hat. Durch die Massenträgheit der angeregten Saite erfolgt am Ende der Einschwingphase ein deutlich wahrnehmbares Überschwingen, bevor der Klang in die Haltephase übergeht. In der Haltephase kann der Klang je nach Spielart in Lautstärke und Tonhöhe konstant sein, aber auch stark durch Tremolo und Vibrato moduliert werden. Nach dem Absetzen des Bogens klingt ein Streichinstrument mit den hochfrequenten Klanganteilen schneller aus als mit den niederfrequenten. Die Ausschwingphase beträgt je nach Größe des Instruments und damit des Resonanzkörpers ca. 50 bis 250 ms [Brin98] (Bild 3.18).

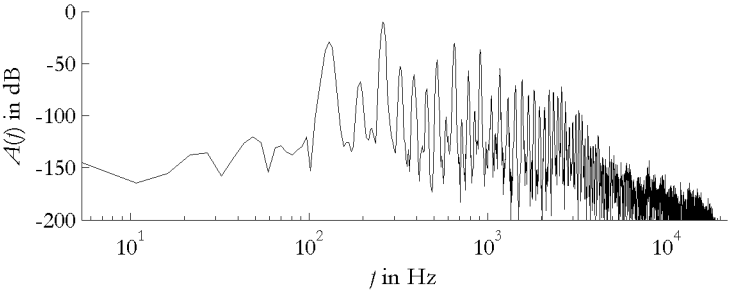


Bild 3.17: Kurzzeitspektrum eines Geigenklanges mit einer Grundfrequenz von 130,81 Hz (SPN: C3)

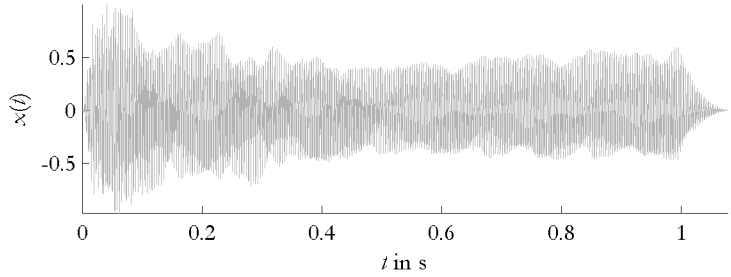


Bild 3.18: Zeitsignal eines Geigenklanges mit einer Grundfrequenz von 130,81 Hz (SPN: C3)

3.2.6 ZUPFINSTRUMENTE

Zupfinstrumente lassen sich in zwei Gruppen fassen. Instrumente der einen Gruppe haben einen ähnlichen Aufbau wie Streichinstrumente, d.h. einen Korpus und einen Hals, auf den die Saiten gespannt sind. Die Tonhöhe der Klänge wird ähnlich wie bei Streichinstrumenten durch Verkürzen der einzelnen Saiten gegriffen. Zur Schwingungsanregung werden die Saiten mit den Fingern oder einem kleinen Plättchen, dem Plektron, gezupft bzw. angeschlagen. Zu dieser Klasse der Zupfinstrumente gehören beispielsweise die Laute und die Mandoline (Bild 3.19) sowie die Ukulele und die Gitarre.

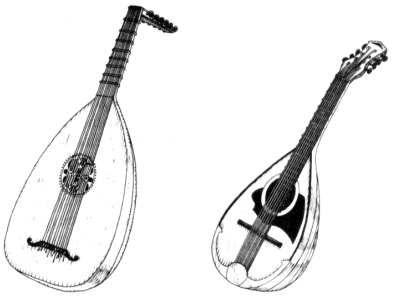


Bild 3.19: Verschiedene Zupfinstrumente: Laute und Mandoline [Brin98]

Die andere Gruppe der Zupfinstrumente sind Instrumente, die für jeden erzeugbaren Ton über eine oder mehrere Saiten verfügen, die nicht über ein Griffbrett manuell verkürzt werden können. Die wichtigsten Instrumente dieser Klasse sind die Harfe, das Klavier und das Cembalo. Da Instrumente dieser Art über sehr viele Saiten verfügen, sind sie naturgemäß relativ groß. Die Saiten sind in einem Rahmen gespannt, der entweder selbst als Resonanzkörper dient (Harfe) oder die Schwingung an einen speziellen Resonanzboden weitergibt (Klavier, Cembalo). Die Saiten werden bei der Harfe direkt vom Spieler angezupft, beim Klavier werden sie über eine Hammermechanik angeschlagen und beim Cembalo über eine Mechanik angezupft [Mich05].

Allen Zupfinstrumenten gemein ist die eigentliche Klangerzeugung, bei der eine Saite durch einen kurzen Impuls beim Zupfen oder Anschlagen um wenige Millimeter ausgelenkt wird. Der Impuls breitet sich in der Saite aus und wird an deren Enden reflektiert, wodurch die Saite in ihren Grundmoden zum Schwingen angeregt wird.

Durch die impulsartige Anregung haben die primär erzeugten Schwingungen der Saiten ein sehr obertonreiches Klangspektrum, vor allem wenn die Anregung dicht am Ende der Saite auftritt, da hierdurch besonders die hohen Schwingungsmoden angeregt werden. Im Gegensatz zu Streichinstrumenten treten bei Zupfinstrumenten keine nennenswerten Rauschanteile auf, da die Saite nach erfolgter Anregung nur in ihren Grundmoden schwingt (Bild 3.20).

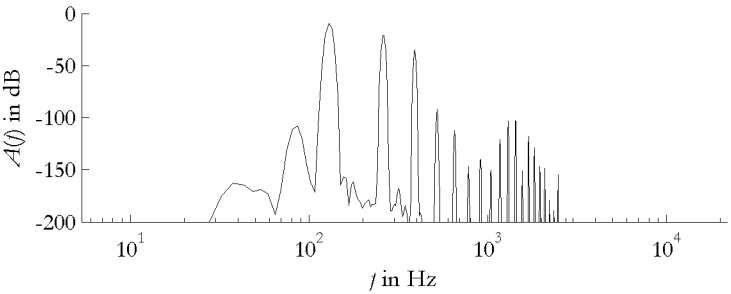


Bild 3.20: Kurzzeitspektrum eines Gitarrenklanges mit einer Grundfrequenz von 130,81 Hz (SPN: C3)

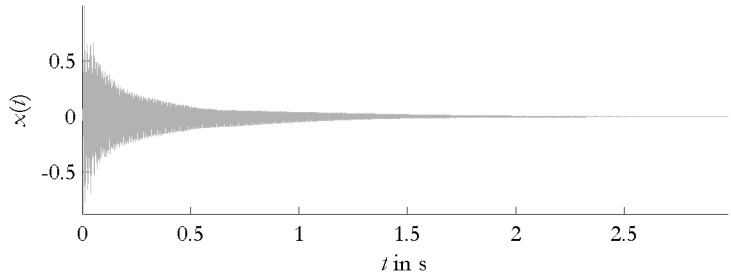


Bild 3.21: Zeitsignal eines Gitarrenklanges mit einer Grundfrequenz von 130,81 Hz (SPN: C3)

Der eigentliche Klang entfaltet sich ähnlich wie bei Streichinstrumenten über den in Resonanz mitschwingenden Korpus, der auch hier als Bandpassfilter auftritt. Somit haben auch die Klänge von Zupfinstrumenten eine charakteristische Formantstruktur, die bei Gitarreninstrumenten recht ähnlich zu Streichinstrumenten ist. Die Formantstruktur bleibt aufgrund des gemeinsamen Resonanzkörpers für alle Saiten und Tonhöhen bestehen [Hemp01].

Klänge von Zupfinstrumenten haben keine Einschwingphase, da sich der Klang direkt nach dem Anzupfen ausbildet (Bild 3.21). Außerdem haben sie keine Haltephase, da die schwingende Saite keine konstante Energiezufuhr erfährt. Vielmehr wird die Energie der schwingenden Saite kontinuierlich in Form von Schallwellen abgestrahlt, was zu einem

exponentiellen Ausklingverhalten führt. Diese sehr charakteristische Eigenschaft haben alle Instrumente, die nur eine impulsartige Anregung erfahren. Die deutlich ausgeprägte Ausklingphase hat abhängig von der Dämpfung der Saite eine Länge von einigen Sekunden, wobei die hochfrequenten Klanganteile schneller ausklingen als die niederfrequenten Klanganteile [Hemp01].

3.2.7 SCHLAGINSTRUMENTE

Schlaginstrumente sind Instrumente, deren Klänge durch impulsartige Anregung in Form von Schlägen erzeugt werden. Sie bilden eine sehr vielfältige Gruppe mit sehr unterschiedlichen Eigenschaften. Als einzige gemeinsame Eigenschaft ergibt sich der zeitliche Verlauf der Lautstärke, der ähnlich wie bei Zupfinstrumenten nur aus einer exponentiellen Ausklingphase besteht.

Es gibt Schlaginstrumente, die Klänge mit einer definierten Tonhöhe erzeugen, und auch Schlaginstrumente, die atonale Geräusche erzeugen und als reine Rhythmusinstrumente verwendet werden.

Innerhalb der Schlaginstrumente gibt es zwei Gruppen (Bild 3.22). Instrumente aus der Gruppe der Trommeln bestehen aus einem zylindrischen Klangkörper aus Holz, Metall oder Kunststoff, auf den ein Fell gespannt ist. Teilweise ist auf die untere Öffnung des Zylinders ein weiteres Fell als Resonanzboden gespannt. Zur Klangerzeugung wird das Fell angeschlagen, das seine Schwingungen an den Resonanzkörper und das Resonanzfell weitergibt. Trommeln können so gestimmt werden, dass sie eine mehr oder weniger ausgeprägte Tonhöhe besitzen. Sie können jedoch auch durch eine ungleichmäßige Spannung der Felle derart verstimmt werden, dass die Tonhöhe während des Klanges variiert oder nicht mehr eindeutig ausgeprägt ist. Weiterhin können Trommeln durch Modifikationen wie bei der Schnarrtrommel, auf deren Resonanzfell kleine Metallketten gespannt sind, überhaupt keine wahrnehmbare Tonhöhe mehr besitzen und sehr rauschartig klingen [Brin98].

Die zweite Gruppe der Schlaginstrumente bilden die Selbstklinger, die nur aus einem Klangkörper bestehen, der zur Klangerzeugung angeschlagen wird. Hierzu zählt eine große Vielzahl von Instrumenten unterschiedlichster Form und Herstellungsart. Einige können Klänge mit

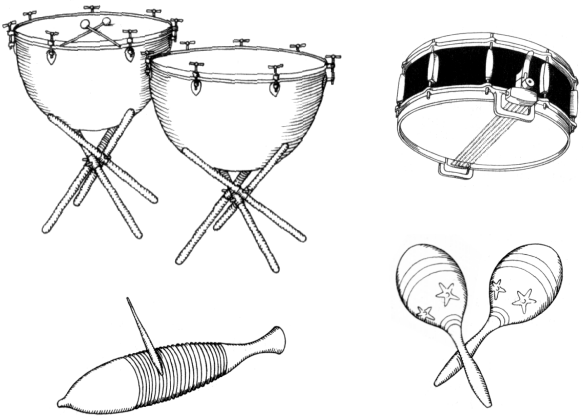


Bild 3.22: Verschiedene Schlaginstrumente: Pauken, Schnarrtrommel, Reibe und Rasseln [Brin98]

einer klar definierten Tonhöhe erzeugen, wie z.B. Xylophone, Marimbas oder Glocken, andere erzeugen Geräusche mit variierender Tonhöhe oder hohem Rauschanteil. Hierzu gehören beispielsweise Becken, Klanghölzer, Rasseln und Reiben [Brin98].

Da sich die Eigenschaften von Schlaginstrumenten aufgrund der großen Vielzahl an verfügbaren Bauarten nicht vereinheitlicht darstellen lassen, werden sie in der vorliegenden Arbeit für eine automatische Identifikation bzw. Klassifikation nicht in Betracht gezogen.

4 MERKMALSEXTRAKTION

Verfahren der Merkmalsextraktion berechnen aus analysierten Audiosignalen Merkmale, die bestimmte Aspekte der Audiosignale gesondert beschreiben. In diesem Kapitel werden der Aufbau sowie die Anforderungen an Merkmale genannt und zusammen mit den verschiedenen, auftretenden Klassifikationsszenarien erläutert. Weiterhin werden die für die automatische Identifikation und Klassifikation von Musikinstrumenten relevanten Merkmale zusammen mit ihrer Extraktion detailliert beschrieben. Abschließend werden Probleme, die durch Merkmalsvektoren mit einer zu großen Dimensionalität entstehen können, zusammen mit Lösungsansätzen der Dimensionsreduktion erklärt.

4.1 MERKMALE

Der generelle Aufbau eines Merkmals sowie die im Rahmen der Musikinstrumentenidentifikation und Klassifikation auftretenden Klassifikationsszenarien werden in den folgenden Abschnitten beschrieben.

4.1.1 AUFBAU

Ein einzelnes Merkmal stellt einen oder mehrere Zahlenwerte dar, die aus einem Audiodatenblock berechnet werden können. Aufgrund der blockbasierten Verarbeitung hat jedes Merkmal eine bestimmte zeitliche Ausdehnung. Die Blöcke werden über ein gleitendes Fenster aus dem zu analysierenden Audiomaterial ausgeschnitten, so dass die Merkmale selbst so genannte Merkmalssignale bilden. Aufgrund der Blockabstände

sind die Merkmalssignale, bezogen auf die Abtastfrequenz des Audiosignals, unterabgetastet.

Werden im Rahmen der Merkmalsextraktion mehrere Merkmale berechnet, so lassen sich diese als ein Merkmalsvektor **a** der Länge R darstellen, bei dem jede Komponente a_r genau ein Merkmal oder einen Aspekt, d.h. eine Komponente, eines mehrdimensionalen Merkmals beschreibt.

Die zu C Zeitpunkten extrahierten Merkmalsvektoren $\mathbf{a}_c$ lassen sich als Spaltenvektoren in einer $R \times C$ großen Merkmalsmatrix **A** zusammenfassen. Die Zeilen der Merkmalsmatrix enthalten die zu einem Merkmal oder Merkmalsaspekt gehörenden Merkmalssignale.

4.1.2 ANFORDERUNGEN

Eine Anforderung an ein Merkmal ist, dass sich mit ihm Daten aus verschiedenen Klassen mit großer Aussagekraft unterscheiden lassen. Hierfür wird zum einen eine geringe Intra-Klassenvarianz, d.h. eine kleine Variation der Merkmalswerte innerhalb einer Klasse, zum anderen eine hohe Inter-Klassenvarianz, d.h. eine große Variation der Merkmalswerte für Daten aus unterschiedlichen Klassen, gefordert. Für ein aussagekräftiges Merkmal ergeben sich somit sehr ähnliche Werte innerhalb einer Klasse und deutlich unterschiedliche Werte zwischen den Datensätzen zweier Klassen [Rohd03].

Ein aussagekräftiges Merkmal sollte weiterhin invariant gegenüber Transformationen sein, die die eigentliche Information des zugehörigen Audiosignals nicht verändern. Dies bezieht sich auf die absolute Lautstärke, auf die Signalqualität in Bezug auf Rauschanteile und auf die Bandbreite des Audiosignals.

4.2 EXTRAKTIONSSZENARIEN

Die Qualität verschiedener Merkmalsextraktionsverfahren ist immer von dem Szenario abhängig, in dem es verwendet wird. Die verschiedenen möglichen Szenarien im Rahmen der automatischen Musikinstrumentenidentifikation und Klassifikation werden in den folgenden Abschnitten beschrieben.

4.2.1 MONOPHONE MUSIKSIGNALE

Bei der Merkmalsextraktion aus monophonen Musiksignalen bestehen die analysierten Musiksignale aus einzelnen, kurzen Klängen oder Geräuschen. Das Szenario ist durch die folgenden Rahmenbedingungen definiert:

- Für jedes zu analysierende Klangereignis sind die Segmentgrenzen, d.h. der Anfang und das Ende, klar definiert und bekannt.
- Die Klangereignisse überlappen sich weder gänzlich noch teilweise.

Aufgrund dieser Rahmenbedingungen können keine Melodien, Akkorde oder ganze Musikstücke direkt verarbeitet werden. Dieses Szenario tritt beispielsweise bei der Analyse von Klangdatenbanken und Multimediaarchiven auf, wobei die verschiedenen Klangereignisse seriell in einer Audiodatei mit den dazugehörigen Segmentierungsinformationen oder jeweils in eigenen Audiodateien gespeichert sind. Die Merkmalsextraktion aus monophonen Musiksignalen hat sich in der Beschränktheit der Aufgabenstellung als eigenständiges Forschungsgebiet entwickelt, zu dem eine Vielzahl an Publikationen existiert [MaMo99, ErKl00, HAB+00, BrHM01, Case01b, Eron01a, KoCz01, HePD02, PeRo02, AgLP03, KiGO03].

4.2.2 EINGESCHRÄNKT POLYPHONE MUSIKSIGNALE

Bei der Merkmalsextraktion aus eingeschränkt polyphonen Musiksignalen werden Musikstücke verarbeitet, die als Einschränkung entweder nur aus harmonischen Klängen oder nur aus rhythmischen Geräuschen bestehen dürfen. Hierbei werden Verfahren verwendet, die das Eingangsmaterial über die Ausnutzung des a-priori Wissens bezüglich der Harmonizität [Mart99, Virt03, LiPR03, LiRo04] bzw. Rhythmik [UhDS03, VTD+04] relativ gut in monophone Einzelklänge zerlegen können, die dann dem Szenario von monophonen Musiksignalen entsprechen.

4.2.3 KOMPLEXE POLYPHONE MUSIKSIGNALE

Bei der Merkmalsextraktion aus komplexen, polyphonen Musiksignalen werden für die verarbeiteten Musikstücke keinerlei Einschränkungen

bezüglich der Harmonizität oder Rhythmik gemacht. Über dieses Szenario wurde bisher relativ wenig publiziert, da es sich hierbei um den komplexesten Forschungsbereich handelt [Mart98, ChCB03, EgBr03a, EgBr03b].

Um die Merkmalsextraktion aus komplexen, polyphonen Musiksignalen auf das Szenario der Merkmalsextraktion aus monophonen Musiksignalen herunter brechen zu können, müssten die Musikstücke vor der eigentlichen Verarbeitung über eine automatische Quellentrennung und Segmentierung in einzelne Klangereignisse zerlegt werden. Die automatische Quellentrennung und Segmentierung stellen allerdings selbst umfangreiche Forschungsgebiete dar, zu denen es noch keine allgemeingültigen Lösungen gibt [Eggi01, BeBi03, Virt03, Zhan03, EiSi06]. Somit können für dieses Szenario nicht die Verfahren der monophonen Merkmalsextraktion verwendet werden, sondern es werden eigenständige Verfahren benötigt.

4.3 ZEITBEREICHSMERKMALE

Die Gruppe der Zeitbereichsmerkmale beschreibt Merkmale, die direkt aus dem Zeitsignal $x(n)$ oder einem L Abtastwerte großen Zeitsignalblock $x_m(n)$ berechnet werden können. Die Hopsize, d.h. der zeitliche Abstand der Zeitsignalblöcke untereinander, beträgt hierbei H Abtastwerte:

$$x_m(n) = x(mH + n), \text{ mit } 0 \le n \le L - 1. \tag{4.1}$$

4.3.1 NULLDURCHGANGSRATE

Die Nulldurchgangsrate a_{ZCR} gibt an, wie viele Nulldurchgänge das Signal $x(n)$ innerhalb des Zeitsignalblocks $x_m(n)$ hat. Hierbei werden entweder nur positive bzw. nur negative Nulldurchgänge gezählt [TzEC01, TzCo02]:

$$a_{ZCR}(m) = \frac{1}{2}\sum_{n=0}^{L-1}\left|\text{sig}(x_m(n)) - \text{sig}(x_m(n-1))\right|. \tag{4.2}$$

Für monophone Signale ergibt die Nulldurchgangsrate eine einfache Näherung der Momentanfrequenz und damit der Tonhöhe und ist für die Musikinstrumentenidentifikation und Klassifikation trotz seiner Einfachheit sehr aussagekräftig.

Bei komplexen, polyphonen Musiksignalen ist das Merkmal eher ein Maß für die Rauschartigkeit und hat für die Musikinstrumentenidentifikation und Klassifikation kaum eine Aussagekraft, weil die Mischung mehrerer Klänge mit unterschiedlichen Tonhöhen die Werte des Merkmals in sehr unterschiedlicher Art und Weise variiert.

4.3.2 EFFEKTIVWERT

Der Effektivwert a_{RMS} eines Signals $x(n)$ gibt den zeitlichen Verlauf der Wurzel aus der Signalenergie an [TzEC01, TzCo02]. Für einen Signalblock x_m der Länge L ergibt sich der Effektivwert über die folgende Formel:

$$a_{RMS}(m) = \sqrt{\frac{1}{L}\sum_{n=0}^{L-1} x_m^2(n)}\,. \tag{4.3}$$

Obwohl die absolute Energie bzw. Lautstärke eines Klangs als Merkmal wenig Aussagekraft hat, ist der Verlauf des Effektivwerts eines der aussagekräftigsten Merkmale für monophone Signale überhaupt. Tests mit monophonen Musiksignalen, bei denen nur dieses eine Merkmal ausgewertet wurde, zeigen für die Musikinstrumentenidentifikation und Klassifikation hohe Erkennungsraten von über 95 % [BKK+05].

Leider lässt sich die hohe Aussagekraft nicht auf komplexe, polyphone Musikinstrumente übertragen. Bei der Mischung und zeitlichen Überlagerung mehrerer Klänge entstehen Verläufe des Effektivwerts, die sich nicht auf einfache Art beschreiben lassen. Das Merkmal lässt sich allenfalls für die Vorverarbeitung in Form der Segmentierung anwenden, hier jedoch auch nur, solange die Klangereignisse sich nicht wesentlich überlappen [KSW+00].

4.3.3 HÜLLKURVE

Die Hüllkurve a_{Env} beschreibt einen ähnlichen Aspekt der Signaldynamik wie der Effektivwert a_{RMS} und hat eine ähnlich starke Aussagekraft für monophone Musiksignale, sie lässt sich jedoch in ihrer direkten Form für einen Signalblock x_m mit deutlich weniger Rechenlast berechnen:

$$a_{Env}(m) = \max\left(\left|x_m(n)\right|\right). \tag{4.4}$$

Neben dieser einfachen Art der Berechnung gibt es Verfahren aus dem Bereich der Synthesizertechnologie, die die Hüllkurve auf komplexere Art berechnen, so dass alle relevanten Hauptspitzen des Signals erfasst sind und zwischen diesen Punkten interpoliert wird. Diese detaillierteren Formen der Hüllkurve bringen jedoch keinen wesentlichen Vorteil bei der Aussagekraft des Merkmals.

Für die Hüllkurve gelten in Bezug auf die Verwendung für komplexe, polyphone Musiksignale die gleichen Einschränkungen wie für den Effektivwert, so dass auch die Hüllkurve für die Musikinstrumentenidentifikation und Klassifikation hier nicht sinnvoll verwendet werden kann.

4.3.4 HÜLLKURVENPARAMETER

Als Hüllkurvenparameter werden die Zeitspannen der verschiedenen Hüllkurvenphasen, also die Anstiegsphase, die Abklingphase, die Haltephase und die Ausschwingphase, eines Klangereignisses bezeichnet (vgl. Abschnitt 2.1.2). Für ihre Berechnung gibt es verschiedene Algorithmen, die sich hinsichtlich ihrer Genauigkeit und Rechenlast unterscheiden. Die einfacheren Modelle sind dreistufig und definieren lediglich die Zeitspannen der Anstiegsphase, der Haltephase und der Ausschwingphase. Diese Vereinfachung ist zulässig, da die Abklingphase hauptsächlich nur für Streicherklänge benötigt wird.

Für die Ermittlung der Phasen wird ein Schwellenwert für den Effektivwert des analysierten Audiosignals festgelegt. Die Anstiegsphase geht vom Beginn des Signals bis zum Überschreiten des Schwellenwertes. Die anschließende Haltephase geht bis zum Unterschreiten des Schwellenwerts, und die Ausschwingphase geht bis zum Ende des Klangereignisses (Bild 4.1) [Eron01a, Eron01b].

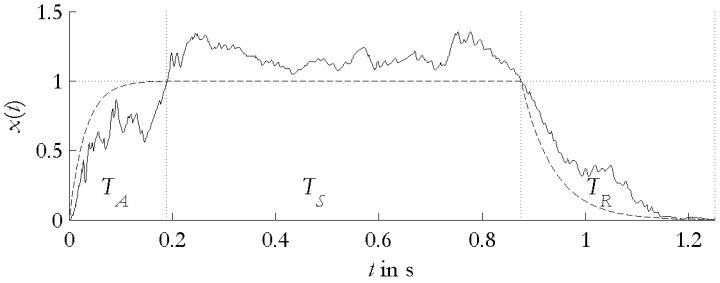


Bild 4.1: Ermittlung der Hüllkurvenparameter: Effektivwert des Signals (schwarz), Schwellenwert bei 1 (horizontal gepunktet), Phasengrenzen (vertikal gepunktet)

Die so gewonnenen Hüllkurvenparameter lassen sich zu einem für monophone Signale sehr aussagekräftigen Merkmalsvektor $\mathbf{a}_{Env}$ zusammenfassen, dessen einzelne Komponenten direkt den ermittelten Zeitspannen der Hüllkurvenphasen entsprechen:

$$a_{Env,1} = T_A\,, \qquad a_{Env,2} = T_S\,, \qquad a_{Env,3} = T_R\,. \tag{4.5}$$

Die Hüllkurvenparameter enthalten in nur drei Komponenten einen Großteil der Information der Hüllkurve bzw. des Effektivwerts. Dies stellt eine enorme Datenreduktion dar, bei annähernd gleicher Aussagekraft. Allerdings gelten in Bezug auf die Verwendung für die Identifikation und Klassifikation von Musikinstrumentenklängen die gleichen Einschränkungen wie für den Effektivwert und die Hüllkurve. Somit lassen sich die Hüllkurvenparameter für die Musikinstrumentenidentifikation und Klassifikation in monophonen Musiksignalen sehr gut verwenden, in komplexen, polyphonen Musiksignalen jedoch nicht.

4.3.5 LINEARE PRÄDIKTION

Bei der linearen Prädiktion wird der aktuelle Abtastwert eines Signals $x(n)$ aus einer Anzahl von P vorangegangenen Abtastwerten vorherbestimmt [Trom95]:

$$x\hat{}(n) = \sum_{k=0}^{P} a_k x(n-k). \tag{4.6}$$

Die Qualität der Prädiktion hängt von der Wahl der Prädiktionskoeffizienten a_k ab und lässt sich über die Größe des Prädiktionsfehlers $e(n)$ quantifizieren:

$$e(n) = x(n) - x\hat{}(n). \tag{4.7}$$

Für die Berechnung der Prädiktionskoeffizienten gibt es verschiedene Verfahren, die darauf abzielen, den Prädiktionsfehler zu minimieren. Die populärste hiervon ist die Autokorrelationsmethode [Makh75]. Die Koeffizienten a_k lassen sich direkt als die Elemente eines Merkmalsvektors $\mathbf{a}_{LP}$ auffassen.

Als weiteres Merkmal a_{LPP} wird aus der linearen Prädiktion die Vorhersagbarkeit des Signals innerhalb eines Audioblocks verwendet. Hierzu wird die mittlere Leistung des Prädiktionsfehlers e_m innerhalb des L Abtastwerte großen Blocks x_m mit ermittelt [Eron01a]:

$$a_{LPP}(m) = \frac{1}{L}\sum_{n=0}^{L-1} {e_m}^2(n). \tag{4.8}$$

Sowohl die lineare Prädiktion als auch ihre Vorhersagbarkeit bilden leistungsfähige Merkmale. Dies gilt vor allem für monophone Signale, aber auch komplexe polyphone Musiksignale lassen sich mit diesen Merkmalen gut beschreiben, wobei die Vorhersagbarkeit sich bei komplexen polyphonen Signalen eher wie die Nulldurchgangsrate als Maß der Rauschartigkeit verhält.

4.4 SPEKTRALE MERKMALE

Die Gruppe der spektralen Merkmale wird aus einem Spektrogramm des Zeitsignals $x(n)$ gewonnen. Für die Berechnung des Spektrogramms über eine Short Term Fourier Transformation (STFT) werden im Abstand der Hopsize H Zeitsignalblöcke $x_m(n)$ mit der Länge von L Abtastwerten mit einem Hamming-Fenster ausgeschnitten.

Die jeweiligen Spektren X_m der Blöcke x_m werden über eine diskrete Fouriertransformation (DFT) berechnet:

$$X_m(k) = \sum_{n=0}^{L-1} x_m(n) \cdot e^{-j\frac{2\pi}{L}kn}, \tag{4.9}$$

wobei die Signalblöcke $x_m(n)$ für die Berechnung mittels einer Fast Fourier Transform (FFT) mit angehängten Nullen auf die Länge der nächst größeren Zweierpotenz verlängert werden (Zero-Padding).

4.4.1 SPECTRAL CENTROID

Der Spectral Centroid a_{SC} stellt den Mittelpunkt des Spektrums dar und beschreibt Eigenschaften der Klanghelligkeit und der Klangschärfe [TzEC01, TzCo02]. Für das Spektrum X_m eines Audioblocks x_m lässt er sich über die folgende Formel berechnen, wobei $f(k)$ die dem Band k entsprechende Frequenz ist:

$$a_{SC}(m) = \frac{\sum_{k=0}^{L/2} f(k)\left|X_m(k)\right|}{\sum_{k=0}^{L/2} \left|X_m(k)\right|}. \tag{4.10}$$

Sind in dem analysierten Signal viele energiereiche Obertöne und hohe Frequenzen vorhanden, wird es als hell wahrgenommen. Sind bei gleicher Tonhöhe weniger Obertöne und hohe Frequenzen vorhanden, wird es als dumpf oder dunkel wahrgenommen. Der Wert des Spectral Centroid ist bei hellen Signalen größer als bei dumpfen.

Der absolute Wert des Spectral Centroid hängt von der Tonhöhe des zugrunde liegenden Audiomaterials ab, weshalb er oftmals auf die Tonhöhe f_0 normiert wird, sofern diese bekannt ist (vgl. die Abschnitte 4.5.1 und 4.5.2):

$$a_{SC\_}(m) = \frac{M_{SC}(m)}{f_0(m)}. \tag{4.11}$$

Neben der beschriebenen allgemeinen Form des Spectral Centroid definiert der MPEG-7-Standard zwei weitere spezielle Varianten [ISO15938]. Der AudioSpectrumCentroidType basiert auf einem linearen Leistungsdichtespektrum mit logarithmischer Frequenzskalierung mit

1 kHz als Mittelpunkt, und der SpectralCentroidType basiert auf einem linearen Leistungsdichtespektrum mit linearer Frequenzskalierung. Zusätzlich gibt es noch den HarmonicSpectralCentroidType, der einen Spectral Centroid der Harmonischen berechnet (vgl. Abschnitt 4.5).

Für die generelle Analyse des Klangs von Musiksignalen konnte sich der Spectral Centroid als aussagekräftiges Merkmal etablieren, das von vielen Algorithmen ausgewertet wird. Bei der Analyse von monophonen Musiksignalen enthält der Spectral Centroid direkt Informationen über die Klangfarbe der einzelnen Klangereignisse.

Bei komplexen, polyphonen Musiksignalen hingegen verschiebt schon die Mischung zweier Klänge den Spectral Centroid zwischen die Werte der Einzelklänge. Somit ist auch dieses Merkmal für die direkte Anwendung im Zusammenhang mit der Musikinstrumentenidentifikation und Klassifikation in komplexen, polyphonen Musikstücken nicht sinnvoll anwendbar.

4.4.2 SPECTRAL SPREAD

Der Spectral Spread a_{SSp} gibt die momentane effektive Bandbreite an und bewertet die Form des Spektrums hinsichtlich der Breitenverteilung um den Spectral Centroid. Der Spectral Spread beschreibt zusätzlich zum Spectral Centroid weitere Eigenschaften der Klanghelligkeit und der Klangschärfe.

Berechnet wird der Spectral Spread für das Spektrum X_m eines Audioblocks x_m über die folgende Formel, wobei $f(k)$ die dem Band k entsprechende Frequenz ist und $a_{SC}(m)$ der zu dem Audioblock gehörende Spectral Centroid:

$$a_{SS}(m) = \sqrt{\frac{\sum_{k=0}^{L/2} \left(f(k) - a_{SC}(m)\right)^2 \left|X_m(k)\right|}{\sum_{k=0}^{L/2} \left|X_m(k)\right|}} \,. \tag{4.12}$$

Der Spectral Spread stellt die Wurzel aus dem zweiten zentralen Moment der Statistik dar und entspricht somit der Standardabweichung. Kleine

Werte des Spectral Spreads bedeuten, dass sich die spektrale Energie eng um den Spectral Centroid verteilt, wohingegen große Werte darauf hindeuten, dass die spektrale Energie breit verteilt ist.

Der MPEG-7-Standard führt passend zu den MPEG-7 Spectrum Centroid Typen den AudioSpectrumSpreadType ein, der von der oben genannten Definition leicht abweicht [ISO15938].

Für die generelle Analyse des Klangs von Musiksignalen liefert der Spectral Spread wertvolle Zusatzinformationen zum Spectral Centroid. Bei der Analyse von monophonen Musiksignalen gibt er beispielsweise an, wie groß die Bandbreite eines Geräuschs oder die Anzahl der Obertöne eines Klangs ist, was wertvolle Informationen sind.

Bei komplexen, polyphonen Musiksignalen hingegen liefert der Spectral Spread ausschließlich Informationen über die momentane Bandbreite, was für die automatische Musikinstrumentenidentifikation und Klassifikation irrelevant ist.

4.4.3 SPECTRAL SKEWNESS

Die Spectral Skewness a_{SSk} stellt ein Maß für die Symmetrie des Spektrums um den Spectral Centroid dar und gibt an, ob das Spektrum höhenlastig oder basslastig ist [Peet04].

Die Berechnung für das Spektrum X_m eines Audioblocks x_m erfolgt über die Hilfsgröße m_3, die das dritte zentrale Moment der Statistik darstellt:

$$m_3(m) = \frac{\sum_{k=0}^{L/2} \left(f(k) - a_{SC}(m)\right)^3 \left|X_m(k)\right|}{\sum_{k=0}^{L/2} \left|X_m(k)\right|} . \qquad (4.13)$$

Hierbei entspricht a_{SC} dem zum Audioblock gehörenden Spectral Centroid und $f(k)$ der dem Band k entsprechenden Frequenz.

Auffällig an (4.13) ist die große Ähnlichkeit zur Berechnung des Spectral Spread a_{SS} (4.12), dessen Wert potenziert mit der berechneten Hilfsgröße m_3 ins Verhältnis gesetzt wird und so das eigentliche Merkmal der Spectral Skewness $a_{SSk}(m)$ bildet:

$$a_{SSk}(m) = \frac{a_3(m)}{a_{SS}^3(m)}. \tag{4.14}$$

Das Vorzeichen der Spectral Skewness beschreibt die Asymmetrie des Spektrums, wobei der Betrag angibt, wie ausgeprägt das Ausmaß der Asymmetrie ist:

- $a_{SSk} > 0$ beschreibt ein Spektrum mit mehr Energie im hochfrequenten Bereich,
- $a_{SSk} = 0$ entspricht einem symmetrischen Spektrum,
- $a_{SSk} < 0$ beschreibt ein Spektrum mit mehr Energie im tieffrequenten Bereich.

Die Spectral Skewness hat zusammen mit dem Spectral Centroid und dem Spectral Spread eine gute Aussagekraft im Zusammenhang mit der Analyse von monophonen Musiksignalen. Sie transportiert hier wertvolle Informationen über den Obertongehalt bzw. die Klangfarbe der analysierten Spektren und trifft mit nur einem Zahlenwert eine wesentliche Aussage über die Form des zugehörigen Spektrums.

Da die Spectral Skewness, wie auch der Spectral Centroid und der Spectral Spread, eine Aussage über das gesamte Spektrum trifft, ist sie für die Analyse von komplexen, polyphonen Musiksignalen im Zusammenhang mit der automatischen Musikinstrumentenidentifikation und Klassifikation nicht aussagekräftig, da sich mit ihr keine Informationen über die einzelnen Klangereignisse gewinnen lassen.

4.4.4 SPECTRAL KURTOSIS

Die Spectral Kurtosis a_{SK} ist ein Maß dafür, wie steil das Spektrum um den Wert des Spectral Centroids abfällt [Peet04]. Die Berechnung der Spectral Kurtosis hat große Ähnlichkeit mit der Berechnung der Spectral Skewness und erfolgt über die Hilfsgröße m_4, die das vierte zentrale Moment der Statistik darstellt. Für das Spektrum X_m eines Audioblocks x_m ergibt sich die Hilfsgröße m_4 aus der folgenden Formel:

$$m_4(m) = \frac{\sum_{k=0}^{L/2} \left(f(k) - a_{SC}(m)\right)^4 \left|X_m(k)\right|}{\sum_{k=0}^{L/2} \left|X_m(k)\right|}. \qquad (4.15)$$

Hierbei entspricht a_{SC} dem zum Audioblock gehörenden Spectral Centroid und $f(k)$ der dem Band k entsprechenden Frequenz.

Das eigentliche Merkmal der Spectral Kurtosis a_{SK} wird über das Verhältnis der Hilfsgröße zum potenzierten Spectral Spread a_{SS} gebildet:

$$a_{SSk}(m) = \frac{m_4(m)}{a_{SSk}^4(m)}. \qquad (4.16)$$

Der Wert der Spectral Kurtosis beschreibt nun die Steilheit des Spektrums:

- $a_{SK} > 3$ beschreibt ein Spektrum mit einem steilen Abfall um den Spectral Centroid,
- $a_{SK} = 3$ entspricht einer Normalverteilung,
- $a_{SK} < 3$ beschreibt ein Spektrum mit einem flachen Abfall um den Spectral Centroid.

Auch die Spectral Kurtosis hat ähnlich wie die Spectral Skewness eine gute Aussagekraft im Zusammenhang mit der Analyse von monophonen Musiksignalen, da sie mit nur einem Zahlenwert eine wesentliche Aussage über die Form des zugehörigen Spektrums und den Obertongehalt bzw. die Klangfarbe der analysierten Klangereignisse trifft.

Da jedoch auch die Spectral Kurtosis ähnlich wie die Spectral Skewness eine Aussage über das gesamte Spektrum trifft, gelten die gleichen Einschränkungen für die Analyse von komplexen, polyphonen Musiksignalen im Zusammenhang mit der automatischen Musikinstrumentenidentifikation und Klassifikation. Somit ist sie hier nicht sinnvoll einsetzbar.

4.4.5 SPECTRAL FLATNESS

Die Spectral Flatness a_{SFl} ist ein Maß der Rauschartigkeit bzw. Tonalität eines Spektrums. Sie wird für das Spektrum X_m eines Audioblocks x_m über das Verhältnis des geometrischen Mittels zum arithmetischen Mittel der Amplitudenkoeffizienten der einzelnen Frequenzbänder berechnet:

$$a_{SFl}(m) = \frac{\left(\prod_{k=0}^{L/2} |X_m(k)|\right)^{\frac{2}{L}}}{\frac{2}{L}\sum_{k=0}^{L/2} |X_m(k)|}. \qquad (4.17)$$

Kleine Werte der Spectral Flatness weisen auf eine hohe Tonalität hin, wohingegen große Werte bedeuten, dass das Signal rauschartig ist. Oftmals wird die Spectral Flatness auch nicht über alle Bänder eines Spektrums berechnet, sondern nur für einzelne Teilbänder, um die Tonalität innerhalb einzelner Bänder zu ermitteln.

Der MPEG-7-Standard definiert den AudioSpectrumFlatnessType, der ebenfalls eine Unterteilung des Spektrums in einzelne Teilbänder erlaubt.

Für die Analyse von monophonen Musiksignalen ist die Spectral Flatness ein sehr aussagekräftiges Merkmal, vor allem, wenn es um die Unterscheidung zwischen Klängen und Geräuschen geht, da Klänge in der Regel eine hohe Tonalität aufweisen, wohingegen Geräusche eher rauschartig sind.

Für die Analyse von komplexen, polyphonen Musiksignalen ist die Spectral Flatness in ihrer direkten Form wenig aussagekräftig, da Mischungen von tonalen Klangereignissen oftmals für die Spectral Flatness einen großen Wert erzeugen und somit fälschlicherweise eine geringe Tonalität anzeigen. Lediglich die Analyse der Spectral Flatness in einzelnen Teilbändern hat eine gewisse Aussagekraft, wenn sich so unterschiedliche Klangereignisse, wie z.B. unterschiedliche Trommelgeräusche, aufteilen lassen.

4.4.6 SPECTRAL CREST FACTOR

Der Spectral Crest Factor a_{SCF} ist sehr eng mit der Spectral Flatness verknüpft. Das Merkmal bildet das Verhältnis des maximalen Amplitudenkoeffizienten zum arithmetischen Mittel der Amplitudenkoeffizienten der einzelnen Frequenzbänder und wird für das Spektrum X_m eines Audioblocks x_m über die folgende Formel berechnet [Peet04]:

$$a_{SCF}(m) = \frac{\max\left(\left|X_m(k)\right|\right)}{\frac{2}{L}\sum_{k=0}^{L/2}\left|X_m(k)\right|}. \tag{4.18}$$

Die Werte der Spectral Flatness und des Spectral Crest Factor sind für normale Musiksignale miteinander korreliert, so dass bezüglich der Aussagekraft für den Spectral Crest Factor Ähnliches wie für die Spectral Flatness gilt.

Die Unterschiede zwischen dem Spectral Crest Factor und der Spectral Flatness sind, dass ersterer bei der Berechnung weniger Rechenlast erzeugt, letztere jedoch etwas genauer ist, da Ausreißer nicht so stark ins Gewicht fallen.

4.4.7 SPECTRAL ROLLOFF

Der Spectral Rolloff a_{SR} geht direkt aus dem Spectral Centroid hervor und beschreibt eine Grenzfrequenz, die die spektrale Gesamtenergie in einen hohen und einen tiefen Frequenzbereich teilt [TzEC01, TzCo02]. Zur Berechnung des Spectral Rolloff wird ein festgelegtes Verhältnis r zwischen der Energie des tiefen Bandes und der Gesamtenergie des Spektrums X_m eines Audioblocks x_m approximiert, indem die Bandgrenze K variiert wird:

$$\frac{\sum_{k=0}^{K}\left|X_m(k)\right|}{\sum_{k=0}^{L/2}\left|X_m(k)\right|} = r. \tag{4.19}$$

Als gängiges Verhältnis zwischen dem hohen und dem tiefen Band hat sich ein Verhältnis von ca. 4:1 etabliert, was einem Wert von $r = 80\ \%$ entspricht.

Die der Bandgrenze entsprechende Frequenz $f(K)$ ergibt den eigentlichen Wert des Spectral Rolloff $a_{SR}(m)$:

$$a_{SR}(m) = f(K). \tag{4.20}$$

Auch der Wert des Spectral Roloff wird häufig auf die Tonhöhe f_0 normiert, sofern diese bekannt ist (vgl. die Abschnitte 4.5.1 und 4.5.2):

$$a_{SR\_}(m) = \frac{a_{SR}(m)}{f_0(m)}. \tag{4.21}$$

Der Spectral Rolloff ist ähnlich dem Spectral Centroid ein Maß für die spektrale Form des Signals. Im Rahmen der Spracherkennung wird der Spectral Roloff sehr erfolgreich für die Unterscheidung zwischen stimmhaften und stimmlosen Phonemen verwendet, wobei ein Wert von $r = 85\ \%$ verwendet wird.

Er hat weiterhin eine große Aussagekraft im Zusammenhang mit der Analyse von monophonen Musiksignalen, da er wertvolle Informationen über den Obertongehalt von Klängen bzw. die Bandbreite von Geräuschen transportiert.

Je komplexer jedoch das analysierte Musiksstück hinsichtlich seiner Polyphonie ist, desto mehr ist der Spectral Rolloff mit dem Spectral Centroid korreliert. Somit gelten für den Spectral Rolloff bei der Analyse von komplexen, polyphonen Musiksignalen im Zusammenhang mit der Musikinstrumentenidentifikation und Klassifikation die gleichen Einschränkungen wie beim Spectral Centroid, so dass der Spectral Rolloff in diesem Szenario nicht sinnvoll anwendbar ist.

4.4.8 SPECTRAL FLUX

Der Spectral Flux a_{SFx} beschreibt die spektrale Konsistenz eines Signals, indem die Veränderungen des Spektrums von einem Block zum nächsten betrachtet werden [TzEC01, TzCo02]. Für das Spektrum X_m eines

Audioblocks x_m wird er über die folgende Formel berechnet, wobei $f(k)$ die dem Band k entsprechende Frequenz darstellt:

$$a_{SFx}(m) = \sum_{k=0}^{L/2} \left(|X_m(k)| - |X_{m-1}(k)| \right)^2 . \qquad (4.22)$$

Der Spectral Flux liefert sehr gute Ergebnisse im Zusammenhang mit der automatischen Segmentierung verschiedener Audiosignale (Sprache, Musik, Geräusch) oder der Genreklassifikation [BuLe03, BuLe04].

Bei der Analyse von monophonen Musiksignalen lassen sich über den Spectral Flux wertvolle Informationen, wie beispielsweise die Abgrenzung von Klängen, die ein starkes Vibrato oder Tremolo aufweisen, gewinnen. Im Zusammenhang mit der automatischen Musikinstrumentenidentifikation und Klassifikation lässt sich das Merkmal somit sehr gut verwenden.

Bei komplexen, polyphonen Musiksignalen ergeben sich beim Einsetzen oder Ausklingen neuer, überlappender Klänge im polyphonen Gemisch unbrauchbar hohe Werte für den Spectral Flux. Wenn überhaupt, ist der Spectral Flux in diesem Szenario nur als Transientenerkennung zu gebrauchen.

4.4.9 AUDIO SPECTRUM ENVELOPE

Der Audio Spectrum Envelope (ASE) ist ein im MPEG-7-Standard spezifiziertes Verfahren, das eine komprimierte Darstellung des berechneten Spektrums im Sinne einer Filterbank definiert [ISO15938]. Zur Berechnung des ASE wird ein Spektrogramm berechnet, bei dem die Länge der Blöcke standardmäßig drei mal größer als die Hopsize H ist:

$$L = 3 \cdot H . \qquad (4.23)$$

Die berechneten DFT-Frequenzbänder haben eine lineare Verteilung. Da das menschliche Gehör eher eine logarithmische Frequenzwahrnehmung aufweist, wird das Leistungsdichtespektrum (LDS) $S_m(k)$ der Bänder in eine logarithmische Frequenzverteilung überführt. Somit ergeben sich die Komponenten des Merkmalsvektors $\mathbf{a}_{ASE}$ aus den ASE-Koeffizienten $a_{ASE,m}(k)$. Diese werden über die folgende Formel berech-

net, wobei n_1 und n_2 die Bandgrenzen des logarithmischen Frequenzbandes sind:

$$a_{ASE,m}(k) = \sum_{n=n_1}^{n_2} |S_m(n)| . \tag{4.24}$$

Fällt ein LDS-Frequenzband zwischen die Bandgrenzen der ASE-Koeffizienten, so wird der entsprechende LDS-Koeffizient mit einer linearen Überblendung auf die beiden benachbarten ASE-Koeffizienten aufgeteilt. Standardmäßig wird der Bereich zwischen 62,5 Hz und 16 kHz in acht Oktavbänder eingeteilt, aus denen sich acht ASE-Koeffizienten ergeben. Zusätzlich gibt es jeweils einen Koeffizienten für alle LDS-Frequenzbänder über bzw. unter diesen Grenzen. Somit ergeben sich standardmäßig zehn Koeffizienten.

Die Berechnung der ASE-Koeffizienten verringert deutlich die Anzahl der Datenelemente pro Block. So werden ursprünglich L Werte aus dem zu analysierenden Signal ausgeschnitten, aus denen ein DFT-Spektrum mit $L/2$ Betragsamplituden berechnet wird. Bei einem Standardwert für L von 1024 Werten ergibt sich für ein ASE mit zehn Koeffizienten eine Dimensionsreduktion um ca. zwei Größenordnungen. Aufgrund der enormen Dimensionsreduktion ist das ASE nicht umkehrbar. Es handelt sich somit um ein reines Analyseinstrument.

Das ASE ist ein sehr aussagekräftiges Merkmal für die monophone Geräuschklassifikation, die Ergebnisse lassen sich jedoch leider nicht auf die Identifikation von Musikinstrumentenklängen in komplexen Musikstücken übertragen, da die charakteristischen Muster der Harmonischen der Musikinstrumentenklänge bei der Filterung mit der logarithmischen Filterbank zerstört werden. Dies ist der Fall, da im hohen Frequenzbereich mehrere Harmonische in einem Band zusammengefasst werden und somit nicht mehr zu unterscheiden sind.

4.4.10 MEL-SCALE FREQUENCY CEPSTRUM COEFFICIENTS

Die Mel-scale Frequency Cepstrum Coefficients (MFCC) lassen sich aus einem Audiospektrum berechnen und beschreiben in kompakter Form die wesentliche Form sowie die Formantstruktur des Spektrums.

Durch die Verwendung der Mel-Skala berücksichtigen die MFCC Aspekte der nichtlinearen Tonhöhenempfindung im Gehör. Die Tonhöhenempfindung und die Grundfrequenz eines Tones stehen näherungsweise in einem logarithmischem Verhältnis zueinander. Dies gilt tatsächlich jedoch nur für hohe Frequenzen, im tieffrequenten Bereich ist das Verhältnis der beiden Größen linear. Die Mel-Skala bildet dieses Verhalten ab, indem sie eine durch psychoakustische Untersuchungen heuristisch ermittelte Skala in mathematisch geschlossener Form approximiert. Die Tonhöhenempfindung z mit der Einheit mel wird folgendermaßen aus einer Frequenz f in Hz berechnet [Rege88]:

$$z(f) = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right). \tag{4.25}$$

Die MFCC werden für das Spektrum X_m eines Audioblocks x_m berechnet, indem die Bänder des LDS ähnlich wie bei der Berechnung der ASE-Koeffizienten entsprechend einer Bandpass-Filterbank zusammengefasst werden. Die Bänder der Filterbank sind hierbei an der Mel-Skala ausgerichtet. Als Standardwert werden 40 Filterbänder verwendet, die einen dreieckigen Frequenzgang haben. Die Spitzen der Bandpässe befinden sich jeweils bei der Mittenfrequenz $f_{cj}(k)$, wobei j den Index des Bandpasses angibt, die Bandbreite der dreieckigen Bandpässe geht jeweils von $f_{cj}(k\text{-}1)$ bis $f_{cj}(k+1)$. Die Mittenfrequenzen der 13 tiefsten Bänder sind linear verteilt und haben einen Abstand von 133,33 Hz, die Mittenfrequenzen der 27 verbleibenden hohen Bänder sind exponentiell über die Frequenz mit dem Faktor 1,0711703 verteilt (Bild 4.2).

Die eigentlichen Koeffizienten der MFCC, die den Merkmalsvektor $\mathbf{a}_{MFCC}$ bilden, werden gewonnen, indem die logarithmisch komprimierten Ausgangsamplituden der Filterbank m_k mittels einer diskreten Cosinustransformation (DCT) transformiert werden, wobei k den Index des Bandpasses und K die Gesamtanzahl der Filterbänder darstellen [OpSc95]:

$$a_{MFCC,l} = \sum_{k=1}^{K} \log_{10}(m_k) \cdot \cos\left(\frac{\pi}{K} l\left(k - \frac{1}{2}\right)\right). \tag{4.26}$$

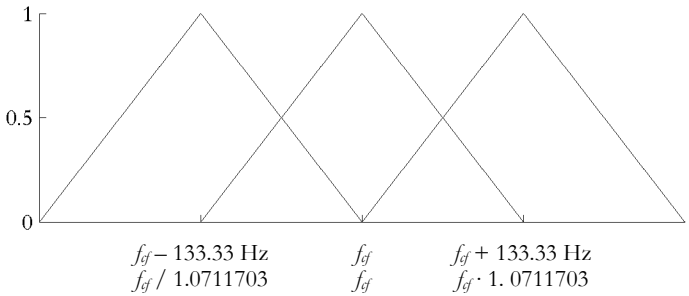


Bild 4.2: Drei Filterbänder einer MFCC-Filterbank, obere Beschriftung: linearer Bereich, untere Beschriftung: logarithmischer Bereich

Durch die Transformation werden die unterschiedlichen spektralen Formen in den berechneten Koeffizienten zusammengefasst. Die Koeffizienten mit kleinem Index enthalten Informationen über quasikontinuierliche Bereiche des Spektrums, die durch die Formanten geprägt sind, wohingegen die Koeffizienten mit großem Index Informationen zur Tonhöhe und den Feinstrukturen des Spektrums enthalten. Somit können diese Aspekte des Spektrums getrennt betrachtet werden. Der nullte Koeffizient ist eine Funktion der Lautstärke und ist somit mit dem Effektivwert bzw. der Hüllkurve korreliert.

Für die Sprach- und Sprechererkennung spielen die Formanteigenschaften des Rachenraumes eine der Sprachgrundfrequenz übergeordnete Rolle, und es reicht aus, nur die ersten fünf Koeffizienten zu verwenden. In diesem Bereich liefern die MFCC sehr gute Ergebnisse und haben sich als quasi Standard entwickelt [DaMe80, Camp97, CaCh99].

Neben der Sprach- und Sprechererkennung wurden MFCC in jüngster Zeit auch sehr erfolgreich für die Analyse von monophonen Musiksignalen verwendet, wobei hier in der Regel deutlich mehr Koeffizienten als für die Sprach- und Sprechererkennung berücksichtigt werden müssen. In diesem Zusammenhang bilden die MFCC sehr gut die Resonanzeigenschaften der verschiedenen Musikinstrumente über deren Formantbereiche ab [Brow97, BrHM01].

Die MFCC sind immer dann leistungsfähig, wenn die analysierten Signale einer statischen Formantstruktur unterliegen und die Grundfrequenz

sowie das genaue Muster der Obertöne irrelevant sind. Dies trifft bei der Sprach- und Sprechererkennung sowie der Analyse von monophonen Musiksignalen zu.

Bei der Analyse von komplexen, polyphonen Musiksignalen gibt es jedoch keine übergeordnete Formantstruktur, sondern es gilt, die Grundfrequenzen und die zugehörigen Obertonmuster der verschiedenen Klangereignisse zu ermitteln [Brow97, MaMo99, Eron01b]. Diese gehen jedoch im Zuge der Mel-Filterung durch die Logarithmierung des Spektrums verloren, da hierdurch in hohen Frequenzbereichen mehrere Obertöne in einem Band zusammengefasst werden. Dies gilt nicht nur für die Mel-Skala, sondern grundsätzlich für alle gehörangepassten Modelle, wie auch die Bark-Skala und die ERB-Skala [Zwic82, SmAb99]. Aus diesem Grunde lassen sich die positiven Eigenschaften der MFCC bei der Analyse von monophonen Musiksignalen leider nicht auf komplexe, polyphone Musiksignale übertragen.

4.5 HARMONISCHE MERKMALE

Die Gruppe der harmonischen Merkmale bezieht sich auf die Lautstärke und Lage der Harmonischen von Klangereignissen. Damit die Merkmale für alle Musiksignale universell und somit auch für Geräusche ohne ausgeprägte Harmonische verwendet werden können, werden bei Geräuschen anstelle der Harmonischen die Hochpunkte des Spektrums und somit die stärksten Formantregionen verarbeitet.

Um die Lage und Lautstärke der Harmonischen aus einem Zeitsignal $x(n)$ extrahieren zu können, werden harmonische Modelle benutzt. Die wichtigste Information hierbei ist die Lage der Grundfrequenz.

In den folgenden Abschnitten werden verschiedene Verfahren zur Extraktion der Harmonischen sowie die Bildung eines harmonischen Modells und anderer harmonischer Merkmale vorgestellt.

4.5.1 MONOPHONE GRUNDFREQUENZERKENNUNG

Die Grundfrequenz f_0 eines Klangs stellt ein wesentliches Merkmal dar, auf dem verschiedene andere Merkmale aufbauen. Für die Erkennung der Grundfrequenz gibt es unterschiedliche Verfahren, die im Zeit- oder

Frequenzbereich arbeiten [Peet04]. Für monophone Klänge ist die einfachste und effizienteste Methode eine auf der Autokorrelation $r(k)$ basierende Erkennung:

$$r(k) = \sum_{n=0}^{L-1} x(n) \cdot x(n+k) . \tag{4.27}$$

Die Grundfrequenz f_0 mit dem dazugehörigen Merkmal a_{FF} ergibt sich aus dem Maximum der Autokorrelation, wobei f_{SR} der Abtastfrequenz entspricht:

$$k_{max} = \arg\max_{k}(r(k)), \tag{4.28}$$

$$a_{FF} = f_0 = f_{SR} / k_{max} . \tag{4.29}$$

Bei der Verarbeitung des Audiosignals $x(n)$ in gefensterten Blöcken $x_m(n)$, deren Abtastwerte Null außerhalb der Blockgrenzen sind, werden für zunehmende Werte von k weniger Abtastwerte miteinander korreliert, was zwangsläufig zu einem größer werdenden Abfall der Autokorrelation führt, selbst wenn die Signale prinzipiell stark miteinander korreliert sind. Diesem Verhalten lässt sich mit einer Normierung entgegenwirken, die aus der Autokorrelation der verwendeten Fensterfunktion gewonnen wird. Da allerdings mit zunehmendem k weniger Werte in die Korrelation einbezogen werden, wird das Ergebnis in seiner relativen Genauigkeit weniger aussagekräftig. Deshalb sollten die Blöcke maximal nur bis zum halben Versatz $k = L/2$ gegeneinander verschoben werden [Boer93].

Die Grundfrequenz eines monophonen Musiksignals ist als einzelnes Merkmal nur bedingt aussagekräftig. Für Identifikations- und Klassifikationsaufgaben ist es einzeln quasi wertlos, allerdings lassen sich viele Algorithmen speziell parametrisieren, wenn die Grundfrequenz bekannt ist (vgl. die Abschnitte 4.4.1 und 4.4.7). Gleiches gilt selbstverständlich für Transkriptionsaufgaben [BEW+04b, Batk06].

4.5.2 POLYPHONE GRUNDFREQUENZERKENNUNG

Die Grundfrequenzerkennung von mehreren Klängen in einer polyphonen Mischung sollte im Frequenzbereich erfolgen, da hier die Grundfrequenzen der einzelnen Klänge in verschiedenen Frequenzbändern abge-

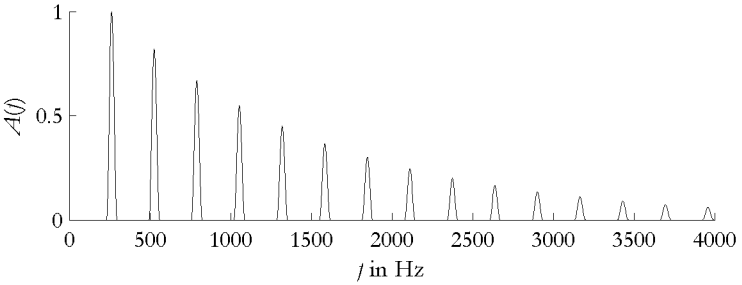


Bild 4.3: Schablone für die Grundfrequenz f_0 = 261,63 Hz (SPN: C4)

bildet werden. Die einfachste Methode zur Grundfrequenzerkennung stellt die Wahl der stärksten Amplitudenwerte des Spektrums dar (engl. Peak-Picking), diese funktioniert allerdings nur für den stärksten Klang zuverlässig. Für die Erkennung der Grundfrequenzen von weiteren Klängen können Oktavverwechselungen auftreten, wenn laute Obertöne eines anderen Klangs fälschlicherweise für Grundfrequenzen von neuen Klängen gehalten werden [Alle87, MaFe02, Naga03].

Ein verlässliches Verfahren für die Erkennung von mehreren Obertönen stellt das Schablonen-Verfahren dar, bei dem für jede in Frage kommende Grundfrequenz ein Kammfilter als Schablone erzeugt wird, die mit dem Spektrum des Klangs multipliziert wird [EgBr04a, EgBr04b]. Bild 4.3 zeigt eine für die polyphone Grundfrequenz verwendbare Schablone. Die Keulen um jede Harmonische sind Hann-Fenster mit der Breite von vier Halbtönen bezogen auf die Grundfrequenz. Da die tiefen Harmonischen bei der Tonhöhenermittlung eine größere Relevanz haben als die hohen Harmonischen, werden die in der Schablone enthaltenen Fenster für jede Harmonische unterschiedlich gewichtet. Der Scheitelwert $w(h)$ der zur Harmonischen h gehörenden Keule wird über die folgende Gleichung berechnet, wobei die Konstante von 0,2 im Exponenten heuristisch ermittelt wurde:

$$w(h) = e^{-0,2\cdot(h-1)}. \tag{4.30}$$

Das Filter mit der größten Ausgangsleistung bestimmt die Grundfrequenz f_0 jeden Klangs, die direkt dem dazugehörigen Merkmal a_{FF} entspricht:

$$a_{FF} = f_0 \,. \tag{4.31}$$

Auch die polyphone Grundfrequenzerkennung ist wie die monophone Grundfrequenzerkennung als einzelnes Merkmal nur bedingt aussagekräftig. Die Relevanz liegt auch hier im Zusammenhang mit der speziellen Parametrisierung anderer Merkmalsextraktionsverfahren sowie der Transkription [BEW+04b, Batk06].

4.5.3 HARMONIC PEAK SPECTRUM

Das Harmonic Peak Spectrum ist ein Merkmal, das für einen Klang die Frequenzen f_h und die zugehörige Leistung p_h der ersten H Harmonischen angibt. Der zugehörige Merkmalsvektor $\mathbf{a}_{HPS}$ hat eine Länge von $2H$ und stellt sich folgendermaßen zusammen:

$$\mathbf{a}_{HPS} = \begin{pmatrix} f_1 \\ \vdots \\ f_H \\ p_1 \\ \vdots \\ p_H \end{pmatrix}. \tag{4.32}$$

Berechnet wird das Harmonic Peak Spectrum, indem aus einem Audioblock x_m über eine STFT ein Kurzzeitspektrum X_m gewonnen wird. Anschließend wird das Spektrum in ein Sinusmodell überführt, indem nur die Spitzen des Spektrums als einzelne Sinustöne erhalten bleiben. Die extrahierten Sinustöne stehen hierbei stellvertretend für die sie umgebende Gruppe.

Um nur relevante Spitzen des Spektrums in das Sinusmodell zu übernehmen, wird das Amplitudenspektrum vor der Ermittlung der Spitzen mit einem Gaußfenster gefaltet. Für das Gaußfenster wurde eine optimale Breite von 20 Hz heuristisch ermittelt. Nach der Faltung werden alle lokalen Hochpunkte als Sinustöne für das Sinusmodell übernommen (Bild 4.4).

Der Klangeindruck des Sinusmodells ist in der Regel sehr gut, da der stärkste Sinuston einer Gruppe benachbarter Töne innerhalb der Gruppe

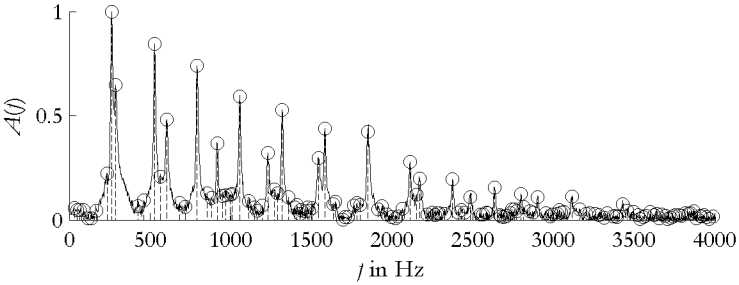


Bild 4.4: Betragsfrequenzgang eines Klanggemischs (schwarz) mit dem daraus abgeleiteten Sinusmodell (gestrichelt)

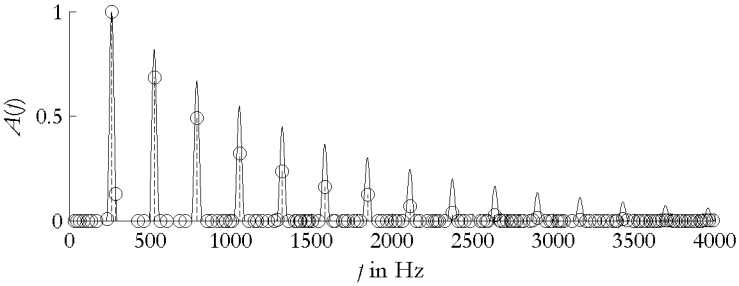


Bild 4.5: Polyphone Grundfrequenzermittlung für ein Sinusmodell (gestrichelt) mit einer Schablone (schwarz) der Grundfrequenz f_0 = 261,63 Hz (SPN: C4)

ohnehin psychoakustisch maskiert und somit nicht wahrgenommen werden kann [Zwic82, JaNo84, McQu85].

Ausgehend von dem berechneten Sinusmodell und dem hierdurch stark ausgedünnten Spektrum wird die Grundfrequenz des dominierenden Klangereignisses über eine polyphone Grundfrequenzerkennung ermittelt (vgl. Abschnitt 4.5.2). Die Schablonen werden hierbei nicht nur zur reinen Ermittlung der Gesamtleistung verschiedener Grundfrequenzen verwendet, sondern auch für die Ermittlung der einzelnen Harmonischen (Bild 4.5). Hierzu wird ausgehend von der ermittelten Grundfrequenz des betrachteten Klangereignisses die größte Sinuskomponente innerhalb jeder Keule als eine Harmonische des Klangs festgelegt. Die so

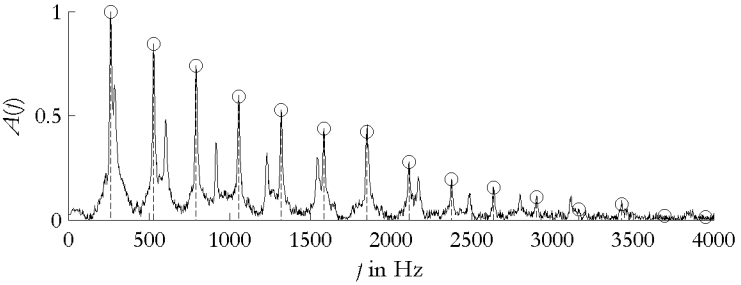


Bild 4.6: Harmonic Peak Spectrum (gestrichelt) eines Klanges mit dem Betragsfrequenzgang des dazugehörigen Klanggemischs (schwarz)

ermittelte, zu der Harmonischen gehörende Frequenz f_h und Leistung p_h können direkt in den Merkmalsvektor $\mathbf{a}_{HPS}$ übernommen werden. Konnte für eine Harmonische h keine zugehörige Sinuskomponente detektiert werden, so wird in den Merkmalsvektor eine virtuelle Komponente mit der Leistung Null und der Frequenz $h \cdot f_0$ übernommen (Bild 4.6).

Das Harmonic Peak Spectrum stellt ein besonders aussagekräftiges Merkmal sowohl für die Analyse von monophonen Musiksignalen als auch für die Analyse von komplexen, polyphonen Musiksignalen dar. Dies liegt daran, dass das Mischen von mehreren Klangereignissen das Merkmal nicht wie bei vielen anderen Merkmalen verfälscht. Vielmehr werden verschiedene Klangereignisse als eigene Instanzen, d.h. mit eigenen Merkmalsvektoren, abgebildet. Die Informationen innerhalb des Merkmalsvektors sind sehr robust gegenüber Störeinflüssen wie Rauschen oder die Mischung mit anderen Klangereignissen und haben somit eine gute Konstanz.

Tatsächlich hat sich das Harmonic Peak Spectrum als das mit Abstand leistungsfähigste Merkmal für die Analyse von polyphonen, komplexen Musiksignalen erwiesen.

4.5.4 HARMONIZITÄT

Die Harmonizität a_H beschreibt, wie genau die Frequenzen der H Harmonischen f_h eines Klangs auf den Vielfachen der Grundfrequenz $h \cdot f_0$ liegen [Peet04]:

$$a_H(m) = 1 - \frac{2}{f_0} \frac{\sum_{k=0}^{K} \left| f_k - k \cdot f_0 \right| \cdot A_m{}^2(k)}{\sum_{k=0}^{K} A_m{}^2(k)}. \tag{4.33}$$

Hierbei entspricht $A_m(h)$ der Amplitude der h-ten Harmonischen aus einem Satz von H Harmonischen. Die Werte der Harmonizität decken den Bereich [0,1] ab, da

$$f_k - k \cdot f_0 \leq f_0 \tag{4.34}$$

gilt. Ist die Harmonizität groß, klingt der Klang einerseits sehr rein, andererseits auch statisch. Bei einer kleinen Harmonizität hingegen enthält der Klang mehr Schwebung.

Eng verknüpft mit der Harmonizität ist die oftmals verwendete Inharmonizität a_{IH}, die sich direkt aus der Harmonizität berechnen lässt und ebenfalls den Bereich [0,1] abdeckt:

$$a_{IH}(m) = 1 - a_H(m). \tag{4.35}$$

Die Harmonizität und die Inharmonizität geben eine gute Auskunft über die interne Struktur der Harmonischen. Dies gilt für Klänge ganz allgemein, egal, ob das Merkmal aus monophonen oder komplexen, polyphonen Musiksignalen extrahiert wurde.

Für die Musikinstrumentenidentifikation und Klassifikation lässt sich das Merkmal gut verwenden, da sich die unterschiedlichen Instrumentenklassen hinsichtlich ihres primär schwingenden Bauteils und auch ihrer Resonanzeigenschaften in ihrer Harmonizität bzw. Inharmonizität deutlich unterscheiden.

4.5.5 RAUSCHARTIGKEIT

Die Rauschartigkeit a_N beschreibt das Verhältnis der Energie der nicht Harmonischen E_N im Verhältnis zur Gesamtenergie E_{tot} [Peet04]

$$a_N(m) = \frac{E_N(m)}{E_{tot}(m)}. \tag{4.36}$$

Für einen Audioblock x_m mit dem Spektrum X_m und H Harmonischen mit den Amplituden A_h ergeben sich die Gesamtenergie E_{tot}, die Energie der Harmonischen E_H und die Energie der nicht Harmonischen E_N aus den folgenden Formeln:

$$E_{tot}(m) = 2\sum_{k=0}^{L/2} \left| X_m(k) \right|^2 , \tag{4.37}$$

$$E_H(m) = \sum_{h=0}^{H} A_h^{\,2}(m), \tag{4.38}$$

$$E_N(m) = E_{ges}(m) - E_H(m). \tag{4.39}$$

Ist ein Klang sehr geräuschhaft, nimmt die Rauschartigkeit a_N einen großen Wert an, ist der Klang sehr tonal, ist die Rauschartigkeit hingegen klein.

Die Rauschartigkeit liefert im Vergleich zur Harmonizität bzw. Inharmonizität weiterhin Informationen über die Form des Spektrums außerhalb der Harmonischen. Dies ist vor allem dann nützlich, wenn entschieden werden muss, ob der durch ein harmonisches Modell wie das Harmonic Peak Spectrum beschriebene Klang noch mit weiteren Klängen oder Geräuschen im analysierten Musiksignal gemischt ist. Ist dies der Fall, nimmt die Rauschartigkeit einen hohen Wert an.

Für die Analyse von monophonen Musiksignalen ist die Rauschartigkeit häufig mit der Inharmonizität korreliert, so dass hier keine neue Aussagekraft entsteht.

4.5.6 HARMONIC SPECTRAL DEVIATION

Die Harmonic Spectral Deviation a_{HSD} gibt an, wie stark die Amplituden der einzelnen Obertöne von einer globalen, spektralen Betragshüllkurve abweichen [Peet04]. Für einen Audioblock x_m mit H Harmonischen mit den Amplituden A_h lässt sich die Harmonic Spectral Deviation über die folgende Formel berechnen:

$$a_{HSD}(m) = \frac{1}{H}\sum_{h=0}^{H}\left(A_h(m) - w(h)\right). \tag{4.40}$$

Die spektrale Betragshüllkurve wird hierbei durch $w(h)$ angegeben und an den Frequenzen der Harmonischen $f(h)$ ausgewertet. Für die spektrale Betragshüllkurve wird eine geglättete Version des Betragsspektrums verwendet, so dass überprüft werden kann, ob die einzelnen Harmonischen stark über den umliegenden Betragsgang hinausragen.

Kleine Werte der Harmonic Spectral Deviation deuten auf einen Klang hin, bei dem die tonalen und atonalen Klangelemente in einem ausgewogenen Verhältnis zueinander stehen. Bei großen Werten sind die Amplituden der Harmonischen entweder deutlich leiser oder deutlich lauter als die restlichen Klangkomponenten. Der Klang als ganzes wirkt hierdurch unausgewogen und enthält eine innere Spannung.

Die Harmonic Spectral Deviation liefert Informationen über die Form des Spektrums außerhalb der Harmonischen. Dies ist wie bei der Rauschartigkeit immer dann nützlich, wenn entschieden werden muss, ob neben dem durch ein harmonisches Modell beschriebenen Klang noch weitere Klänge oder Geräusche im analysierten Musiksignal gemischt sind.

Für die Analyse von monophonen Musiksignalen ist die Harmonic Spectral Deviation ebenfalls sehr nützlich, da sich die unterschiedlichen Instrumentenklassen hinsichtlich ihrer Klangerzeugung in ihrer Tonalität und somit in ihrer Harmonic Spectral Deviation deutlich unterscheiden.

4.5.7 HARMONIC INNER-RATIO

Das Harmonic Inner-Ratio a_{HIR} beschreibt, ob in einem Klang die Amplituden der geraden und ungeraden Harmonischen in einem ausgewogenen Verhältnis zueinander stehen [Peet04]. Für einen Audioblock x_m mit H Harmonischen mit den Amplituden A_h lässt sich das Harmonic Inner-Ratio über die folgende Formel berechnen:

$$a_{HIR}(m) = \frac{\sum_{h=0}^{H/2} A_{2h+1}{}^2(m)}{\sum_{h=0}^{H/2} A_{2h}{}^2(m)}. \tag{4.41}$$

Ist das Harmonic Inner-Ratio groß, hat der Klang Ähnlichkeit mit einer Pulswelle und klingt klarinettenartig. Geht es gegen 1, so hat der Klang Ähnlichkeit mit einer Dreieck- oder Sägezahnwelle und klingt trompetenartig.

Das Harmonic Inner-Ratio gibt eine sehr gute Auskunft über die interne Struktur der Harmonischen. Dies gilt für Klänge ganz allgemein, unabhängig davon, ob das Merkmal für die Analyse von monophonen oder komplexen, polyphonen Musiksignalen extrahiert wurde.

Das Harmonic Inner-Ratio lässt sich sehr gut für die Musikinstrumentenidentifikation und Klassifikation verwenden, da sich die unterschiedlichen Instrumentenklassen hinsichtlich ihrer Harmonischen und dem daraus entstehenden Verhältnis der geraden und ungeraden Harmonischen deutlich unterscheiden.

4.6 DIMENSIONSREDUKTION

Werden für Identifikations- und Klassifikationsaufgaben Merkmalsvektoren mit einer zu großen Dimensionalität verwendet, entstehen Probleme, die unter dem Namen Fluch der Dimensionalität bekannt sind. Um diese Probleme zu umgehen, gibt es eine Reihe von Verfahren zur Dimensionsreduktion, wie die Bildung einer Merkmalsuntermenge (engl. Feature Subset Selection), die Diskriminanzanalyse (engl. Linear Discriminant Analysis) oder Verfahren der Matrixfaktorisierung (engl. Matrix Decomposition) [DuHS01].

Aufgrund der bereits veröffentlichten Ergebnisse von Forschungen zu artverwandten Themen wurden im Rahmen dieser Arbeit verschiedene Verfahren der Matrixfaktorisierung als Dimensionsreduktion untersucht, die in den folgenden Abschnitten ausführlich im Zusammenhang mit der

Identifikation und Klassifikation von Musikinstrumenten beschrieben werden.

4.6.1 FLUCH DER DIMENSIONALITÄT

Ein Merkmalsvektor **a**, der R Merkmale zusammenfasst, lässt sich eindeutig im R-dimensionalen Hyperraum als Punkt beschreiben. Die Aussagekraft des Merkmalsvektors hängt hierbei stark von der Anzahl R und der Art der gewählten Merkmale ab, aber auch ihre Kombination untereinander spielt eine wesentliche Rolle.

Die Merkmalsvektoren verschiedener Klassen treten bei der Wahl aussagekräftiger Merkmale im Hyperraum als Wolken auf. Es ist offensichtlich, dass die Verwendung von zu wenigen Merkmalen dazu führt, dass sich die Wolken verschiedener Klassen im Merkmalsraum nicht mehr trennen lassen, da sie sich berühren oder gar überlappen.

Die Verwendung von sehr vielen Merkmalen hingegen bringt ebenso eine Reihe von Problemen mit sich. So potenziert sich die Anzahl der darstellbaren Zustände mit jeder neuen Dimension. Kann jedes Merkmal beispielsweise mit einer Quantisierung auf Q Zustände zufrieden stellend abgebildet werden, so ergibt sich für die Anzahl der Gesamtzustände Q_{tot} im Hyperraum eine Anzahl von:

$$Q_{tot} = Q^R . \tag{4.42}$$

Diese Eigenschaft hat weitreichende Folgen und wurde von Bellman als Fluch der Dimensionalität (engl. Curse of Dimensionality) bezeichnet [Bell61, Koep00].

Für Identifikations- und Klassifikationsverfahren steigt die für das Training der Modelle benötigte Anzahl der Trainingsdatensätze überproportional mit der Anzahl der möglichen Zustände im Hyperraum. Wird die Anzahl der Trainingsdatensätze bei steigender Dimensionsanzahl nicht erhöht, so führt dies zu einer drastischen Verschlechterung der Klassifikationsergebnisse, was gleichzeitig eine Form der Überanpassung darstellt (vgl. auch Abschnitt 5.2.4).

Weiterhin steigt auch der Rechenaufwand sowohl beim Training der Modelle als auch bei der Klassifikation proportional mit der Anzahl der Dimensionen R oder sogar mit der Anzahl der Zustände Q_{tot}.

Neben der Anzahl der Dimensionen spielt, wie bereits erwähnt, auch die Art der Merkmale eine große Rolle. Werden beispielsweise Merkmale verwendet, die untereinander linear abhängig sind, d.h. sich ineinander überführen lassen, so spannen sie aus mathematischer Sicht im R-dimensionalen Raum lediglich einen kleineren Z-dimensionalen Raum auf, wobei Z die Anzahl der linear unabhängigen Merkmale ist und $Z < R$ gilt.

Aus den vorhergehenden Ausführungen folgt, dass nicht alle Merkmale sinnvoll zusammen verwendet werden können. Deshalb sollte schon bei der Auswahl der Merkmale ein sinnvoller Merkmalssatz gebildet werden, dessen Merkmale unkorreliert sind, d.h. einzigartige Informationen enthalten.

Verfahren der Dimensionsreduktion überführen den R-dimensionalen Raum der Merkmalsvektoren in einen Z-dimensionalen Raum mit weniger Dimensionen ($Z < R$), wobei von der in den Merkmalsvektoren enthaltenen Information maximal viel erhalten bleiben soll.

4.6.2 MATRIXFAKTORISIERUNG

Mit der Matrixfaktorisierung lässt sich die Merkmalsmatrix in eine Form transformieren, in der sich neue Merkmale und neue Merkmalssignale ergeben. Das Ziel der Matrixfaktorisierung ist zum einen die Transformation in statistisch unabhängige oder zumindest unkorrelierte Merkmale, zum anderen die Trennung der Merkmale nach ihrem Informationsgehalt. Merkmale mit geringem Informationsgehalt können anschließend im Zuge der Dimensionsreduktion fallen gelassen werden [Alle87].

Es existieren verschiedene Verfahren der Matrixfaktorisierung, die grundsätzlich dem folgenden Schema entsprechen:

$$\mathbf{A} = \mathbf{X}_1 \mathbf{X}_2 \cdots \mathbf{X}_K , \tag{4.43}$$

wobei $\mathbf{A}_{(R \times C)}$: die Merkmalsmatrix ist und $\mathbf{X}_k$ verschiedene zur Faktorisierung benutzte Matrizen sind. Bei den in den folgenden Abschnitten

betrachteten Verfahren ist $K = 2$ oder $K = 3$. Für jeden Merkmalsvektor, d.h. für jede Spalte der Merkmalsmatrix, kann die Matrixfaktorisierung auch als eine mit den Koeffizienten c_m gewichtete Summe von M normierten Basisvektoren $\mathbf{x}_m$ verstanden werden:

$$\mathbf{a} = \sum_{m=1}^{M} c_m \mathbf{x}_m \,. \tag{4.44}$$

Da die Basisvektoren für alle Merkmalsvektoren fix sind, ergibt sich aus den Koeffizienten c_m ein neuer Merkmalsvektor mit der Dimension M. Die neuen Merkmalsvektoren werden spaltenweise in einer neuen Merkmalsmatrix, der so genannten Signalmatrix mit der Dimension $M \times C$ zusammengefasst. Als Maß für die Relevanz der Basisvektoren kann die Varianz ihrer Koeffizienten herangezogen werden. Basisvektoren, deren zugehörige Koeffizienten eine geringe Varianz aufweisen, können im Zuge der Dimensionsreduktion ohne einen großen Informationsverlust fallen gelassen werden.

Durch die Berechnung neuer Basisvektoren führt die Matrixfaktorisierung eine Projektion des ursprünglichen Merkmalsraums in einen neuen Merkmalsraum durch. Es ist oftmals sinnvoll, den so entstandenen dimensionsärmeren Merkmalsraum durch orthogonale Basisvektoren zu beschreiben, was eine zweite Transformation des Merkmalsraums nach sich zieht.

Ein Nachteil der Matrixfaktorisierung liegt darin, dass die transformierten Merkmale keine physikalische Entsprechung mehr haben und somit nicht mehr direkt gedeutet werden können.

4.6.3 HAUPTKOMPONENTENANALYSE

Über die Hauptkomponentenanalyse (engl. Principal Component Analysis, PCA) können die in der Merkmalsmatrix enthaltenen Merkmalssignale untereinander dekorreliert werden. Von den neu gewonnenen Merkmalssignalen können anschließend Signale mit geringer Varianz im Zuge der Merkmalsreduktion fallen gelassen werden [Alle87, Joll02] (Bild 4.7).

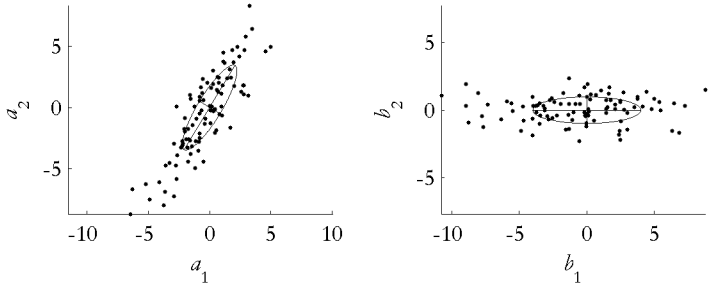


Bild 4.7: Anwendung der PCA auf zweidimensionale Merkmalsvektoren, links: Originalmatrix **A**, rechts: Transformationsergebnis **B**, Standardabweichung mit ihren Hauptachsen (Ellipsen)

Das Ziel der PCA ist die Faktorisierung der Merkmalsmatrix $\mathbf{A}_{(R \times C)}$ durch eine neue Signalmatrix $\mathbf{B}_{(R \times C)}$ und eine orthogonale Transformationsmatrix $\mathbf{P}_{(R \times R)}$:

$$\mathbf{A} = \mathbf{PB} . \tag{4.45}$$

Die neue Signalmatrix **B** soll dekorreliert sein, was bedeutet, dass ihre Kovarianzmatrix eine Diagonalmatrix **D** ist:

$$\mathbf{C_B} = \frac{1}{C-1}\mathbf{B}\mathbf{B}^T = \mathbf{D} . \tag{4.46}$$

Da die Transformationsmatrix **P** orthogonal ist, lässt sich Gleichung 4.45 umformen zu:

$$\mathbf{B} = \mathbf{P}^{-1}\mathbf{A} = \mathbf{P}^T\mathbf{A} . \tag{4.47}$$

Wird nun Gleichung 4.47 in die Formel der Kovarianzmatrix (4.46) eingesetzt, ergibt sich:

$$\mathbf{C_B} = \frac{1}{C-1}\left(\mathbf{P}^T\mathbf{A}\right)\left(\mathbf{P}^T\mathbf{A}\right)^T ,$$

$$= \frac{1}{C-1}\mathbf{P}^T\mathbf{A}\mathbf{A}^T\mathbf{P} ,$$

$$= \mathbf{P}^T \left(\frac{1}{C-1} \mathbf{A}\mathbf{A}^T \right) \mathbf{P} . \tag{4.48}$$

Mit der Formel der Kovarianzmatrix der Merkmalsmatrix

$$\mathbf{C_A} = \frac{1}{C-1} \mathbf{A}\mathbf{A}^T , \tag{4.49}$$

wird Gleichung 4.48 zu:

$$\mathbf{C_B} = \mathbf{P}^T \mathbf{C_A} \mathbf{P} . \tag{4.50}$$

Diese Gleichung lässt sich weiter vereinfachen, indem für die Kovarianzmatrix $\mathbf{C_A}$ der Merkmalsmatrix eine Eigenwertzerlegung durchgeführt wird:

$$\mathbf{C_A} \mathbf{U} = \mathbf{U} \Sigma . \tag{4.51}$$

Da Kovarianzmatrizen immer reell und symmetrisch sind, ergibt die Eigenwertzerlegung eine orthogonale Matrix **U**, deren Spalten die normierten Eigenvektoren enthält. Dies bedeutet umformuliert, dass die Eigenvektoren eine Orthonormalbasis bilden. Die Matrix **Σ** ist eine Diagonalmatrix, deren Diagonale die entsprechenden Eigenwerte enthält. Umgeformt nach:

$$\mathbf{C_A} = \mathbf{U} \Sigma \mathbf{U}^{-1} = \mathbf{U} \Sigma \mathbf{U}^T \tag{4.52}$$

ergibt sich durch Einsetzen in Gleichung 4.50:

$$\mathbf{C_B} = \mathbf{P}^T \mathbf{U} \Sigma \mathbf{U}^T \mathbf{P} . \tag{4.53}$$

Wird nun für die gesuchte Transformationsmatrix

$$\mathbf{P} = \mathbf{U} \tag{4.54}$$

gewählt, so ergibt sich zum einen für die Faktorisierung (4.45):

$$\mathbf{A} = \mathbf{U}\mathbf{B} , \tag{4.55}$$

zum anderen ergibt sich die über die Diagonalisierung der Kovarianzmatrix geforderte Dekorrelation der Signalmatrix und Gleichung 4.53 wird unter Verwendung der Identitätsmatrix **I** zu:

$$\mathbf{C_B} = \left(\mathbf{U}^T\mathbf{U}\right)\Sigma\left(\mathbf{U}^T\mathbf{U}\right) = \mathbf{I}\Sigma\mathbf{I}$$

$$\mathbf{C_B} = \Sigma\,. \tag{4.56}$$

Aus den Gleichungen 4.54 und 4.56 lassen sich nun wesentliche Erkenntnisse ableiten. So wird die Merkmalsmatrix durch die orthogonale Eigenwertmatrix ihrer Kovarianzmatrix faktorisiert. Die orthonormalen Eigenvektoren bilden die Spalten der Transformationsmatrix **P** und werden Hauptkomponenten genannt. Die Matrix **B** enthält in ihren Spalten die neuen Merkmalsvektoren und als Zeilenvektoren die neuen Merkmalssignale. Da die Transformationsmatrix orthogonal ist, handelt es sich bei der PCA um eine Rotation im Raum. Die Varianzen der neuen Merkmalssignale, die auf der Hauptdiagonalen der Kovarianzmatrix $\mathbf{C_B}$ stehen, ergeben sich direkt aus den zugehörigen Eigenwerten. Für eine Dimensionsreduktion können nun die Signale mit geringer Varianz fallen gelassen werden, ohne das Ergebnis maßgeblich zu verfälschen. Über die PCA lassen sich somit die Hauptkeulen der Varianz bestimmen. Dies gilt zwar streng genommen nur für normalverteilte, mittelwertbefreite Daten, ist jedoch auch für abweichend verteilte Daten hilfreich [Case01, KiGO03, KiBS04].

Werden die Merkmalssignale vor der PCA nicht von ihrem Mittelwert befreit, so zeigt in der Regel die erste Hauptkomponente in Richtung des Mittelwertes, was sie für die Auswertung unbrauchbar macht. Zusätzlich werden die Richtungen der weiteren Hauptkomponenten leicht verfälscht, da das Kriterium der Orthogonalität eingehalten werden muss. Diese Verfälschungen treten besonders bei weniger als sechs Dimensionen auf.

Die Vorteile der PCA sind, dass es immer eine eindeutige Lösung gibt, die ohne Parametrisierung auskommt und dass sie mit wenig Aufwand zu berechnen ist. Ein Nachteil hingegen ist, dass die entstehenden Merkmalssignale nicht auf Einheitsvarianz normiert sind, was die relative Aussagekraft vermindert. Dieses Manko wird durch die Singulärwertzerlegung (vgl. Abschnitt 4.6.4) behoben. Ein weiterer Nachteil ergibt sich aus der erzwungenen Orthogonalität der Transformationsvektoren. Werden auch nicht orthogonale Vektoren zugelassen, so kann die Merkmalsmatrix sogar in unabhängige Komponenten zerlegt werden, was durch Verfahren der Analyse der unabhängigen Komponenten (vgl.

Abschnitt 4.6.5) oder die nicht-negative Matrixfaktorisierung (vgl. Abschnitt 4.6.6) erreicht wird.

Bei einer Anwendung der PCA im Zusammenhang mit Merkmalen von Musikinstrumentenklängen können Merkmalssignale mit einer kleinen Varianz und folglich mit einem kleinen Singulärwert fallen gelassen werden, ohne das Ergebnis maßgeblich zu verfälschen, was zu einer Dimensionsreduktion führt.

Die PCA liefert genau dann gute Ergebnisse, wenn die Daten versteckte Korrelationen enthalten und prinzipiell normalverteilt sind. Dies ist im Zusammenhang mit Merkmalen von Musikinstrumentenklängen in der Regel der Fall, wenn Zeitbereichsmerkmale oder einfache spektrale Merkmale verwendet werden, die sich stetig im Merkmalsraum verteilen. Bei der Verwendung von komplexen Merkmalen wie harmonischen Merkmalen hingegen kann es zu Sprüngen im Merkmalsraum kommen, die durch die PCA nicht aufgelöst werden können [CaWe01, Case01a, Case01b, KiBS04].

4.6.4 SINGULÄRWERTZERLEGUNG

Die Singulärwertzerlegung (engl. Singular Value Decomposition, SVD) enthält die Hauptkomponentenanalyse als Spezialfall, stellt aber im Vergleich zu ihr ein mächtigeres Werkzeug dar [Joll02, WaRR03] (Bild 4.8). Auch die SVD dekorreliert die im Merkmalsvektor enthaltenen Merkmalssignale untereinander, liefert jedoch neben den auf Einheitslänge normierten Hauptkomponenten und den zugehörigen Varianzen auch die normierten Merkmalssignale.

Das Ziel der SVD ist die Faktorisierung der Merkmalsmatrix $\mathbf{A}_{(R \times C)}$ durch die orthogonalen Matrizen $\mathbf{U}_{(R \times R)}$ und $\mathbf{V}_{(C \times C)}$ sowie der Diagonalmatrix $\mathbf{S}_{(R \times C)}$:

$$\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^T . \tag{4.57}$$

Für die Berechnung von $\mathbf{U}$ wird eine Hilfsmatrix $\mathbf{X}_{(R \times R)}$ definiert:

$$\mathbf{X} = \mathbf{A}\mathbf{A}^T , \tag{4.58}$$

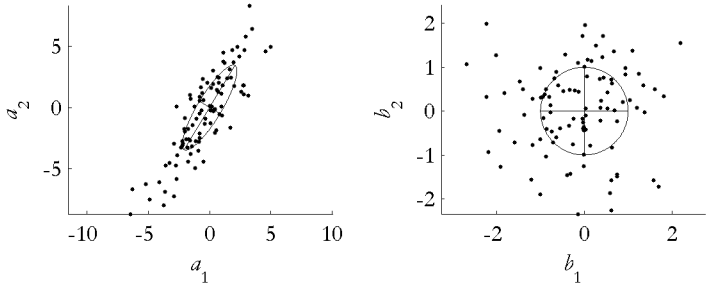


Bild 4.8: Anwendung der SVD auf zweidimensionale Merkmalsvektoren, links: Originalmatrix **A**, rechts: Transformationsergebnis **B**, Standardabweichung mit ihren Hauptachsen (Ellipsen)

die mit den noch unbekannten Matrizen der Faktorisierung (4.57) ausformuliert werden kann zu:

$$\begin{aligned}\mathbf{X} &= \mathbf{U}\mathbf{S}\mathbf{V}^T\left(\mathbf{U}\mathbf{S}\mathbf{V}^T\right)^T \\ &= \mathbf{U}\mathbf{S}\mathbf{V}^T\mathbf{V}\mathbf{S}\mathbf{U}^T \\ &= \mathbf{U}\mathbf{S}^2\mathbf{U}^T . \end{aligned} \tag{4.59}$$

Gleichung 4.59 kann erfüllt werden, indem mit der Hilfsmatrix **X** eine Eigenwertzerlegung durchgeführt wird:

$$\mathbf{X}\mathbf{U} = \mathbf{U}\boldsymbol{\Delta} \tag{4.60}$$

$$\Leftrightarrow \mathbf{X} = \mathbf{U}\boldsymbol{\Delta}\ \mathbf{U}^T . \tag{4.61}$$

Da die Hilfsmatrix **X** aufgrund ihrer Konstruktion zwangsläufig symmetrisch ist, liefert eine Eigenwertzerlegung orthonormale Eigenvektoren, die die Spalten der Matrix **U** bilden. Die Werte der Matrix **S** werden Singulärwerte genannt und ergeben sich durch Gleichsetzen mit Gleichung 4.59 zu:

$$\mathbf{S} = \boldsymbol{\Delta}^{\frac{1}{2}} . \tag{4.62}$$

Die Berechnung von **V** findet analog statt. Es wird eine Hilfsmatrix $\mathbf{Y}_{(C\times C)}$ definiert:

$$\mathbf{Y} = \mathbf{X}^T\mathbf{X}\,, \tag{4.63}$$

die mit Gleichung 4.57 ausformuliert werden kann zu:

$$\mathbf{Y} = \left(\mathbf{USV}^T\right)^T\mathbf{USV}^T$$

$$\mathbf{VSU}^T\mathbf{USV}^T$$

$$= \mathbf{VS}^2\mathbf{V}^T\,. \tag{4.64}$$

Gleichung 4.64 kann erfüllt werden, indem mit der Hilfsmatrix **Y** eine Eigenwertzerlegung durchgeführt wird:

$$\mathbf{YV} = \mathbf{V\Delta} \tag{4.65}$$

$$\Leftrightarrow \mathbf{Y} = \mathbf{V\Delta}\ \mathbf{V}^T\,. \tag{4.66}$$

Auch die Hilfsmatrix **Y** ist aufgrund ihrer Konstruktion zwangsläufig symmetrisch, und ihre Eigenwertzerlegung liefert orthonormale Eigenvektoren, die die Spalten der Matrix **V** bilden. Aufgrund der analogen Formeln 4.59 und 4.64 können die Werte der Matrix **S** auch aus den durch die Zerlegung der Hilfsmatrix **Y** gewonnenen Eigenwerten berechnet werden. Formel 4.62 gilt somit auch hier.

Die Hilfsmatrix **X** ist eine skalierte Version der Kovarianzmatrix $\mathbf{C_A}$, beide haben aber die gleichen Eigenvektoren, da die Skalierung keinen Einfluss auf die Richtung oder Länge der Eigenvektoren hat (vgl. hierzu die Formeln 4.49 und 4.58). Für den Beweis der Identität wird

$$\mathbf{X} = (C-1)\mathbf{C_A} \tag{4.67}$$

in Gleichung 4.61 eingesetzt und umgeformt:

$$(C-1)\mathbf{C_X} = \mathbf{U}\,\Delta\,\mathbf{U}^T$$

$$\mathbf{C_A} = \frac{1}{(C-1)}\mathbf{U\Delta U}^T \text{ mit } \Delta\hat{} = \frac{1}{(C-1)}\Delta$$

$$\mathbf{C_A} = \mathbf{U\Delta\hat{}\,U}^T\,. \tag{4.68}$$

Gleichung 4.68 beschreibt die Eigenwertzerlegung der Kovarianzmatrix $\mathbf{C_A}$. Somit beweist die Herleitung, dass die Eigenvektoren der Hilfsmatrix

X und die Eigenvektoren der Kovarianzmatrix $\mathbf{C_A}$ identisch sind. Der Vergleich mit Gleichung 4.52 zeigt weiterhin, dass die Eigenvektoren identisch zu den aus der PCA gewonnenen Hauptkomponenten sind.

Somit erklärt sich die SVD als eine Faktorisierung der Merkmalsmatrix, wobei die Matrix **U** als Spaltenvektoren die Hauptkomponenten enthält. Die Matrix **V** enthält als Spaltenvektoren die neuen Merkmalsvektoren und als Zeilenvektoren die neuen Merkmalssignale. Folglich sind bei der SVD nicht nur die Hauptkomponenten orthonormal, wie es bei der PCA der Fall ist, sondern auch die Merkmalssignale. Somit sind die Merkmalssignale in ihrer Energie normiert und können leichter verglichen werden. Die Relevanz der einzelnen Merkmalssignale für eine Dimensionsreduktion ergibt sich direkt aus der Größe der Singulärwerte, wobei die Spalten der Matrizen so angeordnet werden, dass die Singulärwerte in absteigender Reihenfolge in der Diagonalmatrix **S** stehen. Insgesamt können bei der Eigenwertzerlegung der Hilfsmatrizen nur eine dem Rang K entsprechende Anzahl an von Null verschiedenen Singulärwerten entstehen, wobei gilt:

$$K = \min(R, C)\,. \tag{4.69}$$

Die Überführung der SVD in die PCA erfolgt, indem der Term $\mathbf{SV}^T$ ausmultipliziert wird und so die Faktorisierung auf zwei Matrizen reduziert wird (vgl. auch Gleichung 4.55):

$$\mathbf{A} = \mathbf{USV}^T = \mathbf{UZ} \;\text{ mit }\; \mathbf{Z} = \mathbf{SV}^T\,. \tag{4.70}$$

Bei einer Anwendung der PCA im Zusammenhang mit Merkmalen von Musikinstrumentenklängen können Merkmalssignale mit einer kleinen Varianz und folglich mit einem kleinen Singulärwert fallen gelassen werden, ohne das Ergebnis maßgeblich zu verfälschen, was zu einer Dimensionsreduktion führt.

Auch die SVD liefert streng genommen nur bei normalverteilten Daten optimale Ergebnisse, sie liefert jedoch auch für abweichend verteilte Daten aufschlussreiche Ergebnisse. Die Merkmalssignale sollten vor der Analyse mittelwertbefreit sein, da sonst die erste Hauptkomponente in Richtung des Mittelwertes zeigt, was sie für die Auswertung unbrauchbar macht und die restlichen Hauptkomponenten leicht verschiebt.

Auch die SVD liefert wie die PCA genau dann gute Ergebnisse, wenn die Daten versteckte Korrelationen enthalten. Dies ist im Zusammenhang mit Merkmalen von Musikinstrumentenklängen der Fall, wenn Zeitbereichsmerkmale oder einfache spektrale Merkmale verwendet werden, die sich stetig im Merkmalsraum verteilen. Bei der Verwendung von komplexen Merkmalen wie harmonischen Merkmalen hingegen kann es zu Sprüngen im Merkmalsraum kommen, die durch die SVD nicht aufgelöst werden können.

4.6.5 ANALYSE DER UNABHÄNGIGEN KOMPONENTEN

Die Analyse der unabhängigen Komponenten (engl. Independent Component Analysis, ICA) ist eine Methode der Quellentrennung, mit der die im Merkmalsvektor enthaltenen Merkmalssignale in neue Signale transformiert werden [HyOj00]. Diese neuen Merkmalssignale sind nicht nur wie im Falle der PCA oder SVD untereinander dekorreliert, sondern sogar statistisch unabhängig (Bild 4.9).

Das Ziel der ICA ist die Faktorisierung der Merkmalsmatrix $\mathbf{A}_{(R \times C)}$ durch eine neue Signalmatrix $\mathbf{B}_{(R \times C)}$ und eine (in der Regel nicht orthogonale) Mischungsmatrix $\mathbf{Q}_{(R \times R)}$:

$$\mathbf{A} = \mathbf{QB} \,. \tag{4.71}$$

Aus dieser Gleichung lässt sich direkt das der ICA zugrunde liegende Modell erklären. Die in den Zeilen der Signalmatrix enthaltenen Signale $b_r(c)$ sind statistisch unabhängige Signale mit dem Zeitindex c, die als Quellsignale aus unabhängigen Quellen aufgefasst werden. Diese werden über die in den Spalten der Mischungsmatrix $\mathbf{Q}$ enthaltenen Vektoren $\mathbf{q}_c$, den so genannten unabhängigen Komponenten, miteinander gemischt. Somit besteht jedes Merkmalssignal $a_r(c)$ aus einer Mischung der R unabhängigen Quellsignale, wobei $q_{r,c}$ das r-te (Zeilen-)Element des c-ten Spaltenvektors $\mathbf{q}_c$ der Matrix $\mathbf{Q}$ ist:

$$a_r(c) = \sum_{r=1}^{R} q_{r,c} b_r(c) \,. \tag{4.72}$$

Wird Gleichung 4.71 unter der Verwendung der Inversen der Mischungsmatrix $\mathbf{W} = \mathbf{Q}^{-1}$ nach $\mathbf{B}$ umgestellt:

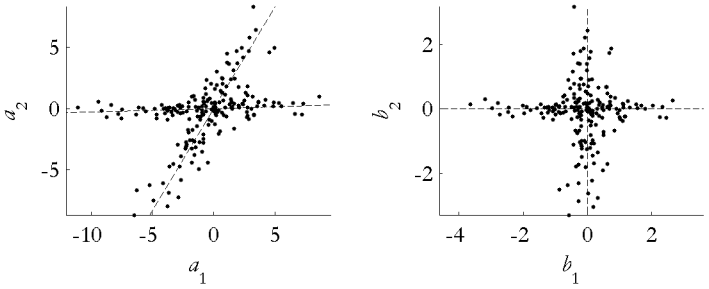


Bild 4.9: Anwendung der ICA auf nicht normalverteilte, zweidimensionale Merkmalsvektoren, links: Originalmatrix **A**, rechts: Transformationsergebnis **B**, Achsen der unabhängigen Komponenten (gestrichelt)

$$\mathbf{B} = \mathbf{Q}^{-1}\mathbf{A} = \mathbf{W}\mathbf{A}\,, \tag{4.73}$$

so lässt sich umgekehrt auch jedes Quellsignal als eine Summe der Merkmalssignale berechnen, wobei $w_{r,c}$ das r-te (Zeilen-)Element des c-ten Spaltenvektors $\mathbf{w}_c$ der Matrix **W** ist:

$$b_r(c) = \sum_{r=1}^{R} w_{r,c}\, a_r(c)\,. \tag{4.74}$$

Der Zentrale Grenzwertsatz der Stochastik besagt, dass die Summe zweier unabhängiger Zufallsvariablen eine Amplitudendichteverteilung hat, die einer Gaußverteilung ähnlicher und somit gaußartiger ist als die Amplitudendichteverteilungen der beiden Summanden es sind. Da in Gleichung 4.72 die Merkmalssignale $a_r(c)$ aus den gewichteten Summen der statistisch unabhängigen Quellsignale $b_r(c)$ bestehen, sind die Amplitudendichteverteilungen der Merkmalssignale gaußartiger als die der Quellsignale. Als Maß für die Gaußartigkeit kann die Kurtosis verwendet werden, die aus der normierten Form des vierten zentralen Moments der Statistik hervorgeht:

$$\beta_a = \frac{\frac{1}{C}\sum_{c=1}^{C}\left(a(c)-\mu_a\right)^4}{{\sigma_a}^4} - 3\,. \tag{4.75}$$

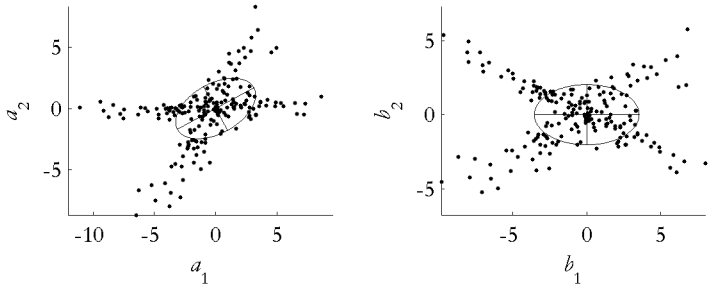


Bild 4.10: Anwendung der PCA auf nicht normalverteilte, zweidimensionale Merkmalsvektoren, links: Originalmatrix **A**, rechts: Transformationsergebnis **B**, Standardabweichung mit ihren Hauptachsen (Ellipsen)

Der Wert der Kurtosis gibt nicht nur Aufschluss über die Gaußartigkeit, sondern auch über die grundsätzliche Form der Amplitudendichteverteilung:

- $\beta_a = 0$ entspricht einer Gaußverteilung,
- $\beta_a < 0$ entspricht einer Verteilung, die flacher als eine Gaußverteilung ist,
- $\beta_a > 0$ entspricht einer Verteilung, die spitzer als eine Gaußverteilung ist.

Für die Berechnung der optimalen Werte von **W** muss die Gaußartigkeit und somit der Betrag der Kurtosis der einzelnen Quellsignale in Abhängigkeit jeweils eines R-dimensionalen Spaltenvektors $\mathbf{w}_c$ minimiert werden. Hierbei treten im R-dimensionalen Raum $2R$ lokale Extrema auf, wobei jeweils zwei Extrema einer unabhängigen Komponente zugeordnet sind (mit positivem und negativem Vorzeichen von $\mathbf{w}_c$). Werden die Merkmalssignale der Merkmalsmatrix **A** im Vorfeld über die SVD in ihrer Energie normiert, so kann die Suche auf die Einheitskugel im R-dimensionalen Hyperraum beschränkt werden. Die hieraus resultierenden Quellsignale haben dann ebenfalls Einheitsvarianz. Für die Berechnung der ICA gibt es sehr effiziente Algorithmen wie Infomax, FastICA und JADE [HyOj00].

Die ICA liefert gute Ergebnisse, wenn das zugrunde liegende Modell die Realität ausreichend genau abbildet und der Merkmalsmatrix verdeckte,

nicht gaußartige Merkmalssignale zugrunde liegen. Dann lassen sich die Merkmale in unabhängige Signale zerlegen, die einen sehr gut separierbaren Merkmalsraum aufspannen. Hierbei werden Signalmatrizen erzeugt, die deutlich aussagekräftiger sind als die Ergebnisse der PCA oder SVD. Ein Beispiel hierzu zeigt Bild 4.10, in dem die PCA auf nicht normalverteilte zweidimensionale Merkmalsvektoren angewendet wird, im Vergleich zu Bild 4.9. Die durch die PCA erkannten Hauptkomponenten liegen in Bild 4.10 genau zwischen den Keulen der Merkmalsverteilung und sind somit nicht aussagekräftig.

Ist die Modellierung der ICA hingegen nicht zutreffend, d.h. sind die Merkmale nicht über eine annähernd lineare Mischungsmatrix miteinander verknüpft, bringt die Transformation mittels der ICA keinen Vorteil gegenüber der PCA, ist jedoch deutlich aufwändiger in der Berechnung. Im schlimmsten Fall können sich die Merkmalssignale sogar verschlechtern, weil durch die Transformation einzelne Merkmalspunkte aus den zu ihnen gehörenden Klassen herausgerissen werden. Weiterhin ist die ICA im Gegensatz zur PCA und SVD nicht robust im Umgang mit mittelwertbehafteten Daten. Die Merkmalssignale müssen also vor der Verarbeitung mittelwertbefreit sein und sollten in ihrer Energie normiert sein.

Für die Anwendung der ICA im Zusammenhang mit Merkmalen von Musikinstrumentenklängen ergeben sich gute Ergebnisse, wenn die Daten untereinander verknüpft sind und sich aus versteckten, zugrunde liegenden Signalen ergeben. Dies ist ähnlich wie bei der PCA und SVD dann der Fall, wenn Zeitbereichsmerkmale oder einfache spektrale Merkmale verwendet werden, die sich stetig im Merkmalsraum verteilen. Je komplexer der Zusammenhang der Merkmale untereinander bzw. zu versteckten Quellsignalen ist, desto schlechter werden die mit der ICA erzielbaren Ergebnisse [CaWe01, Case01a, Case01b, KiBS04].

4.6.6 NICHT-NEGATIVE MATRIXFAKTORISIERUNG

Die nicht-negative Matrixfaktorisierung (engl. Non-negative Matrix Factorization, NMF) ist eine besondere Art der Matrixfaktorisierung, bei der die Merkmalsmatrix $\mathbf{A}_{(R \times C)}$ durch die zwei Matrizen $\mathbf{W}_{(R \times M)}$ und $\mathbf{H}_{(M \times C)}$ so faktorisiert wird, dass die Differenz $\mathbf{U}_{(R \times C)}$ minimiert wird:

$$\mathbf{A} = \mathbf{WH} + \mathbf{U} \,. \tag{4.76}$$

Die namensgebende Einschränkung der Faktorisierung ist, dass die drei Matrizen **A**, **W** und **H** ausschließlich positive Elemente enthalten dürfen. Die Signale einer Merkmalsmatrix mit positiven und negativen Werten müssen also vor der Verarbeitung mit einem Gleichanteil belegt werden.

Die Matrix **W** heißt Basismatrix und enthält als Spalten die normierten Basisvektoren. Die Signalmatrix **H** enthält als Zeilen die für die Rekonstruktion der Merkmalsmatrix aus den Basisvektoren benötigten Summationskoeffizienten. In der Regel wird $M < R$ gewählt, so dass $(R + C)M < RC$ ist. Somit ist die Signalmatrix gegenüber der Merkmalsmatrix dimensionsreduziert. Durch die Einhaltung der folgenden drei Einschränkungen kann das Ergebnis der Faktorisierung optimiert werden:

- Die Anzahl der Basisvektoren in **W** sollte minimiert werden.
- Unterschiedliche Basisvektoren sollten maximal orthogonal sein.
- Die Signale mit dem größten Informationsgehalt sollten als Dimensionen beibehalten werden.

Für die Berechnung der eigentlichen Faktorisierung werden Approximationsverfahren verwendet, die ein Distanzmaß zwischen **A** und **WH** minimieren. Ein Distanzmaß, das gute Ergebnisse liefert, ist die Kullback-Leibler-Divergenz, die die Unterschiedlichkeit zweier Wahrscheinlichkeitsverteilungen beschreibt, wobei $\dddot{\times}$ die elementweise Multiplikation und $\dddot{\div}$ die elementweise Division ausdrücken:

$$D = \mathbf{A} \dddot{\times} \ln\left(\mathbf{A} \dddot{\div} (\mathbf{WH})\right) - \mathbf{A} + (\mathbf{WH}). \tag{4.77}$$

Mit der Minimierung der Kullback-Leibler-Divergenz D können die optimalen Werte für **W** und **H** ermittelt werden. Details zur Berechnung der NMF finden sich in zahlreichen Veröffentlichungen [LeSe99, LeSe00, Hoye04, BKK+05].

Die NMF liefert im Zusammenhang mit Merkmalen von Musikinstrumentenklängen sehr gute Ergebnisse, wenn den in der Merkmalsmatrix beobachteten Merkmalssignalen ähnlich wie bei der ICA unabhängige Quellsignale durch versteckte Merkmalsquellen zugrunde liegen. So wurde die NMF mit Erfolg für die Zerlegung von Amplitudenspektrogrammen [SmBr03, Smar04] und für die Analyse von monophonen Musiksig-

nalen im Zusammenhang mit der Identifikation und Klassifikation von Musikinstrumenten [ChCB03, BKK+05] verwendet.

Wenn im Vorfeld der NMF bekannt ist, wie viele Basisvektoren für eine korrekte Modellierung verwendet werden müssen, so ergeben sich sehr aussagekräftige Basisvektoren, die den unabhängigen Komponenten des analysierten Aufbaus entsprechen. Diese spannen dann ähnlich der ICA einen sehr gut separierbaren Merkmalsraum auf, so dass eine Klassifikation recht einfach durchführbar ist. Auch die NMF liefert die besten Ergebnisse bei Zeitbereichsmerkmalen oder einfachen spektralen Merkmalen, die sich stetig im Merkmalsraum verteilen. Mit der NMF können allerdings im Vergleich zu den andern vorgestellten Verfahren der Dimensionsreduktion auch sehr gut Merkmalsmatrizen bearbeitet werden, die die Verknüpfung von komplexen Merkmalen abbilden. Ein Nachteil der NMF ist, dass das Ergebnis von dem gewählten Verfahren und dem Optimierungsalgorithmus abhängt und somit zu einer Merkmalsmatrix unterschiedliche Lösungen existieren.

5 KLASSIFIKATION

Die automatische Klassifikation oder Mustererkennung beschreibt den Vorgang der Bewertung eines Zustandes und die aus der Bewertung als Ergebnis folgende Zuordnung zu einer von mehreren klar definierten Klassen. Die automatische Identifikation eines Zustandes stellt hierbei einen Sonderfall dar, bei dem lediglich die Entscheidung, ob der Zustand zu einer der bekannten Klassen gehört, interessiert. Der Zustand des zu klassifizierenden Vorgangs wird zu jedem Zeitpunkt über einen R-dimensionalen Merkmalsvektor **a** beschrieben, der den Zustand über R verschiedene Merkmale beschreibt. Für den in dieser Arbeit behandelten Fall der Identifikation und Klassifikation von Musikinstrumenten aus Musiksignalen wurde die Merkmalsextraktion für die Zustandsbeschreibung ausgiebig im vorangegangenen Kapitel beschrieben.

Aus Sicht der Signalverarbeitung stellt ein Klassifikationsverfahren eine Funktion dar, die R-dimensionale Vektoren auf eine Menge von D Musikinstrumentenklassen ω_d abbildet. Hierbei kann der zur Klassifikation verwendete Klassifikator gedächtnisbehaftet sein, so dass er einen inneren Zustand hat, der mit in das Klassifikationsergebnis eingeht.

In diesem Kapitel werden die Grundlagen der automatischen Klassifikation sowie die verschiedenen untersuchten Klassifikationsverfahren beschrieben, die verwendet werden können, um Merkmalsvektoren **a** einer von D Musikinstrumentenklassen ω_d zuzuordnen.

5.1 KLASSIFIKATIONSVERFAHREN

Ein Merkmalsvektor **a**, der *R* Merkmale zusammenfasst, lässt sich eindeutig als Punkt im *R*-dimensionalen Hyperraum beschreiben. Die Merkmalsvektoren verschiedener Klassen treten bei der Wahl aussagekräftiger Merkmale im Hyperraum als Wolken auf (Bild 5.1). Hierbei können je nach Wahl der Merkmale die Form und Position der Wolken zeitvariant sein, was den Klassifikationsvorgang erschwert, da das Klassifikationsverfahren den Zeitpunkt mitberücksichtigen muss.

Um eine Klassifikation durchführen zu können, muss der Klassifikationsalgorithmus auf Erfahrungen über die Eigenschaften der zu klassifizierenden Daten zurückgreifen können. Diese Erfahrungen werden entweder durch einen expliziten oder impliziten Trainingsvorgang gewonnen oder als Satz von Regeln festgelegt. Über die Art, wie diese Erfahrungen gewonnen werden, lassen sich die verschiedenen Klassifikationsverfahren in drei Kategorien einteilen, die in den folgenden Abschnitten beschrieben werden.

5.1.1 SYNTAKTISCHE VERFAHREN

Syntaktische Verfahren zeichnen sich dadurch aus, dass das Verfahren auf einem Satz von Regeln basiert, der im Vorfeld bekannt und implementiert sein muss. Syntaktische Verfahren haben somit keine Trainingsphase in dem Sinne, dass dem Klassifikator Trainingsdaten präsentiert werden. Vielmehr müssen die Regeln im Vorfeld heuristisch oder analytisch gewonnen werden und können dann über die klassische Aussagenlogik oder Prädikatenlogik angewendet werden.

Syntaktische Verfahren benötigen Merkmalsvektoren, die mit einer sehr hohen Genauigkeit komplexe Zustände beschreiben, da die Regeln direkt auf die gewonnenen Merkmalsvektoren angewendet werden. Weiterhin erfordern selbst überschaubare Klassifikationsaufgaben eine sehr große Vielzahl von Regeln, so dass syntaktische Klassifikationsverfahren in dieser Arbeit nicht weiter untersucht wurden.

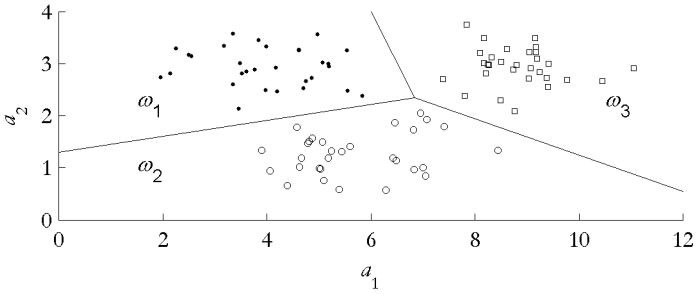


Bild 5.1: Drei Klassenwolken (ω_1, ω_2, ω_3) im zweidimensionalen Merkmalsraum mit den dazugehörigen Merkmalsausprägungen

5.1.2 STATISTISCHE VERFAHREN

Statistische Verfahren erlernen die statistischen Eigenschaften der zu klassifizierenden Merkmalsvektoren in einer Trainingsphase. Hierbei wird dem Klassifikator ein Trainingsdatensatz präsentiert, der eine Vielzahl von Merkmalsvektoren für jede Klasse enthält.

Im Klassifikationsbetrieb berechnet der Klassifikator für jeden Trainingsdatensatz die Wahrscheinlichkeit, mit der er zu den jeweiligen Klassen gehört, und trifft eine Klassifikationsentscheidung über die größte Einzelwahrscheinlichkeit.

Statistische Verfahren sind für Merkmalsvektoren, die eine gewisse Ungenauigkeit enthalten, so wie sie bei der Klassifikation von Musikinstrumentenklängen auftreten, sehr gut anwendbar. Im Rahmen dieser Arbeit wurden verschiedene statistische Klassifikationsverfahren untersucht, die in den Abschnitten 5.3 und 5.4 beschrieben sind.

5.1.3 NEURONALE VERFAHREN

Neuronale Verfahren basieren auf künstlichen Neuronalen Netzen, die durch biologische Neuronalen Netzen inspirierte Netzwerke darstellen. Neuronale Verfahren erlernen die Eigenschaften der zu klassifizierenden Merkmalsvektoren ebenfalls in einer Trainingsphase. Hierbei werden die internen Parameter des künstlichen Neuronalen Netzes so angepasst, dass der Ausgang des Netzes einen gewünschten Wert annimmt, der der

Klassifikationsentscheidung entspricht. Die verschiedenen im Rahmen dieser Arbeit untersuchten neuronalen Klassifikationsverfahren sind in Abschnitt 5.5 beschreiben.

5.2 LERNVERFAHREN

Der während der Trainingsphase eines Klassifikationsverfahrens verwendete Trainingsdatensatz besteht generell aus einer oder mehreren Merkmalsmatrizen. Über den genauen Aufbau des Trainingsdatensatzes und die Art, wie er von dem Klassifikator erlernt wird, lassen sich verschiedene Lernverfahren unterscheiden. Ein generelles Problem, das mit allen Lernverfahren auftreten kann, ist das Problem der Überanpassung, das im Anschluss an die unterschiedlichen Lernverfahren in den folgenden Abschnitten beschrieben wird.

5.2.1 ÜBERWACHTES LERNEN

Beim überwachten Lernen (engl. Supervised Learning) werden dem Klassifikator als Trainingsdaten zu jedem Merkmalsvektor **a** das korrekte Klassifikationsergebnis ω_d als Klassenname oder Klassenindex d angegeben. Wenn die zeitliche Abfolge der Merkmalsvektoren keine Rolle spielt, wird häufig für jede Klasse eine eigene Merkmalsmatrix erstellt, für die während des Trainings einmalig das Klassifikationsergebnis ω_d angegeben wird.

5.2.2 UNÜBERWACHTES LERNEN

Beim unüberwachten Lernen (engl. Unsupervised Learning) werden dem Klassifikator als Trainingsdaten lediglich Merkmalsvektoren **a** ohne das dazugehörige Klassifikationsergebnis ω_d angegeben. Unüberwachte Lernverfahren werden oft auch clusterbildende Verfahren genannt, weil die Aufteilung in verschiedene Cluster dem Algorithmus überlassen wird. Häufig wird dem Verfahren als Parameter die Anzahl der zu bildenden Cluster gegeben.

5.2.3 KONTINUIERLICHES LERNEN

Beim kontinuierlichen Lernen (engl. Online Learning) kann der Klassifikator über die während des Klassifikationsbetriebs präsentierten Merkmalsvektoren kontinuierlich lernen. Bei Verfahren des überwachten Lernens bedeutet dies, dass das Klassifikationsergebnis von außen bewertet werden muss, beim unüberwachten Lernen hingegen können neue Klassen eröffnet werden, wenn die Abweichung von den bisherigen, zu den verschiedenen Klassen typischerweise gehörenden Merkmalsvektoren einen gewissen Schwellenwert überschreitet.

Prinzipiell lässt sich jedes Klassifikationsverfahren für den kontinuierlichen Lernbetrieb nutzen, für viele Verfahren ist es jedoch nicht sinnvoll, weil alle dem Klassifikator präsentierten Mermalsvektoren intern gespeichert werden müssten und nach jedem neuen Merkmalsvektor eine komplette Neuberechnung des Modells erforderlich ist. Sinnvoll ist das kontinuierliche Lernen nur für Verfahren, die ein inkrementelle Veränderung der Modellparameter zulassen.

5.2.4 ÜBERANPASSUNG

Wird ein Klassifikationssystem zu exakt auf einen bestimmten Datensatz trainiert, so kann es zu einer Überanpassung (engl. Overfitting) kommen (Bild 5.2). Hierbei hat das System zwar eine sehr hohe Klassifikationsrate für den Trainingsdatensatz, unbekannte Daten werden jedoch sehr fehlerhaft klassifiziert. Bei einer Überanpassung ist das System nicht mehr in der Lage zu generalisieren, was bedeutet, dass Merkmalsvektoren, die bis auf sehr kleine Abweichungen identisch sind, unterschiedlichen Klassen zugeordnet werden können.

Überanpassung entsteht, wenn die Anzahl der freien Modellparameter den Informationsgehalt der Trainingsdaten übersteigt. Dieser Fall tritt beispielsweise dann auf, wenn zu wenig Trainingsdaten vorhanden sind oder ein zu komplexes Modell trainiert werden soll. Weiterhin kann es zu einer Überanpassung kommen, wenn Merkmalsvektoren mit einer zu großen Dimensionalität gewählt werden, da dies unweigerlich zu einem Modell mit einer großen Anzahl von Freiheitsgraden führt (vgl. hierzu auch Abschnitt 4.6.1).

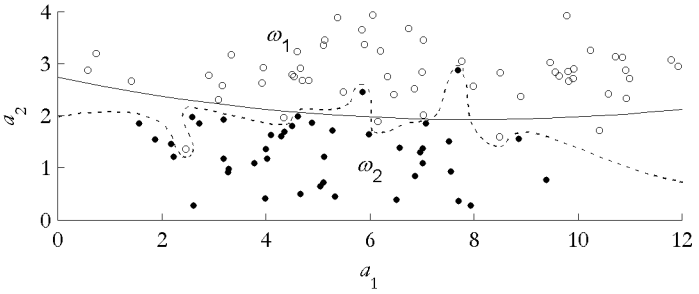


Bild 5.2: Zwei Klassenwolken im zweidimensionalen Merkmalsraum (ω_1: Kreise, ω_2: Punkte) mit überangepasster Klassengrenze (gestrichelt) und generalisierender Klassengrenze (durchgezogen)

Um eine Überanpassung zu verhindern gibt es verschiedene heuristische Verfahren. Eine generell anwendbare Methode ist die Kreuzvalidierung (engl. Cross Validation), bei der für das Training des Systems zusätzlich zur Trainingsmatrix auch eine Validierungsmatrix benötigt wird. Diese Validierungsmatrix enthält ähnliche Merkmalsvektoren wie die Trainingsmatrix, ist dem System allerdings unbekannt. Während des Trainings werden in bestimmten Abständen die Klassifikationsraten des Systems sowohl für die Trainingsmatrix als auch für die Validierungsmatrix berechnet. Das System gilt dann als optimal trainiert, wenn die Klassifikationsrate für die Trainingsmatrix und die Klassifikationsrate für die Validierungsmatrix nicht mehr korrelieren. In der Regel bedeutet dies, dass die Klassifikationsrate für die Trainingsmatrix weiter steigt oder stagniert und die Klassifikationsrate für die Validierungsmatrix zu fallen beginnt (Bild 5.3). Eine zusätzliche Verbesserung der Kreuzvalidierung ergibt sich durch das parallele Training mehrerer Klassifikationssysteme mit leicht unterschiedlichen Trainings- und Validierungsmatrizen bzw. leicht variierten Parametern. Das System mit der besten Klassifikationsrate wird anschließend verwendet [LiRo03].

Neben der Kreuzvalidierung hat sich der frühe Abbruch des Trainings schon nach wenigen Trainingszyklen (engl. Early Stopping) als sinnvolle Methode gegen Überanpassung erwiesen. Hierbei kann ein günstiger Abbruchzeitpunkt jedoch nicht generell festgelegt werden, sondern hängt von den individuellen Gegebenheiten des Systems ab.

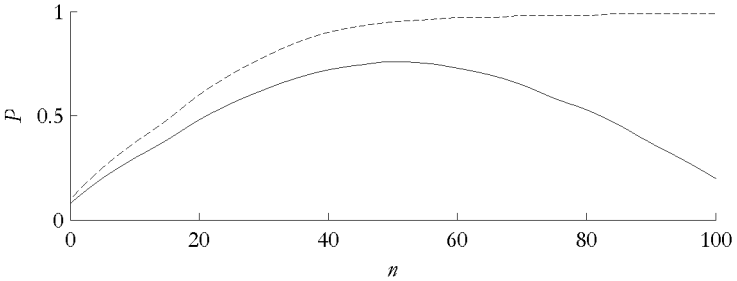


Bild 5.3: Mittlere Klassifikationsrate *P* abhängig vom Trainingszyklus *n*, Klassifikationsrate der Trainingsmatrix (gestrichelt), Klassifikationsrate der Validierungsmatrix (durchgezogen)

5.3 DISTANZBASIERTE MODELLE

Bei distanzbasierten Modellen wird die Klassifikationsentscheidung getroffen, indem die Distanz zwischen dem zu klassifizierenden Merkmalsvektor **a** und einem oder mehreren anderen Vektoren **b** ermittelt wird [DaMe80, HAB+00, DuHS01]. Als Distanzmaß lassen sich verschiedene Vektornormen verwenden, in der Regel wird jedoch die euklidische Distanz verwendet, die dem einfachen Abstand der beiden Vektoren entspricht:

$$l(\mathbf{a},\mathbf{b}) = \sqrt{\sum_{r=1}^{R} (a_r - b_r)^2} \quad . \tag{5.1}$$

In den folgenden Abschnitten werden drei häufig verwendete distanzbasierte Verfahren im Zusammenhang mit der Musikinstrumentenidentifikation und Klassifikation beschrieben.

5.3.1 HIERARCHISCHE CLUSTERBILDUNG

Die hierarchische Clusterbildung (engl. Hierarchical Clustering oder Average-Linkage Clustering) ist ein relativ einfacher und dennoch leistungsfähiger Algorithmus, bei dem aus den Merkmalsvektoren der zum Training verwendeten Merkmalsmatrix eine als Parameter gegebene An-

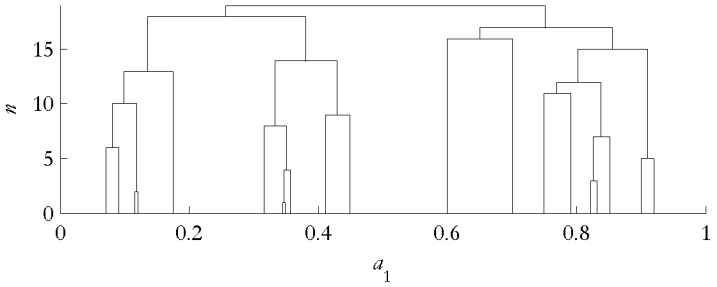


Bild 5.4: Hierarchische Clusterbildung eines skalaren Merkmals a_1 in 19 Iterationsschritten n

zahl von D Clustern gebildet wird. Die Mittelpunkte der Cluster werden durch die Clustervektoren $\mathbf{w}_d$ beschrieben [John67, D'an78].

Das Verfahren arbeitet als unüberwachtes Lernverfahren und ist gut für das kontinuierliche Lernen geeignet, wobei eine zu lange kontinuierliche Lernphase bei einer ungünstigen Abfolge von Merkmalsvektoren die Klassenbildung beeinträchtigen kann. Deshalb sollten für gute Ergebnisse die präsentierten Merkmalsvektoren in Gruppen protokolliert werden, um in Abständen eine komplette Neuberechnung des Modells vornehmen zu können.

Bei einer Merkmalsmatrix mit C Merkmalsvektoren $\mathbf{a}_c$ erfolgt die Clusterbildung in D Klassen über den folgenden iterativen Prozess (Bild 5.4):

1. Als Initialisierung werden C Cluster mit den Clustervektoren $\mathbf{w}$ aus den Merkmalsvektoren $\mathbf{a}$ gebildet, wobei jedes Cluster lediglich einen Vektor enthält:

$$\mathbf{w}_c = \mathbf{a}_c . \tag{5.2}$$

2. Aus den Clustervektoren $\mathbf{w}$ wird eine Distanzmatrix $\mathbf{L}$ gebildet, die die euklidischen Distanzen aller Vektoren $\mathbf{w}$ untereinander enthält.

3. Innerhalb der Distanzmatrix $\mathbf{L}$ wird der kleinste Wert und somit die kleinste Distanz ermittelt. Werden die zugehörigen Clustervektoren mit $\mathbf{w}_x$ und $\mathbf{w}_y$ bezeichnet, so wird aus ihrem Mittelwert ein

neues Cluster mit dem Clustervektor $\mathbf{w}_z$ gebildet, so dass sich die Anzahl der ermittelten Cluster um eins verringert:

$$\mathbf{w}_z = \frac{1}{R}\left(\mathbf{w}_x + \mathbf{w}_y\right). \tag{5.3}$$

4. Ist die vorgegebene Anzahl von D Clustern erreicht, bricht die Iteration ab und die D Clustervektoren $\mathbf{w}_d$ beschreiben die Mittelpunkte der Klassen. Ansonsten wird mit dem aktuellen Satz von Clustern bei Schritt 2 fortgefahren.

Im Klassifikationsbetrieb wird für einen zu klassifizierenden Merkmalsvektor **a** die euklidische Distanz zu allen Clustervektoren **w** berechnet. Der Merkmalsvektor wird in die Klasse eingeordnet, für deren Clustervektor sich die kleinste Distanz ergibt.

Die hierarchische Clusterbildung lässt sich im Zusammenhang mit der Musikinstrumentenidentifikation und Klassifikation gut verwenden, wenn aufgrund des gewählten Extraktionsszenarios und der gewählten Merkmale sichergestellt ist, dass die Merkmalsvektoren sich im Merkmalsraum in deutlich unterscheidbare Wolken verteilen. Dies ist beispielsweise bei monophonen Musiksignalen und der Verwendung von harmonischen Merkmalen der Fall, hier liefert der Algorithmus trotz seiner Einfachheit gute Ergebnisse. Weiterhin lässt das Verfahren sich gut für die Vorbereitung der Trainingsdaten für das Training von komplexeren Verfahren wie Gaussian Mixture Models oder Hidden Markov Models verwenden (vgl. die Abschnitte 5.4.2 und 5.4.3).

Nachteile des Algorithmus sind, dass eine einmal entschiedene Clusterzugehörigkeit sich nicht mehr umkehren lässt, was die Clusterbildung anfällig gegenüber Ausreißern macht. Weiterhin lassen sich keine einfachen Merkmalsvektoren verarbeiten, wenn diese eine gewisse Ungenauigkeit beinhalten.

5.3.2 VEKTORQUANTISIERUNG

Die Vektorquantisierung (engl. Vectorquantization oder K-Means) stellt ebenfalls ein Verfahren des unüberwachten Lernens dar, das den Merkmalsraum in eine als Parameter vorgegebene Anzahl von D Clustern aufteilt. Jedes Cluster wird durch einen Vektor $\mathbf{w}_d$ beschrieben, der auf

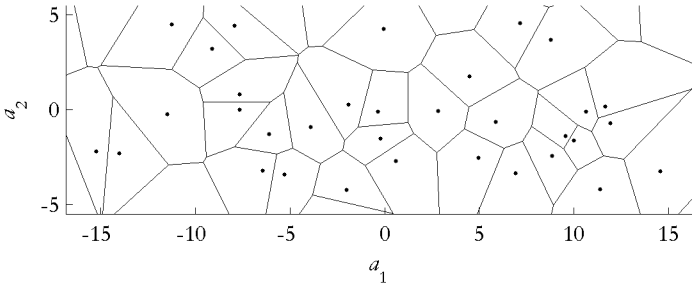


Bild 5.5: Vektorquantisierung im zweidimensionalen Merkmalsraum, die Codebuchvektoren sind als Punkte eingetragen, die ebenfalls eingetragenen Clustergrenzen bilden so genannte Voronoizellen

den Mittelpunkt der zu diesem Cluster gehörenden Datenpunkte zeigt. Die Vektoren $\mathbf{w}_d$ werden Codebuchvektoren genannt und so gewählt, dass alle Merkmalsvektoren $\mathbf{a}$ der Trainingsdatenmenge mit möglichst geringer Verzerrung durch einen der D Codebuchvektoren ersetzt werden können. Ein Beispiel zur Vektorquantisierung im zweidimensionalen Merkmalsraum ist in Bild 5.5 gegeben.

Der optimale Satz von Codebuchvektoren ist der, der die mittlere euklidische Distanz $l\_$ jedes Vektors $\mathbf{a}_c$ zum nächstgelegenen Codebuchvektor $\mathbf{w}_c$ minimiert:

$$l\_ = \frac{1}{C}\sum_{c=1}^{C} l\left(\mathbf{a}_c, \mathbf{w}_c\right) \tag{5.4}$$

und kann in der Regel nur approximiert werden. Hierzu stehen Verfahren wie der Lloyd-Algorithmus zur Verfügung, bei dem die folgende Iteration ausgeführt wird [Lloy82]:

1. Aus Initialisierungsvektoren mit prinzipiell beliebigen Werten wird ein Anfangssatz von Codebuchvektoren $\mathbf{w}_d$ gebildet.
2. Die C Trainingsvektoren $\mathbf{a}_c$ werden jeweils dem Cluster $\mathbf{W}_d$ zugeordnet, dessen Codebuchvektor $\mathbf{w}_d$ die geringste euklidische Distanz zu $\mathbf{a}_c$ aufweist.
3. Der Mittelpunkt $\mathbf{w}_d$ jedes Clusters $\mathbf{W}_d$ wird durch die ihm zugeordneten C_d Vektoren bestimmt:

$$\mathbf{w}_d = \frac{1}{C_d} \sum_{c=1}^{C_d} \mathbf{a}_c \Big|_{\mathbf{a}_c \in \mathbf{W}_d} \tag{5.5}$$

und als neuer Codebuchvektor $\mathbf{w}_d$ gespeichert.

4. Die mittlere euklidische Distanz $\bar{l}$ der C Trainingsvektoren $\mathbf{a}_c$ zu den neuen Codebuchvektoren $\mathbf{w}_d$ wird berechnet. Ist die auf den letzten Iterationsdurchlauf bezogene Verringerung des mittleren Abstandes kleiner als ein vorgegebener Grenzwert oder wurde eine vorgegebene Anzahl von Iterationsdurchläufen erreicht, bricht der Algorithmus ab. Ansonsten wird mit den neuen Codebuchvektoren bei Schritt 2 fortgefahren.

Prinzipiell konvergiert der Algorithmus immer, es kann bei den erreichten Codebuchvektoren jedoch nicht zwischen lokalen und globalen Minima der mittleren Distanz unterschieden werden. Das Laufzeitverhalten und das Ergebnis hängen stark von der Wahl guter Initialisierungsvektoren ab. In der Praxis wird der Algorithmus während der Trainingsphase mehrfach auf eine Datenmenge angewendet. Das Ergebnis, das die kleinste mittlere euklidische Distanz $\bar{l}$ liefert, wird dann als Endergebnis verwendet.

Um optimale Initialisierungsvektoren zu erhalten, werden für mittelwertbefreite Daten die Richtungen der Vektoren so gewählt, dass die Winkel untereinander maximal groß sind. Die Längen der Vektoren werden so gewählt, dass sie der Standardabweichung der gewählten Richtung entsprechen. Für Daten mit einem globalen Mittelwert muss dieser vor der Berechnung der Vektoren abgezogen werden und anschließend auf die ermittelten Initialisierungsvektoren addiert werden.

Als weitere Möglichkeit für die Bildung guter Initialisierungsvektoren kann im Vorfeld mit den Trainingsdaten eine hierarchische Clusterbildung auf D Cluster durchgeführt werden. Die Clusterzentren können dann als Initialisierungsvektoren für die Vektorquantisierung verwendet werden.

Eine Erweiterung des Lloyd-Verfahrens bildet das Linde-Buzo-Gray-Verfahren (LBG), das die folgende Iteration durchführt [LiBG80]:

1. Als Initialisierung wird ein einziger Codebuchvektor $\mathbf{w}_d$ entsprechend dem globalen Mittelpunkt der Trainingsdaten gewählt.

2. Alle Codebuchvektoren $\mathbf{w}_d$ werden mit einem zufällig erzeugten Störvektor in zwei neue Codebuchvektoren $\mathbf{w}_{d+}$ und $\mathbf{w}_{d-}$ aufgeteilt, indem der Störvektor einmal addiert und einmal subtrahiert wird.
3. Das Lloyd-Verfahren wird durchgeführt, wobei alle bereits ermittelten Codebuchvektoren $\mathbf{w}_{d+}$ und $\mathbf{w}_{d-}$ als Initialisierungsvektoren verwendet werden.
4. Ist die Anzahl D der geforderten Codebuchvektoren erreicht, bricht der Algorithmus ab, ansonsten wird bei Schritt 2 fortgefahren.

Obwohl das LBG-Verfahren die mehrmalige Durchführung des Lloyd-Verfahrens beinhaltet, läuft es trotzdem effizienter ab, da die einzelnen Lloyd-Durchläufe schneller konvergieren. Weiterhin liefert das LBG-Verfahren in der Regel bessere Ergebnisse als das Lloyd-Verfahren.

Im Klassifikationsbetrieb wird für einen zu klassifizierenden Merkmalsvektor **a** die euklidische Distanz zu allen Codebuchvektoren **w** berechnet. Der Merkmalsvektor wird in die Klasse eingeordnet, für deren Codebuchvektoren sich die kleinste Distanz ergibt.

Die Vektorquantisierung liefert im Zusammenhang mit der Musikinstrumentenidentifikation und Klassifikation bessere Ergebnisse als die hierarchische Clusterbildung. Besonders in Bezug auf die Stabilität und das Konvergierungsverhalten ist die Vektorquantisierung entschieden besser.

Die besten Ergebnisse ergeben sich auch für die Vektorquantisierung, wenn sich durch das Extraktionsszenario und die gewählten Merkmale Merkmalsvektoren ergeben, die sich im Merkmalsraum in deutlich unterscheidbare Wolken verteilen [SoRo88]. Dies ist beispielsweise der Fall bei der Verwendung von monophonen Musiksignalen und harmonischen Merkmalen. Die Vektorquantisierung liefert im Gegensatz zur hierarchischen Clusterbildung allerdings auch bei der Verwendung einfacherer Merkmale, die eine gewisse Ungenauigkeit beinhalten, gute Ergebnisse. Sie wird wie die hierarchische Clusterbildung ebenfalls häufig für die Vorbereitung der Trainingsdaten für das Training von komplexeren Verfahren wie Gaussian Mixture Models oder Hidden Markov Models verwendet (vgl. hierzu die Abschnitte 5.4.2 und 5.4.3).

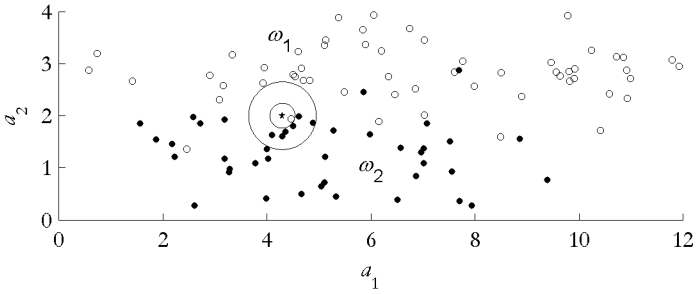


Bild 5.6: Klassifikation eines neuen Merkmalsvektors (Stern) mittels KNN. Für $K = 1$ wird der neue Merkmalsvektor der Klasse ω_1 zugeordnet (Kreise), für $K = 7$ hingegen der Klasse ω_2 (Punkte).

5.3.3 K-NEAREST NEIGHBOUR

Der K-Nearest-Neighbour (KNN) Algorithmus ist ein Klassifikationsverfahren aus dem Bereich des überwachten Lernens, mit dem sich gute Klassifikationsergebnisse erzielen lassen, die oft auch als Benchmark für komplexere Verfahren verwendet werden. Das KNN-Verfahren stellt weiterhin eine effiziente und gute Schätzung der Wahrscheinlichkeitsdichteverteilung der Merkmalsvektoren dar [Alle87, DuHS01, Nabn03].

Während der Trainingsphase werden lediglich alle Merkmalsvektoren $\mathbf{a}_\ell$ und die dazugehörigen Klassen ω_ℓ gespeichert, wobei die zeitliche Abfolge der Merkmalsvektoren irrelevant ist. Somit ist das Trainingsverfahren sehr einfach, und ein nachträgliches Lernen von weiteren Merkmalsvektoren während des Betriebs als kontinuierliches Lernen ist jederzeit problemlos möglich. Nachteilig bei diesem Vorgehen ist jedoch, dass eine große Speicherkapazität benötigt wird.

Für die Klassifikation eines unbekannten Merkmalsvektors werden die euklidischen Distanzen zu allen gespeicherten Merkmalsvektoren $\mathbf{a}_\ell$ berechnet und die K nächstgelegenen Nachbarvektoren zu einer Gruppe zusammengefasst. Aus der Klassenzugehörigkeit der Vektoren dieser Gruppe wird nun die Klassifikation des unbekannten Merkmalsvektors vorgenommen, indem er der Klasse zugeordnet wird, der die meisten Vektoren der Gruppe bereits zugeordnet sind. Bild 5.6 zeigt hierzu ein Beispiel im zweidimensionalen Merkmalsraum.

Im Zusammenhang mit der Musikinstrumentenidentifikation und Klassifikation ist der KNN-Algorithmus wegen seiner Einfachheit und seines Benchmarkcharakters weit verbreitet und liefert gute Ergebnisse [ErKl00, AgLP01, AgLP03, HePD02, KiGO03].

5.4 WAHRSCHEINLICHKEITSMODELLE

Wahrscheinlichkeitsmodelle beschreiben die Statistik des zu klassifizierenden Prozesses, indem die Verteilungsdichtefunktionen der Merkmalsvektoren modelliert werden. Die Klassifikationsentscheidung ergibt sich hierbei direkt über die Bayes'sche Klassifikation, die in den folgenden Abschnitten zusammen mit leistungsfähigen Wahrscheinlichkeitsmodellen beschrieben wird.

5.4.1 BAYES'SCHE KLASSIFIKATION

Die Bayes'sche Klassifikation gibt für einen Merkmalsvektor $\mathbf{a}$ die Wahrscheinlichkeit $P(\omega|\mathbf{a})$ an, mit der er als Klassifikationsergebnis der Klasse ω zugeordnet werden kann. Die Wahrscheinlichkeit $P(\omega|\mathbf{a})$, die auch a-posteriori Wahrscheinlichkeit (engl. Posterior) genannt wird, lässt sich über die Regel von Bayes berechnen [Alle87, DuHS01]:

$$P(\omega|\mathbf{a}) = \frac{p(\mathbf{a}|\omega)P(\omega)}{p(\mathbf{a})}. \tag{5.6}$$

Hierbei beschreibt die so genannte Likelihood $p(\mathbf{a}|\omega)$, mit welcher Wahrscheinlichkeitsdichte die Merkmalsvektoren $\mathbf{a}$ innerhalb der Klasse ω verteilt sind. Für die Klassifikation von Musikinstrumenten ist dies die Verteilung der Merkmalsvektoren für jedes einzelne Instrument. Die a-priori Wahrscheinlichkeit $P(\omega)$ (engl. Prior) gibt die Wahrscheinlichkeit für das generelle Auftreten der Klasse ω unabhängig vom Merkmalsvektor $\mathbf{a}$ an. Im Zusammenhang mit Musikinstrumenten ist diese Größe wichtig, da gewisse Instrumente grundsätzlich häufiger auftreten als andere. So findet man beispielsweise innerhalb der klassischen Musik deutlich mehr Streichinstrumente als elektronische Gitarren. Die so genannte Evidence $p(\mathbf{a})$ gibt an, mit welcher Wahrscheinlichkeitsdichte die Merk-

malsvektoren **a** generell, d.h. unabhängig von einer Klassenzugehörigkeit, verteilt sind.

Die eigentliche Klassifikationsentscheidung für eine Klasse ω_d wird bei der Bayes'schen Klassifikaiton über die maximale a-posteriori Wahrscheinlichkeit $P(\omega_d|\mathbf{a})$ getroffen, weshalb die Bayes'sche Klassifikation oft auch Maximum A-Posteriori (MAP) Klassifikation genannt wird. Likelihood, Prior und Evidence lassen sich innerhalb eines überwachten Lernverfahrens direkt beobachten und somit abschätzen. Da ein Merkmalsvektor nur genau einer Klasse zugeordnet werden kann, bilden die Klassenzuordnungen sich gegenseitig ausschließende Ereignisse. Somit kann die Evidence zusätzlich über die Summe der totalen Wahrscheinlichkeit auch direkt aus der Likelihood und den Priors der D unterschiedlichen Klassen berechnet werden:

$$p(\mathbf{a}) = \sum_{d=1}^{D} p(\mathbf{a}|\omega_d) P(\omega_d). \tag{5.7}$$

Da die Evidence für alle Klassen gleich ist, stellt sie für die Klassifikationsentscheidung lediglich einen Skalierungsfaktor dar, weshalb sie für praktische Anwendungen fallen gelassen werden kann, um Rechenkapazität zu sparen. Weiterhin treten in vielen Klassifikationsszenarien die einzelnen Klassen mit gleicher a-priori Wahrscheinlichkeit auf, so dass die Klassifikationsentscheidung ausschließlich vom Wert der Likelihood abhängt. In diesem Zusammenhang spricht man neben dem MAP auch vom Maximum Likelihood Criterion.

Sind die realen Werte von Likelihood, Prior und Evidence bekannt, so handelt es sich bei der Bayes'schen Klassifikation um eine optimale Klassifikation mit der bestmöglichen Klassifikationsrate, weshalb sie in vielen Szenarien als Benchmark für die Klassifikationsraten anderer Klassifikationsverfahren verwendet wird.

5.4.2 GAUSSIAN MIXTURE MODELS

Das Gaussian Mixture Model (GMM) ist ein Wahrscheinlichkeitsmodell, das eine Clusterbildung im Merkmalsraum sowie die Modellierung von komplexen Verteilungsdichtefunktionen $p(\mathbf{a})$ erlaubt [Alle87, DuHS01, Nabn03, Paal04].

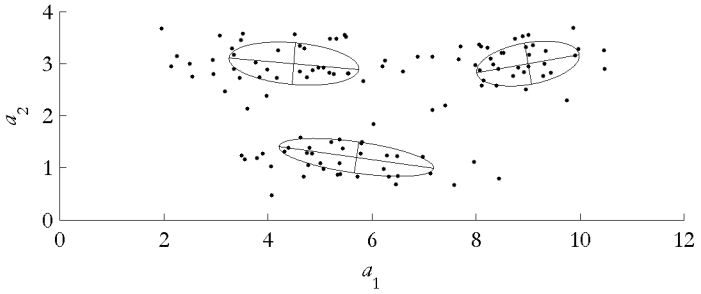


Bild 5.7: Verteilungsdichtemodellierung im zweidimensionalen Merkmalsraum durch ein GMM mit drei Komponenten, Standardabweichung der einzelnen Komponenten mit ihren Hauptachsen (Ellipsen)

Ein GMM teilt den Merkmalsraum in K Cluster ein, die auch Mixturkomponenten genannt werden. Hierbei wird jedes Cluster durch einen Mixturkoeffizienten $P(k)$ und eine Likelihood $p(\mathbf{a}|k)$ beschrieben. Als Summe ergeben die einzelnen Komponenten die zu modellierende Verteilungsdichtefunktion (Bild 5.7):

$$p(\mathbf{a}) = \sum_{k=1}^{K} P(k)\, p(\mathbf{a}|k). \tag{5.8}$$

Die Likelihood entspricht einer multidimensionalen Gaußverteilung und wird durch den Mittelwertvektor $\mathbf{c}_k$ und die Kovarianzmatrix $\mathbf{C}_k$ beschrieben:

$$p(\mathbf{a}|k) \sim N(\mathbf{c}_k, \mathbf{C}_k). \tag{5.9}$$

Als Verteilungsdichtefunktion muss Gleichung 5.8 überall nicht-negativ sein und sich über den gesamten Merkmalsraum zu 1 integrieren lassen. Da die Gaußverteilungen $p(\mathbf{a}|k)$ der einzelnen Cluster nicht-negativ sind und sich zu 1 integrieren lassen:

$$\int p(\mathbf{a}|k)\, d\mathbf{a} = 1, \tag{5.10}$$

wird diese Bedingung erfüllt durch die Einschränkung der Mixturkoeffizienten $P(k)$ zu:

$$\sum_{k=1}^{K} P(k) = 1 \text{ und } 0 \le P(k) \le 1. \tag{5.11}$$

Das Training eines GMMs besteht aus der Bestimmung der optimalen Modellparameter für eine gegebene Merkmalsmatrix bei einer festgelegten Anzahl der Cluster *K*. Die Modellparameter bestehen für jede Komponente k aus dem Mixturkoeffizienten $P(k)$, dem Mittelwertvektor $\mathbf{c}_k$ und der Kovarianzmatrix $\mathbf{C}_k$. Gute Modellparameter maximieren die Verbundverteilungsdichte χ, die das sequentielle Auftreten aller *C* in der Merkmalsmatrix enthaltenen Merkmalsvektoren $\mathbf{a}_c$ beschreibt:

$$\chi = \prod_{c=1}^{C} p(\mathbf{a}_c). \tag{5.12}$$

Für das Training hat sich der Expectation Maximisation (EM) Algorithmus bewährt, der die vorgegebenen Modellparameter iterativ optimiert, so dass die Verbundverteilungsdichte χ maximiert wird [Moon96, Nabn03]. Er wird durch die folgende Iteration beschrieben:

1. Die Modellparameter ($P(k)$, $\mathbf{c}_k$, $\mathbf{C}_k$) jeder Komponente k werden mit beliebigen Anfangswerten initialisiert.
2. Für jedes Cluster k wird für alle *C* Trainingsvektoren $\mathbf{a}_c$ jeweils die a-posteriori Wahrscheinlichkeit $P(k \mid \mathbf{a}_c)$ berechnet, die angibt, mit welcher Wahrscheinlichkeit der Vektor $\mathbf{a}_c$ zum Cluster k gehört.
3. Die neuen Modellparameter ($P^1(k)$, $\mathbf{c}^1{}_k$, $\mathbf{C}^1{}_k$) jeder Komponente werden über die berechnete a-posteriori Wahrscheinlichkeit bestimmt und gespeichert:

$$P^1_k = \frac{1}{C}\sum_{c=1}^{C} P(k|\mathbf{a}_c), \tag{5.13}$$

$$\mathbf{c}^1_k = \frac{\sum_{n=1}^{N} P(k|\mathbf{a}_c)\mathbf{x}_n}{\sum_{n=1}^{N} P(k|\mathbf{a}_c)}, \tag{5.14}$$

$$\mathbf{C}^1_k = \frac{\sum_{c=1}^{C} P(k|\mathbf{a}_c)\mathbf{a}_c\mathbf{a}_c^T}{\sum_{c=1}^{C} P(k|\mathbf{a}_c)} - \mathbf{a}^1_c\mathbf{a}^{1\,T}_c. \tag{5.15}$$

4. Die Verbundverteilungsdichte χ des neu berechneten Modells wird ermittelt. Ist der auf den letzten Iterationsdurchlauf bezogene Zuwachs der Verbundverteilungsdichte kleiner als ein vorgegebener Grenzwert oder wurde eine vorgegebene Anzahl von Iterationsdurchläufen erreicht, bricht der Algorithmus ab. Ansonsten wird mit den Modellparameter bei Schritt 2 fortgefahren.

Für die Berechnung der Verbundverteilungsdichte in Schritt 4 wird in der Praxis die logarithmische Verbundverteilungsdichte $\chi_{\log}$ verwendet:

$$\chi_{\log} = \log\chi = \sum_{c=1}^{C} p(\mathbf{a}_c), \tag{5.16}$$

da hier die rechenintensiveren Multiplikationen durch Additionen ersetzt werden können. Auf die Optimierung des Modells hat dies jedoch keinen Einfluss.

Ähnlich wie beim Training eines Vektorquantisierers kann bei den berechneten Modellparametern nicht zwischen lokalen und globalen Maxima unterschieden werden. Auch hier hängen das Laufzeitverhalten und das Ergebnis stark von der Wahl guter Initialisierungsparameter ab. In der Praxis wird deshalb vor dem eigentlichen Training des GMMs mit den Trainingsdaten ein Vektorquantisierer trainiert. Für die Initialisie-

rungswerte der Mittelwertvektoren $\mathbf{c}_k$ werden dann die Codebuchvektoren des Vektorquantisierers verwendet.

Wenn bekannt ist, dass die zum Training verwendete Merkmalsmatrix unkorrelierte Werte enthält, da sie beispielsweise im Vorfeld über eine PCA dekorreliert wurde, so müssen die Kovarianzmatrizen nicht vollständig trainiert werden, da alle Werte außerhalb der Hauptdiagonale null oder vernachlässigbar klein sind. Dies führt zu einer erheblichen Reduktion des Rechenaufwandes.

GMMs lassen sich, wie eingangs beschrieben, zur Clusterbildung im Merkmalsraum sowie zur Modellierung von komplexen Verteilungsdichtefunktionen $p(\mathbf{a})$ benutzen. Wird ein GMM zur reinen Clusterbildung verwendet, so kann ein einzelnes Modell in einem unüberwachten Lernverfahren trainiert werden. Für den Klassifikationsbetrieb entspricht jedes Cluster einer eigenen Klasse. Die Klassenzuordnung eines Merkmalsvektors $\mathbf{a}$ zu einem Cluster k erfolgt hierbei aus den Modellparametern entsprechend der Bayes'schen Klassifikation über die maximale a-posteriori Wahrscheinlichkeit:

$$P(k \mid \mathbf{a}) = \frac{p(\mathbf{a} \mid k) P(k)}{p(\mathbf{a})}, \tag{5.17}$$

die sich zusammen mit Formel 5.8 umformen lässt zu:

$$P(k \mid \mathbf{a}) = \frac{p(\mathbf{a} \mid k) P(k)}{\sum_{d=1}^{K} p(\mathbf{a} \mid d) P(d)}. \tag{5.18}$$

Werden GMMs zur Modellierung von komplexen Verteilungsdichtefunktionen verwendet, so lassen sich Klassifikationen mit einem höherem Detailgrad durchführen. Hierbei wird für jede Klasse ω ein eigenes GMM trainiert, das die komplexe Verteilungsdichte $p(\mathbf{a} \mid \omega)$ der Klasse modelliert. Dies geschieht in einem überwachten Lernverfahren, in dem jedes zu trainierende GMM mit einer eigenen Merkmalsmatrix trainiert wird. Die Abfolge der einzelnen Merkmalsvektoren ist hierbei für das Trainingsergebnis irrelevant. Aus dem Verhältnis der Anzahl der in den jeweiligen Matrizen enthaltenen Merkmalsvektoren zu der Gesamtanzahl

aller Merkmalsvektoren lässt sich weiterhin für jede Klasse ω die a-priori Auftretenswahrscheinlichkeit $P(\omega)$ als relative Häufigkeit schätzen.

Im Klassifikationsbetrieb erfolgt auch hier die Klassenzuordnung eines Merkmalsvektors **a** zu einer Klasse ω entsprechend der Bayes'schen Klassifikation über die maximale a-posteriori Wahrscheinlichkeit:

$$P(\omega|\mathbf{a}) = \frac{p(\mathbf{a}|\omega)P(\omega)}{p(\mathbf{a})}, \tag{5.19}$$

die sich zusammen mit der Summe der totalen Wahrscheinlichkeit umformen lässt zu:

$$P(\omega|\mathbf{a}) = \frac{p(\mathbf{a}|\omega)P(\omega)}{\sum_{d=1}^{D} p(\mathbf{a}|\omega_d)P(\omega_d)}. \tag{5.20}$$

Die Klassifikationsergebnisse von GMMs hängen stark von der Anzahl ihrer Komponenten ab. Die Wahl von zu vielen Komponenten kann hierbei leicht zu einer Überanpassung führen, so dass sich das Modell nicht mehr ausreichend generalisieren lässt, zu wenige Komponenten hingegen führen zu einer ungenauen Modellierung. Als sinnvoll für die meisten Klassifikationsaufgaben haben sich drei bis fünf Komponenten erwiesen. Um eine Überanpassung zu verhindern, wurde weiterhin von Figueiredo und Jain ein erweitertes Trainingsverfahren entwickelt, das mit einer relativ großen Anzahl von Komponenten (>15) beginnt und im Laufe einer Iteration durch eine permanente leichte Variation der Merkmalsvektoren überprüft, ob eine Überanpassung droht. Ist dies der Fall, so wird das Modell um eine der Komponenten reduziert und weitertrainiert. Der Figueiredo-Jain-Algorithmus liefert in vielen Fällen ein zufriedenstellendes Trainingsergebnis, konvergiert jedoch nicht immer und ist somit nicht allgemeingültig [FiJa02].

GMMs berücksichtigen nicht die zeitliche Abfolge der auftretenden Merkmalsvektoren und lassen sich mit Erfolg anwenden, wenn die Information über den zu klassifizierenden Prozess hauptsächlich in der globalen Verteilungsdichte und nicht im zeitlichen Verlauf enthalten ist. Dies ist entweder dann der Fall, wenn der zu modellierende Prozess in seinem inneren Zustand konstant ist, oder wenn der hypothetische inne-

re Zustand von extrem vielen Faktoren mit fließenden Übergängen geprägt ist und somit nur noch in seiner globalen Statistik beschrieben werden kann.

Beispiele für Szenarien mit (relativ) konstantem inneren Zustand sind die Sprechererkennung und die Klassifikation von Holzblasinstrumenten. Werden bei der Sprechererkennung für die Darstellung der Sprache spektrale Merkmale verwendet, so lassen sich die charakteristischen Eigenschaften des Rachenraumes eines Sprechers durch eine statische Verteilungsdichte gut abbilden. Hier haben sich GMMs zum Standard entwickelt [ReRo95, Camp97]. Bei der Klassifikation von Holzblasinstrumenten kommt zum Tragen, dass der produzierte Klang bis auf die Lautstärke kaum seinen Charakter ändert und somit dauerhaft relativ ähnliche Merkmale produziert [Brow97, BrHM01].

Beispiele für Szenarien, deren Beschreibung zu viele innere Zustände verlangen würde, sind die Musikgenreerkennung oder Gesangsdetektion. In beiden Bereichen wurden GMMs erfolgreich verwendet [BuLe03, BuLe04]. Für das in dieser Arbeit implementierte Polyphone Klassifikationssystem wurden ebenfalls GMMs verwendet, da es sich auch bei der Klassifikation innerhalb von komplexen, polyphonen Musiksignalen um einen Prozess handelt, der nur noch in seiner globalen Statistik beschrieben werden kann (vgl. Abschnitt 6.4).

Obwohl das Training von GMMs relativ rechenintensiv werden kann, sind sie im Klassifikationsbetrieb sehr effizient, was sie für viele Echtzeitaufgaben interessant macht. Ein weiterer Vorteil ist, dass GMMs auch dann verwendet werden können, wenn einzelne Komponenten des zu klassifizierenden Merkmalsvektors nicht oder zu ungenau beobachtet werden können (engl. Missing Features). Hierfür wird die entsprechende Dimension ausgeblendet, und die Klassifikationsentscheidung findet anhand der restlichen verfügbaren Merkmale statt [EgBr03a, EgBr03b]. Aufgrund der positiven Eigenschaften von GMMs wurden sie im Zusammenhang mit der Musikinstrumentenidentifikation und Klassifikation häufig mit guten Ergebnissen verwendet [Brow97, MaMo99, BrHM01, LiPR03].

5.4.3 HIDDEN MARKOV MODELS

Das Hidden Markov Model (HMM) ist ein Wahrscheinlichkeitsmodell, das dem GMM sehr ähnlich ist und ebenfalls die Modellierung von komplexen Verteilungsdichtefunktionen $p(\mathbf{a})$ erlaubt [Rabi89, DuHS01]. Im Gegensatz zum GMM hängt die Form der Verteilungsdichtefunktionen allerdings von einem inneren Zustand s des Modells ab. Das Zustandsdiagramm eines HMM entspricht einer Markov-Kette mit S Zuständen, die prinzipiell beliebig ineinander übergehen können (Bild 5.8). Markov-Ketten sind gedächtnisfrei, d.h. dass jeder Zustandswechsel lediglich vom aktuellen Zustand abhängt, nicht aber von der vorhergehenden Abfolge der Zustände.

Die Wahrscheinlichkeit, dass die Markov-Kette vom Zustand s_i in den Zustand s_j wechselt, wird als Übergangswahrscheinlichkeit a_{ij}, bezeichnet. Alle Übergangswahrscheinlichkeiten lassen sich in der Übergangsmatrix $\mathbf{A}$ zusammenfassen, wobei das Element a_{ij} in der i-ten Zeile und j-ten Spalte die Übergangswahrscheinlichkeit des Zustands s_i in den Zustand s_j beschreibt.

Neben der Übergangswahrscheinlichkeit wird für jeden Zustand s_i eine Anfangswahrscheinlichkeit π_i definiert, die angibt, mit welcher Wahrscheinlichkeit das Modell sich nach einer Initialisierung in diesem Zustand befindet.

Der innere Aufbau des HMMs in Bezug auf seine Markov-Kette und der Zustand, in dem es sich befindet, sind für einen Beobachter verborgen (engl. hidden) und können lediglich über das Verhalten des Modells abgeschätzt werden. Wird das HMM als generatives Modell betrachtet, bedeutet dies, dass der Aufbau der Markov-Kette durch die von ihm emittierten Merkmalsvektoren als Observationsdaten abgeschätzt werden muss.

Da die von dem HMM modellierte Verteilungsdichte $p(\mathbf{a})$ direkt und ausschließlich vom inneren Zustand s_i des Modells abhängt, kann jedem Zustand eine eigene, im zeitlichen Ablauf statische Zustandsverteilungsdichte $p(\mathbf{a}|i)$ zugeordnet werden. Die globale Verteilungsdichte $p(\mathbf{a})$ nimmt dann jeweils die Form der Zustandsverteilungsdichte $p(\mathbf{a}|i)$ des jeweiligen Zustands s_i an.

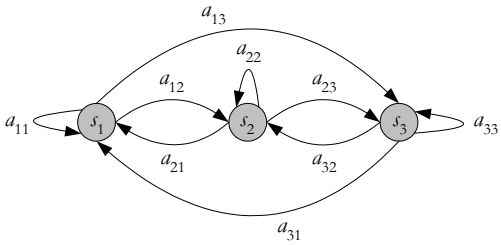


Bild 5.8: Markov-Kette mit drei Zuständen

Im einfachsten Fall entspricht die Zustandsverteilungsdichte $p(\mathbf{a}|i)$ jedes Zustands einer multidimensionalen Gaußverteilung, deren Mittelwertvektor $\mathbf{c}_i$ und Kovarianzmatrix $\mathbf{C}_i$ direkt vom Zustand s_i abhängig sind:

$$p(\mathbf{x}|i) \sim N(\mathbf{c}_i, \mathbf{C}_i). \tag{5.21}$$

In diesem Fall hat das HMM eine große Ähnlichkeit zum GMM, da der Merkmalsraum in gaußförmige Cluster aufgeteilt wird, wobei jeder Zustand einem Cluster entspricht. Im Unterschied zum GMM gibt es jedoch nicht für jedes Cluster eine im Ablauf statische a-priori Wahrscheinlichkeit P_i, mit der ein zu klassifizierender Merkmalsvektor dem Cluster i zugeordnet werden kann, sondern es gibt zu jedem Zeitpunkt nur die dem aktuellen Zustand s_i entsprechende Zustandsverteilungsdichte $p(\mathbf{a}|i)$, der die globale Verteilungsdichte $p(\mathbf{a})$ entspricht. Die anderen Cluster werden somit quasi ausgeblendet. Ein HMM lässt sich in diesem einfachsten Fall auch als GMM verstehen, bei dem die a-priori Wahrscheinlichkeit P_i der einzelnen Cluster binär vom inneren Zustand s_i des Modells abhängt. Dies bedeutet, dass zu jedem Zeitpunkt die a-priori Wahrscheinlichkeit P_i des dem Zustandes s_i entsprechenden Clusters 1 ist und alle anderen a-priori Wahrscheinlichkeiten 0 sind.

Obwohl die Modellierung der Zustandsverteilungsdichten $p(\mathbf{a}|i)$ der einzelnen Zustände mit einfachen Gaußverteilungen häufig und mit guten Ergebnissen in Klassifikationsszenarien verwendet wird, sind HMMs nicht darauf beschränkt. Prinzipiell können die einzelnen Zustandsverteilungsdichten $p(\mathbf{a}|i)$ beliebig komplex sein, so werden sie in der Praxis häufig selbst wieder durch GMMs modelliert.

Das Training eines HMMs geschieht über das Baum-Welch-Verfahren, das eine spezielle Form des EM-Algorithmus darstellt [BaEa67, Rabi89, Blim97].

Für den Einsatz von HMMs in Klassifikationsszenarien ergeben sich prinzipiell zwei Verwendungsarten. Ist im Vorfeld bekannt, dass die zu klassifizierende Merkmalsmatrix Merkmalsvektoren von unterschiedlichen Klassen enthält, kann ein einzelnes HMM verwendet werden, das jede Klasse durch einen oder mehrere Zustände s_i repräsentiert und die Klassifikationsentscheidung direkt aufgrund seines inneren Zustands fällt. Das Training eines derartigen Modells kann theoretisch in einem unüberwachten Lernverfahren geschehen, in der Praxis erfolgt das Training aufgrund der besseren Ergebnisse allerdings in einem überwachten Lernverfahren. Hierbei wird die zum Training verwendete Merkmalsmatrix um den Klassenindex der jeweiligen Klasse ω erweitert.

Ist im Vorfeld der Klassifikationsaufgabe bekannt, dass die zu klassifizierenden Merkmalsvektoren jeweils nur von einer Klasse stammen, so kann in einem überwachten Lernverfahren für jede Klasse ω ein eigenes HMM trainiert werden. Hierzu wird für jedes zu trainierende HMM eine eigene Merkmalsmatrix zusammengestellt, bei der die Abfolge der Merkmalsvektoren relevant ist. Die für ein derartiges Szenario trainierten Modelle können deutlich einfacher ausfallen als ein allgemeines Modell, das alle Klassen abbilden kann.

Die Anzahl der Zustände eines zu trainierenden HMMs und die Parameter der Zustandsverteilungsdichten müssen vor dem Training festgelegt werden. Ähnlich wie beim GMM kann die Wahl von zu vielen Zuständen und zu komplexen Zustandsverteilungsdichten zu einer Überanpassung führen, so dass sich das Modell nicht mehr ausreichend generalisieren lässt. Bei der Verwendung von einfachen Gaußverteilungen als Zustandsverteilungsdichten haben sich 3 bis 5 Zustände für die meisten Klassifikationsaufgaben als sinnvoll erwiesen [CaCh99, HePD02].

Für einen Klassifikationsvorgang ist es bei bekannten Modellparametern theoretisch möglich, die Wahrscheinlichkeit für das Auftreten einer gegebenen Folge von Merkmalsvektoren zu berechnen. Hierfür müssten jedoch alle theoretisch möglichen Zustandsketten erstellt werden und die zugehörige Verbundwahrscheinlichkeit für die Emittierung der beobach-

teten Merkmalsvektoren berechnet werden. Diese Verbundwahrscheinlichkeit ist jedoch selbst für kurze Folgen von Merkmalsvektoren praktisch nicht berechenbar, da zu viele Rechenoperationen durchzuführen wären. So wären beispielsweise bei $S = 5$ Zuständen und $C = 100$ auftretenden Merkmalsvektoren ca. $2C \cdot S^C \approx 10^{72}$ Rechenoperationen durchzuführen. Für den Klassifikationsvorgang wird deshalb der Viterbi-Algorithmus verwendet [Vite67], der für ein HMM mit bekannten Parametern die Auftretenswahrscheinlichkeit einer bestimmten Folge von Merkmalsvektoren sowie die zugehörige Zustandsfolge berechnet. Über diese Informationen lassen sich jedem Merkmalsvektor bzw. jeder Merkmalsmatrix eine eindeutige Klasse zuordnen.

Da HMMs im Gegensatz zu GMMs die zeitliche Abfolge der auftretenden Merkmalsvektoren berücksichtigen, lassen sie sich mit Erfolg anwenden, wenn die Information über den zu klassifizierenden Prozess zu großen Anteilen im zeitlichen Verlauf enthalten ist. Dies ist z.B. bei der Spracherkennung der Fall, da Sprache durch eine eindeutige Phonemabfolge geprägt ist. Hier haben sich HMMs zum Standard entwickelt [Alle87, Rabi89, Haue93, Wolf97, Rott00, Sigm03]. Ähnlich verhält es sich bei der Klassifikation von Geräuschen und monophonen Musikinstrumentenklängen, insbesondere, wenn diese durch spektrale Merkmale repräsentiert werden. Hier lässt sich die immer wiederkehrende Abfolge von Einschwingphase, Haltephase und Ausklingphase nutzen. In diesem Zusammenhang wurden HMMs sehr erfolgreich für die allgemeine Geräuschklassifikation von einzelnen Klangereignissen verwendet [Case01a, Case01b, KiBS04]. So bildet ein HMM auch den Kern des in dieser Arbeit implementierten monophonen Echtzeitsystems (vgl. Abschnitt 6.3).

Für eine direkte Identifikation von Musikinstrumentenklängen in einem polyphonen Szenario sind HMMs nicht geeignet, da sie hierfür massiv parallel und sehr komplex ausgelegt werden müssen. Derartige Modelle benötigen übermäßig große Mengen an Trainingsdaten und neigen aufgrund der Komplexität zur Überanpassung.

Im Vergleich zu GMMs erzeugen HMMs sowohl im Training durch den Baum-Welch-Algorithmus als auch im Klassifikationsbetrieb durch den Viterbi-Algorithmus einen höheren Rechenaufwand. Dies trifft vor allem auf komplexe HMMs zu, die aufgrund ihres Aufbaus bessere Klassifikationsergebnisse liefern können als GMMs. Ist die Information über den

zu klassifizierenden Prozess zu großen Anteilen im zeitlichen Verlauf enthalten, lohnt sich der Mehraufwand, da bessere Klassifikationsraten erzielt werden können.

Wird ein HMM, dessen Zustandsverteilungsdichten einfache Gaußverteilungen sind, hingegen mit Daten trainiert, die keine zeitliche Struktur haben, so wird das fertig trainierte Modell eine Übergangsmatrix haben, bei der für jeden Zustand die Übergangswahrscheinlichkeiten in die anderen Zustände statisch sind. Die Spalten der Übergangsmatrix enthalten hierbei jeweils konstante Werte und es gilt:

$$a_{ij} = a_{fj} = a_j \text{ für alle } i,f. \tag{5.22}$$

Diese degenerierten Übergangswahrscheinlichkeiten a_j können als Mixturkoeffizienten $P(j)$ eines GMMs interpretiert werden und das HMM geht in das GMM über:

$$a_j = P(j). \tag{5.23}$$

Somit kann ein HMM, das mit Daten ohne zeitliche Struktur trainiert wurde, mindestens die Klassifikationsraten eines GMMs liefern, dessen Mixturkoeffizienten den Übergangswahrscheinlichkeiten des HMM entsprechen. Allerdings sollte das HMM nicht generell als ein besseres GMM verstanden werden, da mit der Komplexität des Modells auch die Menge der benötigten Trainingsdaten wächst und komplexere Modelle eher zur Überanpassung neigen. Weiterhin sind die Klassifikationsergebnisse von HMMs häufig nicht signifikant besser als die von GMMs, so dass der Mehraufwand oftmals nicht gerechtfertigt ist .

5.5 NEURONALE NETZE

Eine der größten Intelligenzleistungen von Lebewesen ist die fehlertolerante Mustererkennung und Klassifikation. Mit künstlichen Neuronalen Netzen (engl. Artificial Neural Nets) versucht man, den Teil des Gehirns zu imitieren, der für die Verarbeitung und Bewertung von Sinneseindrücken verwendet wird und somit für die Mustererkennung und Klassifikation relevant ist [DuHS01, Nabn03].

In den folgenden Abschnitten werden Neuronale Netze sowohl von der biologischen als auch von der theoretischen Seite einleitend erläutert sowie leistungsfähige künstliche Neuronale Netze im Zusammenhang mit der Musikinstrumentenidentifikation und Klassifikation beschrieben.

5.5.1 BIOLOGISCHE NEURONALE NETZE

Biologische Neuronale Netze wie das menschliche Gehirn bestehen aus einer Vielzahl von vernetzten Gehirnzellen, den so genannten Neuronen. So enthält das menschliche Gehirn ca. 100 Milliarden Neuronen, zwischen denen Informationen über elektrische Impulse ausgetauscht werden. Ein Neuron empfängt hierbei Impulse über die Dendriten, die quasi seine Eingänge darstellen. Wird der Zellkörper eines Neurons, das so genannte Soma, ausreichend durch die über die Dendriten aufgenommenen Impulse stimuliert, generiert er einen eigenen elektrischen Impuls. Hierbei gilt, je stärker die Eingangsstimulation ist, desto mehr eigene Impulse generiert das Neuron. Diese werden durch das Axon, das quasi den Ausgang des Neurons darstellt, zu den Dendriten anderer Neuronen transportiert. Die Verknüpfung zwischen Axon und Dendrit wird hierbei Synapse genannt (Bild 5.9).

Neben diesen einfachen Neuronen gibt es spezialisierte Zellen wie beispielsweise Sinneszellen und Motoneuronen, die Schnittstellen zum Körper darstellen. Sinneszellen stellen Eingangsschnittstellen zum Neuronalen Netz dar und beziehen ihre Information von äußeren Vorgängen. Für optische Sinneszellen ist dies die Stimulation durch Lichtwellen, bei akustischen Sinneszellen hingegen eine durch Schallwellen ausgelöste feine Resonanzschwingung. Motoneuronen stellen Ausgangsschnittstellen des Neuronalen Netzes zur Muskulatur dar und bewirken hier bei Stimulation das Entspannen oder Kontrahieren eines Muskels. Durch diesen Aufbau des Gehirns ist der Mensch in der Lage, äußere Eindrücke über die Sinneszellen aufzunehmen, im Gehirn entsprechend zu verarbeiten und körperlich darauf zu reagieren. Dies reicht von einfachen Reflexen bis hin zum Bewusstsein [Cofe08].

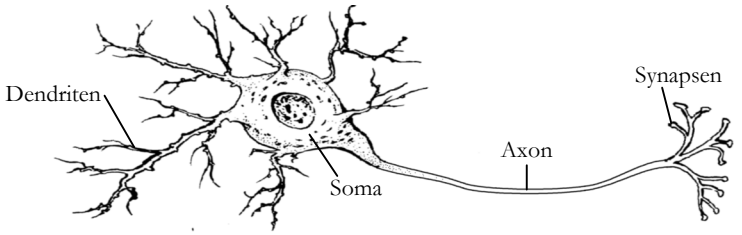


Bild 5.9: Biologisches Neuron eines Wirbeltieres [Cofe08]

5.5.2 KÜNSTLICHE NEURONALE NETZE

Für die Modellierung eines biologischen Neurons durch ein künstliches wird das Generieren der elektrischen Impulse dahingehend vereinfacht, dass es durch einen einfachen Zahlenwert repräsentiert wird, der die mittlere Impulsrate darstellt. Die Dendriten des Neurons werden durch eine beliebige Anzahl M von Eingängen e_m modelliert, deren Werte jeweils mit einem Gewicht g_m gewichtet werden. Diese Gewichte modellieren die Stärke der Verbindung zum Zellkörper in einem biologischen Neuron. Die dem biologischen Stimulus innerhalb des Zellkerns entsprechende Größe s ergibt sich als Summe der gewichteten Eingänge (Bild 5.10):

$$s = \sum_m g_m e_m \,. \tag{5.24}$$

Der Ausgangswert des Neurons a ergibt sich, indem der Stimulus über eine Schwellenfunktion f ausgewertet wird (Bild 5.11). Hierbei kann als harte Entscheidungsschwelle eine Variante der Sprungfunktion verwendet werden, die bei Überschreitung eines bestimmten Stimuluswerts c auf einen definierten Ausgangswert schaltet. Um eine Ungewissheit in der Entscheidung des Neurons auszudrücken und um effiziente Lernalgorithmen wie das Back-Propagation-Verfahren zu ermöglichen (vgl. Abschnitt 5.5.3), wird jedoch häufig als weiche Entscheidungsschwelle eine Sigmoid-Funktion verwendet:

$$a = f(s) = \frac{1}{1 + \exp(-s + c)} \,. \tag{5.25}$$

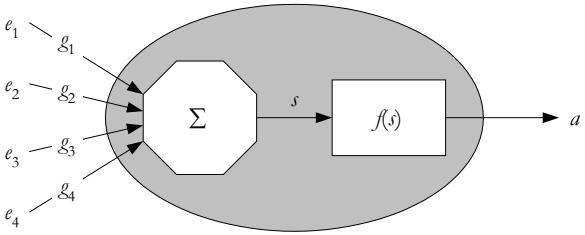


Bild 5.10: Schematischer Aufbau eines künstlichen Neurons

Somit stellt ein künstliches Neuron einen sehr einfachen mathematischen Prozessor dar, der lediglich die beschränkte Aufgabe der gewichteten Summenbildung mit anschließendem Schwellenwertvergleich ausführen kann. Werden viele Neuronen allerdings zu einem künstlichen Neuronalen Netz verknüpft, verhalten sie sich als eine künstliche Intelligenz, mit der bei entsprechendem Training beispielsweise die komplexe Aufgabe der Mustererkennung und Klassifikation erfolgreich bewältigt werden kann. Die gelernte künstliche Intelligenz wird hierbei durch die Verbindungen der Neuronen untereinander und die jeweiligen Eingangsgewichte der einzelnen Neuronen repräsentiert [DuHS01, Nabn03, Cofe08].

Für die Mustererkennung werden die Werte des Merkmalsvektors an die Eingänge des Netzes angelegt. Hierbei muss für jeden Wert eines Merkmalsvektors mindestens ein Eingangsneuron vorhanden sein. Es kann allerdings auch von Vorteil sein, für jede Komponente des Merkmalsvektors mehrere Eingangsneuronen zu verwenden, die den Wertebereich des jeweiligen Merkmals unter sich aufteilen. Dies ist insbesondere bei diskreten Merkmalen sinnvoll. Die Klassifikationsentscheidung kann an den Ausgängen abgelesen werden. Hierbei ist es sinnvoll, für jede Klasse einen eigenen Ausgang zur Verfügung zu stellen, dessen Wert ähnlich einer a-posteriori Klassenwahrscheinlichkeit gedeutet werden kann [Trom95, Rott00].

Der Aufbau und die Anzahl der verwendeten Neuronen lassen sich nicht allgemeingültig angeben und hängen stark von der Klassifikationsaufgabe ab. Generell gilt nur, dass das Netz bei einer zu geringen Anzahl von Neuronen nicht in der Lage ist, alle Aspekte der Klassifikationsaufgabe zu erlernen, bei einer zu großen Anzahl von Neuronen hingegen zur Überanpassung neigt.

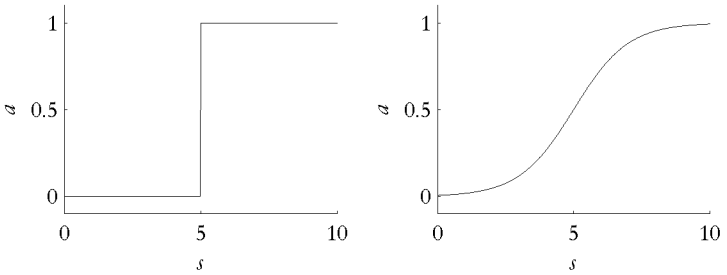


Bild 5.11: Verschiedene Schwellenfunktionen mit dem Schwellenwert $c = 5$, links einfache Sprungfunktion, rechts Sigmoidfunktion

5.5.3 VORWÄRTSGERICHTETE NETZE

Vorwärtsgerichtete Netze (engl. Feed Forward Networks) bestehen aus einer oder mehreren Schichten von Neuronen, wobei die Neuronen ihre Informationen jeweils nur an die Neuronen der folgenden Schicht weitergeben. Somit propagiert ein an die Eingangsneuronen angelegtes Muster unidirektional durch das Netz (Bild 5.12).

Der einfachste Netzaufbau mit nur einer Schicht wird auch Perzepton genannt. Mit ihm lassen sich bereits viele Klassifikationsprobleme lösen, wenn sie gut konditioniert (engl. well-conditioned) sind, was bedeutet, dass ähnliche Merkmalsvektoren zu ähnlichen Klassifikationsergebnissen führen müssen. Ferner müssen sich die verschiedenen Klassen im R-dimensionalen Merkmalsraum durch eine (R-1)-dimensionale Entscheidungsebene trennen lassen. Ein Beispiel für ein gut konditioniertes Klassifikationsproblem sind die logischen Funktionen AND oder OR. Ist das Klassifikationsproblem hingegen schlecht konditioniert (engl. ill-conditioned), wie beispielsweise die logische Funktion XOR, so lässt es sich mit einem einschichtigen Neuronalen Netz nicht mehr lösen. Um beliebige Klassifikationsprobleme lösen zu können, werden mindestens zweischichtige Neuronale Netze benötigt [DuHS01, Nabn03, Cofe08].

Für die Berechnung der Ausgabe eines vorwärtsgerichteten Netzes werden im sogenannten Forward-Pass die Eingangswerte an die entsprechenden Eingangsneuronen angelegt. Ausgehend von diesen Eingangsneuronen werden die Ausgangswerte der jeweiligen Neuronen für jede

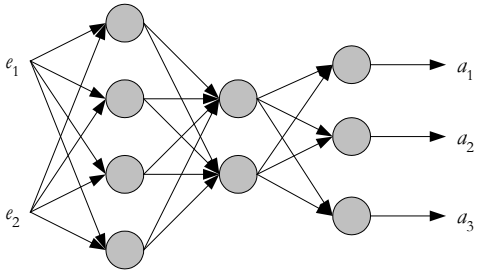


Bild 5.12: Vorwärtsgerichtetes Netz mit drei Schichten

Schicht jeweils aus den Eingangswerten der vorhergehenden Schicht berechnet.

Trainiert werden vorwärtsgerichtete Netze durch das überwachte Lernverfahren der Back-Propagation. Hierbei werden alle zu trainierenden Merkmalsvektoren zusammen mit den korrekten Ergebniswerten zyklisch dem zu trainierenden Netz zugeführt. Das eigentliche Lernen erfolgt über die Anpassung der Eingangsgewichte jedes Neurons.

Zur Initialisierung eines untrainierten Netzes werden die Eingangsgewichte mit zufälligen Werten belegt. Die Werte der Trainingsvektoren werden jeweils an die Eingänge des Netzes angelegt und die zugehörigen Ausgangswerte über den Forward-Pass berechnet. Nun wird rückwärts, beginnend mit den Ausgangsneuronen, für jedes Neuron der vorhandene Ausgangswert mit dem erwarteten Ergebniswert verglichen. Das eigentliche Lernen erfolgt über die Anpassung der Eingangsgewichte. Hierbei gelten die folgenden intuitiven Regeln:

- Ist ein Ausgangswert groß, obwohl er klein sein sollte, werden die Eingangsgewichte, deren Eingangswerte groß sind, reduziert.
- Ist ein Ausgangswert klein, obwohl er groß sein sollte, werden die Eingangsgewichte, deren Eingangswerte klein sind, vergrößert.
- Ist ein Ausgangswert korrekt, werden die zugehörigen Eingangsgewichte nicht verändert.

Diese Art der Rückwärtsberechnung wird Reverse Pass genannt. Der genaue Wert der Veränderung lässt sich aus der Differenz zwischen dem Soll- und Istwert des Ausgangswerts ermitteln und ist abhängig vom

konkreten Lernverfahren. Als Abbruchkriterium wird der mittlere Klassifikationsfehler von allen zum Training verwendeten Merkmalsvektoren verwendet. Das Lernverfahren konvergiert grundsätzlich immer, es kann jedoch vorkommen, dass sich für den mittleren Klassifikationsfehler ein suboptimales lokales Minimum ergibt.

Besonders große Netze neigen in der Praxis zur Überanpassung, was konkret bedeutet, dass das Netz nicht die Aspekte der Klassifikationsaufgabe erlernt, sondern die Trainingsdaten abspeichert. Um dies zu verhindern, sollten beim Training Verfahren wie die Kreuzvalidierung verwendet werden. Neuere Algorithmen zum Training von vorwärtsgerichteten Netzen reduzieren das Risiko der Überanpassung, indem während des Trainings die Anzahl der Neuronen so lange reduziert wird, bis ein stabiles Verhalten des Netzes erreicht ist.

Vorwärtsgerichtete Netze werden für Aufgaben der Musikinstrumentenidentifikation und Klassifikation zunehmend populärer [HAB+00, KoCz01, HePD02, LiRo03]. Die besten Klassifikationsergebnisse ergeben sich hierbei, wenn die gewählten Merkmalsvektoren im Merkmalsraum deutlich unterscheidbare Wolken bilden. Die Wolken können aufgrund der Eigenschaften von Neuronalen Netzen beliebig komplex ausfallen, so lange sie in ihrer Form konstant bleiben. Dies ist beispielsweise bei spektralen und harmonischen Merkmalen oft der Fall.

5.5.4 WETTBEWERBSNETZE

Wettbewerbsnetze (engl. Winner Takes All Networks) sind vorwärtsgerichtete Netze, die nicht mittels Back-Propagation trainiert werden, sondern in einem unüberwachten Lernverfahren, das problemlos auch ein kontinuierliches Lernen erlaubt. Wettbewerbsnetze werden selten alleine benutzt, sie bilden jedoch die Hauptstütze einiger komplexer Netzwerke. Eine der Hauptanwendungen für Wettbewerbsnetze sind neben der Clusterbildung sich selbst organisierende Karten (engl. Self Organizing Maps, SOM), die selbstständig Muster und Gruppen in Daten erkennen und diese im zwei- oder dreidimensionalen Raum anordnen können. Dies ist auch ihre Hauptanwendung im Zusammenhang mit der Musikinstrumentenidentifikation und Klassifikation [HePD02, CrWe03].

In der Trainingsphase werden alle Eingangsgewichtungen des Netzes mit zufälligen Werten initialisiert. Für jeden Merkmalsvektor, der an das Netz angelegt wird, werden alle Verbindungen, die Ein- und Ausgänge mit großen Werten miteinander verbinden, vergrößert bzw. alle Verbindungen, die Ein- und Ausgänge mit kleinen Werten miteinander verbinden, verkleinert. Durch diese Vorgehensweise prägen sich im zunehmenden Trainingsverlauf für bestimmte Klassen von Merkmalsvektoren innerhalb der Ausgänge Gewinner heraus, die einen großen Ausgangswert zugewiesen bekommen, während alle anderen Ausgangswerte klein sind. Über die Ausgangswerte kann nun die Klassifikationsentscheidung erfolgen, indem die Ausgangswerte ähnlich einer a-posteriori Klassenwahrscheinlichkeit gedeutet werden.

5.5.5 HOPFIELD-NETZE

Hopfield-Netze können im Gegensatz zu vorwärtsgerichteten Netzen Merkmalsvektoren abspeichern, die als Clustervektoren jeweils ein Clusterzentrum im Merkmalsraum beschreiben. Hopfield-Netze haben genau so viele Eingänge wie Ausgänge und sind in der Lage, einen degenerierten, fehlerhaften Merkmalsvektor, der an den Eingängen anliegt, auf einen der gespeicherten Clustervektoren abzubilden und diesen auf die Ausgänge zu legen [Hopf84].

Im Gegensatz zu vorwärtsgerichteten Netzen enthalten Hopfield-Netze Rückkopplungen, wobei jeder Ausgang auf einen Eingang zurückgekoppelt ist. Diese Rückkopplungen ermöglichen das Speichern der Clustervektoren und verleihen dem Netz sein Gedächtnis. Das Netzwerk besitzt eine einzige Schicht, die für jedes Ein- und Ausgangspaar ein eigenes Neuron enthält (Bild 5.13).

Nach dem Anlegen eines Merkmalsvektors an den Eingängen schwingt sich ein Hopfield-Netz in einem Vorgang, der paralleles Einschwingen (engl. Parallel Relaxation) genannt wird, auf einen stabilen Zustand ein. Hierbei bildet sich an den Ausgängen der Clustervektor aus, der die größte Ähnlichkeit mit dem angelegten Merkmalsvektor hat.

Das Training eines Hopfield-Netzes kann aufgrund der Rückkopplungen nicht wie bei vorwärtsgerichteten Netzen mittels Back-Propagation erfolgen. Da Hopfield-Netze allerdings nur eine Neuronenschicht besitzen,

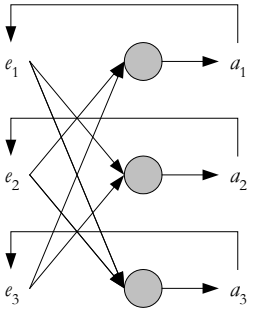


Bild 5.13: Hopfield-Netz mit drei Ein- und Ausgängen

können in einem überwachten Lernverfahren die Eingangsgewichte in geschlossener Form berechnet werden. Für Merkmalsvektoren, deren Werte diskret sind und nur -1 oder +1 sein können, ergibt sich das Gewicht $g_{m,n}$ zwischen dem m-ten Eingang und dem n-ten Ausgangsneuron aus der einfachen Summe:

$$g_{m,n} = \sum e_m e_n \tag{5.26}$$

über die Werte e aller Trainingsvektoren. Damit sich für die Eingangsgewichte während des Trainings keine Triviallösungen ergeben, werden die Gewichte für die direkten Verbindungen zwischen den Eingängen und dem dahinterliegenden Neuron grundsätzlich auf Null gesetzt:

$$g_{m,m} = 0\,, \tag{5.27}$$

was gleichbedeutend mit dem Auftrennen der Verbindung ist. Somit werden alle für das Training verwendeten Muster über alle Neuronen und die dazugehörigen Gewichtungen verteilt.

Hopfield-Netze werden im Zusammenhang mit der Musikinstrumentenidentifikation und Klassifikation hauptsächlich für eine Clusterbildung im Merkmalsraum als Vorverarbeitung für Verfahren wie GMMs oder HMMs verwendet. Ein weiteres Anwendungsgebiet ist die Rekonstruktion fehlerhafter oder unvollständiger Merkmalsvektoren (engl. Missing Features) [HePD02].

6 IMPLEMENTIERUNG

Für die vorliegende Arbeit wurden drei Klassifikationssysteme implementiert, mit denen sich unterschiedliche theoretische und technische Aspekte erforschen und demonstrieren lassen.

Das erste System ist ein monophones Klassifikationssystem, mit dem als Experimentiersystem eine Vielzahl von verschiedenen Merkmalsextraktionsverfahren und Klassifikationsverfahren sehr einfach gesondert und im Zusammenspiel untersucht werden kann.

Das zweite System ist ein monophones Klassifikationssystem, das als echtzeitfähiges VST-Plugin implementiert ist. Es verwendet Merkmalsextraktions- und Klassifikationsverfahren, die sehr gute Klassifikationsergebnisse erzielen, sich aber dennoch effizient in Echtzeit berechnen lassen. Weiterhin unterstützt das System den MPEG-7-Standard und dient somit auch als Technikstudie für den industriellen Gebrauch.

Das dritte System ist ein polyphones Klassifikationssystem, das eine Klassifikation und Identifikation von Musikinstrumentenklängen in komplexen Musiksignalen erlaubt. Hierbei werden für den Aufbau oder den Inhalt der Musiksignale oder der Klangereignisse keinerlei Einschränkungen gemacht. Das System verwendet einen neuartigen Algorithmus, der durch seine viel versprechenden Ergebnisse eine wertvolle Bezugsmarke für die weitere Erforschung des Gebietes liefert.

In diesem Kapitel werden die bisherigen Ansätze auf dem Gebiet der Klassifikation und Identifikation von Musikinstrumentenklängen zusammen mit den drei im Rahmen dieser Arbeit implementierten Systemen beschrieben. Eine im Detailgrad über dieses Kapitel hinausgehende

Beschreibung der einzelnen implementierten Module findet sich weiterhin in der jeweiligen digitalen Hilfe.

6.1 BISHERIGE ANSÄTZE

Die bisherigen in der Literatur vorgestellten Ansätze lassen sich in die drei bereits im Zusammenhang mit der Merkmalsextraktion in Abschnitt 4.1 vorgestellten Szenarien unterteilen. Hierbei handelt es sich um die monophone Klassifikation, die eingeschränkt polyphone Klassifikation und die komplexe polyphone Klassifikation. In den folgenden Abschnitten werden die verschiedenen Szenarien zusammen mit den für sie vorgestellten bisherigen Ansätzen beschrieben.

6.1.1 MONOPHONE KLASSIFIKATION

Bei der monophonen Klassifikation bestehen die analysierten Musiksignale aus einzelnen kurzen Klängen oder Geräuschen. Hierbei sind für jedes zu analysierende Klangereignis die Segmentgrenzen, d.h. der Anfang und das Ende, klar definiert und bekannt. Weiterhin überlappen sich die Klangereignisse weder gänzlich noch teilweise, so dass keine Melodien, Akkorde oder ganze Musikstücke direkt verarbeitet werden können.

Eine der ersten Arbeiten zur monophonen Klassifikation wurde von Brown vorgestellt [Brow97]. Als Merkmale werden hierbei Cepstrum-Koeffizienten und ein auf der Constant Q Transformation basierendes Spektrogramm verwendet [Brow91, BrPu92]. Zur Modellierung wurden GMMs mit einem MAP-Klassifikator verwendet. Die Rate der falsch klassifizierten Klangereignisse beläuft sich in dieser Arbeit auf 15,9 %, was einer Klassifikationsrate von 84,1 % entspricht. Die Klassifikationsergebnisse werden weiterhin in einem Expertenhörtest mit der Klassifikationsfähigkeit eines Menschen verglichen.

In einer weiteren Arbeit untersucht Brown zusammen mit Houix und McAdams die Anwendung von Techniken zur Sprechererkennung im Zusammenhang mit der monophonen Musikinstrumentenklassifikation [BrHM01]. Als Klangereignisse werden die Klänge von Oboen, Saxophonen, Klarinetten und Flöten verwendet. Als Merkmale werden neben

Cepstrum-Koeffizienten und dem Spektrogramm der Constant Q Transformation weiterhin auch der Spectral Centroid, die Autokorrelation und verschiedene Zeitbereichsmerkmale verwendet. Die Klassifikation erfolgt ebenfalls über GMMs und Bayes'sche MAP-Klassifikatoren. Das System erzielt Klassifikationsraten zwischen 79 % und 84 %.

Marques und Moreno stellen in ihrer Arbeit ein monophones Klassifikationssystem vor, das die Klangereignisse in kleine 0,2 s lange Segmente unterteilt und eine Klassifikation über die größte Verbundwahrscheinlichkeit durchführt [MaMo99]. Für die Untersuchungen werden Klangereignisse von acht verschiedenen Instrumentenarten verwendet. Das vorgestellte System extrahiert als Merkmale lineare Prädiktionskoeffizienten und MFCCs und führt eine Klassifikation über GMMs und Support Vector Machines (SVM) durch [Burg98, HsCL03]. Die Klassifikationsrate liegt bei 70 %.

Eronen und Klapuri beschreiben ein System zur grundfrequenzunabhängigen Klassifikation von 30 verschiedenen Orchesterinstrumenten, das mit 1498 monophonen Klangereignissen ausgewertet wird [ErKl00]. Es wird eine Vielzahl von unterschiedlichen spektralen Merkmalen und Zeitbereichsmerkmalen extrahiert, deren Aussagekraft bezogen auf das gewählte Szenario verglichen wird. Die Klassifikation der Klangereignisse erfolgt hierarchisch, wobei in jeder Gliederungsebene ein GMM oder das KNN-Verfahren verwendet wird. Die Klassifikationsrate beträgt 94 % für einzelne Instrumente und 80 % für Instrumentenklassen.

In einer weiteren Arbeit geht Eronen eingehend auf das Timbre von Musikinstrumenten und die hieraus resultierenden Merkmale ein [Eron01a]. Es werden verschiedene Klassifikationssysteme vorgestellt und implementiert. Die Systeme unterscheiden zwischen 29 Instrumenten und werden mit der großen Anzahl von 5286 Klangereignissen ausgewertet. Die Klassifikationsrate für einzelne Instrumente liegt bei 35 % und für Instrumentenfamilien bei 77 %. Die Klassifikationsrate wird mit der Klassifikationsrate der menschlichen Wahrnehmung verglichen, die bei 46 % für einzelne Instrumente und bei 92 % für Instrumentenfamilien liegt [Mart99].

Eine zu den Untersuchungen von Eronen und Klapuri sehr ähnliche Forschung beschreiben die von Agostani, Longari und Pollastri vorge-

stellten Arbeiten [AgLP01, AgLP03]. Ihr System ist in der Lage, monophone Klangereignisse 27 unterschiedlichen Instrumenten zuzuordnen und wird mit 1007 Klangereignissen ausgewertet. Die Merkmale werden um harmonische Merkmale erweitert, und als Klassifikationsverfahren werden die Diskriminanzanalyse und Support Vector Machines verwendet. Die Klassifikationsrate für das beste untersuchte System beträgt 94,7 % für einzelne Instrumente und bis zu 96,8 % für Instrumentenklassen.

Die Arbeit von Kostek und Czyzewski hat ihren Schwerpunkt in den unterschiedlichen Darstellungsarten von Musikinstrumentenklängen und ihren Merkmalen [KoCz01]. Es werden unterschiedliche Merkmale dargestellt, die ein Minimum an Redundanz untereinander aufweisen. Weiterhin wird ein System beschrieben, das eine Klassifikation auf der Basis von Neuronalen Netzen durchführt.

Auch Herrera-Boyer, Peeters und Dubnov konzentrieren sich in ihrer Arbeit auf Merkmale, mit der sich Klangereignisse für Klassifikationsaufgaben besonders effizient darstellen lassen [HePD02]. Sie präsentieren zwei Systeme zur monophonen Klassifikation, das erste verfolgt einen wahrnehmungsbasierten Ansatz, das zweite eine hierarchische Klassifikation. Es wird eine Vielzahl von Klassifikatoren untersucht, darunter das KNN-Verfahren, die Bayes'sche MAP-Klassifikation, die Diskriminanzanalyse, Neuronale Netze, Support Vector Machines und HMMs.

Peeters und Rodet beschreiben ein System zur monophonen Klassifikation, das auf einer automatischen Auswahl relevanter Merkmale aus einem großen Merkmalsraum beruht [PeRo02]. Das System ist im Rahmen des CUIDADO-Projekts entstanden und stellt eine Vielzahl hochwertiger Merkmale vor [ViHP02, Peet04]. Für die Klassifikation wird zum einen eine Diskriminanzanalyse, zum anderen eine hierarchische Klassifikation verwendet. Ihr System ist in der Lage, 14 Instrumente zu unterscheiden und wird mit 1400 Klangereignissen ausgewertet. Die Klassifikationsrate liegt bei 86 % für einzelne Instrumente und 89 % für Instrumentenklassen.

Die Abhängigkeit des Timbres und der daraus resultierenden Merkmale steht in der Arbeit von Kitahara, Goto und Okuno im Vordergrund

[KiGO03]. In ihrem System wird in einem ersten Schritt die Grundfrequenz der Klangereignisse ausgewertet. Alle dann extrahierten Merkmale werden durch multidimensionale Gaußverteilungen repräsentiert, die als Kontrollparameter die ermittelte Grundfrequenz haben. Das System unterscheidet zwischen 19 Musikinstrumenten und wird mit 6247 Klangereignissen trainiert bzw. ausgewertet, wobei eine Kreuzvalidierung verwendet wird. Die Klassifikation erfolgt über einen Bayes'schen MAP-Klassifikator, der eine Klassifikationsrate für einzelne Instrumente von bis zu 79,7 % und für Instrumentenfamilien von bis zu 90,7 % liefert.

Einen der populärsten Ansätze zur monophonen Geräusch- bzw. Musikinstrumentenklassifikation liefert Casey, dessen Verfahren auch im MPEG-7-Standard als Beispielanwendung zitiert wird [Case01a, Case01b, ISO15938, KiMS05]. Das System extrahiert als Merkmal den Audio Spectrum Envelope (ASE) entsprechend dem MPEG-7-Standard. Die Koeffizienten werden über eine PCA oder ICA in ihrer Dimension reduziert und zur Klassifikation mittels HMMs verwendet. Das System unterscheidet Klangereignisse aus 19 verschiedenen Klassen und wird mit mehr als 2000 Klangereignissen ausgewertet. Die durchschnittliche Klassifikationsrate liegt bei 92,6 %. Aufgrund seiner Popularität stellt das von Casey vorgestellte System eine wertvolle Benchmark für andere Systeme dar und wurde deshalb im Rahmen dieser Arbeit als eines von drei implementierten Systemen in einer für die Musikinstrumentenklassifikation optimierten Version als echtzeitfähiges VST-Plugin umgesetzt (vgl. Abschnitt 6.3).

Eine Übersicht über weitere monophone Klassifikationssysteme und deren Aufbau zusammen mit der jeweils verwendeten Datenbankgröße und erzielten Klassifikationsrate liefert die Arbeit von Herrera et al. [HAB+00].

6.1.2 EINGESCHRÄNKT POLYPHONE KLASSIFIKATION

Bei der eingeschränkt polyphonen Klassifikation werden Musikstücke verarbeitet, die als Einschränkung entweder nur aus harmonischen Klängen oder nur aus rhythmischen Geräuschen bestehen dürfen.

Eine der ersten und bedeutendsten Veröffentlichungen zum Thema der eingeschränkten polyphonen Klassifikation ist die oft zitierte Arbeit von

Martin [Mart99]. Obwohl auch in dieser Arbeit der Fokus auf der monophonen Klassifikation liegt, behandelt ein wesentlicher Teil die Erweiterung für die Klassifikation von harmonischen Klangereignissen als eingeschränkt polyphone Klassifikation über ein harmonisches Modell. Das vorgestellte System kann einen kontinuierlichen Audiostrom mit Klängen von harmonischen Orchesterinstrumenten verarbeiten und in 25 verschiedene Instrumentenklassen klassifizieren. Der Klassifikationsvorgang baut auf einem auditorischen Modell des Gehörs auf und wird mit den Klassifikationsfähigkeiten des Menschen in einem Expertenhörtest verglichen. Die Klassifikationsrate des präsentierten Systems beträgt 45,9 % für einzelne Instrumente und 91,7 % für Instrumentenklassen.

Virtanen beschreibt in seiner Arbeit die Erzeugung eines harmonischen Modells aus polyphoner Musik, mit dem direkt eine eingeschränkt polyphone Klassifikation durchgeführt werden kann [Virt03]. Als Merkmale können hierbei die Parameter des harmonischen Modells verwendet werden. Die ebenfalls vorgestellte Modellierung von Schlagzeugklängen kann aufgrund der speziellen Parameter hingegen nicht für eine Klassifikation verwendet werden.

Die Arbeiten von Livshin, Rodet und Peeters beschreiben ein System zur Klassifikation einer Soloaufnahme eines Instruments [LiPR03, LiRo04]. Das System kann zwischen sieben Instrumentenklassen unterscheiden und wurde mit 108 verschiedenen Soloaufnahmen trainiert bzw. getestet. In einem ersten Schritt wird aus den Soloaufnahmen die große Anzahl von 62 verschiedenen Merkmalen extrahiert, die dann über einen neu entwickelten Algorithmus der Dimensionsreduktion reduziert werden. Das System erzielt eine Klassifikationsrate von 85,2 %.

Ein Verfahren für die Quellentrennung von perkussiven Klängen und Geräuschen, mit dem auch eine eingeschränkt polyphone Klassifikation durchgeführt werden kann, wird von Uhle, Dittmar und Sporer beschrieben [UhDS03]. Das System führt mit dem Spektrum des Audiomaterials eine ICA durch, bei der die perkussiven Anteile als unabhängige Komponenten erhalten bleiben. Diese können dann über ihre Spektren klassifiziert werden.

Ein weiterer Ansatz zur Klassifikation der nach einer Trennung verbleibenden perkussiven Klänge und Geräusche wird von Van Steelant et al.

beschrieben [VTD+04]. Das vorgestellte System ist in der Lage, in einer eingeschränkt polyphonen Klassifikation Klangereignisse fünf Idiophonen zuzuordnen. Hierzu wird eine Vielzahl verschiedener Merkmale extrahiert und über eine SVM klassifiziert. Die Klassifikationsraten liegen bei 78 % bis 90,3 %.

6.1.3 KOMPLEXE POLYPHONE KLASSIFIKATION

Bei der komplexen, polyphonen Klassifikation werden für die verarbeiteten Musikstücke keinerlei Einschränkungen bezüglich der Harmonizität oder Rhythmik gemacht. Über dieses Szenario wurde bisher relativ wenig publiziert, da es sich hierbei um den komplexesten Forschungsbereich handelt.

Einer der ersten Ansätze wurde auch für dieses Gebiet von Martin in einer Studie veröffentlicht, in der Merkmale für ein komplexes, polyphones Szenario vorgestellt werden [Mart98]. Die Merkmale beziehen sich einerseits auf die physikalischen Vorgänge der Klangerzeugung von Orchesterinstrumenten (vgl. Abschnitt 3.2), andererseits wird das log-lag Corellogram vorgestellt, bei dem es sich um eine besondere Form der Autokorrelation handelt. Das System wird für verschiedene Orchesterinstrumente ausgewertet, zeigt jedoch bei der Klassifikation noch deutliche Schwächen.

In der von Cho, Choi und Bang vorgestellten Arbeit geht es primär nicht um die Identifikation und Klassifikation von Musikinstrumentenklängen in polyphoner Musik sondern um generelle Geräusche [ChCB03]. Das vorgestellte System beruht auf einer Transformation eines kontinuierlichen Audiostroms in NMF-Blöcke und funktioniert mit einer Klassifikationsrate von 95,2 % zufriedenstellend für Geräusche, die sich stark unterscheiden. Für die Unterscheidung von Musikinstrumentenklängen, die teilweise sehr ähnliche, grundfrequenzabhängige Klangeigenschaften haben, liefert das System hingegen schlechte Ergebnisse, weil die NMF-Darstellung hier mit einer zu geringen Detailgenauigkeit erfolgt.

Eggink und Brown stellen in ihren Arbeiten ein Verfahren zur komplexen, polyphonen Klassifikation vor, das als Merkmale ausschließlich spektrale Merkmale verwendet [EgBr03a, EgBr03b]. Besonders das Spektrum selbst wird erfolgreich verwendet, da sich in ihm verschiedene

Klangereignisse in unterschiedlichen Bins wiederfinden lassen. Die Klassifikation erfolgt über die Verwendung von GMMs, die so ausgelegt sind, dass Merkmale, die aufgrund einer Überlagerung zweier Klänge nicht verlässlich ausgewertet werden können, während der Klassifikation ausgeblendet werden können, um das Ergebnis nicht zu verfälschen. Die Auswertung des Systems erfolgt für zweifach polyphone Musik, in der das System Musikinstrumentenklänge von fünf Instrumentenklassen automatisch mit einer mittleren Klassifikationsrate von 46,7 % erkennt.

In einer weiteren Arbeit von Eggink und Brown wird ein leicht abgewandeltes System vorgestellt, das zwar höhere Klassifikationsraten von bis zu 86 % erzielt, aber für das zu verarbeitende Material stärkere Einschränkungen bezüglich des Läutstärkeunterschieds zwischen den zu klassifizierenden Klängen und der Begleitmusik macht [EgBr04a].

Aufgrund der Ausdruckskraft des Spektrogramms in polyphonen Szenarien stellen die von Eggink und Brown vorgestellten Systeme die Grundlage für das polyphone Klassifikationssystem dar, das im Rahmen dieser Arbeit als eines von drei Systemen implementiert wurde (vgl. Abschnitt 6.4). Das hier implementierte System verwendet allerdings anstatt des direkten Spektrums ein spezielles harmonisches Modell, wodurch bessere Klassifikationsergebnisse bzw. eine größere Polyphonie ermöglicht werden.

6.2 MONOPHONES EXPERIMENTIERSYSTEM

Das monophone Experimentiersystem stellt den Rahmen für die Erforschung der in dieser Arbeit betrachteten Algorithmen dar. Aufgrund seines modularen Aufbaus lassen sich die verschiedenen Verfahren der Merkmalsextraktion und Klassifikation sehr einfach sowohl gesondert als auch im Zusammenspiel untersuchen. Die in den Kapiteln 4 und 5 beschriebenen Eigenschaften der verschiedenen untersuchten Algorithmen wurden über eine Vielzahl qualitativer Experimente mit dem monophonen Experimentiersystem gewonnen. Weiterhin wurden so die leistungsfähigsten Algorithmen ermittelt, die dann für das monophone Echtzeitsystem und das polyphone System in optimierter Form implementiert wurden [BKK+05]. In den folgenden Abschnitten werden der Aufbau

und die Verwendung des monophonen Experimentiersystems ausgiebig beschrieben.

6.2.1 AUFBAU

Das System ist in Matlab implementiert und arbeitet als Offline-Prozess. Es ist in der Lage, monophone Klangereignisse mit klar definierten Anfangs- und Endzeitpunkten zu verarbeiten. Im Trainingsbetrieb lassen sich Modelle für das monophone Klassifikationsszenario erstellen, die dann im Klassifikationsbetrieb zur Klassifikation verwendet werden können. Die Klassifikationsentscheidung erfolgt hierbei für jedes Klangereignis als Ganzes.

Der grundlegende Aufbau des Systems im Trainingsbetrieb bzw. im Klassifikationsbetrieb ist in Bild 6.1 und Bild 6.2 dargestellt. Es besteht sowohl für den Trainings- als auch für den Klassifikationsbetrieb im Wesentlichen aus den drei Prozessschritten Merkmalsextraktion, Merkmalsaufbereitung und Modell-Training/Klassifikation. Das System ist modular ausgeführt, so dass die in den einzelnen Prozessschritten verwendeten Verfahren schnell und einfach ausgetauscht werden können. Somit können unterschiedliche Verfahren sehr einfach und effizient im Zusammenspiel untersucht werden.

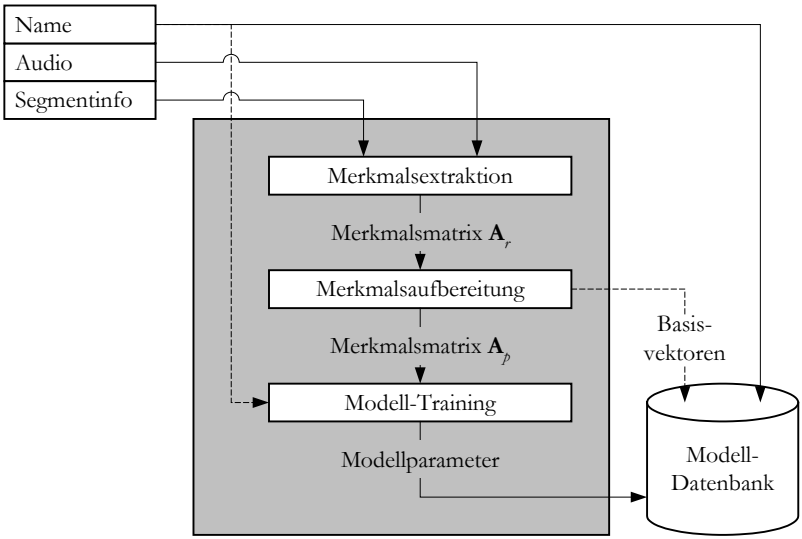


Bild 6.1: Schematischer Aufbau des monophonen Experimentiersystems im Trainingsbetrieb

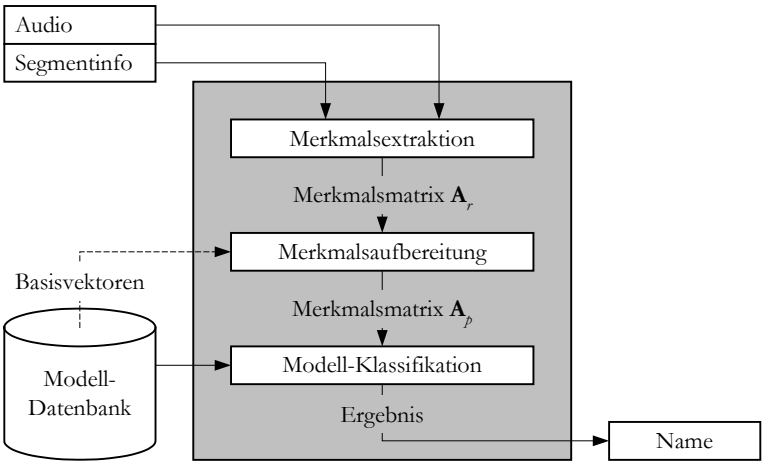


Bild 6.2: Schematischer Aufbau des monophonen Experimentiersystems im Klassifikationsbetrieb

6.2.2 MERKMALSEXTRAKTION

Das Modul zur Merkmalsextraktion wendet die in Kapitel 4 beschriebenen Verfahren an und erzeugt eine Folge von Merkmalsvektoren **a** mit konstanter Abtastrate. Erzeugt ein Verfahren Merkmale mit variablem zeitlichen Abstand oder wird ein Verfahren mit abweichender Abtastrate verwendet, so werden die zu dem entsprechenden Merkmal gehörenden Werte des Merkmalsvektors über ein Sample-and-Hold Modul gehalten, um den konstanten Strom der Vektoren aufrecht zu halten. Alle erzeugten Merkmalsvektoren werden in einer Merkmalsmatrix $\mathbf{A}_r$ zusammengefasst.

Um das Ausgabeformat der erzeugten Merkmalsvektoren zu vereinheitlichen, wurde im Rahmen dieser Arbeit zusammen mit Lerch und Tanghe die in der Multimediaforschung mittlerweile weit verbreitete Feature Extracion API (FEAPI) entwickelt [LeET05]. Hierbei handelt es sich um eine API, die das technische Format von Merkmalsvektoren sowie die Schnittstelle zu Analyseplugins und Klassifikationsprogrammen vereinheitlicht.

6.2.3 MERKMALSAUFBEREITUNG

Im Modul der Merkmalsaufbereitung wird die Merkmalsmatrix $\mathbf{A}_r$ in ihrem Informationsgehalt aufbereitet und wenn möglich in ihrer Dimension reduziert. Die wichtigsten Schritte der Aufbereitung sind die Normierung, das Entfernen von zu kleinen, nicht aussagekräftigen Werten sowie die Transformation in eine andere Skalierung wie beispielsweise die Transformation in eine Dezibel-Skala.

Weiterhin lassen sich in der Merkmalsaufbereitung kleine Rauschanteile zu gewissen Merkmalen hinzuaddieren. Dies ist beispielsweise für einige Klassifikationsverfahren wie das GMM sinnvoll, wenn bei der Merkmalsextraktion Merkmale erzeugt werden, die durch ein Sample-and-Hold-Modul repliziert werden. Dies verhindert, dass das Modell während des Trainings auf mehrfach belegten Punkten im Merkmalsraum kollabiert und zu singulärem Verhalten übergeht.

Für die Dimensionsreduktion verwendet die vorliegende Implementierung Verfahren der Faktorisierung, wie sie in Abschnitt 4.6 beschrieben sind. Im Trainingsbetrieb werden hierbei Basisvektoren für eine Projek-

tion der Merkmalsmatrix in einen dimensionsreduzierten Merkmalsraum erzeugt, die anschließend im Musikinstrumentenmodell gespeichert werden. Im Klassifikationsbetrieb werden die Basisvektoren hingegen aus dem Musikinstrumentenmodell geladen. Verfahren wie die Auswahl eines Subraums im Merkmalsraum (engl.: Feature Subset Selection) wurden für die vorliegende Arbeit nicht implementiert [BuLe03, PeRo03, BuLe04].

6.2.4 TRAINING

Das letzte Modul im Trainingsbetrieb stellt die Modellerzeugung dar. Im Modul für die Modellerzeugung werden die Modelle mit den aufbereiteten Merkmalsmatrizen $\mathbf{A}_p$ trainiert. Wird hierfür ein überwachtes Lernverfahren verwendet, muss dem Training neben der Merkmalsmatrix weiterhin auch der jeweilige Klassenname bzw. Index zu der verarbeiteten Merkmalsmatrix $\mathbf{A}_p$ präsentiert werden.

Die erzeugten Modellparameter werden anschließend zusammen mit dem Instrumentennamen und eventuell in der Merkmalsaufbereitung erzeugten Basisvektoren im Musikinstrumentenmodell gespeichert.

Der folgende Beispielaufruf zeigt die Erzeugung eines Geigenmodells, dessen Verarbeitungskette als Merkmale ASE-Koeffizienten verwendet, die über eine PCA aufbereitet und dekorreliert werden und anschließend ein GMM trainieren:

```
sViolinModel = Train(...
'name', 'Violin',...
'container', 'C:\sounds\violin1',...
'feature', 'ase',...
'distiller', 'pca',...
'model', 'gmm');
```

6.2.5 KLASSIFIKATION

Im Klassifikationsbetrieb werden die Modellparameter der verschiedenen Musikinstrumentenmodelle sowie der jeweilige Instrumentenname in das Modul der Modell-Klassifikation geladen. Weiterhin werden eventuell benötigte Basisvektoren in das Modul der Merkmalsaufbereitung geladen. Die Klassifikationsentscheidung erfolgt dann entsprechend dem

internen Aufbau des Klassifikators, und der Name des erkannten Musikinstruments wird für jeden Klang ausgegeben.

Der folgende Beispielaufruf zeigt die Klassifikation einiger Audioaufnahmen durch ein Geigenmodell und ein Gitarrenmodell:

```
Classify(...
'model', sViolinModel, ...
'model', sGuitarModel, ...
'container', 'C:\sounds\demo_violin_guitar');
```

Die Ausgabe des Systems könnte für den obigen Aufruf folgendermaßen aussehen:

```
Processing 1/5: 'violin1.wav' ... done
Result: Violin
Certainty: 86.6%

Processing 2/5: 'violin2.wav' ... done
Result: Violin
Certainty: 92.4%

Processing 3/5: 'guitar1.wav' ... done
Result: Guitar
Certainty: 81.2%

Processing 4/5: 'guitar2.wav' ... done
Result: Guitar
Certainty: 85.6%

Processing 5/5: 'guitar_distorted.wav' ... done
Result: Guitar
Certainty: 32.8%
```

6.3 MONOPHONES ECHTZEITSYSTEM

Das monophone Echtzeitsystem verwendet für die Merkmalsextraktion und die Klassifikation Verfahren, die effizient berechnet werden können und sehr gute Klassifikationsergebnisse erzielen. Es unterstützt die Industriestandards VST und MPEG-7 und dient somit auch als Technikstudie für den industriellen Gebrauch. In den folgenden Abschnitten

werden der Aufbau sowie die Verwendung des monophonen Echtzeitsystems ausgiebig beschrieben.

6.3.1 AUFBAU

Das System wurde aufbauend auf dem gemeinsam mit Roth entwickelten System [Roth05] in C/C++ programmiert und in einem echtzeitfähigen VST-Plugin implementiert. Aus den Audiodaten wird als Merkmal die Hüllkurve extrahiert, die sich in den Voruntersuchungen mit dem monophonen Experimentiersystem als besonders aussagekräftig herausgestellt hat [BKK+05]. Weiterhin wird als MPEG-7 kompatibles, spektrales Merkmal der Audio Spectrum Envelope extrahiert, da dieser sich als sehr robust gegenüber Rauscheinflüssen erwiesen hat. Beide Merkmale lassen sich effizient berechnen, was für ein Echtzeitsystem unerlässlich ist.

Sowohl im Trainings- als auch im Klassifikationsbetrieb besteht das vom System verarbeitete Eingangsmaterial aus aufeinander folgenden monophonen Klangereignissen, die über den VST-Host eingespielt werden. Der kontinuierliche Audiostrom wird automatisch in einzelne monophone Klangereignisse segmentiert, die dann als Ganzes zum Training bzw. zur Klassifikation weiterverarbeitet werden.

6.3.2 PLUGINBETRIEB

Die für die Implementierung verwendete Virtual Studio Technology (VST) wurde von der Firma Steinberg entwickelt und ermöglicht eine virtuelle Studioumgebung innerhalb eines PCs. Hierzu wurde eine Plug-in-Schnittstelle entwickelt, die den Dialog zwischen Signalverarbeitungsprogrammen oder Sequenzern als Hosts und virtuellen Effekten oder Instrumenten als Plugins ermöglicht. Host-Programme können Plugins laden, die selbstständige Module zur Signalverarbeitung darstellen. Zwischen Host und Plugin sind hierbei Schnittstellen für Audiodaten und MIDI-Signale definiert. VST-Plugins können eine eigene grafische Benutzeroberfläche besitzen und stellen sich innerhalb der Umgebung des Hosts autonom dar. Die VST-Schnittstelle ist im Studiobereich mittlerweile ein Quasi-Standard geworden und wird von vielen Programmen unterstützt.

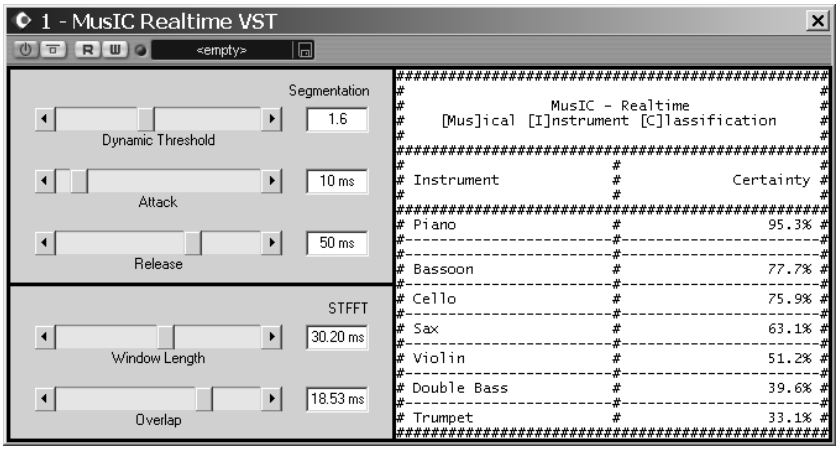


Bild 6.3: Benutzeroberfläche des monophonen Echtzeitsystems als VST-Plugin im Betrieb mit dem Programm Cubase SX 2 der Firma Steinberg, links: Parametereingabe, rechts: Ergebnisliste

Als VST-Plugin kann das entwickelte monophone Echtzeitsystem plattformübergreifend in beliebigen VST-Hosts verwendet werden. Die Benutzeroberfläche des Systems zeigt Bild 6.3.

Aus technischer Sicht stellen VST-Plugins Laufzeitbibliotheken dar, dies sind beispielsweise dll-Dateien unter Windows oder so-Dateien unter Linux und MacOS. Die Plugin- Laufzeitbibliotheken werden von Host-Programmen dynamisch zur Laufzeit geladen und verwendet.

6.3.3 TRAINING

Der grundlegende Aufbau des Systems im Trainingsbetrieb ist in Bild 6.4 dargestellt. Das Audiosignal besteht aus mehreren nacheinander folgenden monophonen Instrumentenklängen, die im Trainingsbetrieb alle von einem Instrument sein müssen. Da für die Berechnung der Basisvektoren im Rahmen der Faktorisierung alle Trainingsbeispiele bekannt sein müssen, kann das System kein Online-Learning durchführen. Aus diesem Grunde kann das Training im Gegensatz zur Klassifikation nicht in Echtzeit erfolgen, weshalb es als Offline-Prozess in Matlab durchgeführt wird. Das Ergebnis des Trainings ist für jedes Musikinstrument eine

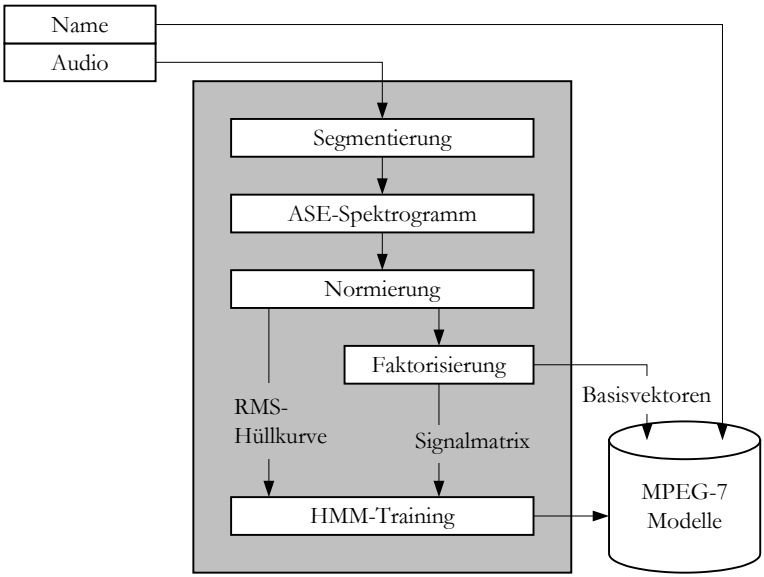


Bild 6.4: Schematischer Aufbau des monophonen Echtzeitsystems im Trainingsbetrieb

MPEG-7 kompatible XML-Datei, in der die Modellparameter gespeichert werden.

Die einzelnen für das Training relevanten Module und der Verarbeitungsprozess des Systems sowie die Erzeugung der MPEG-7 Modelldatei werden in den folgenden Abschnitten eingehend beschrieben.

6.3.4 SEGMENTIERUNG

Das Audiosignal $x(n)$ wird automatisch über seinen Kurzzeiteffektivwert $x\sim(n)$ in einzelne Klangereignisse segmentiert, indem der Kurzzeiteffektivwert mit einem adaptiven Schwellenwert $a(n)$ verglichen wird (Bild 6.5). Der Effektivwert des Audiosignals wird über die nachstehende Gleichung berechnet:

$$x\sim(n)=\sqrt{\frac{1}{N}\sum_{k=-K/2}^{K/2-1}x^2(k+n)}\,. \tag{6.1}$$

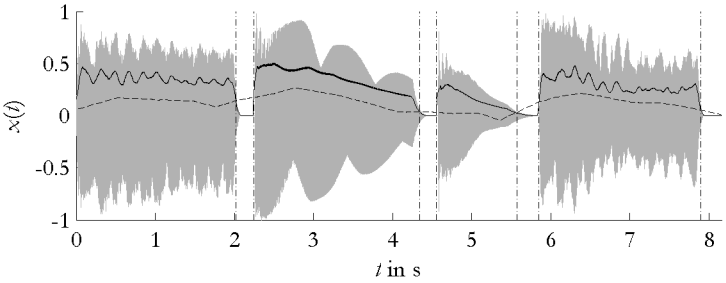


Bild 6.5: Audiosignal mit vier Klangereignissen, Effektivwert (schwarz), adaptive Schwelle (gestrichelt) und Segmentgrenzen (strichpunktiert)

Die Anzahl der für die Berechnung des Kurzzeiteffektivwerts betrachteten Abtastwerte N ergibt sich direkt aus der als Parameter festgelegten Fenstergröße. Die besten Segmentierungsergebnisse konnten mit einer Fensterbreite von 25 ms erzielt werden.

Die adaptive Schwelle $a(n)$ stellt den mit einem Faktor c gewichteten Langzeiteffektivwert des Audiosignals dar und lässt sich über die folgende Gleichung berechnen:

$$a(n) = c\sqrt{\frac{1}{L}\sum_{k=-L/2}^{L/2-1} x^2(k+n)}\,. \tag{6.2}$$

Die Anzahl der für die Berechnung betrachteten Abtastwerte L ergibt sich hierbei aus einer Fensterbreite von 5 s.

Der Beginn eines neuen Segments wird immer dann markiert, wenn der Kurzzeiteffektivwert die adaptive Schwelle für mindestens 50 ms überschreitet:

$$\tilde{x}(n) > a(n). \tag{6.3}$$

Diese Zeitspanne soll verhindern, dass kleine Signalpeaks falsche Segmentanfänge auslösen. Umgekehrt wird das Ende eines Segments markiert, wenn der Kurzzeiteffektivwert die adaptive Schwelle für mindestens 50 ms unterschreitet:

$$\tilde{x}(n) < a(n). \tag{6.4}$$

Sollte die Länge eines erkannten Segments kürzer als 250 ms sein, so wird es verworfen um zu verhindern, dass ein Klangereignis fälschlicherweise in mehrere kurze Klangereignisse zerlegt wird.

Der Schwellenwertfaktor c aus Gleichung 6.2 verhindert, dass in Perioden mit lang anhaltender Stille die Werte von $x\sim(n)$ und $a(n)$ konvergieren und durch das Grundrauschen des Signals fälschlicherweise neue Segmente markiert werden. Für die vorliegende Implementierung wurde ein Wert von 1,5 als Optimalwert ermittelt. Weiterhin bestimmt der Faktor, ob die im Audiosignal enthaltenen Klangereignisse tendenziell zu großzügig ($c \approx 1$) oder zu knapp ($c \geq 1{,}5$) segmentiert werden. Für die vorliegende Implementierung ist eine tendenziell zu knappe Segmentierung zu bevorzugen (vgl. hierzu die Abschnitte 6.3.7 und 6.3.8).

Nach der Segmentierung enthält jedes Segment einen eigenen Instrumentenklang. Die Bereiche, die zwischen den Segmenten liegen, werden als Stille oder Bereiche mit geringen Signalanteilen verworfen. Somit wird den weiteren Prozessschritten eine Anzahl von Segmenten mit jeweils einem Klangereignis, das einen klar definierten Anfang und ein klar definiertes Ende hat, zugeführt.

Obwohl die Segmentierung sehr zuverlässig arbeitet, sollte zwischen den Klangereignissen ein kurzer Moment Stille mit einer Dauer von mindestens 0,2 s sein, damit die Segmentierung in einzelne Klangereignisse fehlerfrei erfolgen kann. Ist die Pause zwischen zwei Klangereignissen kürzer, kann es passieren, dass zwei Klangereignisse als ein einziges segmentiert und weiterverarbeitet werden.

6.3.5 SPEKTROGRAMM

Die Spektrogramme jedes einzelnen Klangereignisses werden entsprechend dem MPEG-7 Audio Spectrum Envelope (ASE) berechnet (vgl. Abschnitt 4.4.9). Als Parameter der STFT werden eine Hopsize von 10 ms und eine Fensterbreite von 30 ms verwendet. Das verwendete Hamming-Fenster ist so skaliert, dass die Kurzzeitleistung erhalten bleibt. Für die Umrechnung des Linearen FFT-Spektrums in einen logarithmischen ASE wird der Bereich zwischen 62,5 Hz und 8 kHz in 28 Bänder eingeteilt. Jedes Band deckt den Bereich einer kleinen Terz ab, dies entspricht einer Auflösung von vier Bändern pro Oktave. Zusam-

men mit den Koeffizienten für alle FFT-Frequenzbänder über bzw. unter den Grenzfrequenzen ergeben sich somit 30 ASE-Koeffizienten, die den Merkmalsvektor $\mathbf{a}_{ASE}$ bilden (vgl. Abschnitt 4.4.9).

6.3.6 NORMIERUNG

Da die ASE-Koeffizienten aus dem Leistungsdichtespektrum berechnet werden, bleibt die Kurzzeitleistung p des analysierten Blocks als Summe aller ASE-Koeffizienten erhalten. Die ermittelte Kurzzeitleistung wird einerseits verwendet, um alle ASE-Koeffizienten auf Einheitsleistung zu normieren:

$$\mathbf{a}_{ASE-} = \frac{\mathbf{a}_{ASE}}{p}, \tag{6.5}$$

andererseits beschreibt sie als Approximation die Hüllkurve des Signals und wird als wertvolles zusätzliches Merkmal verwendet. Für die weitere Verarbeitung werden die ASE-Koeffizienten in eine Dezibel-Skala umgewandelt:

$$\mathbf{a}_{ASE\_dB} = 10 \lg \mathbf{a}_{ASE-}. \tag{6.6}$$

Somit ergibt sich nach der Normierung ein vorläufiger Merkmalsvektor $\mathbf{a}_{Norm}$ mit 31 Elementen, genauer 30 ASE-Koeffizienten und der Kurzzeitleistung p als Hüllkurve.

6.3.7 FAKTORISIERUNG

Die normierten ASE-Koeffizienten $\mathbf{a}_{ASE\_dB}$ aller aus dem Klangereignis analysierten Blöcke werden im nächsten Schritt zu einer Merkmalsmatrix $\mathbf{A}_{ASE}$ zusammengefasst. Diese Merkmalsmatrix wird nun über die PCA oder ICA faktorisiert, wobei die einzelnen Merkmalssignale untereinander dekorreliert (PCA) oder statistisch unabhängig (ICA) werden. Prinzipiell sollte die ICA aufgrund der besseren Ergebnisse der PCA vorgezogen werden, allerdings ist sie vor allem im Training rechenintensiver. Aus dieser Faktorisierung ergibt sich zum einen eine neue Signalmatrix $\mathbf{A}_{Fak}$, zum anderen die Transformationsmatrix $\mathbf{P}$, deren Spalten die Hauptkomponenten (PCA) oder die unabhängigen Komponenten (ICA) als Basisvektoren enthält:

$$\mathbf{A}_{ASE-} = \mathbf{P}\mathbf{A}_{Fak}. \tag{6.7}$$

Die neue Signalmatrix $\mathbf{A}_{Fak}$ lässt sich durch Umstellen der Gleichung berechnen:

$$\mathbf{A}_{Fak} = \mathbf{P}^{-1}\mathbf{A}_{ASE-} = \mathbf{B}\mathbf{A}_{ASE-}, \tag{6.8}$$

wobei die Merkmalsvektoren in einer Basistransformation auf die Projektionsvektoren in den Zeilen der Matrix **B** projiziert werden.

Für die Faktorisierung spielen die Qualität der einzelnen Klangereignisse hinsichtlich der korrekten Start- und Endzeitpunkte und der Signal-Rauschabstand eine große Rolle. Werden die Klangereignisse zu großzügig segmentiert, so werden zu viele leise bzw. stille Blöcke und somit zu viel Grundrauschen in die Merkmalsextraktion einbezogen. Da die ASE-Blöcke jedoch in ihrer Kurzzeitleistung normiert werden, werden aus diesen leisen Blöcken ASE-Blöcke mit quasi zufälligem Inhalt. Dies führt zu einer starken Verfälschung der berechneten Basisvektoren und damit der Faktorisierung an sich. Werden die Klangereignisse hingegen, wie in Abschnitt 6.3.4 beschrieben, tendenziell zu knapp segmentiert, fehlen lediglich kurze Bereiche des Ein- und Ausschwingens. Diese fehlenden Informationen verfälschen die berechneten Basisvektoren kaum, weshalb eine zu knappe Segmentierung einer zu großzügigen vorzuziehen ist.

Merkmalssignale mit einer geringen Varianz werden nach der Faktorisierung im Zuge einer Dimensionsreduktion fallen gelassen. Hierzu werden die den Merkmalssignalen mit den kleinsten Varianzen σ_r^2 entsprechenden Zeilen der Matrix $\mathbf{A}_{Fak}$ sowie die dazugehörigen Projektionsvektoren der Matrix **B** entfernt. Hierbei gilt, dass die Gesamtvarianz σ_{neu}^2, die sich als Summe der Einzelvarianzen ergibt, mindestens 90 % der Ursprungsvarianz σ^2 ergeben muss, so dass nur K Merkmalssignale verbleiben:

$$\sigma^2{}_{neu} = \sum_{r=1}^{K} \sigma_r{}^2 \geq 0{,}9 \cdot \sigma^2 = 0{,}9 \cdot \sum_{r=1}^{R} \sigma_r{}^2 . \tag{6.9}$$

Bei dieser Dimensionsreduktion bleiben in der Regel drei bis zehn Merkmalssignale und die dazugehörigen Basisfunktionen bestehen. Einige ähnliche Verfahren führen die Dimensionsreduktion während der Berechnung der PCA durch, indem die PCA als SVD berechnet wird

[Case01a, Case01b, KiBS04]. Dies scheint sich aufgrund der effizienteren Berechnung in einem Schritt anzubieten, allerdings hat sich herausgestellt, dass dieses Vorgehen mit einigen Nachteilen verbunden ist. Zum einen sind die Merkmalssignale nach der Berechnung der SVD in ihrer Varianz normiert, was ihnen für das nachfolgende Training und vor allem für die Klassifikation wesentliche Informationen entzieht, zum anderen kann eine der SVD nachgeschaltete ICA nur noch geringe Effekte erzielen, da durch die mit der SVD verbundene Projektion wesentliche Merkmale der Verteilungsdichte verloren gehen.

6.3.8 HIDDEN MARKOV MODEL

Für das Training des HMMs werden die Vektoren $\mathbf{a}_{Fak}$ der faktorisierten, dimensionsreduzierten Merkmalsmatrix $\mathbf{A}_{Fak}$ zusammen mit den jeweiligen Kurzzeitleistungswerten p verwendet. Somit ergibt sich ein Merkmalsraum mit vier bis elf Dimensionen.

Die HMMs besitzen eine Bakis-Topologie, bei der jeder Zustand nur in sich selbst, seinen Folgezustand oder seinen Folge-Folgezustand übergehen kann. Bakis-Topologien stellen spezielle links-rechts-Topologien dar, die sich im Feld der Spracherkennung sehr bewährt haben. Diese Art der Topologie bietet sich für den Echtzeitbetrieb an, da sie zum einen sehr gute Klassifikationsergebnisse liefert, sich zum anderen aber vor allem während des Klassifikationsbetriebs deutlich effizienter berechnen lässt als allgemeine HMMs. Bild 6.6 zeigt ein Beispiel für eine Bakis-Topologie mit fünf Zuständen.

Die Anzahl der Zustände hängt von der Dimension des Merkmalsraums ab. Hierbei wurde heuristisch ermittelt, dass halb so viele Zustände wie Merkmalsdimensionen optimal sind. Das System erzeugt deshalb dynamisch HMMs mit halb so vielen Zuständen wie Merkmalsdimensionen, verwendet aber mindestens drei Zustände. Die Verteilungsdichten der einzelnen Zustände sind einfache, multidimensionale Gaußverteilungen.

Das im Training gewonnene HMM hängt in seinem Aufbau von den durch die Segmentierung ermittelten Start- und Endzeitpunkten der einzelnen Klangereignisse ab. Werden die Klangereignisse zu großzügig segmentiert, so wird der erste und letzte Zustand des HMMs Stille bzw. Grundrauschen beschreiben und die Anfangswahrscheinlichkeit π_1 des

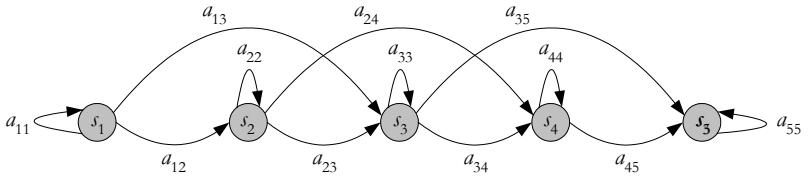


Bild 6.6: HMM mit fünf Zuständen in einer Bakis-Topologie

ersten Zustands wird tendenziell gegen 1 gehen, so dass das Modell fast immer im ersten Zustand beginnt. Werden die Klangereignisse, wie in Abschnitt 6.3.4 beschrieben, tendenziell zu knapp segmentiert, beschreiben alle Zustände des HMMs echte Signalzustände. Allerdings führen Klangereignisse, bei denen die Einschwingphasen abgeschnitten sind, dazu, dass das Modell nicht zwangsläufig im ersten Zustand startet, sondern auch in einem der mittleren Zustände starten kann. Dies führt dazu, dass die Anfangswahrscheinlichkeiten π der einzelnen Zustände tendenziell alle größer Null sind, jedoch monoton abfallen, was die Ergebnisse kaum verfälscht, weshalb ein zu knappes Segmentieren einem zu großzügigen vorzuziehen ist.

6.3.9 MPEG-7 MODELL

Jedes trainierte Musikinstrument wird in einer eigenen MPEG-7 konformen XML-Datei gespeichert. In der Datei werden der Name des Instruments, die Basisvektoren, die den dimensionsreduzierten Merkmalsraums beschreiben, sowie die Modellparameter des HMMs gespeichert. Die so erzeugten Modelle lassen sich aufgrund der Verwendung des MPEG-7-Standards nicht nur für die Klassifikation in der vorliegenden Implementierung verwenden, sondern auch in anderen MPEG-7 konformen Klassifikationssystemen. Weiterhin können für den Klassifikationsbetrieb auch Modelle verwendet werden, die mit anderen MPEG-7 konformen Systemen erzeugt wurden.

6.3.10 KLASSIFIKATION

Der grundlegende Aufbau des Systems im Klassifikationsbetrieb ist in Bild 6.7 dargestellt. Die MPEG-7-Dateien aller Modelle, die für die Klassifikation in Betracht gezogen werden sollen, müssen vor Beginn der

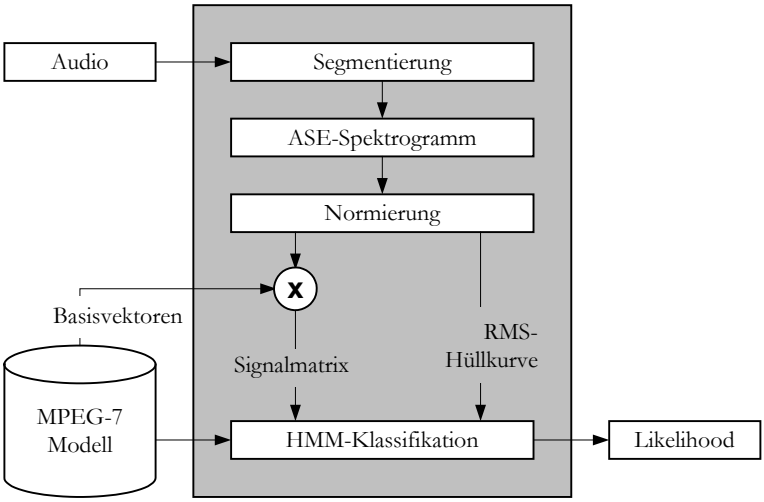


Bild 6.7: Schematischer Aufbau des monophonen Echtzeitsystems im Klassifikationsbetrieb

Klassifikation von dem System geladen werden. Die ersten drei Bearbeitungsschritte im Klassifikationsbetrieb sind identisch mit denen im Trainingsbetrieb. Der kontinuierliche Eingangsaudiostrom mit monophonen Klangereignissen wird durch die automatische Segmentierung in einzelne Klangereignisse zerlegt. Hieraus werden für jedes Klangereignis zunächst das normierte ASE-Spektrogramm sowie die Kurzzeitleistung als Approximation der Hüllkurve berechnet.

Die weiteren Verarbeitungsschritte weichen von den im Training durchgeführten Schritten ab und werden in den folgenden Abschnitten beschrieben.

6.3.11 BASISPROJEKTION

Die aus dem normierten ASE-Spektrogramm gewonnene Merkmalsmatrix $\mathbf{A}_{ASE-}$ muss, um eine nachfolgende Klassifikationsentscheidung treffen zu können, für jedes dem System bekannte Musikinstrument in den dazugehörigen Merkmalsraum projiziert werden. Dies geschieht für jeden Merkmalsvektor $\mathbf{a}_{ASE-}$ über ein Skalarprodukt mit den in den Zeilen der Matrix **B** enthaltenen Basisvektoren:

$$\mathbf{A}_{Fak} = \mathbf{B}\mathbf{A}_{ASE-} . \tag{6.10}$$

Die durch diese Projektion erzeugte Merkmalsmatrix $\mathbf{A}_{Fak}$ hat einerseits Merkmalsvektoren mit einer reduzierten Dimension, andererseits fungiert die Projektion als Filter für die Merkmalsvektoren. Entstammen die Merkmalsmatrix $\mathbf{A}_{Fak}$ und die Basismatrix **B** von Klangereignissen aus unterschiedlichen Musikinstrumentenklassen, so wird die Projektion der ASE-Koeffizienten sehr kleine, tendenziell zufällig verteilte Merkmalsvektoren in der Merkmalsmatrix $\mathbf{A}_{Fak}$ erzeugen. Entstammen die Merkmalsmatrix $\mathbf{A}_{Fak}$ und die Basismatrix **B** hingegen von Klangereignissen der gleichen Musikinstrumentenklasse, so ergibt die Projektion der ASE-Koeffizienten charakteristische Muster in der Merkmalsmatrix $\mathbf{A}_{Fak}$.

6.3.12 KLASSENZUORDNUNG

Für jedes Klangereignis erfolgt die eigentliche Klassenzuordnung als Ganzes, indem für jedes HMM über den Viterbi-Algorithmus eine a-posteriori Wahrscheinlichkeit $P(\omega|\mathbf{A}_{Fak},, p)$ berechnet wird. Diese gibt an, mit welcher Wahrscheinlichkeit die in den Merkmalsraum des Instruments projizierte Merkmalsmatrix $\mathbf{A}_{Fak}$ zusammen mit dem Kurzzeitleistungssignal p dem HMM der Klasse ω zugeordnet werden kann.

Hat die Projektion der ASE-Koeffizienten zu kleinen, tendenziell zufällig verteilte Merkmalsvektoren in der Merkmalsmatrix $\mathbf{A}_{Fak}$ geführt, weil die Merkmalsmatrix $\mathbf{A}_{Fak}$ und die Basismatrix **B** zu unterschiedlichen Musikinstrumentenklassen gehören, ergibt sich eine geringe a-posteriori Klassenwahrscheinlichkeit, so dass die überprüfte Klasse nicht als Ergebnis in Frage kommen kann.

Nur wenn die Merkmalsmatrix $\mathbf{A}_{Fak}$ und die Basismatrix **B** zu der gleichen Musikinstrumentenklasse gehören, ergibt die Projektion der ASE-Koeffizienten charakteristische Muster in der Merkmalsmatrix $\mathbf{A}_{Fak}$, die mit einer hohen a-posteriori Wahrscheinlichkeit bewertet werden und zu einer positiven Klassifikationsentscheidung führen.

Die gesamte Klassifikation erfolgt in Echtzeit, und die zu jedem Musikinstrumentenmodell gehörende a-posteriori Wahrscheinlichkeit wird in einer Tabelle mit den zugehörigen Instrumentennamen ausgegeben. Die Tabelle wird nach jedem klassifizierten Klangereignis aktualisiert, so dass

entsprechend der Bayes'schen Klassifikation das Instrument mit der höchsten a-posteriori Wahrscheinlichkeit als Klassifikationsergebnis an erster Stelle steht.

6.4 POLYPHONES SYSTEM

Das polyphone Klassifikationssystem ist in der Lage, monophone Klangereignisse in polyphoner, komplexer Musik zu identifizieren und einer der im Vorfeld trainierten Klassen zuzuordnen. Für den Aufbau oder Inhalt der Musiksignale oder der verwendeten Klangereignisse werden keinerlei Einschränkungen gemacht. Das System verwendet einen völlig neuartigen Algorithmus, der nicht auf die fehlerträchtige Zerlegung von polyphoner Musik in Einzelklänge bzw. ihre Entmischung angewiesen ist. Die folgenden Abschnitte beschreiben ausgiebig den Aufbau sowie die Verwendung des polyphonen Systems.

6.4.1 AUFBAU

Die Merkmalsextraktion des in Matlab implementierten Offline-Systems basiert ausschließlich auf dem Harmonic Peak Spectrum. Aus den zum Training verwendeten Klangereignissen werden Musikinstrumentenmodelle erstellt, die im Kern auf GMMs aufgebaut sind. Diese Modelle können nicht nur zur Klassifikation verwendet werden, sondern sind auch in der Lage, die trainierten Klangereignisse aus den verwendeten Merkmalen zu resynthetisieren. Hierbei wird für jedes Musikinstrument ein eigenes Modell erstellt, das seinerseits für jeden Halbton, den das Instrument erzeugen kann, ein eigenes GMM enthält.

Aus dem analysierten Audiomaterial wird im Klassifikationsbetrieb zu jedem Zeitpunkt das in seiner Lautstärke dominierende Klangereignis identifiziert und klassifiziert. Somit wird in der Regel das Soloinstrument identifiziert, kurze Pausen des Soloinstruments reichen allerdings aus, um auch die Begleitinstrumente zu identifizieren.

6.4.2 MERKMALSEXTRAKTION

Als Merkmal wird das Harmonic Peak Spectrum aus den ersten 15 Harmonischen des lautesten Klangereignisses extrahiert, so dass sich Merk-

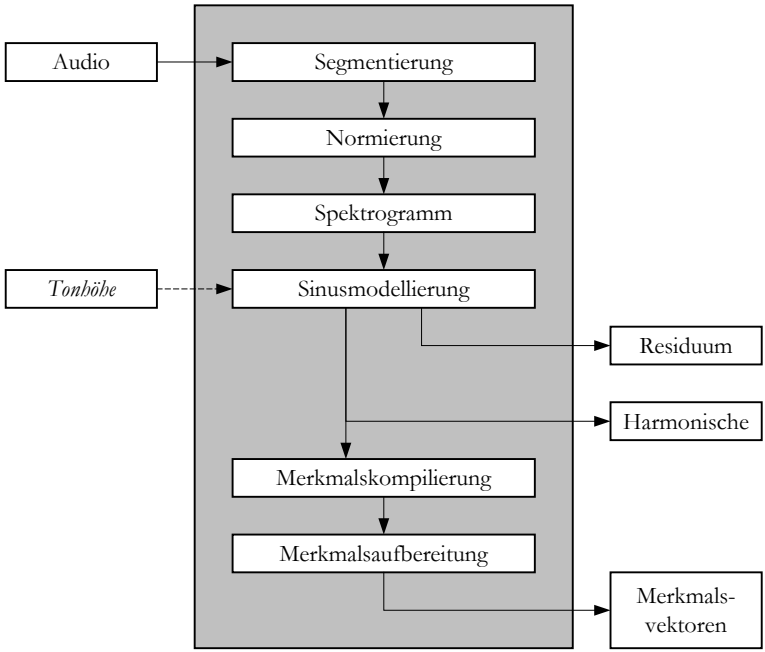


Bild 6.8: Schematischer Aufbau der Merkmalsextraktion des polyphonen Systems

malsvektoren $\mathbf{a}_{HPS}$, mit 30 Werten ergeben (vgl. Abschnitt 4.5.3 und Formel 4.32). Der schematische Aufbau der Merkmalsextraktion ist in Bild 6.8 dargestellt.

Im Trainingsbetrieb wird das zu verarbeitende Audiosignal $x(n)$ im ersten Schritt über den bereits in Abschnitt 6.3.4 beschriebenen Prozess segmentiert. Hierzu wird der Kurzzeiteffektivwert $\tilde{x}(n)$ berechnet, der mit einem adaptiven Schwellenwert $a(n)$ verglichen wird. Die Fenstergröße des Kurzzeiteffektivwerts beträgt hierbei 25 ms, die der adaptiven Schwelle 5 s. Der zum Auslösen eines Segmentes erforderliche Signalabstand beträgt 10 dB. Dieser Schritt entfällt im Klassifikationsbetrieb, damit ein durchgehender Strom von Merkmalsvektoren sichergestellt ist.

Im nächsten Schritt wird der berechnete Kurzzeiteffektivwert benutzt, um das Eingangssignal auf Einheitsleistung zu normieren:

$$x_{-}(n) = \frac{x(n)}{x_{\sim}(n)}. \tag{6.11}$$

Aus dem so gewonnenen Signal $x_{-}(n)$ wird nun das Harmonic Peak Spectrum extrahiert. Als Parameter der STFT werden für die Extraktion eine Hopsize von 20 ms und eine Fensterbreite von 40 ms verwendet. Die Fensterung erfolgt hierbei über ein die Kurzzeitleistung erhaltendes Hann-Fenster. Um eine feine Frequenzauflösung zu erreichen, werden die gefensterten Signalausschnitte mittels Zero-Padding auf eine FFT-Länge von $2^{14} = 16384$ Abtastwerten gebracht. Für eine Signalfrequenz von 44,1 kHz entspricht dies einer FFT-Länge von 371,52 ms.

Für die Faltung zur Glättung des Amplitudenspektrums wird ein 20 Hz breites Gaußfenster verwendet. Die Ermittlung der Grundfrequenz erfolgt über eine Schablone mit Hann-Fenstern als Keulen mit der Breite von vier Halbtönen. Die Schablonen werden für alle Grundfrequenzen zwischen 65,41 Hz und 523,25 Hz (SPN: C2-C5) im Abstand von einem Halbton angewendet.

Für einen optimalen Trainingsbetrieb sollten die Tonhöhen und somit die Grundfrequenzen der präsentierten Klangereignisse als a-priori Informationen zur Verfügung gestellt werden, um evtl. auftretende Fehler bei der automatischen Grundfrequenzerkennung zu umgehen.

Die nach der Extraktion des Harmonic Peak Spectrum verbleibenden Sinuskomponenten bilden ein Residuum. Es beschreibt das ursprüngliche Spektrum, aus dem das analysierte Klangereignis mittels spektraler Subtraktion entfernt wurde. Es kann zur Ermittlung von weiteren Klangereignissen weiterverarbeitet werden [EiSi06].

Im letzten Schritt der Merkmalsextraktion erfolgt eine Aufbereitung der Merkmalsvektoren. Im Trainingsbetrieb werden ihre Werte über einen kleinen Rauschanteil zufällig variiert, wobei die Standardabweichung des verwendeten normalverteilten Rauschens 10 % des jeweiligen Wertes beträgt. Diese Verrauschung führt zu GMMs, die einerseits besser generalisieren, andererseits schneller über den EM-Algorithmus trainiert werden können, da die Komponenten keine Gefahr laufen, zu Singularitäten auf mehrfach belegten Merkmalsvektoren zu kollabieren. Die Verrauschung kann während des Klassifikationsbetriebs entfallen. Abschließend

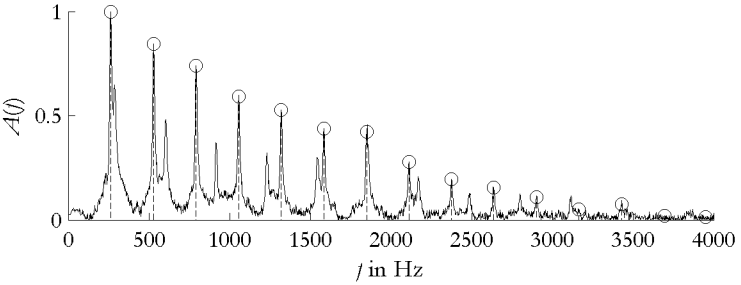


Bild 6.9: Resynthese eines Klangereignisses in einer Klangmischung (schwarz) aus den ersten 15 Harmonischen (gestrichelt)

werden die Leistungen der Harmonischen in jedem Merkmalsvektor derart normiert, dass die Summe aller Werte 1 ergibt.

6.4.3 RESYNTHESEMODELL

Neben der Extraktion des Harmonic Peak Spectrum als Merkmalsvektor wird das analysierte Klangereignis zusätzlich in ein harmonisches Modell überführt. Hierzu wird zu den für das Harmonic Peak Spectrum bereits ermittelten Werten für jede harmonische Sinuskomponente die Phasenlage aus dem Spektrogramm übernommen. Somit müssen für jede Sinuskomponente des Modells drei Werte gespeichert werden: die Frequenz, die Amplitude und die Phasenlage.

Das so gewonnene harmonische Modell bildet ein im Musikinstrumentenmodell enthaltenes Resynthesemodell und ist in der Lage, die analysierten Klangereignisse über eine Additive Resynthese zu reproduzieren (Bild 6.9). Bemerkenswert hierbei ist, dass lediglich die Informationen der ersten 15 Harmonischen des Klangereignisses benötigt werden, um ein Klangereignis in sehr guter Qualität zu resynthetisieren, dies zeigen Hörtests mit musikalischen Experten [EiSI06]. Die gute Qualität der Resynthese belegt, dass die extrahierten Merkmale ein Klangereignis ausreichend beschreiben und dass der verwendete Ansatz zur Merkmalsextraktion aus polyphoner Musik sinnvoll ist.

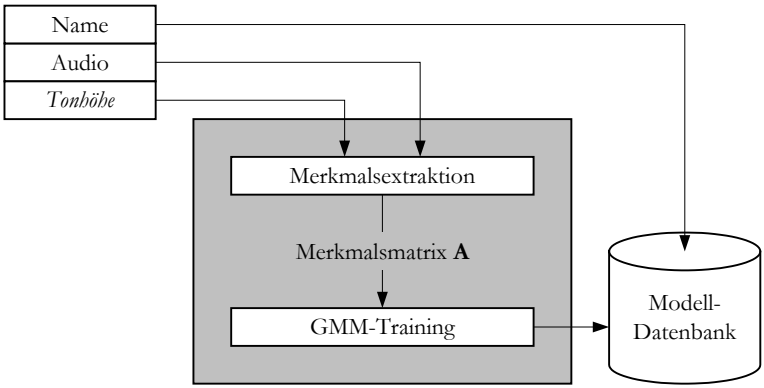


Bild 6.10: Schematischer Aufbau des polyphonen Systems im Trainingsbetrieb

6.4.4 TRAINING

Der grundlegende Aufbau des Systems im Trainingsbetrieb ist in Bild 6.10 dargestellt. Das Audiosignal besteht optimalerweise aus mehreren nacheinander folgenden monophonen Instrumentenklängen, die im Trainingsbetrieb alle von einem Instrument sein müssen. Weiterhin können leistungsfähigere Modelle trainiert werden, wenn die Tonhöheninformation der einzelnen Klangereignisse nicht vom Modell selbst ermittelt werden muss, sondern als zusätzliche a-priori Information angegeben wird. Hierzu kann zusätzlich eine MIDI-Datei eingelesen werden, die die Tonhöheninformation sowie die Segmentgrenzen jedes Klangereignisses enthält. Dies erleichtert und beschleunigt den praktischen Trainingsablauf deutlich.

Aus den Klangereignissen werden im Abstand von 20 ms Merkmalsvektoren über den in Abschnitt 6.4.2 beschriebenen Prozess erzeugt, die in einer Merkmalsmatrix **A** zusammengefasst werden. Hierbei wird für jede Tonhöhe eine eigene Merkmalsmatrix zusammengestellt.

Anschließend wird für jede Tonhöhe mit der dazugehörigen Merkmalsmatrix ein GMM mit drei Komponenten trainiert. Die Modellparameter werden zusammen mit dem Namen des Musikinstruments und der Tonhöhe in der Modelldatenbank gespeichert.

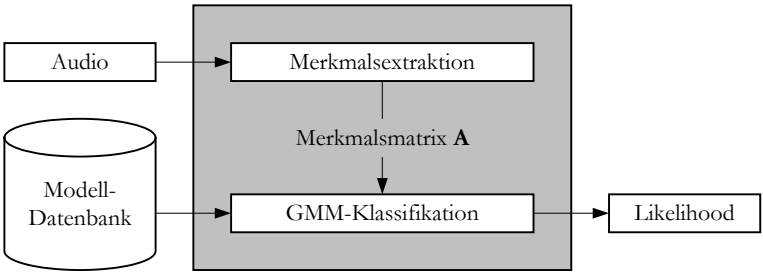


Bild 6.11: Schematischer Aufbau des polyphonen Systems im Klassifikationsbetrieb

Der folgende Beispielaufruf zeigt die Erzeugung eines Geigenmodells mit der MIDI-Tonhöhe 48 (130,81 Hz, SPN: C3) mittels des beschriebenen Prozesses.

```
sViolinModel = Train(...
'name', 'Violin',...
'feature', 'pol_harmonics',...
'wave', 'C:\sounds\violin1\violin1.wav',...
'midi', 'C:\sounds\violin1\violin1.mid',...
'pitch', 48,...
'model', 'gmm', 3);
```

6.4.5 KLASSIFIKATION

Der grundlegende Aufbau des Systems im Klassifikationsbetrieb ist in Bild 6.11 dargestellt. Aus dem polyphonen Audiosignal werden wie beim Training im Abstand von 20 ms Merkmalsvektoren extrahiert. Hiervon werden jeweils die letzten 50 Vektoren in einer Merkmalsmatrix **A** zusammengefasst. Somit enthält die Merkmalsmatrix **A** jeweils Informationen über eine Sekunde Audiomaterial.

Für die aktuelle Merkmalsmatrix **A** wird jeweils eine Klassifikationsentscheidung mit allen dem System bekannten Modellen getroffen. Für jedes Modell wird für alle Merkmalsvektoren eine Verbundauftretenswahrscheinlichkeit berechnet. Das Modell mit der größten Verbundauftretenswahrscheinlichkeit klassifiziert die aktuelle Merkmalsmatrix und ordnet ihr die in der Modell-Datei gespeicherte Tonhöhe sowie den Instrumentennamen zu. Somit erfolgt die Klassifikationsentscheidung als

fortlaufender Prozess und nicht für einzelne Klangereignisse. Diese Vorgehensweise hat den Vorteil, dass das System fehlertoleranter arbeiten kann, da es nicht auf die korrekte Segmentierung bzw. Ermittlung des Anfangs und des Endes eines jeden Klangereignisses angewiesen ist. Hierdurch wird eine fehlerträchtige Zerlegung der polyphonen Musik in Einzelklänge bzw. ihre Entmischung umgangen [Eggi01, BeBi03].

Für die Untersuchung des Systems können die Tonhöheninformation sowie die korrekte Klassenzugehörigkeit der einzelnen in dem Audiosignal vorkommenden Klangereignisse als Ground-Truth angegeben werden. Hierzu kann, ähnlich wie im Trainingsbetrieb, zusätzlich eine MIDI-Datei eingelesen werden, die die Tonhöheninformation und die Segmentgrenzen sowie die korrekte Klassenzugehörigkeit jedes Klangereignisses enthält.

Der folgende Beispielaufruf zeigt die Klassifikation einer Audioaufnahmen durch Geigenmodelle und Gitarrenmodelle mit unterschiedlichen Tonhöhen:

```
Classify(...
'model', sViolinModel36, ...
'model', sViolinModel38, ...
'model', sViolinModel52, ...
'model', sGuitarModel37, ...
'model', sGuitarModel39, ...
'wave', 'C:\sounds\polyphone\demo1.wav',...
'midi', 'C:\sounds\polyphone\demo1.mid',...
'truth', 'C:\sounds\polyphone\demo1.gtr');
```

Die Ausgabe des Systems für den obigen Aufruf könnte folgendermaßen aussehen:

```
Processing 1/55: ... done
Truth:  Pitch: 36 - Violin
Result: Pitch: 36 - Violin
Certainty: 87.7%

Processing 2/55: ... done
Truth:  Pitch: 37 - Guitar
Result: Pitch: 37 - Guitar
Certainty: 81.2%
```

```
Processing 3/55: ... done
Truth:  Pitch: 38 - Violin
Result: Pitch: 38 - Guitar
Certainty: 79.6%

Processing 4/55: ... done
Truth:  Pitch: 39 - Guitar
Result: Pitch: 39 - Guitar
Certainty: 85.9%

Processing 5/55: ... done
Truth:  Pitch: 40 - Violin
Result: Pitch: 52 - Guitar
Certainty: 33.1%

...
```

7 AUSWERTUNG

Die im Rahmen dieser Arbeit entwickelten Systeme wurden in umfangreichen Tests untersucht. Die Testabläufe sowie die erzielten Ergebnisse werden in diesem Kapitel beschrieben und ausgewertet. Im Vordergrund stehen hierbei die von den Systemen erzielten Klassifikationsraten sowie die auftretenden Klassifikationsfehler.

7.1 TESTUMGEBUNG

Für die Evaluierung der entwickelten Systeme wurde durchgehend einkanaliges, PCM-codiertes Audiomaterial mit einer Abtastrate von 44,1 kHz und 16 Bit Auflösung verwendet.

7.1.1 TRAININGSMATERIAL

Das Trainingsmaterial besteht aus monophonen Klängen der folgenden Instrumente: Klavier, Geige, Saxophon, Kontrabass, Fagott, Cello und Trompete. Es hat eine Gesamtlänge von 22 Stunden und 41 Minuten und eine Gesamtgröße von 6,71 GB. Für die Zusammenstellung wurden die folgenden vier Quellen verwendet:

- die Klangbibliothek „Musical Instrument Samples“ von der University of Iowa [Iowa08],
- die Klangbibliothek „Sample Cell II - Library“ der Firma Digidesign [Digi08],

- der General MIDI (GM) kompatible Synthesizer „Bandstand“ der Firma Native Instruments [Nati08],
- der General MIDI (GM) kompatible Synthesizer „Universal Sound Module“ der Firma Steinberg [Stei08].

Der Umfang und die Qualität des Trainingsmaterials sind entscheidend für die mit dem trainierten System erzielbaren Klassifikationsraten. So kann ein Klang nur dann richtig klassifiziert werden, wenn dem System während des Trainings bereits ein ähnlicher Klang präsentiert wurde. Dies zieht auch nach sich, dass die Trainingsdatenmenge ausreichend groß sein muss, was in der vorliegenden Auswertung durch den Umfang des verwendeten Trainingsmaterials jedoch sicher gestellt ist.

7.1.2 EVALUIERUNG

Für die Evaluierung wurde das Trainingsmaterial mit dem Verhältnis 70 : 30 in einen Trainingsdatensatz und einen Testdatensatz unterteilt. Mit dem größeren Trainingsdatensatz wurden die verschiedenen Systeme in einer Trainingsphase trainiert, um anschließend den kleineren, dem System unbekannten Testdatensatz automatisch klassifizieren zu können. Hierbei wurden als Ergebnisse neben der Klassifikationsrate auch die Klassifikationsfehler ermittelt.

Für die Evaluierung der monophonen Systeme wurden die Klänge hierbei direkt zur Klassifikation verwendet. Für die Evaluierung des polyphonen Systems hingegen wurden die Klänge des Testdatensatzes in polyphone Musik überführt. Hierzu wurden sie in einen der verwendeten GM-kompatiblen Synthesizer geladen, dessen Ausgangssignal aufgenommen wurde, während mit ihm verschiedene MIDI-Sequenzen abgespielt wurden. Die so entstandenen polyphonen Sequenzen haben eine Gesamtlänge von 22 Stunden und 17 Minuten und eine Gesamtgröße von 6,59 GB.

Aufgrund der im Training der Modelle vorkommenden Zufallsprozesse und der zufälligen Trennung des Trainingsmaterials im Verhältnis 70 : 30 können die entstehenden Modelle und auch das Klassifikationsergebnis variieren. Um zu überprüfen, ob die Klassifikationsrate der Systeme bei unterschiedlichen Testdurchläufen konsistente Werte annimmt, wurden

die in diesem Abschnitt beschriebenen Testdurchläufe unter Beibehaltung der restlichen Parameter jeweils 30 mal wiederholt.

Die Ergebnisse eines Tests werden jeweils in Prozent in einer Confusion-Matrix dargestellt (vgl. als Beispiel Tabelle 7.1). Hierbei werden die Klassenzugehörigkeiten der analysierten Klänge jeweils durch die Zeilen (E – Eingabe) und das ermittelte Klassifikationsergebnis des Systems jeweils durch die Spalten (A – Ausgabe) angegeben. Neben den detaillierten Angaben der Matrizen werden weiterhin die mittlere Klassifikationsrate sowie als Vergleichswert der so genannte Random Guess aufgeführt, der der Klassifikationsrate bei einer rein zufälligen Wahl des Klassifikationsergebnisses entspricht.

7.2 MONOPHONE KLASSIFIKATION

Für die Auswertung des monophonen Echtzeitsystems (vgl. Abschnitt 6.3) wurde mit den Klängen des Trainingsmaterials für jedes Musikinstrument jeweils ein eigenes HMM trainiert. Hierbei wurden alle zu dem jeweiligen Instrument gehörenden Klänge unabhängig von ihrer Tonhöhe verwendet, was sehr allgemeine Modelle zur Folge hat. Die Eigenschaften des Systems hinsichtlich Parametervariationen wurden in verschiedenen Tests untersucht, deren mittlere Klassifikationsraten zum Überblick in Bild 7.1 aufgeführt sind. Der genaue Aufbau und die detaillierten Ergebnisse der Tests werden in den folgenden Abschnitten beschrieben.

7.2.1 STANDARDPARAMETER

Den Ausgangspunkt für die Auswertung des monophonen Echtzeitsystems bilden die in Abschnitt 6.3 genannten Standardparameter, die heuristisch als Optimum ermittelt wurden (Test 1). Die Fenstergröße des Kurzzeiteffektivwerts beträgt hierbei 25 ms, die der adaptiven Schwelle 5 s. Der Schwellenwertfaktor c aus Gleichung 6.2 hat den Wert 1,5 mit Attack- und Release-Zeiten von jeweils 50 ms. Die STFT wird mit einer Hopsize von 10 ms und einer Fensterbreite von 30 ms berechnet, was einem Overlap von 66,7 % entspricht. Das ASE wird im Bereich zwischen 62,5 Hz und 8 kHz für 28 Bänder berechnet, so dass sich 30 ASE-Koeffizienten ergeben. Aus diesen gewählten Parametern ergibt sich für

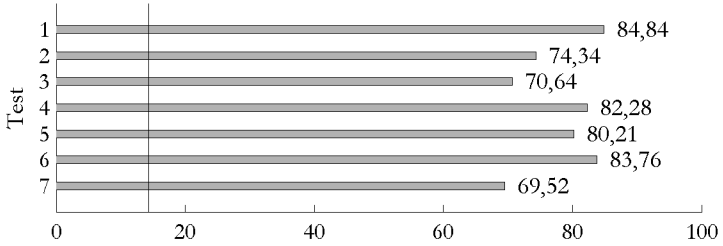


Bild 7.1: Mittlere Klassifikationsraten der mit dem monophonen Echtzeitsystem durchgeführten Tests in Prozent, Random Guess: 14,28 % (durchgezogene Linie)

die mittlere Anzahl der ICA-Komponenten ein Wert von 9,3 und für die mittlere Anzahl der HMM-Zustände ein Wert von 4,8.

Die Klassifikationsergebnisse zeigt Tabelle 7.1. Mit einer mittleren Klassifikationsrate von 84,84 % liefert das System bezogen auf die große Trainingsdatenmenge und den damit verbundenen Umfang der klassifizierbaren Klänge sehr gute Ergebnisse.

Die meisten falsch klassifizierten Klänge werden dem Saxophon zugeordnet. Dies betrifft vor allem die Klänge des Cellos mit einem Klassifikationsfehler von 27,47 %. Die Ursache hierfür ist, dass das Spektrum des Saxophons sehr breit ist und dass Saxophonklänge sehr variationsreich sind. Somit verhält sich die Saxophonklasse von allen Klassen am ehesten wie eine allgemeine Geräuschklasse, die Ausreißer von anderen Klassen aufnimmt.

7.2.2 SPEKTROGRAMMPARAMETER

Um den Einfluss der Frequenzauflösung des Spektrogramms auf das Verhalten des Systems und die Klassifikationsergebnisse zu untersuchen, wurden neben dem Standardwert für die Fenstergröße von 30 ms ebenfalls Werte von 40 ms und 20 ms verwendet, während den restlichen Parametern Standardwerte zugeordnet wurden. Dies bedeutet insbesondere für den Overlap einen gleichbleibenden Wert von 66,7 %.

Durch den gleichbleibenden Overlap ändert sich die Hopsize bei einer Fensterbreite von 40 ms auf einen Wert von 13,3 ms (Test 2). Durch die

Tabelle 7.1: Test 1, Klassifikationsraten des monophonen Echtzeitsystems in Prozent bei Standardparametern, mittlere Klassifikationsrate: 84,84 %, Random Guess: 14,28 %

↓E\A→	Klavier	Geige	Cello	Kontra-bass	Saxo-phon	Fagott	Trom-pete
Klavier	**76,07**	0	9,73	0	12,97	1,23	0
Geige	0	**75,53**	0,37	0,90	16,63	6,37	0,20
Cello	0	0	**71,93**	0,60	27,47	0	0
Kontrabass	0	0	0	**96,27**	3,73	0	0
Saxophon	0	0	1,50	0,60	**97,90**	0	0
Fagott	1,15	0	1,00	0	12,04	**85,81**	0
Trompete	0	0,30	4,93	0	1,70	2,67	**90,40**

längeren Fenster vergrößert sich auch die Frequenzauflösung, bei sinkender Zeitauflösung. Die Klassifikationsergebnisse zeigt Tabelle 7.2. Die mittlere Klassifikationsrate sinkt hierbei gegenüber dem System mit Standardparametern um 10,50 % auf 74,34 %. Das Sinken der Klassifikationsrate in diesem Test beruht auf der schlechteren Zeitauflösung, die dazu führt, dass die extrahierten Merkmale insgesamt geräuschartiger werden und vor allem sehr kurze Klänge ohne Haltephase sich nicht mehr ausreichend abbilden lassen. So zeigen die Fehlklassifikationen, dass der Charakter des Saxophones als allgemeine Geräuschklasse weiter angewachsen ist, was vor allem die Fehlklassifikationen des Cellos betrifft.

Bei einer Fensterbreite von 20 ms ändert sich die Hopsize durch den gleichbleibenden Overlap auf einen Wert von 6,67 ms (Test 3). Durch die kürzeren Fenster verkleinert sich auch die Frequenzauflösung, wohingegen die Zeitauflösung ansteigt. Die mittlere Klassifikationsrate sinkt in diesem Test gegenüber dem System mit Standardparametern noch stärker um 14,20 % auf 70,64 %. Das Sinken der Klassifikationsrate lässt sich darauf zurückführen, dass sich die Obertöne im Spektrogramm nicht mehr ausreichend abbilden lassen und alle Klänge dadurch für das System ähnlicher werden, was Fehlklassifikationen nach sich zieht.

Die gegenüber Test 1 schlechteren Klassifikationsergebnisse der Tests 2 und 3 zeigen, dass von einer Hopsize von 10 ms und einer Fensterbreite von 30 ms nicht abgewichen werden sollte. Diese Werte stellen auch die

Tabelle 7.2: Test 2, Klassifikationsraten des monophonen Echtzeitsystems in Prozent bei einer Fensterbreite von 40 ms, mittlere Klassifikationsrate: 74,34 %, Random Guess: 14,28 %

↓E\A→	Klavier	Geige	Cello	Kontra-bass	Saxo-phon	Fagott	Trom-pete
Klavier	**76,67**	3,33	6,67	6,67	3,33	6,67	0
Geige	0	**76,67**	3,33	6,67	10	3,33	0
Cello	0	0	**50,00**	3,33	46,67	0	0
Kontrabass	0	0	0	**80,00**	20,00	0	0
Saxophon	0	0	0	3,33	**96,67**	0	0
Fagott	0	7,41	3,70	14,81	3,70	**70,37**	0
Trompete	0	0	3,33	10	16,67	0	**70,00**

Empfehlung des MPEG-7-Standards für das Erstellen von Spektrogrammen dar [ISO15938].

7.2.3 ICA-PARAMETER

In weiteren Tests wurde der Einfluss der verwendeten ICA-Komponenten untersucht. Hierbei wurde die Anzahl der Komponenten nicht wie in den anderen Tests so ermittelt, dass die Summe der Merkmalssignale 90 % der Gesamtvarianz ergeben (vgl. Abschnitt 6.3.7 und Formel 6.9), sondern die Anzahl der Komponenten wurde im Vorfeld auf einen festen Wert von 15 bzw. 6 festgelegt. Im Vergleich dazu ergab sich in Test 1 bei einer varianzabhängigen Wahl der ICA-Komponenten eine mittlere Anzahl von 9,3.

Bei einer festen Anzahl von 15 ICA-Komponenten (Test 4) sinkt die mittlere Klassifikationsrate gegenüber dem System mit Standardparametern geringfügig um 2,56 % auf 82,28 % (Tabelle 7.3). Die für das Training zur Verfügung stehenden Daten können aufgrund der erhöhten Anzahl der ICA-Komponenten detaillierter als im varianzabhängigen Betrieb dargestellt werden, allerdings enthalten die meisten der zusätzlichen ICA-Komponenten nur wenig Informationen und transportieren größtenteils Rauschen, was den Rückgang der mittleren Klassifikationsrate erklärt. Besonders wirkt sich dies wieder auf die Unterscheidung von Cello und Saxophon aus, wobei sich für das Cello ein Klassifikationsfehler von 60,00 % ergibt.

Tabelle 7.3: Test 4, Klassifikationsraten des monophonen Echtzeitsystems in Prozent bei 15 ICA-Komponenten, mittlere Klassifikationsrate: 82,28 %, Random Guess: 14,28 %

↓E\A→	Klavier	Geige	Cello	Kontra-bass	Saxo-phon	Fagott	Trom-pete
Klavier	**86,67**	0	0	0	13,33	0	0
Geige	0	**96,67**	3,33	0	0	0	0
Cello	0	0	**40,00**	0	60,00	0	0
Kontrabass	0	0	0	**90,00**	10,00	0	0
Saxophon	0	0	0	3,33	**96,67**	0	0
Fagott	0	0	0	0	7,41	**92,60**	0
Trompete	0	0	6,67	0	20,00	0	**73,33**

Bei einer festen Anzahl von 6 ICA-Komponenten (Test 5) sinkt die mittlere Klassifikationsrate gegenüber dem System mit Standardparametern um 4,63 % auf 80,21 %. Eine Verringerung der ICA-Komponenten stellt grundsätzlich eine Reduktion des Informationsgehalts dar, so dass bei einem Betrieb mit weniger ICA-Komponenten als sich im varianzabhängigen Fall ergeben, ein Sinken der Klassifikationsrate zu erwarten ist. Dass die Klassifikationsrate jedoch nicht deutlich stärker absinkt, ergibt sich aus der Tatsache, dass ein Großteil der Informationen in der Regel bereits von den ersten drei ICA-Komponenten transportiert wird.

Wird beispielsweise aus technischen Gründen ein System gefordert, dass eine konstante Anzahl von ICA-Komponenten aufweist, kann aufgrund der Ergebnisse von Test 5 eine Festlegung auf 6 Komponenten sinnvoll sein, da sich die mittlere Klassifikationsrate hierbei nur um 4,63 % verringert, der benötigte Speicherplatz und der Rechenaufwand sich allerdings um ca. 35 % reduzieren, was den echtzeitfähigen Betrieb erleichtert. Eine Erhöhung auf mehr als 9 Komponenten ist hingegen nicht sinnvoll, wie aus den Ergebnissen von Test 4 hervorgeht, da sich hierbei suboptimale Modelle ergeben und die mittlere Klassifikationsrate wieder sinkt, Speicherplatz und Rechenaufwand allerdings um 60 % zunehmen, was den Echtzeitbetrieb erschwert.

7.2.4 HMM-Parameter

Um den Einfluss der Anzahl der HMM-Zustände auf das Verhalten des Systems zu untersuchen, wurden in weiteren Tests die Anzahl der Zu-

Tabelle 7.4: Test 6, Klassifikationsraten des monophonen Echtzeitsystems in Prozent bei 8 HMM-Zuständen, mittlere Klassifikationsrate: 83,76 %, Random Guess: 14,28 %

↓E\A→	Klavier	Geige	Cello	Kontra-bass	Saxo-phon	Fagott	Trom-pete
Klavier	**70,00**	13,33	0	3,33	13,33	0	0
Geige	3,33	**96,67**	3,33	0	0	0	0
Cello	0	13,33	**70,00**	3,33	13,33	0	0
Kontrabass	0	0	0	**83,33**	16,67	0	0
Saxophon	0	0	0	6,67	**93,33**	0	0
Fagott	3,70	0	0	0	0	**96,30**	0
Trompete	0	0	3,33	0	20,00	0	**76,67**

stände nicht dynamisch aus der Anzahl der ICA-Komponenten ermittelt (vgl. Abschnitt 6.3.8), sondern im Vorfeld auf einen festen Wert von 8 bzw. 4 festgelegt. Im Vergleich dazu ergab sich in Test 1 bei einer von den ICA-Komponenten abhängigen, dynamischen Wahl der HMM-Zustände eine mittlere Anzahl von 4,8.

Bei einer festen Anzahl von 8 HMM-Zuständen (Test 6) sinkt die mittlere Klassifikationsrate gegenüber dem System mit Standardparametern geringfügig um 1,08 % auf 83,76 % (Tabelle 7.4). Durch die im Vergleich zum dynamischen Betrieb erhöhte Anzahl der HMM-Zustände lassen sich komplexere Modelle erzeugen. Diese neigen teilweise allerdings schon zur Überanpassung, was sich in einer sinkenden Klassifikationsrate ausdrückt.

Bei einer festen Anzahl von 4 HMM-Zuständen (Test 7) sinkt die mittlere Klassifikationsrate gegenüber dem System mit Standardparametern deutlich um 15,32 % auf 69,52 %. Diese Verschlechterung ergibt sich aus der Tatsache, dass die meisten Modelle mit nur vier Zuständen nicht in der Lage sind, alle Aspekte der Klassifikationsaufgabe im Merkmalsraum abzubilden.

Aufgrund der Ergebnisse der Tests 6 und 7 wird deutlich, dass es nicht sinnvoll ist, von einer dynamischen Wahl der HMM-Zustände abzuweichen, da dies grundsätzlich zu schlechteren Klassifikationsergebnissen führt. Werden allerdings beispielsweise aus technischen Gründen Modelle mit einer konstanten Anzahl von HMM-Zuständen gefordert, so sollte aufgrund der Ergebnisse von Test 6 eine Festlegung auf 8 HMM-

Zustände gewählt werden, da sich hierbei fast die gleichen Klassifikationsergebnisse wie in einem System mit einer dynamischen Wahl der HMM-Zustände ergeben.

7.3 POLYPHONE KLASSIFIKATION

Für die Auswertung des polyphonen Systems (vgl. Abschnitt 6.4) wurden mit den Klängen des Trainingsmaterials für jedes Musikinstrument ein eigenes Musikinstrumentenmodell trainiert. Die jeweiligen Musikinstrumentenmodelle enthalten hierbei für jeden Halbton ein eigenes GMM-Modell. Die Eigenschaften des Systems für verschiedene Klassifikationsszenarien wurden in verschiedenen Tests untersucht, deren mittlere Klassifikationsraten zum Überblick in Bild 7.2 aufgeführt sind. In einigen Tests wurden dem System die korrekten Tonhöhen der zu klassifizierenden Klänge als a-priori Information über MIDI-Dateien zur Verfügung gestellt. Hierdurch konnten Fehler der automatischen Tonhöhenerkennung ausgeschlossen werden, und das eigentliche Klassifikationssystem wurde getrennt ausgewertet. In den Tests in denen die Tonhöhe automatisch vom System ermittelt werden musste, wurde sie mit einer mittleren Genauigkeit von 57,88 % richtig erkannt. Der genaue Aufbau und die detaillierten Ergebnisse der Tests werden in den folgenden Abschnitten beschrieben.

7.3.1 TRAININGSPARAMETER

Das für die Tests verwendete System wurde mit den in Abschnitt 6.4 genannten Standardparametern trainiert. Die Fenstergröße des Kurzzeiteffektivwerts beträgt hierbei 25 ms, die der adaptiven Schwelle 5 s. Der ein Segment auslösende Signalabstand beträgt 10 dB. Die STFT wird mit einer Hopsize von 20 ms und einer Fensterbreite von 40 ms berechnet, was einem Overlap von 50 % entspricht. Die Glättung des Amplitudenspektrums erfolgt über ein 20 Hz breites Gaußfenster. Für die Schablonen zur Tonhöhenerkennung werden Hann-Fenster mit einer Breite von vier Halbtönen verwendet, wobei die korrekten Tonhöhen der Trainingsklänge über eine MIDI-Datei zur Verfügung gestellt werden. Das extrahierte Harmonic Peak Spectrum wird aus 15 Harmonischen gebildet und erzeugt Merkmalsvektoren mit 30 Werten. Pro Instrumentenmodell

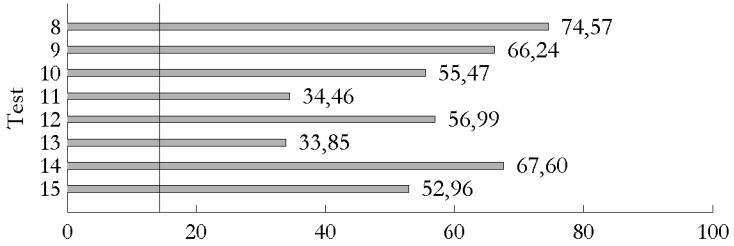


Bild 7.2: Mittlere Klassifikationsraten der mit dem polyphonen System durchgeführten Tests in Prozent, Random Guess: 14,28 % (durchgezogene Linie)

wird für jeden Halbton im Bereich von 65,41 Hz bis 523,25 Hz (SPN: C2-C5) ein eigenes GMM mit drei Komponenten trainiert.

7.3.2 EINZELKLÄNGE

Als Vergleichsbasis wurde das polyphone System in einem ersten Test unter optimalen Bedingungen untersucht, um eine Obergrenze für die mittlere Klassifikationsrate zu ermitteln. Hierbei wurden die Klänge des Testdatensatzes direkt als einzelne, monophone Klänge zur Klassifikation verwendet, deren Tonhöhe dem System über eine MIDI-Datei zur Verfügung gestellt wurde (Test 8).

Tabelle 7.5 zeigt die Klassifikationsergebnisse. Mit einer mittleren Klassifikationsrate von 74,57 % liefert das System bezogen auf die große Trainingsdatenmenge und den damit verbundenen Umfang der klassifizierbaren Klänge sehr gute Ergebnisse. Die Fehlklassifikationen der Instrumente untereinander sind relativ ausgewogen, wobei die meisten falsch klassifizierten Klänge dem Kontrabass zugeordnet werden. Dies betrifft insbesondere die Klänge des Klaviers mit einem Klassifikationsfehler von 18,06 %, was vor allem an der großen klanglichen Ähnlichkeit der tiefen Töne liegt. Die zur Klangerzeugung verwendeten gespannten Saiten haben hier teilweise einen sehr ähnlichen Aufbau.

In einem weiteren Test wurden dem System die Tonhöhen der einzelnen Klänge nicht zur Verfügung gestellt, sondern automatisch ermittelt (Test 9). Die mittlere Klassifikationsrate sinkt hierbei gegenüber Test 8

Tabelle 7.5: Test 8, Klassifikationsraten des polyphonen Systems in Prozent für monophone Einzelklänge, mittlere Klassifikationsrate: 74,57 %, Random Guess: 14,28 %

↓E\A→	Klavier	Geige	Cello	Kontra-bass	Saxo-phon	Fagott	Trom-pete
Klavier	**59,52**	0	8,73	18,06	0,40	8,73	4,56
Geige	0,20	**63,69**	15,68	10,91	1,59	3,57	4,37
Cello	0,79	4,17	**74,01**	12,90	0	5,16	2,98
Kontrabass	0,99	1,79	11,51	**82,34**	0	1,39	1,98
Saxophon	0,20	0,79	5,16	4,96	**82,54**	1,39	4,96
Fagott	0,60	0	0,40	0,60	0	**93,45**	4,96
Trompete	1,39	2,18	8,33	8,33	0	13,29	**66,47**

um 8,33 % auf 66,24 %. In diesem Test arbeitet das polyphone System im gleichen Klassifikationsszenario wie das monophone Echtzeitsystem in Test 1. Die mittlere Klassifikationsrate des polyphonen Systems ist hierbei 18,60 % kleiner als die des monophonen Echtzeitsystems. Dies ist darauf zurückzuführen, dass das polyphone System aufgrund seines inneren Aufbaus und der Verwendung von harmonischen anstatt spektralen Merkmalen zwar allgemeiner und vor allem in polyphonen Szenarien verwendet werden kann, im monophonen Fall allerdings Merkmale mit weniger Aussagekraft zur Auswertung hat.

7.3.3 MUSIKSTÜCKE

Für die Klassifikation von Musikinstrumentenklängen aus komplexer Musik wurden die Klänge des Testdatensatzes in polyphone, sechsstimmige Musik überführt. Hierzu wurden sie in einen der verwendeten GM-kompatiblen Synthesizer geladen, dessen Ausgangsignale aufgenommen wurden, während mit ihm eine MIDI-Sequenz abgespielt wurde. Die verwendeten MIDI-Sequenzen stammen aus der von Batke erstellten MIDI-Datenbank [Batk06]. Die in den Musikstücken vorkommenden Klänge wurden anschließend von dem polyphonen System klassifiziert. Hierbei wurden dem System in einem Test die korrekten Tonhöhen der zu klassifizierenden Klänge als a-priori Information über MIDI-Dateien zur Verfügung gestellt, im einem weiteren Test mussten sie vom System automatisch ermittelt werden.

Tabelle 7.6: Test 10, Klassifikationsraten des polyphonen Systems in Prozent für komplexe Musiksignale, mittlere Klassifikationsrate: 55,47 %, Random Guess: 14,28 %

↓E\A→	Klavier	Geige	Cello	Kontra-bass	Saxo-phon	Fagott	Trom-pete
Klavier	**31,07**	47,14	11,79	0,71	4,64	2,14	2,50
Geige	0	**63,22**	16,09	2,30	10,35	0	8,05
Cello	17,58	8,79	**43,96**	5,49	21,98	1,10	1,10
Kontrabass	2,24	13,81	17,91	**38,43**	20,90	4,10	2,61
Saxophon	0	0	2,70	0	**97,30**	0	0
Fagott	1,28	3,85	0	1,28	10,90	**82,05**	0,64
Trompete	4,84	16,13	12,90	0	22,58	11,29	**32,26**

Die Klassifikationsergebnisse bei bekannter Tonhöhe (Test 10) zeigt Tabelle 7.6. Da in diesem Test keine Fehler der automatischen Tonhöhenerkennung auftreten können, lässt sich so das eigentliche Klassifikationssystem getrennt auswerten. Die mittlere Klassifikationsrate beträgt 55,47 %, was ein sehr guter Wert für sechsfach polyphone Musik ist.

Die meisten falsch klassifizierten Klänge werden wie bei dem monophonen Echtzeitsystem dem Saxophon zugeordnet, da es sich aufgrund seines breiten Spektrums und seines Variationsreichtums am ehesten als allgemeine Geräuschklasse verhält, die Ausreißer von anderen Klassen aufnimmt. Ein großer Klassifikationsfehler entsteht weiterhin für Klavierklänge, die mit 47,14 % als Geigenklänge klassifiziert werden. Die Ursache hierfür liegt wieder in der ähnlichen Klangerzeugung durch schwingende Saiten, die vor allem für leise Töne ähnliche spektrale Muster erzeugt. Hierbei liegt die Hauptenergie im Grundton, was sinusartige Klänge nach sich zieht.

Wird die Tonhöheninformation dem System nicht als a-priori Information zur Verfügung gestellt, sondern automatisch ermittelt (Test 11), so sinkt die mittlere Klassifikationsrate auf 34,46 %. Hier machen sich nun verstärkt Fehler in der automatischen Tonhöhenerkennung bemerkbar, so dass teilweise zur Klassifikation eines Klanges völlig falsche Modelle herangezogen werden. Die mittlere Klassifikationsrate ist in diesem Test zwar immer noch 2,4 mal größer als der Random Guess, allerdings zeigt der Test die Grenzen des Systems für eine automatische polyphone Klassifikation auf Notenebene auf. Eine generelle Aussage über die in

Tabelle 7.7: Test 12, Klassifikationsraten des polyphonen Systems in Prozent für komplexe Zufallsmusik, mittlere Klassifikationsrate: 56,99 %, Random Guess: 14,28 %

↓E\A→	Klavier	Geige	Cello	Kontra-bass	Saxo-phon	Fagott	Trom-pete
Klavier	**20,71**	25,71	25,00	13,57	13,57	0	1,43
Geige	4,17	**61,81**	13,89	10,42	6,94	1,39	1,39
Cello	8,89	5,19	**69,63**	8,15	6,67	1,48	0
Kontrabass	4,38	6,57	11,68	**66,42**	10,95	0	0
Saxophon	4,23	4,23	0	0,70	**90,84**	0	0
Fagott	7,97	7,25	5,07	3,62	10,14	**39,13**	26,81
Trompete	7,19	6,47	12,23	10,79	8,63	4,32	**50,36**

einem Lied vorkommenden Instrumente im Sinne einer multimedialen Metadatengewinnung ist allerdings auch mit einer mittleren Klassifikationsrate von 34,46 % noch gut möglich.

7.3.4 ZUFALLSMUSIK

Um eine größere Datenmenge mit polyphoner Musik zum Testen verwenden zu können, wurde das System weiterhin mit sechsstimmiger Zufallsmusik untersucht. Für die verwendete Zufallsmusik wurden MIDI-Sequenzen generiert, die die rhythmischen und harmonischen Strukturen sowie das Noten-/Pausenverhältnis von westlicher Musik besitzen, in der Regel allerdings keine für den menschlichen Hörer angenehme Melodien oder Akkordfolgen enthalten [BaEi06]. Die generierten MIDI-Sequenzen wurden, wie bei der Auswertung mit polyphonen Musikstücken im vorangegangenen Abschnitt, in Musiksignale überführt, indem sie von einem der verwendeten GM-kompatiblen Synthesizer abgespielt wurden, während dieser die Klänge des Testdatensatzes geladen hatte. Für die Klassifikation wurden dem System in einem Test die korrekten Tonhöhen der zu klassifizierenden Klänge als a-priori Information über MIDI-Dateien zur Verfügung gestellt, in einem weiteren Test mussten sie vom System automatisch ermittelt werden.

Die Klassifikationsergebnisse bei bekannter Tonhöhe (Test 12) zeigt Tabelle 7.7. Da auch in diesem Test keine Fehler der automatischen Tonhöhenerkennung auftreten können, trifft das Klassifikationsergebnis direkt eine Aussage über die Qualität des eigentlichen Klassifikationssys-

tems. Die mittlere Klassifikationsrate beträgt 56,99 %, was ein sehr guter Wert für sechsfach polyphone Musik ist. Die meisten falsch klassifizierten Klänge werden erneut dem Saxophon zugeordnet, der größte Klassifikationsfehler tritt mit 25,71 % beim Klavier auf. Da die mittlere Klassifikationsrate nur um 1,52 % von der in Test 10 für real komponierte polyphone Musik erzielten abweicht, ist weiterhin belegt, dass die generierte Zufallsmusik tatsächlich den Charakter von real komponierter Musik beibehält.

Muss die Tonhöheninformation vom System automatisch ermittelt werden (Test 13), sinkt die mittlere Klassifikationsrate auf 33,85 %. Hier machen sich wie bei real komponierter polyphoner Musik in Test 11 verstärkt Fehler in der automatischen Tonhöhenerkennung bemerkbar. Auch die in diesem Test erreichte mittlere Klassifikationsrate weicht nur um 0,61 % von der in Test 11 ab. Somit gilt auch hier, dass die Klassifikation zwar nicht mehr ausreichend genau ist, um eine Klassifikation auf Notenbasis durchzuführen, eine generelle Aussage über die in einem Lied vorkommenden Instrumente im Sinne einer multimedialen Metadatengewinnung allerdings noch gut möglich ist.

7.3.5 ZWEIKLÄNGE

Die schwierigste Aufgabe im Rahmen der Identifikation und Klassifikation von Musikinstrumentenklängen ergibt sich bei der Verarbeitung von Akkorden, da sich hier viele Obertöne der einzelnen Klänge überlagern und so neue eigenständige Klänge entstehen. Um das System mit Zweiklängen testen zu können, wurden zufällig ausgewählte Klänge des Testdatensatzes in ihrer Lautstärke auf ein gleiches Niveau gebracht und gemischt. Diese gemischten Klänge wurden anschließend vom System klassifiziert (Test 14). Für die Klassifikation wurden dem System die korrekten Tonhöhen der zu klassifizierenden Klänge als a-priori Information über MIDI-Dateien zur Verfügung gestellt, da eine Grundfrequenzerkennung bei Akkorden ein nicht triviales Problem darstellt, für das bisher keine umfassende Lösung existiert [MaFe02, Naga03, EgBr04b].

Tabelle 7.8 zeigt die Klassifikationsergebnisse. Die mittlere Klassifikationsrate ist mit 67,60 % nur um 6,97 % kleiner als die mittlere Klassifikationsrate für monophone Einzelklänge aus Test 8, was bezogen auf die

Tabelle 7.8: Test 14, Klassifikationsraten des polyphonen Systems in Prozent für Zweiklänge, mittlere Klassifikationsrate: 67,60 %, Random Guess: 14,28 %

↓E\A→	Klavier	Geige	Cello	Kontra-bass	Saxo-phon	Fagott	Trom-pete
Klavier	**39,58**	25,00	2,08	8,33	20,83	2,08	2,08
Geige	2,17	**89,13**	2,17	0	6,52	0	0
Cello	10,64	12,77	**63,83**	6,38	4,26	0	2,13
Kontrabass	16,67	8,33	8,33	**62,50**	4,17	0	0
Saxophon	0	0	4,17	6,25	**87,50**	0	2,08
Fagott	6,25	6,25	2,08	0	12,50	**60,42**	12,50
Trompete	4,26	8,51	2,13	4,26	4,26	6,38	**70,21**

schwierige Aufgabenstellung ein sehr gutes Ergebnis darstellt. Die Klassifikationsraten der einzelnen Instrumente sind untereinander relativ ausgewogen, wobei die korrekte Klassifikation der Klavierklänge mit einer Klassifikationsrate von 39,58 % die größte Schwierigkeit darstellt. Ein Grund hierfür ist, dass bei der Klangerzeugung neben den direkt angeschlagenen Saiten grundsätzlich auch andere in einem harmonischen Verhältnis zum Grundton stehende Saiten über Resonanzen mitschwingen.

7.3.6 DREIKLÄNGE

Abschließend wurde das System mit künstlich erzeugten Dreiklängen getestet. Hierfür wurden wie beim Test mit Zweiklängen zufällig ausgewählte Klänge des Testdatensatzes in ihrer Lautstärke angepasst und gemischt. Diese gemischten Klänge wurden anschließend vom System klassifiziert (Test 15). Auch in diesem Test wurden dem System die korrekten Tonhöhen der zu klassifizierenden Klänge als a-priori Information über MIDI-Dateien zur Verfügung gestellt.

Die Klassifikationsergebnisse zeigt Tabelle 7.9. Die mittlere Klassifikationsrate sinkt gegenüber Test 14 mit Zweiklängen um 14,64 % auf 52,96 %, was für die Klassifikation von Dreiklängen ein sehr gutes Ergebnis darstellt. Wie im Test mit Zweiklängen stellt die Erkennung der Klavierklänge mit einer Klassifikationsrate von 27,66 % die größte Schwierigkeit dar, aber auch die Erkennung der Fagottklänge ist mit einer Klassifikationsrate von 35,71 % schwierig.

Tabelle 7.9: Test 15, Klassifikationsraten des polyphonen Systems in Prozent für Dreiklänge, mittlere Klassifikationsrate: 52,96 %, Random Guess: 14,28 %

↓E\A→	Klavier	Geige	Cello	Kontra-bass	Saxo-phon	Fagott	Trom-pete
Klavier	**27,66**	31,91	12,77	2,13	23,40	0	2,13
Geige	1,75	**71,93**	7,02	5,26	10,53	0	3,51
Cello	9,33	10,67	**50,67**	14,67	5,33	2,67	6,67
Kontrabass	5,80	13,04	17,39	**49,28**	8,70	0	5,80
Saxophon	5,26	1,32	1,32	7,89	**77,63**	3,95	2,63
Fagott	9,52	9,52	14,29	2,38	15,48	**35,71**	13,10
Trompete	2,41	12,05	7,23	8,43	8,43	3,61	**57,83**

Die mittlere Klassifikationsrate des Tests mit Dreiklängen hat einen ähnlichen Wert wie die für die Verarbeitung von Musikstücken in Test 10 und Zufallsmusik in Test 12, obwohl diese sechsfach polyphon sind und die verarbeiteten Dreiklänge lediglich dreifach polyphon sind. Dieses Ergebnis belegt einerseits die Schwierigkeit bei der Verarbeitung von reinen Akkorden, andererseits aber auch, dass selbst in sechsfach polyphoner Musik genügend Bereiche mit geringerer Polyphonie vorkommen, um gute mittlere Klassifikationsraten zu erzielen.

8 ZUSAMMENFASSUNG

Im Rahmen dieser Arbeit wurden Verfahren der Musikinstrumentenidentifikation und Klassifikation insbesondere in komplexer, polyphoner Musik untersucht. Hierzu wurden alle in diesem Zusammenhang relevanten musikalischen Phänomene und die musikalischen Parameter sowie die physikalischen Eigenschaften von Musikinstrumenten und die daraus resultierenden Klangeigenschaften erläutert. Weiterhin wurde eine Vielzahl von Verfahren zur Extraktion relevanter Merkmale aus Musikstücken und den in ihnen vorkommenden Musikinstrumentenklängen analysiert sowie verschiedene Verfahren der Mustererkennung bezüglich ihrer Verwendbarkeit im Zusammenhang mit der Identifikation und Klassifikation von Musikinstrumentenklängen untersucht. Die im Rahmen dieser Arbeit implementierten Klassifikationssysteme wurden in ihrer Funktionsweise ausführlich beschrieben und abschließend hinsichtlich ihrer Klassifikationsergebnisse umfangreich ausgewertet.

8.1 MUSIKINSTRUMENTENERKENNUNG

Die als ein Ergebnis dieser Arbeit entwickelten Klassifikationssysteme leisten einen wesentlichen Beitrag zur Erforschung und Lösung des Problems der Musikinstrumentenidentifikation und Klassifikation.

Das monophone Experimentiersystem ist aufgrund seines modularen, offenen Charakters und seiner einfachen Anwendbarkeit hervorragend geeignet für weitere Untersuchungen und Experimente im Rahmen der Merkmalsextraktion, Merkmalsaufbereitung und Klassifikation. Mit ihm

lassen sich verschiedenste Verfahren sehr effizient und einfach sowohl einzeln als auch im Zusammenspiel testen.

Das monophone Echtzeitsystem empfiehlt sich durch seine sehr guten Klassifikationsergebnisse mit mittleren Klassifikationsraten von 84,84 % und seine Echtzeitfähigkeit als Benchmark für ähnliche Verfahren. Durch die Kompatibilität zu den Standards MPEG-7 und VST stellt es weiterhin eine direkt umsetzbare Studie für den industriellen Gebrauch dar und kann bereits in der vorliegenden Implementierung direkt für die automatische Metadatengewinnung nach MPEG-7 eingesetzt werden.

Das polyphone Klassifikationssystem liefert einen wesentlichen Beitrag auf einem bisher wenig erforschten Gebiet. Für sechsfach polyphone Musik erzielt das System je nach Klassifikationsszenario mittlere Klassifikationsraten von bis zu 56,99 %, wobei die Tonhöhen der zu klassifizierenden Klänge mit einer mittleren Genauigkeit von 57,88 % richtig erkannt werden. Die Klassifikationsergebnisse hängen noch stark von dem Ergebnis der Tonhöhenerkennung ab, so dass ein verbessertes Verfahren der Tonhöhenerkennung hier direkt eine Erhöhung der mittleren Klassifikationsrate nach sich ziehen würde. Die benutzten Algorithmen zur Merkmalsextraktion und Klassifikation sind sehr robust gegenüber Störeinflüssen durch überlappende oder gleichzeitig auftretende Klänge oder Geräusche. Neben der reinen Klassifikation ist das System weiterhin in der Lage, die während des Trainings erlernten Klänge über ein Sinusmodell nach dem Prinzip der Additiven Synthese zu resynthetisieren.

8.2 ANWENDUNG

Die implementierten Systeme liefern aufgrund der sehr guten Klassifikationsergebnisse einen wertvollen Beitrag für die automatische Metadatengewinnung aus Audiodaten und die damit verbundene automatische Annotation von Klangarchiven und der in einem Musikstück enthaltenen Musikinstrumente.

So könnten durch die Auswertung dieser Metadaten beispielsweise Klang- und Musikarchive direkt über schlagwortbasierte Suchmaschinen

auf semantischer Ebene durchsucht werden, indem als Suchkriterium ein oder mehrere Musikinstrumente gewählt werden.

Im Bereich der Studiotechnologie könnten die extrahierten Informationen über die in einem Musikstück erkannten Musikinstrumente auch dazu genutzt werden, um dem Komponisten weitere zu seiner bisherigen Komposition passende oder ähnlich klingende Musikinstrumentenklänge zu präsentieren.

Weiterhin könnten die gewonnenen Metadaten im Zusammenhang mit Verfahren der Quellentrennung oder der automatischen Transkription verwendet werden, indem die Parameter der Verfahren auf die in dem zu bearbeitenden Audiomaterial enthaltenen Musikinstrumente angepasst werden. So könnten sich beispielsweise auch Vereinfachungen für die in QBH- oder QBT-Systemen durchzuführenden Transkriptionen von komplexen Musikstücken in monophone Melodien oder Rhythmen ergeben [EiBS04a, EiBS04b, BEW+04b, Batk06].

9 LITERATURVERZEICHNIS

[Acke91] ACKERMANN, Phillip: *Computer und Musik - Eine Einführung in die digitale Klang- und Musikverarbeitung*. 1. Aufl. Wien, Österreich : Springer Verlag, 1991

[AgLP01] AGOSTINI, Giulio; LONGARI, Maurizio; POLLASTRI, Emanuele: *Content-Based Classification of Musical Instrument Timbres*. In: *International Workshop on Content-Based Multimedia Indexing, IEEE Multimedia Signal Processing Technical Committee*, Brescia, Italien : 2001

[AgLP03] AGOSTINI, Giulio; LONGARI, Maurizio; POLLASTRI, Emanuele: *Musical Instrument Timbres Classification with Spectral Features*. In: *EURASIP Journal on Applied Signal Processing*, 2003, Nummer 1, S. 1-11

[Alle87] ALLERHAND, Michael: *Knowledge based speech pattern recognition*. 1. Aufl. International Thomson Computer Press, 1987

[Anwa00] ANWANDER, Florian: *Synthesizer - So funktioniert elektronische Klangerzeugung*. 1. Aufl. Bergkirchen : PPV Presse Project Verlags GmbH, 2000

[BaEa67] BAUM, L.E.; EAGON, J.A.: *An inequality with applications to statistical estimation for probalistic functions of Markov processes and to a model for ecology*. In: *American Mathematical Society Bulletin*, 1967, Nummer 73, S. 360-363

[BaEi06] BATKE, Jan-Mark; EISENBERG, Gunnar: *Evaluation of Query-by-Humming Systems using a Random Melody Database*. In: *Proceedings of the 120th AES Convention*, 2006

[Bär 03] BÄR, Frank P.: *Musikinstrumente*. 1. Aufl. Nürnberg : Tessloff Verlag, 2003

[Batk06] BATKE, Johann-Markus: *Untersuchung von Melodiesuchsystemen sowie von Verfahren zu ihrer Funktionsprüfung*. Berlin : Technische Universität Berlin, Dissertation, 2006

[BeBi03] BENAROYA, L.; BIMBOT, F.: *Wiener based source separation with HMM/GMM using a single sensor*. In: *Proceedings of the 4th ICA*, Nara, Japan : 2003

[Bell61] BELLMAN, R.E.: *Adaptive Control Processes*. Princeton, USA : Princeton University Press, 1961

[BEW+04a] BATKE, Jan-Mark; EISENBERG, Gunnar; WEISHAUPT, Philipp; SIKORA, Thomas: *A Query by Humming system using MPEG-7 Descriptors*. In: *Proceedings of the 116th AES Convention*, 2004

[BEW+04b] BATKE, Jan-Mark; EISENBERG, Gunnar; WEISHAUPT, Philipp; SIKORA, Thomas: *Evaluation of Distance Measures for MPEG-7 Melody Contours*. In: *Proceedings of the International Workshop on Multimedia Signal Processing (MMSP)*, 2004

[BKK+05] BENETOS, Emmanouil; KOTTI, Margarita; KOTROPOULOS, Constantine; BURRED, Juan José; EISENBERG, Gunnar; HALLER, Martin; SIKORA, Thomas: *Comparison of Subspace Analysis-based and Statistical Model-based Algorithms for Musical Instrument Classification*. In: *Proceedings of the 2nd Workshop on Immersive Communication and Broadcast Systems (ICOB)*, 2005

[Blim97] BLIMES, Jeff A.: *A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models*. Berkeley, USA : University of Berkeley, Technical Report, 1997

[Boer93] BOERSMA, Paul: *Accurate Short-Term Analysis of the Fundamental Frequency and the Harmonics-to-Noise Ratio of a Sampled Sound*. In: *In Proceedings of the Institute of Phonetic Sciences*, 1993, Nummer 17, S. 97-110

[BrHM01] BROWN, J.C.; HOUIX, O.; MCADAMS, S.: *Feature dependence in the automatic identification of musical woodwind instruments*. In: *J. Acoust. Soc. Am.*, 2001, Nummer 109, S. 1064-1072

[Brin98] BRINER, Ermanno: *Reclams Musikinstrumentenführer - Die Instrumente und ihre Akustik*. 4. Aufl. Stuttgart : Philipp Reclam jun., 1998

[Brow91] BROWN, J.C.: *Calculation of a Constant Q Spectral Transform*. In: *J. Acoust. Soc. Am.*, 1991, Nummer 89, S. 425-434

[Brow97] BROWN, Judith C.: *Computer identification of musical instruments using pattern recognition with cepstral coefficients as features*. In: *J. Acoust. Soc. Am.*, 1997, Nummer 105, S. 1933-1941

[BrPu92] BROWN, J.C.; PUCKETTE, M.S.: *An Efficient Algorithm for the Calculation of a Constant Q Transform*. In: *J. Acoust. Soc. Am.*, 1992, Nummer 92, S. 2698-2701

[BuLe03] BURRED, Juan José; LERCH, Alexander: *A hierarchical approach to automatic musical genre classification*. In: *Proceedings of the 6th International Conference on Digital Audio Effects (DAFx-03)*, 2003

[BuLe04] BURRED, Juan José; LERCH, Alexander: *Hierarchical Automatic Audio Signal Classification*. In: *J. Audio Eng. Soc.*, 2004, Nummer 52, 7/8, S. 724-739

[Burg98] BURGES, Christopher J. C.: *A Tutorial on Support Vector Machines for Pattern Recognition*. In: *Data Mining and Knowledge Discovery*, 1998, Nummer 2-2, S. 121-167

[CaCh99] CARPENTER, Bob; CHU-CARROLL, Jennifer: *Spoken Dialogue Systems*. Lucent Technologies, Bell Labs Innovations, 1999

[Camp97] CAMPBELL, Joseph P.: *Speaker Recognition: A Tutorial*. In: *Proceedings of the IEEE*, 1997, Nummer 85, 9, S. 1437-1462

[Case01a] CASEY, Michael: *General sound classification and similarity in MPEG-7*. In: *Organised Sound*, 2001, Nummer 6-2, S. 153-164

[Case01b] CASEY, Michael: *MPEG-7 Sound-Recognition Tools*. In: *IEEE Transactions on Circuits and Systems for Video Technology*, 2001, Nummer 11,6, S. 737

[CaWe01] CASEY, Michael A.; WESTNER, Alex: *Separation of Mixed Audio Sources by Independent Subspace Analysis*. In: *Proceedings of the International Computer Music Conference, ICMC*, 2001

[ChCB03] CHO, Young-Choon; CHOI, Seungjin; BANG, Sung-Yang: *Non-Negative Component Parts of Sound for Classification*. In: *Proceedings of the IEEE International Symposium on Signal Processing and Information Technology, ISSPIT*, 2003

[Cofe08] COFER, David: *Mind Creators - Neuron Basics*. 01.01.2008 URL http://www.mindcreators.com/NeuronBasics.htm

[CrWe03] CRYSANDT, Holger; WELLHAUSEN, Jens: *Music Classification with MPEG-7*. In: *Proceedings SPIE Storage and Retrieval for Media Databases*, 2003

[DaMe80] DAVIS, S.; MERMELSTEIN, P.: *Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences*. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1980, Nummer 28-4, S. 357-366

[D'an78] D'ANDRADE, R.: *U-Statistic Hierarchical Clustering*. In: *Psychometrika*, 1978, Nummer 4, S. 58-67

[Diet02] DIETEL, Gerhard: *Wörterbuch Musik*. 2. Aufl. München : Deutscher Taschenbuch Verlag, 2002

[Digi08] DIGIDESIGN: *Sample Cell II - Library*. 01.01.2008 URL http://www.digidesign.com

[DIN1320] DIN: *1320: Akustik - Begriffe*. 1997

[DIN13320] DIN: *13320: Akustik; Spektren und Übertragungskurven, Begriffe, Darstellung*. 1979

[DIN61672] DIN: *EN 61672: Elektroakustik - Schallpegelmesser*. 2003

[Dixo06] DIXON, Simon: *Onset Detection Revisited*. In: *Proceedings of the 9th International Conference on Digital Audio Effects (DAFx)*, 2006

[DuHS01] DUDA, Richard O.; HART, Peter E.; STORK, David G.: *Pattern Classification*. 2. Aufl. New York, USA : John Wiley & Sons, 2001

[EgBr03a] EGGINK, Jana; BROWN, Guy J.: *A Missing Feature Approach to Instrument Identification in Polyphonic Music*. In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 2003

[EgBr03b] EGGINK, Jana; BROWN, Guy J.: *Application of Missing Feature Theory to the Recognition of Musical Instruments in Polyphonic Audio*. In: *Proceedings of the International Symposium on Music Information Retrieval, ISMIR*, 2003

[EgBr04a] EGGINK, Jana; BROWN, Guy J.: *Instrument Recognition in Accompanied Sonatas and Concertos*. In: *Proc. International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 2004

[EgBr04b] EGGINK, Jana; BROWN, Guy J.: *Extracting Melody Lines from Complex Audio*. In: *Proceedings of the International Symposium on Music Information Retrieval, ISMIR*, 2004

[Eggi01] EGGINK, Jana: *Wahrnehmungsbasierte Trennung und Gruppierung auditorischer Objekte. Ein Vergleich aktueller computergestützter Modelle*. Universität Hamburg, Musikwissenschaftliches Institut, Magisterarbeit, 2001

[EiBS04a] EISENBERG, Gunnar; BATKE, Jan-Mark; SIKORA, Thomas: *BeatBank - An MPEG-7 compliant Query by Tapping System*. In: *Proceedings of the 116th AES Convention*, 2004

[EiBS04b] EISENBERG, Gunnar; BATKE, Jan-Mark; SIKORA, Thomas: *Efficiently Computable Similarity Measures for Query by Tapping Systems*. In: *Proceedings of the 7th International Conference on Digital Audio Effects (DAFx)*, 2004

[EiHu73] EIMERT, Herbert; HUMPERT, Hans Ulrich: *Das Lexikon der elektronischen Musik*. 1. Aufl. Regensburg : Gustav Bosse Verlag Regensburg, 1973

[EiSi06] EISENBERG, Gunnar; SIKORA, Thomas: *Granular Resynthesis for Sound Unmixing*. In: *Proceedings of the 9th International Conference on Digital Audio Effects (DAFx)*, 2006

[Elli85] ELLIS, Alexander J.: *The Musical Scales of Various Nations*. In: *Journal of the Society of Arts*, 1885

[ErKl00] ERONEN, Antti; KLAPURI, Anssi: *Musical Instrument Recognition using Cepstral Coefficients and Temporal Features*. In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 2000

[Eron01a] ERONEN, Antti: *Automatic Musical Instrument Recognition*. Tampere, Finnland : Tampere University of Technology, Master of Science Thesis, 2001

[Eron01b] ERONEN, Antti: *Comparison of Features for Musical Instrument Recognition*. In: *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA*, 2001

[FiJa02] FIGUEIREDO, M.A.T.; JAIN, A.K.: *Unsupervised learning of finite mixture models*. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002, Nummer 24 (3), S. 381-396

[FlRo91] FLETCHER, Neville H.; ROSSING, Thomas D.: *The Physics of Musical Instruments*. 2. Aufl. Berlin : Springer Verlag, 1991

[Goog08] GOOGLE: 01.01.2008 URL www.google.de

[Grei05] GREISBACH, Reinhold: *Skalierung der Empfindungen*. 02.06.2005
URL http://phonetik.sprachsignale.de/grundlagen/aud18.html

[HAB+00] HERRERA, Perfecto; AMATRIAIN, Xavier; BATLLE, Eloi; SERRA, Xavier: *Towards instrument segmentation for music content description: a critical review of instrument classification techniques*. In: *International Symposium on Music Information Retrieval*, 2000

[Hanu94] HANUS, Bo: *Erfolgreicher Service elektronischer Musikinstrumente - Grundlagen, Fehlerdiagnose, Meßmethoden, Fehlerbeseitigung bei Keyboards, Synthesizern, elektronischen Orgeln und E-Gitarren*. 1. Aufl. Poing : Franzis-Verlag GmbH, 1994

[Harr06] HARRER, Wolfgang: *Was finden Google und Yahoo noch? Suchmaschinen erfassen nur noch Bruchteile des Internet*. In: *ZDF - Heute.de*, 18.01.2006
URL http://www.heute.de/ZDFheute/inhalt/18/0,3672,3712146,00.html

[Haue93] HAUENSTEIN, Alfred: *Optimierung von Algorithmen und Entwurf eines Prozessors für die automatische Spracherkennung*. 1. Aufl. München : Technische Universität München, Dissertation, 1993

[Hemp01] HEMPEL, Christoph: *Neue Allgemeine Musiklehre*. 2. ergänzte Aufl. Mainz : Atlantis Musikbuch-Verlag, 2001

[HePD02] HERRERA-BOYER, Perfecto; PEETERS, Geoffroy; DUBNOV, Shlomo: *Automatic Classification of Musical Instrument Sounds*. In: *MOSART Midterm Meeting*, 2002

[Hoen01] HOENIG, Uwe G.: *Workshop Synthesizer - Klangsynthese und Programmierung für Musiker*. 1. Aufl. Bergkirchen : PPV Presse Project Verlags GmbH, 2001

[Hopf84] HOPFIELD, J.J.: *Neurons with graded response have collective computational properties like those of two-state neurons*. In: *Proceedings Natl. Acad. Sci. USA*, 1984, Nummer 81, S. 3088-3092

[Hoye04] HOYER, Patrik O.: *Non-negative Matrix Factorization with Sparseness Constraints*. In: *Journal of Machine Learning Research*, 2004, Nummer 5, S. 1457-1469

[HsCL03] HSU, Chih-Wie; CHANG, Chih-Chung; LIN, Chih-Jen: *A Practical Guide to Support Vector Classification*. 2003

[HyOj00] HYVÄRINEN, Aapo; OJA, Erkki: *Independent Component Analysis: Algorithms and Applications*. In: *Neural Networks*, 2000, Nummer 13 (4-5), S. 411-430

[Iowa08] THE UNIVERSITY OF IOWA: *The University of Iowa - Musical Instrument Samples*. 01.01.2008
URL http://theremin.music.uiowa.edu

[ISO15938] ISO/IEC JTC: *15938: Information Technology - Multimedia Content Description Interface*. 2002

[JaNo84] JAYANT, N. S.; NOLL, Peter: *Digital Coding of Waveforms: Principles and Applications to Speech and Video*. 1. Aufl. Prentice Hall, 1984

[John67] JOHNSON, S.C.: *Hierarchical Clustering Schemes*. In: *Psychometrika*, 1967, Nummer 2, S. 241-254

[Joll02] JOLLIFFE, I.T.: *Principal Component Analysis*. 2. Aufl. Berlin : Springer Verlag, 2002

[KiBS04] KIM, H.-G.; BERDAHL, E.; SIKORA, T.: *Study of MPEG-7 Sound Classification and Retrieval*. In: *Proceedings of the 5th International ITG Conference on Source and Channel Coding*, 2004

[KiGO03] KITAHARA, Tetsuro; GOTO, Masataka; OKUNO, Hiroshi G.: *Musical Instrument Identification based on F0-dependent Multivariate Normal Distribution*. In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 2003

[KiMS05] KIM, Hyoung-Gook; MOREAU, Nicolas; SIKORA, Thomas: *MPEG-7 Audio and Beyond - Audio Content Indexing and Retrieval*. 1. Aufl. New York, USA : John Wiley & Sons, 2005

[KoCz01] KOSTEK, Bozena; CZYZEWSKI, Andrzej: *Representing Musical Instrument Sounds for Their Automatic Classification*. In: *J. Audio Eng. Soc.*, 2001, Nummer 49-9, S. 768-785

[Koep00] KOEPPEN, Mario: *The Curse of Dimensionality*. In: *Proceedings of the 5th Online World Conference on Soft Computing in Industrial Applications (WSC5)*, 2000

[KSW+00] KAMP, Thomas; SCHMIDT, Michael; WESTPHAL, Martin; ; WAIBEL, Alex: *Strategies for Automatic Segmentation of Audio Data*. In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 2000

[Kueh75] KUEHNELT, Wolf Dieter: *Die Klassifizierung der Musikinstrumente von Michael Praetorius bis Hornbostel/Sachs*. 1. Aufl. Berlin : Technische Universität Berlin, Magisterarbeit, 1975

[LeET05] LERCH, Alexander; EISENBERG, Gunnar; TANGHE, Koen: *FEAPI: A Low Level Feature Extraction Plugin API*. In: *Proceedings of the 8th International Conference on Digital Audio Effects (DAFx)*, 2005

[LeSe99] LEE, Daniel D.; SEUNG, H. Sebastian: *Learning the parts of objects by nonnegative matrix factorization*. In: *Nature*, 1999, Nummer 401, S. 788-791

[LeSe00] LEE, Daniel D.; SEUNG, H. Sebastian: *Algorithms for Non-negative Matrix Factorization*. In: *Advances in Neural Information Processing Systems*, 2000, Nummer 13, S. 556-562

[LiBG80] LINDE, Y.; BUZO, A.; GRAY, R.: *An Algorithm for Vector Quantizer Design*. In: *IEEE Transactions on Communication*, 1980, Nummer 28-1, S. 84-95

[LiPR03] LIVSHIN, Arie A.; PEETERS, Geoffroy; RODET, Xavier: *Studies and Improvements in Automatic Classification of Musical Sound Samples*. In: *Proceedings of the ICMC*, 2003

[LiRo03] LIVSHIN, Arie A.; RODET, Xavier: *The Importance of Cross Database Evaluation in Sound Classification*. In: *Proceedings of the International Symposium on Music Information Retrieval, ISMIR*, 2003

[LiRo04] LIVSHIN, Arie A.; RODET, Xavier: *Musical Instrument Identification in Continuous Recordings*. In: *Proceedings of the International Conference on Digital Audio Effects, DAFx*, 2004

[Lloy82] LLOYD, S.P.: *Least squares quantization in PCM*. In: *IEEE Transactions on Information Theory*, 1982, Nummer IT-28, S. 127-135

[MaFe02] MARTINS, Luís Gustavo P.M.; FERREIRA, Aníbal J.S.: *PCM to MIDI Transposition*. In: *Proceedings of the 112th AES Conventition*, 2002

[Makh75] MAKHOUL, J.: *Linear prediction: A tutorial review*. In: *Proceedings of the IEEE*, 1975, Nummer 63 (5), S. 561-580

[MaMo99] MARQUES, Janet; MORENO, Pedro J.: *A Study of Musical Instrument Classification Using Gaussian Mixture Models and Support Vector Machines*. In: *Compaq & DEC Technical Reports (CRL-99-4)*, 1999

[Mart98] MARTIN, Keith D.: *Toward automatic sound source recognition: identifying musical instruments*. In: *NATO Computational Hearing Advanced Study Institute*, 1998

[Mart99] MARTIN, Keith Dana: *Sound-Source Recognition: A Theory and Computational Model*. Massachusetts, USA : Massachusetts Institute of Technology, PhD Thesis, 1999

[MaSS02] MANJUNATH, B.S.; SALEMBIER, Philippe; SIKORA, Thomas (Hrsg.) *Introduction to MPEG-7 - Multimedia Content Description Interface*. 1. Aufl. New York, USA : John Wiley & Sons, 2002

[McQu85] MCAULAY, R. J.; QUATIERI, T. F.: *Speech analysis-synthesis based on a sinusoidal representation*. In: *Technical Report, Lincoln Laboratory, MIT*, 1985, Nummer 693,

[Meye04] MEYER, Jürgen: *Akustik und musikalische Aufführungspraxis*. 5. aktualisierte Aufl. Bergkirchen : PPV Medien GmbH, 2004

[Mich05] MICHELS, Ulrich: *Dtv-Atlas Musik, Band 1 Systematischer Teil, Musikgeschichte von den Anfängen bis zur Renaissance*. 21. durchgesehene und korrigierte Aufl. München : Deutscher Taschenbuch Verlag, 2005

[Moon96] MOON, Todd K.: *The Expectation-Maximisation Algorithm*. In: *IEEE Signal Processing Magazine*, 1996, Nummer 13-6, S. 47-60

[Nabn03] NABNEY, Ian T.: *Netlab: Algorithms for Pattern Recognition*. 3. Aufl. London, GB : Springer Verlag, 2003

[Naga03] NAGARAJ, Keerthi C.: *Toward Automatic Transcription - Pitch Tracking In Polyphonic Environment*. The University of Texas at Austin, Literature Survey, 2003

[Nati08] NATIVE INSTRUMENTS: *Bandstand*. 01.01.2008 URL http://www.native-instruments.com

[OpSc95] OPPENHEIM, Alan V.; SCHAFER, Ronald W.: *Zeitdiskrete Signalverarbeitung*. 2. überarbeitete Aufl. Oldenbourg Verlag, 1995

[Paal04] PAALANEN, Pekka: *Bayesian Classification using Gaussian Mixture Models and EM Estimation: Implementations and Comparisons*. Lappeenranta University of Technology, Information Technology Project, 2004

[Peet04] PEETERS, Geoffroy: *A large set of audio features for sound description (similarity and classification) in the CUIDADO project*. In: *CUIDADO I.S.T. Project Report*, 2004

[PeRo02] PEETERS, Geoffroy; RODET, Xavier: *Automatically selecting signal descriptors for Sound Classification*. In: *Proceedings of the International Computer Music Conference, ICMC*, 2002

[PeRo03] PEETERS, Geoffroy; RODET, Xavier: *Hierarchical Gaussian tree with inertia ratio maximization for the classification of large musical instrument databases*. In: *Proceedings of the 6th International Conference on Digital Audio Effects (DAFx-03)*, 2003

[Rabi89] RABINER, Lawrence R.: *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*. In: *Proceedings of the IEEE*, 1989, Nummer 77-2, S. 257-286

[Rege88] REGEL, Peter: *Akustisch-phonetische Transkription für die automatische Spracherkennung*. 1. Aufl. Düsseldorf : VDI Verlag, 1988

[ReRo95] REYNOLDS, Douglas A.; ROSE, Richard C.: *Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models*. In: *IEEE Transactions on Speech and Audio Processing*, 1995, Nummer 3-1,

[Rohd03] ROHDENBURG, Thomas: *Klassifikation von Audio-Signalen*. Universität Bremen, Diplomarbeit, 2003

[Roth05] ROTH, Marcel: *Entwicklung eines VST-PlugIns zur echtzeitfähigen Geräuschklassifikation nach MPEG-7*. Technische Universität Berlin, Diplomarbeit, 2005

[Rott00] ROTTLAND, Jörg Mathias: *Ein hybrider Ansatz zur automatischen Spracherkennung und Sprecheradaption für große Wortschätze*. 1. Aufl. Düsseldorf : VDI Verlag, 2000

[Sach65] SACHS, Curt: *Geist und Werden der Musikinstrumente*. 2. Aufl. Hilversum : Knuf, 1965

[ShPr01] SHERMAN, Chris; PRICE, Gary: *The Invisible Web: Uncovering Information Sources Search Engines Can't See*. 1. Aufl. Cyberage Books, 2001

[Sigm03] SIGMUND, Milan: *Voice Recognition by Computer*. 1. Aufl. Marburg : Tectum Verlag, 2003

[SmAb99] SMITH, J.O.; ABEL, J.S.: *Bark and ERB bilinear transforms*. In: *IEEE Transactions on Speech and Audio Processing*, 1999, Nummer 7-6, S. 697-708

[Smar04] SMARAGDIS, Paris: *Non-negative Matrix Factor Deconvolution; Extraction of Multiple Sound Sources from Monophonic Inputs*. In: *Proceedings of the Congress on Independent Component Analysis and Blind Signal Separation, ICA*, 2004

[SmBr03] SMARAGDIS, Paris; BROWN, Judith C.: *Non-negative Matrix Factorization for Polyphonic Music Transcription*. In: *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA*, 2003

[SoRo88] SOONG, F.K.; ROSENBERG, A.E.: *On the use of instantaneous and transitional spectral information in speaker recognition*. In: *IEEE Transactions ASSP*, 1988, Nummer 36, S. 871-879

[Stai75] STAINER, John: *On the Principles of Musical Notation*. In: *Proceedings of the Musical Association, 1st Session*, 1875

[Stei08] STEINBERG: *VST-PlugIn Interface*. 01.01.2008 URL http://www.steinberg.com

[Step68] STEPHAN, Rudolf (Hrsg.) *Das Fischer Lexikon - Musik*. 9. Aufl. Frankfurt am Main : Fischer Bücherei KG, 1968

[Tant06] TANTAU, Till: *Das World Wide Web - Angels and Demons*. Universität zu Lübeck, 2006 URL http://www.tcs.uni-luebeck.de/Lehre/InfoB/SS2006/2006-06-01-beamer_version.pdf

[Thie73] THIEL, Eberhard: *Sachwörterbuch der Musik*. 2. verbesserte Aufl. Stuttgart : Alfred Kröner Verlag, 1973

[Trom95] TROMPF, Michael: *Künstliche neuronale Netzwerke zur adaptiven Geräuschreduktion für robuste Spracherkennung*. Karlsruhe : Universität Fridericiana Karlsruhe, Dissertation, 1995

[TzCo02] TZANETAKIS, George; COOK, Perry: *Musical Genre Classification of Audio Signals*. In: *IEEE Transactions on Speech and Audio Processing*, 2002, Nummer 10-5,

[TzEC01] TZANETAKIS, George; ESSL, Georg; COOK, Perry: *Automatic Musical Genre Classification of Audio Signals*. In: *Proceedings of the International Symposium on Music Information Retrieval, ISMIR*, 2001

[TzEC02] TZANETAKIS, George; ERMOLINSKYI, Andreye; COOK, Perry: *Beyond the Query-By-Example Paradigm: New Query Interfaces for Music Information Retrieval*. In: *Proceedings of the ICMC*, 2002

[UhDS03] UHLE, Christian; DITTMAR, Christian; SPORER, Thomas: *Extraction of Drum Tracks from Polyphonic Music Using Independent Subspace Analysis*. In: *Proceedings of the 4th ICA*, Nara, Japan : 2003

[Veit96] VEIT, Ivar: *Technische Akustik - Grundlagen der physikalischen, physiologischen und Elektroakustik*. 5. durchgesehene Aufl. Würzburg : Vogel, 1996

[ViHP02] VINET, Hugues; HERRERA, Perfecto; PACHET, Francois: *The CUIDADO Project*. In: *Proceedings of the International Symposium on Music Information Retrieval, ISMIR*, 2002

[Virt03] VIRTANEN, Tuomas: *Sound Source Separation Using Sparse Coding with Temporal Continuity Objective*. In: *Proceedings of the ICMC*, 2003

[Vite67] VITERBI, A.: *Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm*. In: *IEEE Transactions on Information Theory*, 1967, Nummer 13-2, S. 260-269

[VoSa14] VON HORNBOSTEL, Erich Moritz; SACHS, Curt: *Systematik der Musikinstrumente - Ein Versuch*. In: *Zeitschrift für Ethnologie*, 1914, Nummer 46, (4-5), S. 53-90

[VTD+04] VAN STEELANT, D.; TANGHE, K.; DEGROEVE, S.; DE BAETS, B.; LEMAN, M.; MARTENS, J.-P.: *Classification of Percussive Sounds Using Support Vector Machines*. In: *Proceedings of the annual machine learning conference of Belgium and The Netherlands*, Brüssel, Belgien : 2004

[WaRR03] WALL, Michael E.; RECHTSTEINER, Andreas; ROCHA, Luis M.: *Singular value decomposition and principal component analysis*. In: *A Practical Approach to Microarray Data Analysis*, 2003 S. 91-109

[WeCr03] WELLHAUSEN, Jens; CRYSANDT, Holger: *Temporal Audio Segmentation Using MPEG-7 Descriptors*. In: *Proceedings SPIE Storage and Retrieval for Media Databases*, 2003

[Wein08] WEINZIERL, Stefan (Hrsg.) *Handbuch der Audiotechnik*. 1. Aufl. Berlin : Springer Verlag, 2008

[Wink98] WINKLER, Klaus: *Die Physik der Musikinstrumente*. 2. Aufl. Heidelberg : Spektrum Verlag, 1998

[WoCa04] WOLF, Matt; CASALIS, Anna: *Fattoria - Tocca et Senti*. 1. Aufl. Mailand, Italien : Dami Editore, 2004

[Wolf97] WOLFERTSTETTER, Franz: *Verallgemeinerte stochastische Modellierung für die automatische Spracherkennung*. 1. Aufl. Aachen : Shaker Verlag, 1997

[Yaho08] YAHOO: 01.01.2008 URL http://www.yahoo.com

[Zhan03] ZHANG, Tong: *Semi-Automatic Approach for Music Classification*. In: *HP Laboratories, Imaging Systems Laboratory (HPL-2003-183)*, Palo Alto, USA : 2003

[Zieg00] ZIEGENRÜCKER, Wieland: *ABC Musik - Allgemeine Musiklehre*. 3. unveränderte Aufl. Wiesbaden : Breitkopf & Härtel, 2000

[Zwic82] ZWICKER, Eberhard: *Psychoakustik*. 1. Aufl. Berlin : Springer Verlag, 1982