Institut für Logik, Komplexität und Deduktionssysteme
der Universität Karlsruhe
Lehrstuhl Prof. A. Waibel

Klassifikation von Musikstilen

Diplomarbeit
von

Hagen Soltau
(soltau@ira.uka.de)

Erstgutachter: Prof. Dr. Alex Waibel
Betreuerin: Dipl.-Inform. Tanja Schultz

Karlsruhe, den 28. Mai 1997

Zusammenfassung

Die Wahrnehmung akustischer Ereignisse wie Sprache und Musik durch Maschinen ist ein Schlüssel auf dem Weg zur Künstlichen Intelligenz. Gegenstand dieser Arbeit ist das Erkennen von Musikstilen. Schwierigkeiten und Lösungsansätze werden anhand der Stilrichtungen Rock, Pop, Techno und Klassik aufgezeigt.

Das Ergebnis ist ein lernendes System zur Musikstilerkennung. Zwei Prinzipien prägen die Repräsentation des Musikwissens und die Informationsverarbeitung. Das Einfachheitsprinzip besagt, daß die Repräsentation des Wissens so einfach wie möglich sein soll. Konsequenterweise sollte ein Modell des Musikstückes im Computer nur relevante Merkmale des Musikstückes enthalten. Probleminvariante Eigenschaften müssen eliminiert werden. Experimente zur Wahrnehmung von Musik beim Menschen erleichtern das Auffinden solcher Eigenschaften. So kann nachgewiesen werden, daß die Wahrnehmung von Musik in einem gewissen Rahmen zeitinvariant ist. Relevante Merkmale für die Erkennung eines Musikstils sind Rhythmus und Klangfarbe.

Das zweite Prinzip betrifft die Art der Informationsverarbeitung. Das Gruppenprinzip besagt, daß ähnliche Informationen ähnlich verarbeitet werden sollen. Die Anwendung dieses Prinzips führt zu modular strukturierten Neuronalen Netzwerken. Zwei Spezialnetze verarbeiten die relevanten Eigenschaften Rhythmus und Klangfarbe. Zeitabhängigkeiten werden auf explizite Weise modelliert. Eine Integration dieser so verarbeiteten Informationen wird auf einer höheren Ebene durchgeführt. Diese Aufgabe wird durch ein Kombinationsnetz gelöst.

Ein Schwerpunkt der Arbeit ist nicht nur die Synthese eines Systems zur Musikstilerkennung, sondern auch dessen Analyse. Dies betrifft die gelernten Netzwerkverbindungen der Neuronalen Netze. Damit kann das Verhalten des Systems auf Plausibilität geprüft werden. Erkennungsleistung und Dominanz einzelner Stile werden nachvollziehbar.

Insgesamt werden mit diesem System 88,9% der Stilrichtungen aller Musikstücke richtig erkannt. Besonders gut ist die Erkennung klassischer Stücke, auch die eher als schwierig eingestuften Stücke, wie die *Bilder einer Ausstellung* von Mussogorsky, werden korrekt erkannt. Die Trennung der ähnlichen Musikstile Rock und Pop durch das System ist ebenfalls sehr gut.

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

Kapitel 1

Einführung

Nicht erst seit dem Zeitalter der Computer untersuchen Forschungsgruppen aus unterschiedlichen Disziplinen die kognitiven und perzeptiven Fähigkeiten des Menschen. Aus Sichtweise der Informatik können das Wahrnehmen und Erkennen, das Denken und Planen als informationsverarbeitende Prozesse verstanden werden. Gegenstand der Künstlichen Intelligenz (KI) ist die Modellierung und Simulation dieser Fähigkeiten. Die rasch fortschreitende Entwicklung immer schnellerer (und eventuell auch besserer) Computer ermöglicht es, solche Systeme der KI auch praktisch anzuwenden. Als aktuelles Beispiel sei da das Spracherkennungs - und übersetzungsystem JANUS [27] erwähnt, welches in einer Kooperation der Universität Karlsruhe mit der Carnegie Mellon University (Pittsburgh,USA) entwickelt wird.

Die rasante Entwicklung zu einer Informationsgesellschaft bringt es hervor, daß heute dem Menschen eine Vielzahl von Informationen zur Verfügung stehen, wie es vor einem Jahrhundert noch undenkbar erschien. Die zielgerichtete Nutzung dieser Möglichkeiten und die Suche bestimmter Informationen wird damit immer schwieriger. Mit Methoden der KI können informationsverarbeitende Systeme zur Unterstützung des Menschen entwickelt werden. Mit die wichtigsten Informationsquellen sind Rundfunk und Fernsehen. Um diese Informationen nutzen zu können, ist es erforderlich, akustische Ereignisse zu erkennen. Gegenstand der vorliegenden Arbeit ist die Erkennung von Musikstilen.

Die Klassifikation von Musikstilen ist nicht nur auf dem Gebiet der Informationsverarbeitung von Bedeutung, sondern kann auch auch im Bereich der Mensch-Maschine-Kommunikation Anwendung finden. So sind Radioap-

parate vorstellbar, die fähig sind, die verfügbaren Sender nach bevorzugten Musikrichtungen abzusuchen. Dazu ist es erforderlich, adaptive Systeme zu entwickeln, die in der Lage sind, den Stil der Musikstücke zu erkennen und zu lernen, welche Musikrichtungen vom Hörer gewünscht beziehungsweise abgelehnt werden. Es sei dahin gestellt, inwieweit solche Geräte Akzeptanz finden.

Ziel dieser Diplomarbeit ist es, verschiedene Musikstile zu erkennen. Genauer: Der Klassifikator soll ein Musikstück in einen der Bereiche Rockmusik, Popmusik, Techno oder klassische Musik einordnen. In diesem Sinne wird auch der Begriff Erkennung verstanden. Aus diesem Grunde werden die Begriffe Erkennung und Klassifikation synonym verwendet. Anforderungen an den Klassifikator sind Zuverlässigkeit und Geschwindigkeit. Es soll ein möglichst kurzer Ausschnitt aus dem Musikstück genügen, um das Stück richtig zu klassifizieren.

Das zur Modellierung des Klassifikators erforderliche Problemwissen wird hauptsächlich durch eine Beispielsammlung von Musikstücken repräsentiert. Diese sind in einer Datenbasis zusammengefaßt, die in dem sich anschließenden Kapitel 2 näher beschrieben wird. Ein (disjunkter) Teil der Datenbasis dient ferner auch zur Evaluierung des Systems.

Aktuelle Ansätze im Bereich der Musikerkennung werden in dem dritten Kapitel diskutiert. Gegenstand dieser Arbeiten ist die Identifikation von Noten sowie Rhythmuswahrnehmung und Klangfarbenerkennung. Fernziel der Musikerkennung ist die automatische Transkription von Musikstücken.

Die Trennung von Musikstilen ist keine eindeutige Aufgabe. So ist die Zuordnung eines Musikstückes zu einem Musikstil auch von den Hörgewohnheiten der Person abhängig. In dem vierten Kapitel werden Experimente zur Untersuchung der Zuverlässigkeit und Schnelligkeit von Versuchspersonen vorgestellt. Geklärt werden sollen auch, welche Informationen zur Erkennung benutzt werden, ob eher Kurzzeit- oder Langzeitinformationen von Bedeutung sind. Wird von der Arbeitshypothese ausgegangen, daß diese Erkenntnisse auf einen maschinellen Klassifikator übertragbar sind, dann lassen sich daraus Aussagen zu einer geeigneten Repräsentation der Musikstücke ableiten.

Eine angemessene Repräsentation der Musikstücke ist für die Informationsverarbeitung von entscheidener Bedeutung. Eine Repräsentation heißt dabei angemessen in Hinblick auf die Nutzung des Wissens. In Abbildung 1.1 ist der Ausschnitt eines Musikstückes dargestellt. Das akustische Signal ist

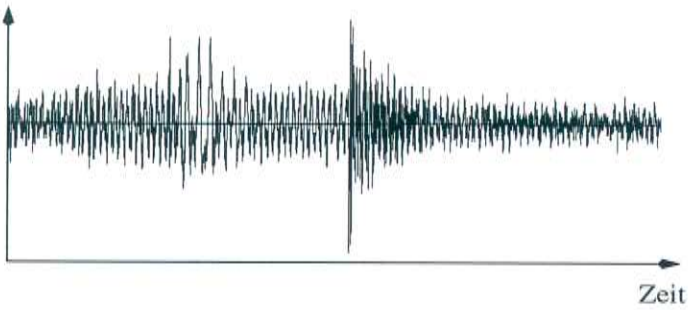

Abbildung 1.1: akustisches Signal im Zeitbereich

die Ausgangsbasis, in der die Musikbeispiele vorliegen. Ganz offensichtlich kann man aus dem akustischen Signal nicht die Stilrichtung erkennen. Das Signal im Zeitbereich ist für die Musikstilerkennung keine geeignete Darstellung.

Das fünfte Kapitel widmet sich daher der Suche nach einer geeigneten Repräsentation der Musikstücke. Es beinhaltet Experimente zur Bestimmung relevanter Merkmale für die Klassifikation der Musikstücke. Berücksichtigung finden dabei Erkenntnisse aus der Psychoakustik, wie sie auch in Spracherkennungsystemen angewendet werden. Desweiteren dienen diese Experimente zur Klärung der Begriffe Rhythmus, Melodie und Klangfarbe. Resultat dieser Untersuchungen ist eine kompakte Repräsentation, wie sie das Einfachheitsprinzip verlangt.

Experimente zum Verarbeiten dieser Informationen werden in dem sich anschließenden Kapitel 6 vorgestellt. Basis der Untersuchung ist die Arbeitshypothese, das konnektionistische Klassifikatoren dem vorhandenen Problemwissen in Form von Beispielen angemessen sind. Es werden dabei mehrere Ansätze verglichen. In Betracht gezogen werden Neuronale Netze. Grundlage für deren modularen Struktur bilden die Erkenntnisse des fünften Kapitels zusammen mit dem Gruppenprinzip. Ein anderer, statistischer Ansatz besteht darin, die klassenbedingten Wahrscheinlichkeiten durch versteckte Markov Modelle (Hidden Markov Models, HMM) zu schätzen. In einem weiteren Ansatz wird versucht, die Unsicherheiten der Zeitinvarianzannahme auszugleichen. Das Ergebnis ist ein diskriminatives Lernverfahren für zeitabhängige Phänomene, genannt *ETM-NN*. Vorteil des Ansatzes ist eine explizite Modellierung der Zeitabhängigkeiten.

Kapitel 2

Datenbasis

Die Experimente zur Wahrnehmung von Musik, die Bestimmung relevanter Merkmale und nicht zuletzt der Bau eines Klassifikators erfordert eine Sammlung von Beispielen der unterschiedlichen Musikstile. Es wurde deshalb eine Datenbasis erstellt, die diese Informationen enthält. Im Abschnitt *Aufnahme der Musikstücke* werden Umfang und Art der Musikstücke erläutert. Gegenstand des zweiten Abschnitts ist die Zuordnung der Musikstücke zu einem Musikstil.

2.1 Aufnahme der Musikstücke

Diese Datenbasis enthält Ausschnitte von Musikstücken, die jeweils etwa von 30 Sekunden Dauer sind. Die Ausschnitte befinden sich in der Mitte des Stückes, da oft Anfang und Ende nicht charakteristisch für das Musikstück sind. Zudem ist bei zufälligem Einschalten oder Senderwechsel die Wahrscheinlichkeit höher, ein Ausschnitt aus dem mittleren Teil zu hören.

Aufgenommen wurden die Daten von Compact Disc's (CD). Die Abtastrate der Daten auf den CD's beträgt 44100 Hz. Als weitere Anforderung für den Klassifikator werden die Daten mit einem Tiefpaßfilter auf 16000 Hz begrenzt und anschließend die beiden Stereosignale zu einem Monosignal zusammengesetzt. Somit wird eine gleiche Ausgangsbasis geschaffen, wie sie bei dem Spracherkennungs- und übersetzungssystem JANUS (Interactive System Laboratories) derzeit verwendet wird.

Zur Repräsentation eines jeden Musikstils enthält die Datenbasis Aufnah-

men von 15 Compact Disc's. Zur Repräsentation klassischer Musik enthält die Datenbasis beispielsweise Aufnahmen von *Beethoven*, *Haydn*, *Mozart* und *Mussorgsky*. Vertreter der Popmusik sind unter anderem *Depeche Mode*, *Madonna*, *Sade* und *Talk Talk*. Von jeder CD wurden 6 Musikstücke entnommen. Die gesamte Datenbasis enthält Aufnahmen von insgesamt 4∗15∗6∗30 Sekunden, dies entspricht 3 Stunden. Jeder Stil ist durch eine gleich große Anzahl von Musikstücken repräsentiert. Zudem zeichnet sich die Datenbasis durch eine hohe Diversität aus, da nur jeweils eine CD einer Musikgruppe verwendet wird. Mit nur 6 Stücken pro Gruppe in der Datenbasis ist die Gefahr einer Spezialisierung auf einzelne Musikgruppen gering.

Die Datenbasis ist in drei Mengen aufgeteilt. Die Trainingsmenge wird zur Modellierung des Klassifikators verwendet, sie enthält jeweils 10 CD's pro Musikstil und damit 67% der gesamten Datenbasis. Die Validierungsmenge dient zur Kontrolle des Trainings und besteht aus je 2 CD's (13%). Die Leistung des Klassifikator wird mit der Testmenge gemessen und enthält je 3 CD's, dies entspricht 20% der Datenbasis. Um eine Verfälschung der Ergebnisse durch Spezialisierung auf einzelne Musikgruppen zu verhindern, sind die Teilmengen bezüglich der Musikgruppen disjunkt gehalten.

2.2 Zuordnung der Musikstücke

Die Datenbasis muß neben den Musikstücken selbst auch die Zuordnung der Stücke zu den Musikstilen enthalten. Dabei stellen sich eine Reihe von Fragen. Znächst ist zu diskutieren, ob es objektive Kriterien für Musikstile gibt und ob eine objektive Zuordnung möglich und sinnvoll ist.

In [22] wird der *off beat* als ein charakteristisches Stilmittel der Rockmusik genannt. Es handelt sich dabei um eine bewußte Störung des Grundschlags (beat). Betonte Melodietöne werden nicht mit dem Grundschlag zusammen gespielt, sondern geringfügig vorgezogen. Es entsteht der Eindruck, das Tempo werde stetig schneller. Beim Hören wird so ein Spannungsverhältnis zwischen Grundschlag und Melodie wahrgenommen. Die Wirkung des *off beats* als Drive oder Swing kann wahrgenommen werden, läßt sich jedoch nicht präzise erklären. Als weiteres Stilmittel der Rockmusik wird in der gleichen Quelle *off pitch* genannt. Dabei wird versucht, die Tonhöhe nicht exakt zu spielen, sondern die Tonhöhe zu verschleiern (dirty intonation).

Entscheidend für diese Stilmittel ist aber die Wirkung auf den Hörer.

Auch die Emotionen, die mit einem Musikstück verbunden sind, spielen eine Rolle. So wird in [22] die Verknüpfung von Musik mit Urlaubserinnerungen als charakteristisch für Popmusik angesehen. Auch in [21] werden die Musikstile eher hinsichtlich ihrer Bedeutung und Wahrnehmung betrachtet.

Wie ordnet man nun die Musikstücke den Stilrichtungen zu? Sollte die Zuordnung aufgrund solcher Stilmittel erfolgen? Die Frage ist, welche Informationen man gewinnen möchte. Ist man interessiert, Informationen über einzelne Stilmittel zu erhalten oder gilt das Interesse nicht viel mehr den vom Menschen wahrgenommenen Informationen. Um die Bedeutung zu erfassen, ist es angebrachter, die Zuordnung wahrnehmungsbasiert vorzunehmen. Eine Möglichkeit dies zu tun, besteht darin, die Zuordnung durch eine Gruppe von Versuchspersonen vornehmen zu lassen. In dieser Arbeit wird aber anders vorgegangen. Die Zuordnung wird durch den Autor allein vorgenommen, anschließend wird diese Zuordnung durch eine emprische Studie überprüft. Die Zuordnung der Musikstücke in der Datenbasis basiert auf die Zuordnung der Interpreten und Musiker zu einer Stilrichtung. So werden alle Stücke von *U96* dem Techno-Stil zugeordnet. In der gleichen Art und Weise werden die Klaviersonaten von *Mozart* der klassischen Musik zugeordnet. Diese Methode scheint bei Klassik und Techno unproblematisch. Wie steht es aber mit Rock und Pop? So spielen Bands sowohl Rockmusik als auch Popmusik. Die Untersuchungen in dem Kapitel *Experimente zur Wahrnehmung von Musik* dienen zur Beantwortung dieser Frage. Die Resultate bestätigen die gewählte Vorgehensweise bei der Zuordnung der Stücke.

Kapitel 3

Stand der Forschung

In dem einführenden Kapitel ist der Gegenstand der vorliegenden Arbeit umrissen worden. Es ist nun interessant zu fragen, in welches Forschungsgebiet das Erkennen von Musikstilen gehört. Aus Sicht der Informatik ist dies sicherlich der Bereich der Künstlichen Intelligenz. Diese Sichtweise ist aber nur unzureichend. Vielmehr handelt es sich um ein interdisziplinäres Gebiet. Neben der Informatik sind auch die Disziplinen Musik und Psychologie beteiligt.

Zwei Gebiete beschäftigen sich mit dem Wahrnehmen, dem Erkennen und dem Verstehen akustischer Ereignisse. Dies sind die Gebiete Spracherkennung und Musikerkennung. Das Ziel der Spracherkennung ist den meisten Lesern wohl bekannt - die Entwicklung eines Systems zum Erkennen und Verstehen gesprochener Sprache. Welches Ziel hat sich nun die Musikerkennung gesetzt? Dies ist die automatische Transkription von Musikstücken. Die Identifikation von Noten ist ein Schwerpunkt der Forschung. Andere Arbeiten beschäftigen sich mit der Erkennung von Musikinstrumenten. Exemplarisch werden einige Arbeiten zu dem Thema Musikerkennnung in den nächsten beiden Abschnitten vorgestellt.

In dem letzten Abschnitt wird schließlich eine Arbeit vorgestellt, die sich nicht direkt den Einzelgebieten Musikerkennung oder Spracherkennung zuordnen läßt. Gegenstand ist die Trennung von Musik und Sprache.

3.1 Notenidentifikation

Ein wichtiger Bestandteil eines Musikerkennungssystems ist das Erkennen der Tonhöhe der gespielten Note. Gemeinsamer Ausgangspunkt der vorgestellten Arbeiten ist das Berechnen des Leistungsspektrums. Die Analyse der Leistungsspektren basiert auf der Annahme, daß die Frequenzanteile harmonisch sind. Die Partialtöne eines Tons mit Tonhöhe f liegen also bei genau $2f$,$3f$,.. usw. Beschäftigen wir uns nun mit den Arbeiten im Detail.

- Die Arbeit von Judith C. Brown in [3, 4] beruht auf einer spektralen Q-Transformation. Diese Transformation ist äquivalent zu einer Filterbankanalyse mit logarithmischer Frequenzskalierung. Vorteil der Q-Transformation gegenüber der gewöhnlichen Kurzzeit-Fouriertransformation ist eine von dem Frequenzband abhängige Frequenzauflösung. Man muß aber einen höheren Rechenaufwand in Kauf nehmen. Durch die logarithmische Frequenzskalierung wird erreicht, daß die Abstände aufeinanderfolgender Partialtöne unabhängig von der Grundtonhöhe f ist (beispielsweise ist $log(3f) - log(2f) = log(3) - log(2)$). Aufgrund dieser Tatsache kann nun zur Feststellung der Grundtonhöhe ein Mustervergleich vorgenommen werden.

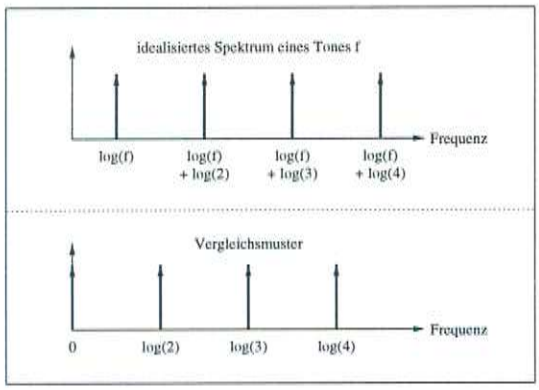

Abbildung 3.1: Identifikation von Noten

Das Spektrum wird kreuzkorreliert mit einem Muster bestehend aus 0'en und 1'en. Die 1'en sind an den Stellen der harmonischen Frequenzkomponenten positioniert, die Abstände der 1'en sind also $0, log(2/1)$

$log(3/2)$ usw. (siehe Abbildung 3.1). Die Amplitude der Kreuzkorrelationsfunktion muß dann bei $log(f)$ eine Spitze haben. In den Experimenten worden jeweils einzelne Noten gespielt. Experimente auf verschiedenen Musikinstrumenten lieferten gute Ergebnisse in der Erkennung. Genaue Erkennungsraten sind aber nicht angegeben.

- Größere Schwierigkeiten ergeben sich, wenn mehrere Noten simultan gespielt werden. Ein Ansatz zur Lösung dieses Problems wird in [6] vorgestellt. Simon Dixon geht dabei in zwei Schritten vor. Nach Berechnen der Kurzzeit-Fouriertransformation werden die Amplitudenspitzen des Spektrums detektiert. Dies geschieht, indem nach lokalen Maxima gesucht werden. In dem zweiten Schritt werden die Frequenzanteile eliminiert, die ein ganzzahliges Vielfache andererer Frequenzanteile sind. Übrig bleiben die Frequenzanteile der Grundtöne. Damit lassen sich die Noten feststellen. Die Experimente basieren auf Gitarrenmusik. Es werden etwa 95% der Noten richtig erkannt.

- Tanguiane [23] versucht die Tonhöhe durch eine Autokorrelataion des Spektrums bezüglich der Frequenz festzustellen. Er erweitert diesen Ansatz auf polyphone Stücke, indem eine mehrdimensionale Autokorrelation berechnet wird. Getestet ist dieser Ansatz jedoch nur auf künstliche Daten, die Spektren bestanden lediglich aus 0'en und 1'en.

3.2 Erkennung von Instrumenten

Publikationen zum Erkennen von Musikinstrumenten sind eher rar gesät. Vorgestellt wird eine Arbeit, mittels Kohonenkarten Musikinstrumente zu erkennen. Ansätze zur Trennung von mehreren gleichzeitig spielenden Musikinstrumenten in einem Orchester konnten nicht gefunden werden. Interessant wäre eine Untersuchung, inwieweit sich Arbeiten auf dem Gebiet der Quellenseparierung nutzen lassen.

- Cosi et al. stellen in [5] ein System zur Klassifikation von 12 Musikinstrumenten vor. Das akustische Signal ist jeweils 300 ms lang. Gespielt wurde jeweils die Note $C4$ (261 Hz). Die Ausschnitte enthalten also immner nur ein 'Ereignis'. Die Vorverarbeitung des Signals besteht

aus mehreren Stufen mit dem Ziel die biologischen Gegebenheiten genauer zu berücksichtigen. Der Rechenaufwand liegt bei etwa 100-fach Echtzeit. Die Vorverarbeitung eines 10 Sekunden langen Aussschnittes würde also fast 17 Minuten dauern.

Das Lernen erfolgt dann mit einem Kohonen-Netz unüberwacht. Es bilden sich Regionen des Netzes aus, die zu den einzelnen Musikinstrumenten korrespondieren. Die Klassifikation erfolgt dann durch die Berechnung des maximalen Erregungszentrums.

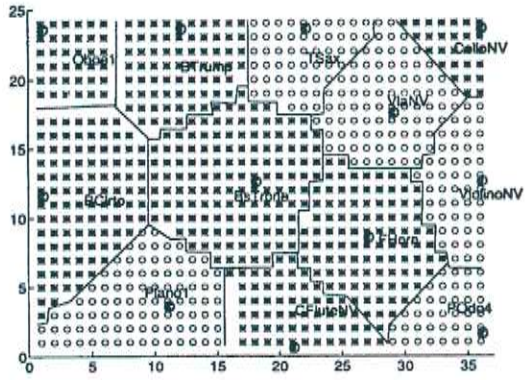

Abbildung 3.2: Kohonen-Karte der Instrumente, von Cosi et al.

Erkennungsleistungen werden leider nicht angegeben. Interessant wäre sicherlich auch ein Vergleich mit einem überwachten Lernverfahren. Auch ein Vergleich der *Advanced Auditory Modelling Analysis* mit einer gewöhnlichen Filterbankanalyse wäre hilfreich, um den Nutzen einer so aufwendigen Vorverarbeitung erkennen zu können.

- In [24] berichten Toiviainen et al. von Experimenten zur Wahrnehmung von Klangfarben. In einem ersten Versuch sollten Versuchspersonen die Ähnlichkeit von Tönen bewerten. Die Grundtonhöhe war jeweils 440 Hz. Unterschiede ergaben sich bei der Ankling- und Abklingzeit und dem Verhältnis der Partialtöne. Hintergrund ist dabei folgender Sachverhalt: Ein Ton mit der Grundfrequenz f hat Partialtöne bei $2f$, $3f$ usw. Die Klangfarbe eines Tones kann nun durch das Verhältnis der Energien der Partialtöne charakterisiert werden. Die Klangfarbe beschreibt das Resonanzverhalten der Musikintrumente.

Es gab insgesamt 27 unterschiedliche Töne. Bei einem paarweisen Vergleich bestand ein Experiment aus der Darbietung von $27 * 26/2 = 351$ Tonpaaren. Auf einer Skala von 0 bis 12 sollte jeweils die Ähnlichkeit der beiden Töne bewertet werden.

In dem zweiten Experiment diente eine Kohonenkarte zur Ähnlichkeitsmesssung der Töne. Mit einer Filterbankanalyse sind die Daten zunächst vorverarbeitet worden. Ein zweidimensionales Gitter bildete die Nachbarschaftsstruktur der Kohonenkarte. Zu jedem Ton gibt es dann ein Zentrum maximaler Erregung in der gitterförmigen Struktur. Der Abstand zweier Töne kann als der Abstand der Erregungszentren aufgefaßt werden.

Festgestellt wurde in der Studie, daß die von den Versuchspersonen angegebene Ähnlichkeit von Tönen zu dem maschinellen Maß korreliert ist. Die Korrelation zwischen den Ähnlichkeitsmaßen der Versuchspersonen und dem Kohonen-Maß beträgt $0,59$. Zum Vergleich sind auch Korrelationen *zwischen* den Ähnlichkeitsmaßen der Versuchspersonen angegeben. Die Korrelationen der durch die Vp definierten Maße schwanken zwischen $0,532$ und $0,69$. Die Autoren schlußfolgern, daß die mentale Repräsentation durch solche topologischen Karten modelliert werden können.

3.3 Trennung von Musik und Sprache

Gegenstand der Arbeit von J. Saunders in [19] ist die Klassifikation eines akustischen Signals in die Klassen Musik oder Sprache. Der Begriff Trennung gibt den Sachverhalt aber nur ungenau wieder. Gemeint ist nicht die Trennung zweier sich überlagernder Teilsignale, wie etwa die Trennung von Sprache und Hintergrundmusik. Einen solchen Ansatz zur Quellenseparierung, der hier aber nicht weiter beschrieben ist, wird von Te-Won Lee et.al. in [11] vorgestellt. Das hier beschriebene Problem ist dagegen ein Klassifikationsproblem. Bei dem zu untersuchenden Signal handelt es sich entweder um Musik oder um Sprache. Zur Klassifikation verwendet J. Saunders einen Gaußklassifikator. Der Eingaberaum wird durch zwei Merkmale definiert.

1. Nulldurchgangsrate
   Ein wichtiges Merkmal für die Klassifikation ist die Nulldurchgangsrate.

Dabei werden die Anzahl der Nulldurchgänge des Signals im Zeitbereich gezählt und man erhält ein grobes Maß für die Frequenz. Die Bandbreite von Musiksignalen ist typischerweise größer als die der Sprache. Eingabe für den Klassifikator bilden Mittelwert und Standardabweichung der Nulldurchgangsrate.

2. Energiekonturen
   Sprache und Musik unterscheiden sich auch typischerweise in den Energiekonturen des Signals. So folgen in der Sprache stimmhafte Abschnitte stimmlose. Der Klassifikator erhält daher als Eingabe die Anzahl der Energieminima des Signals.

Getestet wurde der Klassifikator auf Radiodaten. Als mögliche Anwendung wird das Erkennen von Werbeunterbrechungen genannt. Die Erkennungsrate ist $98,4\%$. Aufgrund der Einfachheit des Systems ist die Klassifikation auch in Echtzeit möglich.

Kapitel 4

Wahrnehmung von Musik

Im Gegensatz zu Problemen der Mathematik und der Physik kann die Wahrnehmung von Sprache und Musik nicht durch eine Struktur von Axiomen und Theoremen korrekt erfaßt werden. Dies gilt insbesondere bei der Zuordnung eines Musikstückes zu einer Stilrichtung. Wird nun ein maschineller Klassifikator für Musikstile gebaut, so stellt sich dann auch die Frage nach der Bewertung der Leistung eines solchen Klassifikators. Wie ist ein Klassifikator einzuschätzen, der 3/4 aller Musikstücke richtig erkennt? Eng verbunden mit dieser Frage ist die Frage nach der menschlichen Wahrnehmung von Musik. Die Übertragkeit der Verhältnisse bei der menschlichen Erkennung auf ein maschinelles System muß dabei allerdings in einem gewissen Rahmen vorausgesetzt werden. Es scheint aber plausibel, anzunehmen, daß, falls die Zuordnung von Musik zu einem Musikstil dem Menschen sehr leicht fällt, dies tendenziell auch für einen maschinellen Klassifikator gilt. So ist es nun folgerichtig, als Maßstab für eine maschinelle Klassifikation die Wahrnehmung des Menschen zu nehmen. Aus diesem Grunde wurden Experimente durchgeführt, die zum Ziel hatten, die Leistungsfähigkeit der Menschen zu untersuchen. Die Resultate dieser Untersuchungen sind in dem Abschnitt *Darbietung nicht verfremdeter Ausschnitte* dargelegt.

Gegenstand der Untersuchungen ist ebenfalls die Relevanz bestimmter Merkmale für die Wahrnehmung eines Musikstückes. Dem diente die Präsentation modifizierter Musikstücke. Die Ergebnisse dieser Versuche sind in dem Abschnitt *Darbietung verfremdeter Ausschnitte* festgehalten.

4.1 Versuchsaufbau und Versuchspersonen

Für die Experimente standen 37 Versuchspersonen (Vp) im Alter von 22 bis 50 Jahren zur Verfügung, davon waren 9 Frauen und 28 Männer. Alle Vp hatten eine abgeschlossene Hochschulausbildung oder befanden sich noch im Studium. Die Versuchsumgebung war für alle Vp gleich. Die Vp waren durchweg kooperationsbereit, es wurde häufig Interesse an der Studie und der Auswertung der Ergebnisse bekundet.

In einem einführenden Gespräch mit den Vp wurden Art und Weise des Experimentes erläutert. Desweiteren sollte so die Konzentration auf das Geschehen gefördert werden. Anschließend ist den Vp ein Fragenkatalog (siehe Abbildung 4.1) vorgelegt worden. Hauptaugenmerk der Fragen galt den Hörgewohnheiten, da als Arbeitshypothese eine Abhängigkeit der Erkennungsleistung von den Hörgewohnheiten zugrunde gelegt wurde. Zudem sollten die Vp schätzen, wie lang ein Ausschnitt eines Musikstückes dauern muß, um die Stilrichtung zu erkennnen.

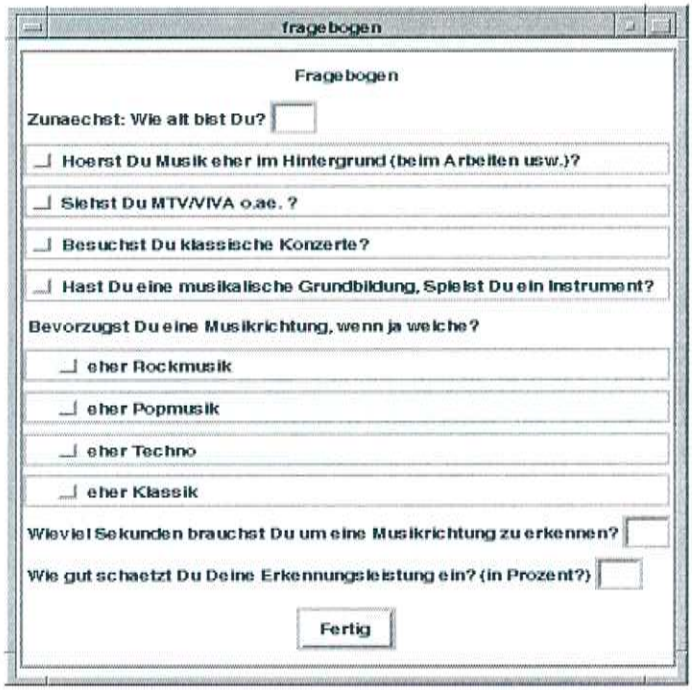

Abbildung 4.1: Fragebogen bei der Untersuchung der Musikwahrnehmung

Die für die Experimente verwendeten Musikstücke entstammen der Testmenge, die für die Evaluation des Gesamtsystems verwendet wird. Der Anteil der Präsentationsmenge an der gesamten Testmenge beträgt 33, 3% und wurde zufällig aus der Testmenge gezogen. Die Musikstücke sind bezüglich der Musiksstile gleichverteilt. Zur Untersuchung der für die Wahrnehmung benötigten Länge von Musikstücken wurden Ausschnitte von 1sec Länge und von 3sec Länge gespielt. Die längeren Ausschnitte wurden zudem zwei Modifikationen unterzogen. Die Modifikationen werden in den folgenden Abschnitten näher erläutert. Zusammenfassend wurden vier Arten von Ausschnitten dargeboten:

- 1sec-Ausschnitte
- 3sec-Ausschnitte
- Modifikation der 3sec-Ausschnitte zur Relevanz kurzzeitiger Merkmale
- Modifikation der 3sec-Ausschnitte zur Relevanz langzeitiger Merkmale

Die Reihenfolge dieser Ausschnitte war zufällig und für jede Vp unterschiedlich, aber unter der Maßgabe, daß die kürzeren Ausschnitte vor den längeren Ausschnitten gespielt wurden. Letzteres erfolgte, um Wiedererkennungseffekte zu vermeiden. Zu Beginn des Experimentes erhielten die Vp Informationen über die Art der Musikstücke, nicht jedoch über die genaue Dauer der Ausschnitte und die Art der Modifikationen.

Gegenstand der Untersuchung ist die Wahrnehmung von Musik, nicht aber das Lernen von Zuordnungen. Aus diesem Grunde fand die Darbietung der Musikstücke unüberwacht statt. Nach Präsentation eines Ausschnittes und anschließender Bewertung wurde die Vp nicht über den Musikstil des soeben gehörten Musikstückes informiert.

Im Anschluß des Exerimentes erfolgte eine Nachbefragung. So wurde wie zu Beginn nach einer Schätzung der benötigten Dauer der Ausschnitte gefragt. Gefragt wurde auch nach einer eventuellen Strategie, ob auf Vokalmusik oder einzelne Instrumente geachtet wurde. Außerdem wurde um einen Gesamteindruck gebeten und insbesondere, ob die Versuchsdauer von 15 bis 25 min als zu lang empfunden wurde.

Eine Unterscheidung nach Männer und Frauen bei der Darstellung der Ergebnisse in den folgenden Abschnitten erfolgt nicht. Es war nicht die Aufgabe

der Studie zu prüfen, ob die Erkennung von Musikstilen geschlechtsspezifisch ist.

4.2 Darbietung nicht verfremdeter Ausschnitte

4.2.1 Experimente

Wie bereits im vorhergehenden Abschnitt beschrieben, wurden nicht verfremdete Ausschnitte mit 1sec und 3sec Dauer dargeboten. Primäres Anliegen der Darbietung nicht verfremdeter Ausschnitte war es, Informationen zu der Genauigkeit und Zuverlässigkeit der Zuordnungen der Vp zu erhalten. Die Ergebnisse zu diesen Punkten beziehen sich auf die 3sec-Ausschnitte. Ebenfalls auf Basis der 3sec-Ausschnitte werden Konfusionen und der Einfluß der Hörgewohnheiten diskutiert. Zur Gewinnung von Informationen über die benötigte Dauer nicht verfremdeter Ausschnitte wurden unterschiedliche lange Ausschnitte gespielt. Ein Vergleich der Erkennungsleistungen der 1sec- und 3sec-Ausschnitte soll Aufschluß darüber geben. Die Experimente dienen also zur Beantwortung folgender Fragen:

- Genauigkeit der Zuordnungen
- Einfluß der Hörgewohnheiten
- benötigte Dauer der Ausschnitte

4.2.2 Ergebnisse

Genauigkeit der Zuordnungen Wenden wir uns zunächst der Klärung der ersten Frage zu. Betrachten wir dazu die Abbildung 4.2. Insgesamt haben die Vp 84,9% der präsentierten unverfremdeten 3sec-Abschnitte richtig zugeordnet. Es fällt auf, das Pop und Klassik mit 92,4% bzw. 95,9% überdurchschnittlich gut erkannt werden, hingegen Rock mit 69,8% Probleme bereitet.

Die Anzahl richtig erkannter Musikstile gibt aber nur ungenau Auskunft über die Erkennungsleistung. Vielmehr muß auch die Anzahl der Stücke berücksichtigt werden, die einem Musikstil zugeordnet werden. So zeigt sich,

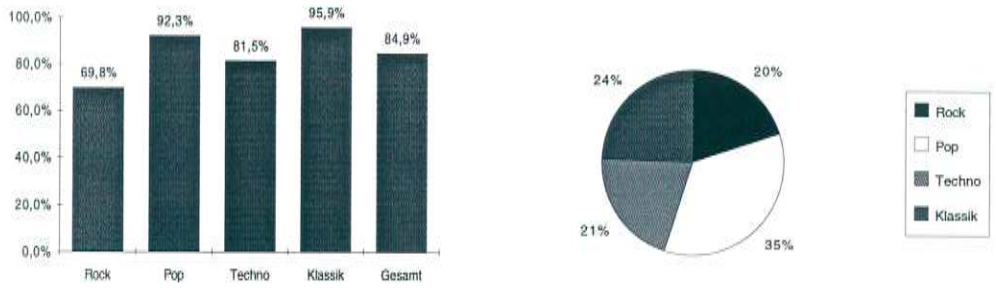

Abbildung 4.2: Erkennung und Dominanz von 3sec-Ausschnitten

daß überdurchschnittlich viele Stücke der Popmusik zugeordnet werden. Es werden 35% aller Stücke diesem Stil zugeordnet. Bedenkt man, daß die dargebotenen Stücke gleichverteilt sind, so kann also von einer Dominanz der Popmusik gesprochen werden. Im Gegenzug sind Rock mit 20% und Techno mit 21% deutlich unterrepräsentiert. Der Anteil zugeordneter klassischer Stücke entspricht mit 24% etwa dem tatsächlichen Anteil.

Fragen\Antworten	Rock	Pop	Techno	Klassik	$\sum$
Rock	155	64	2	1	222
Pop	16	205	0	1	222
Techno	7	34	181	0	222
Klassik	0	8	1	213	222
$\sum$	178	311	184	215	

Tabelle 4.1: Konfusionen bei der Zuordnung von 3sec-Ausschnitten

Eine genauere Analyse ist durch die Betrachtung der Konfusionen möglich. Die Spalten in Tabelle 4.1 geben die Antworten der Vp zu den in den Zeilen gestellte Fragen an. So wurden 64 Rockmusik-Ausschnitte der Popmusik zugeordnet. Bemerkenswert ist aber, daß umgekehrt nur 16 Ausschnitte der Popmusik als Rockmusik angesehen wurden. Wenn die Konfusionen bei Techno und Klassik mit Pop berücksichtigt werden, so wird die Dominanz der Popmusik ersichtlich.

Einfluß der Hörgewohnheiten Eine ebenso interessante Frage ist der Einfluß der Hörgewohnheiten auf die Erkennungsleistung. Zu diesem Zwecke

wurden die Fragebögen der Vp des oberen Quartils mit denen des unteren Quartils verglichen. Dabei werden durch das obere bzw. untere Quartil die Erkennungsleistungen der 25% (= 9 von 37) besten bzw. schlechtesten Vp erfaßt. Die durchschnittliche Erkennungsleistung der 9 besten Vp beträgt 93, 5%, die der schlechtesten 76, 8%. In Abbildung 4.3 sind die Hörgewohnheiten der beiden Gruppen ersichtlich. Die Hauptunterschiede liegen bei der Beantwortung der Frage nach der musikalischen Bildung und der Art des Hörens. So gaben 7 von 9 in der Gruppe mit den besseren Erkennungsleistungen an, über eine musikalische Grundbildung zu verfügen. In der anderen Gruppe waren dies nur 4 von 9. Unterschiede sind ebenfalls bei den Antworten zu dem Punkt Hintergrundmusik vorhanden. In der besseren Gruppe gaben 7 von 9 an, sie hören Musik eher im Hintergrund, weit mehr als in der anderen Gruppe. Dort gaben lediglich 3 Vp an, Musik eher im Hintergrund zu hören.

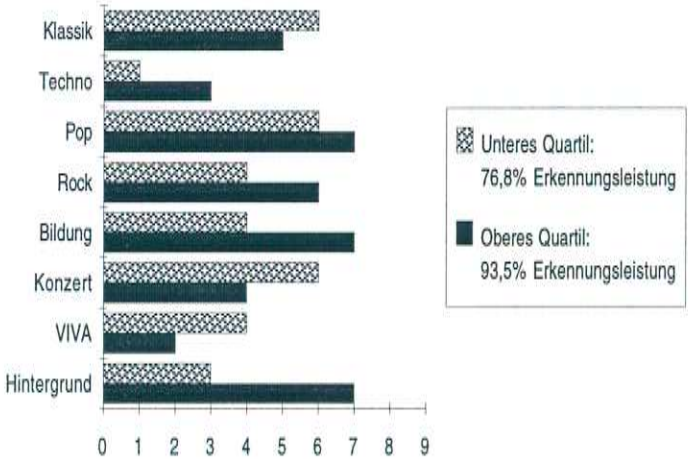

Abbildung 4.3: Vergleich der Hörgewohnheiten in Relation zur Erkennungsleistung

Dauer der Ausschnitte Neben dem Einfluß der Hörgewohnheiten ist ein weiteres, interessantes Resultat die benötigte Dauer der Ausschnitte, um Musikstile zu erkennen. Informationen zu diesem Punkt lassen sich in zwei Teile gliedern.

- Schätzung der Vp in der Vor- und Nachbefragung
  Die Vp sollten schätzen, wie lange der Ausschnitt dauern müßte. Die Vp gaben vor dem Versuch im Durchschnitt 11 Sekunden an, nach dem Versuch waren es 4,7 Sekunden, also deutlich weniger. Den Vp war die tasächliche Dauer der Ausschnitte nicht bekannt.

- Erkennung von 1sec- und 3sec-Ausschnitten
  Die bereits angesprochene Abbildung 4.2 enthält die Erkennungsleistung der 3sec-Ausschnitte. Es werden 84,9% der Stücke korrekt zugeordnet. Bei der Darbietung der 1sec-Ausschnitte sind 82,9% der Stücke richtig erkannt worden.

4.2.3 Interpretation

Genauigkeit der Zuordnungen Die Anzahl richtig erkannter Stücke eines Musikstils muß in Bezug gesetzt werden mit dem Anteil der Stücke, die diesem Musikstil zugeordnet werden. Wenn sich die Vp sehr häufig für den Popstil entscheiden, dann ist die Anzahl richtig erkannter Popstücke hoch. Dies bedeutet aber dann nicht, daß die Vp die Popstücke *erkannt* haben. Es kann also nicht davon gesprochen werden, daß die Erkennung von Pop besonders gut ist. Die 92,4% richtig erkannter Popstücke müssen relativiert werden. Analog kann die Erkennung klassischer Stücke analysiert werden. Es zeigt sich, daß den Vp die Erkennung klassischer Stücke eher leicht fällt. Es werden 95,9% richtig zugeordnet. Auch werden kaum Musikstücke anderer Stilrichtungen dem klassischen Stil zugerechnet. Konfusionen sind hauptsächlich im Bereich Rock/Pop zu finden. Die Verwechselungen sind aber einseitig. Dies findet man auch bei den Konfusionen von Pop- und Technostücken. Auch hier ist die Beziehung eher einseitig. Man kann vermuten, daß die Vp sich im Zweifelsfalle für den Popstil entscheiden.

Einfluß der Hörgewohnheiten Mit der musikalischen Grundbildung ist vermutlich ein besser geschultes Gehör verbunden, welches für die Erkennung sicherlich von Vorteil ist. Diskussionswürdig sind aber auch die Angaben zu dem Punkt Hintergrundmusik. Aufgrund der Tabelle 4.3 kann die Hypothese aufgestellt werden, daß Menschen, die eher Hintergrundmusik hören, besser Musikstile erkennen können. Menschen, die eher nicht Hintergrundmusik

hören, suchen sich die Musik, die sie hören, bewußt aus. Es kann nun vermutet werden, daß das Musikspektrum dieser Vp kleiner ist. Versuchspersonen, die dagegen eher Musik im Hintergrund hören, hören dann auch Musikstücke, die sie nicht mögen. Dies kann die unterschiedlichen Erkennungsleistungen erklären. Zusammenfassend können also zwei Hypothesen aufgestellt werden.

- Hypothese 1: Menschen mit musikalischer Grundbildung sind eher in der Lage, Musikstile zu erkennen.
- Hypothese 2: Menschen, die Hintergrundmusik hören, können besser Musikstile erkennen.

Mit obigem Zahlenmaterial können diese Hypothesen motiviert werden, nicht jedoch aber begründet. Zweifler könnten entgegenhalten, die Unterschiede seien rein zufällig. Oder anders formuliert: Kann nachgewiesen werden, daß die Erkennungsleistungen signifikant von der musikalischen Grundbildung bzw. der Hintergrundmusik abhängig sind? Dies läßt sich mit einem Signifikanztest klären. Betrachten wir zwei Stichproben, bestehend aus der Versuchsgruppe mit musikalischer Grundbildung und ohne Grundbildung. Geprüft wird, ob die Nullhypothese H_0 (Unterschiede der Versuchsgruppen rein zufällig) mit Sicherheitswahrscheinlichkeit $1-\alpha$ abgelehnt werden kann. In diesem Falle ist die Alternativhypothese (Unterschiede der Versuchsgruppen sind nicht zufällig) mit Wahrscheinlichkeit $1-\alpha$ statistisch gesichert. Für einen Test zum Sicherheitsniveau α gilt, daß die Wahrscheinlichkeit für einen Fehler 1. Art (irrtümliche Ablehnung der Nullhypothese) maximal α ist. Die Wahrscheinlichkeit für einen Fehler erster Art hängt unter Normalverteilungsannahme der x_i von der Schätzung von Mittelwert und Varianz ab. In van de Waerden [28] wird nachgewiesen, daß der Quotient der Zufallsvariablen Mittelwert $\bar{X}$ und Streuung S der Student-Verteilung (t-Verteilung) gehorcht. Damit die Wahrscheinlichkeit für eine irrtümliche Ablehnung der Nullhypothese kleiner α ist, wird die Nullhypothese nur dann abgelehnt, wenn:

$$|\mu_1 - \mu_2| \geq \sqrt{\frac{(N_1+1)*\sigma_1 + (N_2+1)*\sigma_2}{N_1+N_2+2} * (\frac{1}{N_1}+\frac{1}{N_2})} * t_{N_1+N_2-2,1-\alpha}$$

Dieser Test wird *T-Test* genannt. Den Konventionen entsprechend ist dabei N die Stichprobengröße, μ der Mittelwert und σ die Varianz. $t_{N_1+N_2-2,1-\alpha}$

bezeichnet das $1-\alpha$ Fraktil der Student-Verteilung zum Freiheitsgrad $N_1 + N_2 - 2$. Das Fraktil (auch Quantil genannt) gibt denjenigen Wert an, so daß die Wahrscheinlichkeit, daß eine Student-verteilte Zufallsvariable kleiner diesen Wert ist, gerade $1-\alpha$ ist. Je größer also die Differenz der Mittelwerte und je kleiner die Varianz der Stichproben ist, desto eher sind die Unterschiede signifikant.

Ein Wort noch zur Gültigkeit des Tests: Neben der Normalverteilung der Erkennungsleistung ist zu fordern, daß die Größe der Stichproben gleich ist, oder daß sich die Varianzen nicht signifikant unterscheiden. Die Signifikanz der Varianzendifferenz läßt sich mit dem *F-Test* überprüfen, dem die Fisher-Verteilung zugrunde liegt. Die Unterschiede der Varianzen sind nicht signifikant, wenn

$$\frac{\max \sigma_1, \sigma_2}{\min \sigma_1, \sigma_2} \leq f_{N_1-1,N_2-1,1-\alpha}$$

Kommen wir nun zum praktischen Teil des Signifikanztests. Geprüft wird zum Signifikanzniveau $\alpha = 0,05$. Mittelwert, Varianz und Stichprobengröße der Versuchsgruppen mit und ohne musikalischer Grundbildung können der Tabelle 4.2 entnommen werden.

Versuchsgruppe	Mittelwert	Varianz	Stichprobe
Vp mit musikl. Grundbildung	87,3	53,6	19
Vp ohne musikl. Grundbildung	82,4	36,4	18
Vp mit Hintergrundmusik	86,7	44,5	22
Vp ohne Hintergrundmusik	82,2	48,5	15

Tabelle 4.2: Signifikanz der Hörgewohnheiten

Anhand dieser Zahlen können die Hypothesen geprüft werden. Demnach besteht ein signifikanter Unterschied der Erkennungsleistungen der Versuchsgruppen. Vp mit musikalischer Grundbildung können die Musikstile signifikant besser erkennen. Das gleiche Bild ergibt sich bei Betrachtung der zweiten Hypothese. Vp, die häufiger Hintergrundmusik hören, sind ebenfalls siginifikant besser. Beide Tests sind auch gültig, da sich mit dem F-Test keine signifikanten Unterschiede der Varianzen feststellen läßt. Somit können beide Hypothesen akzeptiert werden. Allerdings sollte man exakter formulieren: Die Hypothese, daß die Unterschiede rein zufällig sind, kann abgelehnt werden.

- Test der Hypothese 1: Musikalische Bildung
  T-Test: 4,9 $\geq$ 2,23*1,697 $\Rightarrow$ signifikante Unterschiede
  F-Test: 1,47 $\leq$ 2,18 $\Rightarrow$ Test gültig

- Test der Hypothese 2: Hintergrundmusik
  T-Test: 4,5 $\geq$ 2,22*1,697 $\Rightarrow$ signifikante Unterschiede
  F-Test: 1,09 $\leq$ 2,15 $\Rightarrow$ Test gültig

Dauer der Ausschnitte Einen Anhaltspunkt zu der erforderlichen Dauer der Ausschnitte gibt die Schätzung der Vp. Durchschnittlich 11 Sekunden gaben die Vp in der Vorbefragung an. In der Nachbefragung korrigierten die Vp die Schätzung auf 4,7 Sekunden. Diese deutliche Reduzierung kann bereits vermuten lassen, daß nur Abschnitte kurzer Dauer benötigt werden. Ein Vergleich der Erkennungsleistungen der 1sec-Ausschnitte mit den der 3sec-Ausschnitte läßt genauere Schlußfolgerungen zu. Die Erkennungsleistung bei den 1sec-Ausschnitten ist 82,9% und 84,9% bei den längeren Ausschnitten. So gravierend fällt die Verbesserung nicht aus.

Analysiert man, bei welchen Musikstücken Veränderungen auftreten, so findet man, daß die Veränderungen bei einzelnen Musikstücken konzentriert sind. Dies legt den Schluß nahe, daß die Repräsentanz der Ausschnitte der betreffenden Musikstücke variert. Nicht jeder Ausschnitt ist gleich repräsentativ für ein Musikstück. Konkret tritt dies etwa bei einem Stück der Popband *Wet Wet Wet* auf. Diese Vermutung läßt sich allerdings nicht überprüfen, da nicht erfaßt ist, inwieweit ein Ausschnitt repräsentativ für eine Musikrichtung ist.

dargebotene Ausschnitte	Mittelwert	Varianz	Stichprobengröße
1sec Ausschnitte	82,9	89,4	37
3sec Ausschnitte	84,9	50,7	37

Tabelle 4.3: Signifikanz der Ausschnittsdauer

- Hypothese 3: Zur Erkennung des Musikstils werden nur Ausschnitte kurzer Dauer benötigt.

Den Einfluß der Dauer auf die Erkennungsleistung kann in ähnlicher Weise statistisch geprüft werden, wie die Hypothesen 1 und 2 überprüft worden sind. Nun ist aber zu beachten, daß die beiden Stichproben verbunden

(voneinander abhängig) sind, da die 1-sec Ausschnitte Bestandteil der 3sec-Ausschnitte sind. Dadurch reduziert sich das Problem auf einen Student-Test mit einer Stichprobe. Die Stichprobe setzt sich aus den Differenzen d_i der Meßwerte x_i und y_i der 1sec- und 3sec-Stichprobe zusammen. Sei σ_d die Varianz der Differenz und μ_d der Mittelwert der Differenz. Die Unterschiede sind signifikant, wenn:

$$|\mu_d| \geq \sqrt{\frac{\sigma_d}{N}} * t_{N-1,1-\alpha}$$

Unter Berücksichtigung des in Tabelle 4.3 referierten Zahlenmaterials kann kein signifikanter Unterschied der Erkennungsleistungen der 1sec-Ausschnitte und der 3sec-Ausschnitte festgestellt werden. Sofern dieses Resultat auf einen maschinellen Klassifikator übertragbar ist, würde es also genügen, dem Klassifikator hinreichend kurze Ausschnitte zu präsentieren.

- Test der Hypothese 3: benötigte Dauer
  T-Test: 2,0 $\leq$ 1,769*1,697 $\Rightarrow$ nicht signifikante Unterschiede

Zuordnung der Musikstücke Weiterhin rechtfertigen die in den Experimenten gefundenden Ergebnisse die Vorgehensweise bei der Zuordnung der Musikstücke (siehe auch dazu Kapitel *Beschreibung der Datenbasis*). Um dies zu belegen, betrachten wir eine (hypothetische) Zuordnung der Musikstücke gemäß einer Mehrheitsentscheidung. Wären die Erkennungsleistungen nach dieser Zuordnung zugunsten einer Mehrheitsentscheidung signifikant besser, so müßte man das in Kapitel 2 erläuterte Vorgehen bei der Zuordnung in Frage stellen. Dies ist aber nicht der Fall. Die Verbesserung beträgt $0,9\%$, wenn die Ergebnisse auf Basis der 3sec-Ausschnitte berücksichtigt werden. Statt $84,9\%$ werden dann $85,8\%$ bei Anwendung der hypothetischen Mehrheitsentscheidung richtig erkannt.

Musikstück	Zuordnung bei 1sec		Zuordnung bei 3sec	
	Rock	Pop	Rock	Pop
Meat Loaf	23	14	16	21
ZZ-Top	31	5	17	20

Tabelle 4.4: betroffene Musikstücke bei einer Mehrheitsentscheidung

Konkret wären zwei Rockstücke betroffen, die eher der Popmusik zugeordnet werden. Betrachten wir nun aber die Trefferquoten der beiden Musikstücke bei den 1sec-Ausschnitten der Tabelle 4.4. Dort ergibt sich ein völlig anderes Bild. Bei der Darbietung der 1sec-Ausschnitte rechnete die Mehrheit der Vp die beiden Stücke der Rockmusik zu. Es scheint also eher der Fall vorzulegen, daß die Repräsentanz der 3sec-Ausschnitte die Zuordnung beeinflußt.

4.3 Darbietung verfremdeter Ausschnitte

4.3.1 Experimente

Gegenstand dieses Abschnittes ist die Relevanz bestimmter Merkmale der Musikstücke bei der Unterscheidung der Musikstile. Sofern sich die unten diskutierten Resultate auf einen maschinellen Klassifikator übertragen lassen, kann sich die Vorverarbeitung der Daten auf diese Erkenntnisse stützen. Untersucht wurden die Relevanz kurzzeitiger und langzeitiger Merkmale.

- Kurzzeitmerkmale
  Zur Betonung der Kurzzeitmerkmale wurde der Grundverlauf des Musikstückes zerstört. Um dies zu erreichen, wurden 50ms-Abschnitte aus dem Stück herausgeschnitten und in beliebiger Reihenfolge wieder eingeordnet.

- Langzeitmerkmale
  Eine Glättung des Signals im Zeitbereich und die Überlagerung des Signals mit weißem Rauschen soll für die Abschwächung der kurzzeitigen Merkmale sorgen.

4.3.2 Ergebnisse

Zusammengefaßt sind die Ergebnisse der Zuordnungen der verfremdeten Stücke in den Tabellen 4.4 und 4.5. Werden die Stücke durch das Vertauschen von Abschnitten modifiziert, dann werden nur noch 69,6% der Stücke richtig erkannt. 80% der Stücke werden dem richtigen Stil zugeordnet, wenn Langzeitmerkmale betont werden. Insbesondere ist die Anzahl richtig zugeordneter klassischer Stücke mit 92,8% sehr hoch.

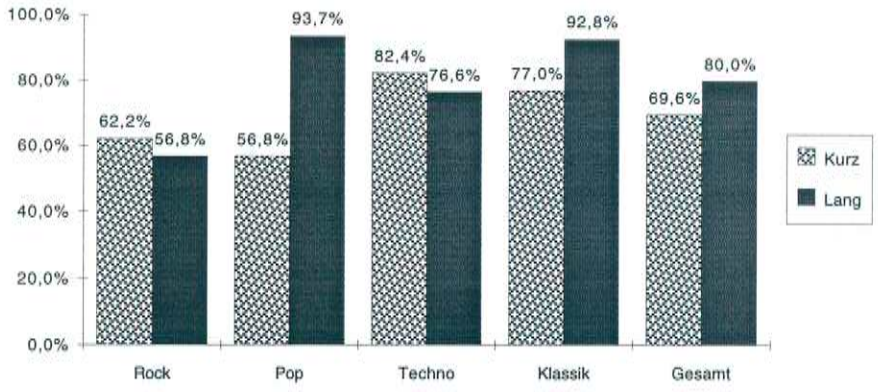

Abbildung 4.4: Erkennung verfremdeter Stücke

Die Erkennung von Rockstücken bei langzeitbetontem Merkmalen ist lediglich 62,2%. Im Gegensatz dazu schneiden die Popstücke mit 93,7% bedeutend besser ab. Eine Bewertung der Ergebnisse ist allerdings nur in Zusammenhang mit dem Dominanzverhalten möglich.

Erhebliche Unterschiede werden bei der Betrachtung der Dominanz einzelner Musikstile sichtbar (Abbildung 4.5). Der Technostil ist bei den kurzzeitbetonten Stücken mit 35,8% deutlich überrepräsentiert. Der Anteil der Stücke, die dem klassischen Stil zugeordnet worden sind, ist nunmehr bei 19,7%. Bei den langzeitbetonten Stücken ergibt sich ein völlig anderes Bild. Dort ist der Popstil dominant. Es werden 37,8% der dargebotenen Stücke als Popstücke klassifiziert. Im Gegenzug ist der Anteil der als Rock angesehenen Stücke lediglich 16,5%.

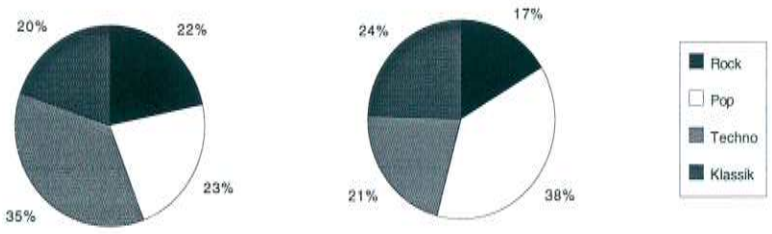

Abbildung 4.5: Dominanz kurz- und langzeitbetonter Stücke

4.3.3 Interpretation

Erkennung Werden die Erkennungsleistungen der Stücke verglichen, so kann die Hypothese aufgestellt werden, daß die Erkennung langzeitbetonter Musikstücke einfacher als die der kurzeitbetonten Musikstücke ist. Der zugehörige Test ist ein Test für verbundene Stichproben. Die beiden Stichproben enthalten die gleichen Musikstücke, nur unterschiedlich modifiziert. Die Testsituation entspricht damit der Situation bei der Prüfung der benötigten Dauer der Ausschnitte.

- Hypothese 4: Langzeitbetonte Merkmale sind für die Erkennung des Musikstils relevanter als kurzzeitbetonte Merkmale.
- T-Test für verbundene Stichproben
  10,4 $\geq$ 1,681*1,697 $\Rightarrow$ signifikante Unterschiede

Dominanz Betrachtet man die Abbildungen 4.4 und 4.5, so fällt auf, daß insbesondere die Kurzzeitmerkmale bei Popstücken unzureichend sind. Über 26, 6% der dargebotenen Popstücke wurden dem Technostil zugeordnet. Die durch das willkürliche Umordnen der 50ms-Abschnitte entstehenden Änderungen der musikalischen Ordnung werden anscheinend mit dem typischen Techno-Sound assoziiert. Ähnlich verhält es sich mit Rockstücken. Dort werden 22, 1% als Techno angesehen. Die Zuordnungen sind also über alle Stile hinweg von dem Techno-Sil dominiert. Der Anteil der dem Techno-Stil zugeordneten Stücke liegt insgesamt bei 35%.

Bei den Zuordnungen der langzeitbetonten Ausschnitte ist eine Dominanz der Popmusik vorhanden, wie sie auch den unverfremdeten Stücken zueigen ist (siehe Abbildung 4.2). Dies gilt hauptsächlich für die Verwechselungen von Rock mit Pop. So werden 35, 1% aller Rockstücke dem Popstil zugerechnet. Erklärbar wird dieser Umstand auch durch das hinzugefügte Rauschen. Die elektrisch verstärkten Gitarren, die in vielen Rockformationen von Bedeutung sind, können von ungeübten Hörern mit einer rauschbehafteten Quelle assoziiert werden. Enthalten alle präsentierten Ausschnitte ein solches Signal, so nimmt natürlich die Schwierigkeit zu, Rockmusik zu erkennen.

Vergleich der Verfremdungen Ein schlechtes Abschneiden der abschnittsvertauschten Ausschnitte kann als Argument angesehen werden, daß ein Klas-

sifikator, der jeweils 50ms-Abschnitte getrennt bewertet und dann die Bewertungen aufsummiert, wenig Aussichten auf Erfolg hat. Andererseits muß berücksichtigt werden, ob die Ursachen für die geringe Erkennungsleistung bei diesem Experiment übertragbar sind. Die Änderungen im Grundverlauf werden ja eher als technotypisch wahrgenommen, eine Übertragung dieses Sachverhalts auf einen maschinellen Klassifikator scheint also weniger angebracht. Dies ist bei einem Vergleich der kurz- und langzeitbetonter Merkmale unbedingt zu berücksichtigen. An der Arbeitshypothese von der Übertragbarkeit der Verhältnisse kann also bei dem Vergleich der Verfremdungen nicht festgehalten werden. Bei der Repräsentation des Musikwissens ist dieser Punkt zu beachten.

Zusammenfassung Insgesamt gesehen, scheinen eher langzeitige Merkmale für die Unterscheidung der Musikstile von Bedeutung zu sein. Aus diesen Untersuchungen kann gleichwohl nicht der Schluß gezogen werden, daß Kurzzeitinformationen keinen Beitrag für die Trennung der Stile leisten. Das wohl wichtigste Resultat der Experimente ist die benötigte Dauer der Ausschnitte. Schätzten die Vp vor dem Experiment, etwa 11 Sekunden zu benötigen, so gaben sie anschließend an, daß 4, 7 Sekunden ausreichend wären. Auch die Resultate der präsentierten 1sec- und 3sec-Ausschnitte zeigen, daß die relevanten Informationen in einem zeitlich lokalen Bereich liegen. Belegt sind diese Ergebnisse durch die Durchführung statistischer Tests.

Kapitel 5

Extraktion relevanter Merkmale

In der Datenbasis wird das Musiksignal im Zeitbereich durch eine Folge von quantisierten Abtastwerten der Welle repräsentiert. Diese Repräsentation ist sicherlich für eine klanggetreue Wiedergabe geeignet, nicht jedoch als Eingabe für einen Musikstilklassifikator. Die Wissensrepräsentation muß in Hinblick auf die Nutzung des Wissens angemessen sein. Die Modelle der Musikstücke sollten nur die Eigenschaften enthalten, die für die Erkennung des Musikstils relvant sind. Probleminvariante Eigenschaften sollen eliminiert werden.

Die Extraktion relevanter Merkmale aus dem Zeitbereichsignal ist auch Gegenstand der Vorverarbeitung eines Spracherkennungsystems. Es liegt deshalb nahe, die Erkenntnisse, die auf dem Gebiet der Spracherkennung gewonnen wurden, mitzuverwenden. Aus diesem Grunde werden in dem folgenden Abschnitt Gemeinsamkeiten und Unterschiede von Musik und Sprache diskutiert.

5.1 Vergleichende Betrachtung von Musik und Sprache

In diesem Abschnitt soll versucht werden, Gemeinsamkeiten und Unterschiede von Musik und Sprache aufzuzeigen. Musik und Sprache lassen sich zunächst einmal als eine Abfolge akustischer Ereignisse ansehen. Als akustische Grundeinheit in der Musik gilt der Klang, charakterisiert durch Tonhöhe,

Lautstärke und Klangfarbe. Auf der Sprachebene wird mit dem Phon als Grundeinheit gearbeitet. Eine Charakterisierung von Musik und Sprache als Abfolge akustischer Ereignisse wird sicherlich nicht allen Aspekten gerecht. Die Erfassung der Bedeutung von Musik und Sprache sprengt aber den Rahmen dieser Arbeit. Die folgende Charakterisierung erfaßt die Erzeugung und Wahrnehmung.

Der Sprechvorgang besteht aus einer Anregung der Stimmbänder und einer Resonanzbildung im Vokaltrakt. Die Anregung der Stimmbänder kann auf zweierlei Arten erfolgen. Bei stimmhafter Betonung verursachen die enganliegenden Stimmbänder ein periodisches Signal. Stehen die Stimmbänder weit auseinander, so bewirken die Luftturbulenzen eine stimmlose Betonung. Die Resonanzbildung in dem Vokaltrakt beeinflußt die Intensität der einzelnen Partialtöne, die die Ausprägung der einzelnen Phone bewirkt. Das Frequenzspektrum der menschl. Stimme beginnt bei etwa 100Hz und endet bei 7500 Hz. Ein Laut dauert etwa 60 msec.

Die Erzeugung von Klängen geschieht in ähnlicher Weise wie bei dem Sprechvorgang. Zunächst wird mittels einer gespannten Saite, Membran o.ä. eine Schwingung hervorgerufen. Das Musikinstrument selbst bildet den Resonanzkörper. Charakteristisch für ein Musikinstrument ist also dessen Klangfarbe, d.h. das Verhältnis der Energien der Partialtöne. Der Frequenzumfang ist, bedingt durch die zahlreichen Instrumente, bei weitem umfangreicher als die der menschlichen Stimme. So liegt der niedrigste Grundton einer Orgel bei 16,35 Hz (Subkontra-c). Mit einer Flöte können Obertöne von 15000 Hz erzeugt werden [15],[21]. Die Dauer eines Tons variert und ist vom Metrum als auch von der Taktart abhängig. Tendenziell ist das Signal einer Musikquelle länger stationär als ein Sprachsignal.

Musik wird im allgemeinen mehrstimmig gespielt. In einem Orchesterwerk werden eine Vielzahl von Instrumenten eingesetzt. Bei Vokalmusik werden menschliche Stimmen (Sopran, Alt, Tenor, Baß) häufig von Instrumenten begleitet. Das Musiksignal setzt sich also aus mehreren korrelierten Teilsignalen zusammen. Die Situation ist beim Sprechen grundlegend anders. Es findet keine gewollte Überlagerung mehrerer Sprachsignale statt. Die Erkennungsleistung des Menschen sinkt rapide, wenn mehrere Personen gleichzeitig sprechen.

Die Wahrnehmung von Musik als auch von Sprache leistet das menschliche Gehör, bestehend aus Ohrmuschel, Gehörgang, Trommelfell und Gehörschnecke. Das Gehör ist in der Lage ein großes Spektrum an Schallinten-

sitäten zu erfassen. Die Hörschwelle liegt frequenzabhängig bei etwa 6dB. Die Lautheit der menschlichen Sprache liegt etwa in dem Bereich von 30dB bis zu 70dB. Bei Konzerten von Rockgruppen werden bis 120dB ereicht [20]. Eine ausführliche Abhandlung zu dem Bereich Wahrnehmung von Musik und Sprache ist auch in [9] zu finden.

5.2 Rhythmus

Charakteristisch für ein Musikstück ist dessen musikalisch-zeitliche Ordnung. Das Verhältnis der Tonlängen gehört ebenso dazu wie die Betonung. Die musikalischen Begriffe Metrum, Taktart und Rhythmus dienen zur Erfassung der musikalisch-zeitlichen Ordnung. Die Anzahl der Grundschläge pro Minute (beats per minute, bpm) unterscheidet sich grundlegend bei verschiedenen Musikstilen. So ist beispielsweise Techno durch harte, schnelle Grundschläge gekennzeichnet. Es werden bis zu 200 bpm erreicht.

Der Rhythmus charakterisiert das zeitliche Verhältnis akustischer Ereignisse, das heißt, das Verhältnis der Tonlängen. Der Takt gibt das Schema für die Betonung wieder. In der Literatur ist der Begriff Rhythmus allerdings nicht einheitlich definiert. In [22] wird unter Rhythmus auch Taktart und Betonung verstanden, während in [21] die Begriffe getrennt werden. Wichtiger erscheint es, den Begriff in physikalischer Art und Weise zu erfassen, insbesondere in Hinblick auf die Relevanz für die Musikstilerkennung.

Abschnitte unterschiedlicher Betonung unterscheiden sich in der Intensität des Signals. Zur Repräsentation eignet sich die Kurzzeitenergie. Das Berechnen der Kurzzeitenergie [20] gestaltet sich folgendermaßen:

$$p(t) = \sum_{\tau=-\triangle}^{\triangle} s(t+\tau)^2 \tag{5.1}$$

Die Energie verläuft proportional zu den Quadratsamplituden des Signals. Zur Berechnung der Kurzzeitenergie wird jeweils ein Fenster betrachtet und dann weiterverschoben (5.1). Die Verschiebung erfolgt mit Überlapp, um Kontext zu berücksichtigen. Die Breite des Fensters sollte entsprechend der Dauer akustischer Ereignisse gewählt werden. Anhaltspunkte dazu liefert das Signal im Zeitbereich. Die Wahl der Fensterbreite bleibt allerdings heuristisch. In Spracherkennungsystemen liegt die Fensterbreite zwischen 10

und 20 msec. Die folgenden Berechnungen basieren auf einer Fensterbreite von 50 msec und einer Verschiebung von 40 msec.

Nach Annahme enthält der Verlauf der Kurzzeitenergie Regelmäßigkeiten, die charakteristisch für den Rhythmus des Musikstückes sind. Sich wiederholende Sequenzen im Signal werden in der Autokorrelationsfunktion sichtbar.

$$a(t) = \sum_{\tau=\triangle_1}^{\triangle_2} p(\tau) * p(t+\tau) \tag{5.2}$$

Zur Berechnung der Kurzzeitautokorrelation (5.2) wird ein Fenster $\triangle_1$ bis $\triangle_2$ aus der Kurzeitenergie $p(t)$ ausgeschnitten. Die Autokorrelationsfunktion $a(t)$ erreicht ihr Maximum bei $t = 0$. Ist das Signal $p(\tau)$ innerhalb des Fensters $\triangle_1$ bis $\triangle_2$ periodisch mit Periode k, so ist auch $a(t)$ periodisch mit Periode k.

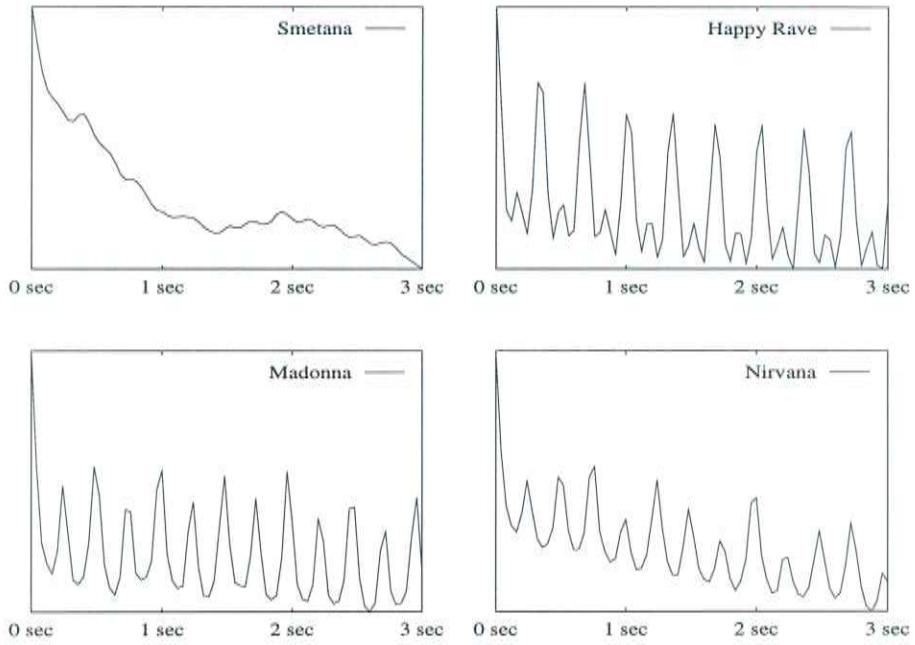

Abbildung 5.1: Autokorrelation der Kurzzeitenergie

Zu diskutieren ist die Wahl des Fensters. Das Fenster muß breit genug sein, um die Regelmäßigkeiten zu erfassen. Andererseits sind Rhythmusänderungen insbesondere bei klassischer Musik denkbar. Eine exakte Festlegung

der Fensterbreite kann an dieser Stelle daher nicht getroffen werden. Dazu ist ein Vergleich der Erkennungsleistungen des Klassifikators mit unterschiedlichen Fensterbreiten erforderlich. In Abbildung 5.1 sind Autokorrelationsfunktionen zu sehen, die auf 3 Sekunden breiten Fenster berechnet wurden. Die vier Ausschnitte entstammen Musikstücken unterschiedlicher Stilrichtungen. Die Regelmäßigkeiten der Kurzzeitenergie bei dem klassischen Stück (Smetana) sind weniger stark ausgeprägt, ganz im Gegensatz zu dem Vertreter des Technostils (Happy Rave). Klar erkennbar sind die harten Grundschläge. In dem 3 Sekunden langen Fenster lassen sich 9 Grundschläge zählen. Dies entsprechen 180 bpm (beats per minute). In der unteren Hälfte der Abbildung sind die Autokorrelationsfunktionen eines Popstückes (Madonna) und eines Rockstückes (Nirvana) dargestellt. Im Vergleich zu dem Techno-Rhythmus varieren die Rhythmen der beiden Rock- und Popstücke langsamer, sind aber denoch ebenso regelmäßig strukturiert.

Ein Klassifikator auf Rhythmus-Basis könnte das Potential haben, klassische Stücke von den übrigen Stilen zu trennen. Weniger aussichtsreich erscheint eine Trennung von Rock, Pop und Techno auf Basis von Rhythmus.

5.3 Klangfarbe

Zur Unterscheidung von Musikstilen tragen sicherlich auch die verwendeten Instrumente bei. So wird ein E-Gitarre wohl in einem klassischen Stück nicht zum Einsatz kommen, in der Rockmusik dagegegen schon. Ebenso gibt die Violine einen Hinweis auf den Musikstil. Meist wird aber nicht ein Instrument allein gespielt, sondern es werden mehrere Instrumente kombiniert. Es resultiert ein Klangbild durch die Überlagerung der Klangfarben der einzelnen Instrumente. Etwas ungenau wird für diese Kombination auch der Begriff Klangfarbe verwendet [22]. Neben der Tonhöhe und der Tonstärke ist also die Klangfarbe für einen Ton charakteristisch.

Die Fourier-Analyse zerlegt einen Ton in dessen Partialtöne. Der Begriff Klangfarbe kann so in physikalischer Art und Weise erfaßt werden: Die Klangfarbe gibt das Verhältnis der Energien der einzelnen Partialtöne an. Das Vorgehen zur Berechnung des Kurzzeitleistungsspektrum gliedert sich in drei Schritte. Dem Ausschneiden mit einer Fensterfunktion $w_t(\tau)$ (Hamming-Fenster) folgt das Berechnen der (komplexen) Fourier-Koeffizienten $S_t(w)$. Anschließend wird der reellwertige Betrag $P_t(w)$ gebildet.

$$\begin{aligned} S_t(w) &= \sum_{\tau=-\infty}^{\infty} w_t(\tau) * s(\tau) * e^{-i*w*\tau} \\ P_t(w) &= Re(S_t(w))^2 + Im(S_t(w))^2 \end{aligned} \tag{5.3}$$

Das Kurzzeitleistungsspektrum (5.3) enthält Tonhöhe und Klangfarbe gleichermaßen. Erwünscht ist aber eine Trennung beider Merkmale. Betrachtet man die Erzeugung von Klängen (vgl. Abschnitt *Vergleichende Betrachtung von Musik und Sprache*), so kann die Kombination aus Grundton und Resonanz durch eine Faltung im Zeitbereich modelliert werden. Ausgehend von diesem Modell kann man zur Extraktion der Klangfarbe einen Ansatz aus der homomorphen Sprachvorverarbeitung verwenden. Durch den Übergang in den Frequenzbereich und anschließender Logarithmierung des Betrages ist die Zusammenhang beider Teilsignale nun additiver Art. Mit der inversen Fouriertransformation wird das so zerlegte Signal wieder in den Zeitbereich gebracht. Das Resultat wird *Cepstrum* genannt.

$$\begin{aligned} s_t(\tau) &= g_t(\tau) * r_t(\tau) \\ S_t(w) &= G_t(w) * R_t(w) \\ log|S_t(w)| &= log|G_t(w)| + log|R_t(w)| \\ FT^{-1}\{log|S_t(w)|\} &= FT^{-1}\{log|G_t(w)|\} + FT^{-1}\{log|R_t(w)|\} \end{aligned} \tag{5.4}$$

Beide Teilsignale lassen sich nun durch einen linearen Filter trennen. Bei diesem auch *Lifterung* genannten Prozeß werden die Cepstrum-Koeffizienten ab Beginn der Grundtondauer weggeschnitten. Es verbleiben nur die Grundanteile des zweiten Teilsignals, welches das Resonanzverhalten des Musikinstrumentes charakterisiert. Die Fouriertransformation der verbleibenen Cepstrum-Koeffizienten liefert ein geglättetes Kurzzeitleistungspektrum, welches nicht mehr die Frequenzanteile des Grundtons enthält. Dieses Verfahren findet auch in der Sprachvorverarbeitung Anwendung [17].

Die Darstellungen in Abbildung 5.2 basieren auf einem 50ms Fenster eines Piano-Ausschnittes. Das logarithmische Leistungsspektrum ist in der linken Hälfte ersichtlich, die ersten 5ms des zugehörigen Cepstrums kann man der rechten Hälfte entnehmen. Nach Logarithmierung ist das Signal mittels inverser Fouriertransformation in den Zeitbereich transformiert worden. Bei der

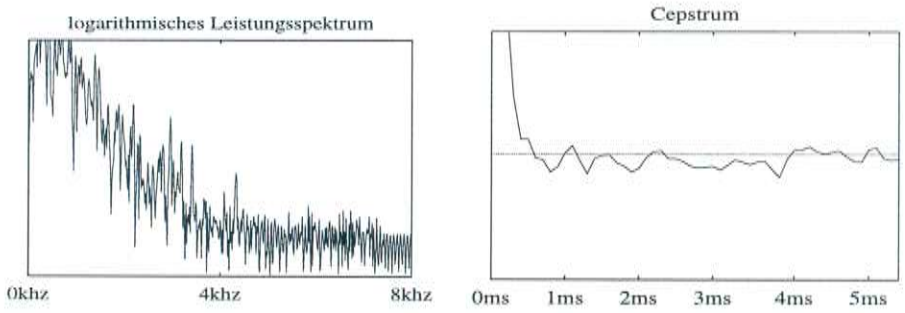

Abbildung 5.2: Spektrum und Cepstrum eines Piano-Ausschnittes

Grundtondauer sollte nach Annahme eine deutliche Signalspitze des Cepstrum vorhanden sein. Dies ist so in der Abbildung nicht erkennbar. Der Amplitudenanstieg des Signals bei 1ms kann der Beitrag des Anregungssignals sein. Dies würde einer Grundfrequenz von 1000 Hz entsprechen. Um die Grundtonanteile zu entfernen, muß man also vor dem Amplitudenanstieg bei 1ms das Cepstrum abschneiden. Eine Rücktransformation des so gefilterten Cepstrums in den Frequenzbereich sollte dann die Klangfarbe ergeben.

Einen Kompromiß muß man bei dem Filter im Cepstral-Bereich eingehen. Einerseits muß unterhalb der Grundtondauer weggeschnitten werden, andererseits werden hinreichend viele Koeffizienten zur Beschreibung der Klangfarbe benötigt. Nimmt man etwa einen Ton mit Grundfrequenz von 3500 Hz als Basis der Überlegung (man denke etwa an eine Harfe), dann verbleiben nur die ersten zwei Cepstrum-Koeffizienten (Durch Zweier-Potenzierung bei der FFT-Berechnung entsprechen 50ms bei 16khz Abtastrate insgesamt 512 Koeffizienten, also hat man bei einer Grundtondauer von 1/3.5 ms 2 Koeffizienten). Es ist zu überlegen, ob mit 2 Koeffizienten eine angemessene Beschreibung der Klangfarbe erreicht werden kann. Werden dagegen die ersten 10 Cepstrum-Koeffizienten für die weitere Verarbeitung in Betracht gezogen, so ist die Filterung für Grundtöne bis 1024 Hz korrekt. Klangfarben bei Grundtönen oberhalb dieser Grenzfrequenz würden nicht korrekt berechnet. Als Kompromiß werden daher die ersten 5 Cepstrum-Koeffizienten vorgeschlagen. Damit ist das Verfahren zumindest bei Tönen mit bis zu 2048 Hz korrekt. Antwort auf die Frage nach der Richtigkeit dieses Kompromisses kann nur die Erkennungsleistung des Klassifikators geben.

In Abbildung 5.3 sind zwei Kurzzeitspektren, die durch den oben beschrie-

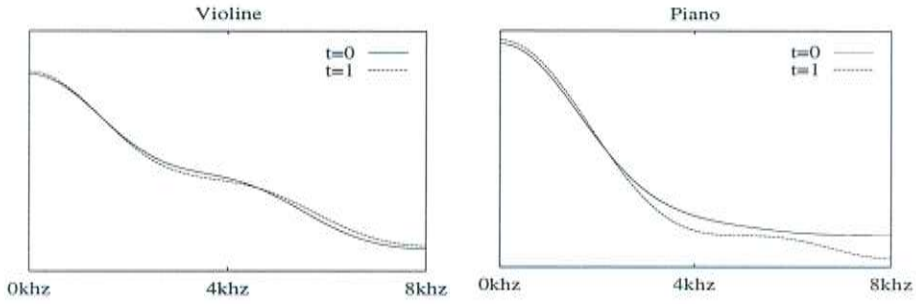

Abbildung 5.3: durch Lifterung geglättete Kurzzeitspektren

benen Prozeß geglättet wurden, ersichtlich. Betrachtet werden zwei aufeinanderfolgende 50ms-Ausschnitte (mit t=0 und t=1 markiert), in der eine Violine bzw. ein Piano zu hören sind. Die Ausschnitte sind außerdem bezüglich der Lautheit normiert, so daß eine Vergleichbarkeit der einzelnen Frequenzanteile gegeben ist. Bei dem Piano nehmen ab 3kHz die Frequenzanteile sehr rasch ab. Demgegenüber bestitzt das Klangfarbenspektrum der Violine auch Anteile höherer Frequenzen. Bei dem Piano überwiegen die Anteile geringerer Frequenzen. Von Interesse ist ebenfalls ein Vergleich des Leistungsspektrum des Piano-Ausschnittes (linke Hälfte der Abbildung 5.2) und dem durch Lifterung geglätteten Leistungspektrum in Abbildung 5.3. Der Grundverlauf des Leistungspektrums wird recht gut durch die Lifterung herausgearbeitet. An der Abbildung wird ebenfalls die Beschränkung auf eine Abtastrate von 16kHz deutlich. Somit sind nach dem Abtasttheorem nur Frequenzanteile bis 8kHz berechenbar.

Zu diskutieren sind die Annahmen, die für die Anwendbarkeit des Verfahrens erforderlich sind. Die Modellannahme von der Erzeugung des Signals aus *einer* Quelle als Konvolution zweier Teilsignale (Anregung und Resonanzverhalten) ist sicherlich berechtigt. Sie wird nicht zuletzt auch durch die Abbildungen gestützt. Welche Probleme entstehen aber bei mehrstimmiger Musik? Es ist zu überlegen, ob dann die Einhüllende des Spektrums ein Klangbild repräsentiert, welches sich durch die Klangfarben der beteiligten Instrumente zusammensetzt. Ein ebenfalls bisher nicht betrachteter Punkt ist die Behandlung polyphoner Musikstücke.

Ein Ansatz, bei dem obige Schwierigkeiten vermieden werden können, besteht darin, nicht von der Erzeugung des Signals auszugehen, sondern von

der Wahrnehmung. In der Psychoakustik haben Untersuchungen folgende Eigenschaften des menschlichen Hörens zutage gefördert [29]:

- Bei steigenden Tonhöhen sinkt die Fähigkeit nichtlinerar, benachbarte Töne unterscheiden zu können.
- Bei steigender Lautstärke nimmt die Genauigkeit der Lautstärkenwahrnehmung nichtlinear ab.
- Die Wahrnehmung von Tönen oberhalb einer Grenzfrequenz nimmt nichtlinear ab.

Diese Eigenschaften können durch eine Filterbankanalyse modelliert werden. Dies bewirkt eine Glättung des Leistungspektrums. Populär in der Spracherkennung ist die *Melscale*-Analyse. In Abbildung 5.4 sind zwei durch Mel-Skalierung geglättete Kurzzeitspektren ersichtlich.

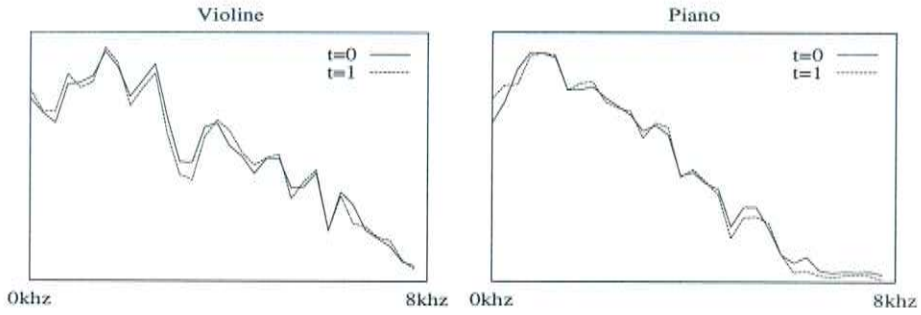

Abbildung 5.4: durch Mel-Skalierung geglättete Kurzzeitspektren

Bei einem Vergleich dieser Abbildung mit den vorherigen Abbildungen ist zu beachten, daß hier eine andere Skalierung der Frequenzachse vorliegt. Die Darstellungen in Abbildung 5.4 basieren auf der Berechnung von 32 Filterbankkoeffizienten.

Ob die Repräsentation der Klangfarbe durch Lifterung besser geeignet ist als die durch Filterbankanalyse kann an dieser Stelle nicht endgültig geklärt werden. Auskunft kann nur ein Vergleich der Erkennungsleistungen eines Klassifikators geben. Die Darstellung mittels Cepstren ist in jedem Fall äußerst kompakt. Die Klangfarbe eines 50ms langen Abschnittes nur durch 5 Koeffizienten zu beschreiben, birgt aber sicherlich auch gewisse Risken in

sich. Andererseits muß bei Verwendung der 32 Filterbankkoeffizienten eine erhebliche Anzahl an Modellparametern des Klassifikators geschätzt werden.

5.4 Melodie

Wie am Anfang des Kapitels erwähnt, ist das Ziel der Vorverarbeitung die Extraktion relevanter Merkmale. Dementsprechend sind in den letzten beiden Abschnitten Methoden zur Bestimmung von Rhythmus und Klangfarbe vorgestellt worden. Die Melodie ist nun aber kein relevantes Merkmal, um Musikstile unterscheiden zu können. So kann die Melodie eines Schlagers sehr wohl auch im Stil eines Rockstückes gespielt werden. Andererseits können Änderungen der Melodie Anfang und Ende akustischer Abschnitte kennzeichnen. Dem Klassifikator können diese Informationen von Nutzen sein, um die Merkmale Rhythmus und Klangfarbe richtig einzuschätzen.

Melodie ist als eine Folge von Tonhöhen charakterisierbar. Ein grobes Maß für die Frequenz ist die Anzahl von Nulldurchgängen des Signals im Zeitbereich. Durch eine Glättung der Nulldurchgangsrate werden die Spitzen entfernt.

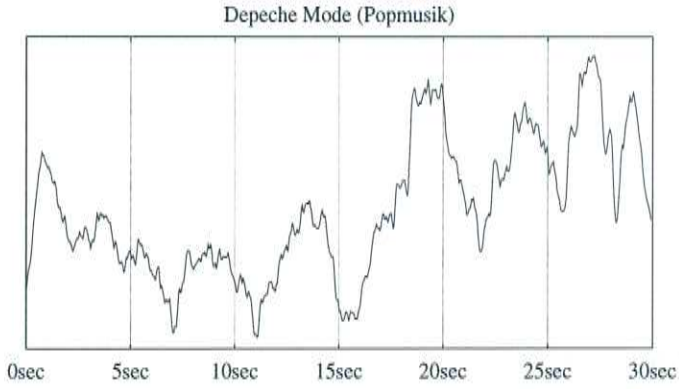

Abbildung 5.5: geglättete Nulldurchgangsrate

Deutlich sind Anfang und Ende akustischer Ereignisse in Abbildung 5.5 zu erkennen. Im Bereich von 11 bis etwa 15 Sekunden ist eine Wiederholung des Abschnittes von 7 bis 11 Sekunden sichtbar. In den Bereichen mit hohen Nulldurchgangsraten (am Anfang, im Bereich von 18-21 sec, von 23

bis 25 sec und im Bereich von 26 bis 27 sec) ist Gesang zu hören. Gleichwohl kann nicht davon gesprochen werden, mit Hilfe der Nulldurchgangsrate verschiedene akustische Ereignisse erkennen zu können. Es ist eher so, daß eine Segmentierung des Signals in Bereiche unterschiedlicher Art möglich sein könnte. Es sind weitere Überlegungen vonnöten, um die in der Nulldurchgangsrate enthaltenen Informationen in den Klassifikator mit einzubinden.

Kapitel 6

Klassifikation

Die Untersuchungen, deren Ergebnisse im vorhergehenden Kapitel präsentiert worden sind, befassen sich mit der Extraktion relevanter Merkmale. Gegenstand der in den folgenden Abschnitten beschriebenen Experimente ist der Bau von Klassifikatoren, die die Stile der Musikstücke erkennen. Bei dem Bau von Klassifikatoren ist im wesentlichen folgenden Gesichtspunkten Aufmerksamkeit zu schenken. Dies ist zum einem der Eingaberaum des Klassifikators. Wie sollen die Musikstücke repräsentiert werden? Aufgrund welcher Merkmale eines Musikstückes soll der Klassifikator seine Entscheidung zugunsten eines Musikstils treffen? Nützlich zum Auffinden geeigneter Eingaberäume sind dabei die Erkenntnisse der Experimente zur Extraktion relevanter Merkmale, Rhythmus und Klangfarbe.

Der andere Gesichtspunkt betrifft die Architektur des Klassifikators. Dies ist die Frage nach der Art der Informationsverarbeitung. Unser vorhandenes Problemwissen ist nicht durch einen Satz von Regeln gegeben, sondern durch eine Menge von Beispielen. Regelbasierte Ansätze aus der Ära der klassischen KI (Künstliche Intelligenz) werden dieser Art an Informationen nicht gerecht. Die Unsicherheit der Informationen kann auch nicht adäquat durch nichtmonotones Schließen beschrieben werden. Der Klassifikator sollte vielmehr aus den Beispielen das richtige Verhalten lernen. Die Informationsverarbeitung liegt dabei auf subsymblischer Ebene. Die Untersuchungen zu diesem Thema beschränken sich daher auf das Gebiet der konnektionistischen Klassifikatoren. Gegenstand des nun folgenden Abschnittes sind vorwärtsgerichtete Neuronale Netze.

6.1 Neuronale Netze

Künstliche Neuronale Netze wurden in vielen Forschungsarbeiten der letzten zwei Jahrzehnte ausführlich untersucht. So entstand eine Vielzahl von Neuronalen Netzen und es wurden auch etliche Bücher zu diesem Gegenstand publiziert. Eine gute Einführung bietet Hecht-Nielsen [7] oder auch Hertz, Krough und Palmer [8]. Minsky und Papert [13] zeigten die Restriktionen Neuronaler Netze auf. Nach Erscheinen des Buches *Perceptrons* war einer massiver Forschungsstop zu verzeichnen. Erst Rumelhart brachte mit der Entwicklung des Backpropagation-Algorithmus wieder Bewegung. Auf eine Einführung in das Gebiet der Neuronalen Netze wird an dieser Stelle verzichtet.

Netzstruktur Die in den folgenden Unterabschnitten präsentierten Ergebnisse basieren auf dreischichtigen vorwärtsgerichten Netzen. Die Neuronen der Eingabeschicht werden mit den Neuronen der verdeckten Schicht verbunden. Für jede zu trennende Klasse gibt es ein korrespondierendes Neuron in der Ausgabeschicht. Jedes Neuron der Ausgabeschicht ist mit jedem Neuron der verdeckten Schicht verbunden. Es gibt keine Direktverbindungen von der Eingabeschicht zu der Ausgabeschicht.

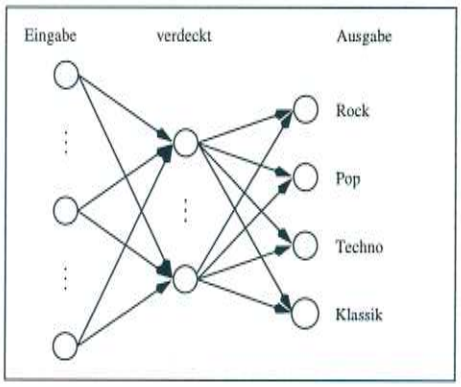

Abbildung 6.1: Netz-Struktur

Gelernt werden die Gewichte der Neuronenverbindungen mittels eines Gradientenabstiegsverfahrens. Unter der Voraussetzung einer geeigneten Fehlerfunktion (Summe der Fehlerquadrate) läßt sich zeigen, daß die Aktivierungen der Ausgabeneuronen die a-posteriori-Wahrscheinlichkeit approximieren.

Wenn also die Eingabeschicht des Neuronalen Netzes mit dem Modell x eines Musikstücks mit der Stilrichtung ω aktiviert wird, dann ist das zu dem Stil ω korrespondierende Ausgabeneuron mit dem Wert $P(\omega|x)$ aktiviert. Einsichtigerweise erfolgt die Entscheidung zugunsten einer Klasse nach der höchsten Aktivierung des korresondierenden Ausgabeneurons. In Abbildung 6.1 ist eine typische Netzstruktur schematisch wiedergegeben.

Zeitinvarianz Wie können nun Neuronale Netze zur Klassifikation von Musikstilen verwendet werden? Die Länge der Musikstücke variert, der Eingaberaum eines Neuronalen Netzes ist aber fix. Dieses Problem kann man lösen, indem man ein Musikstück in kleine Abschnitte fester Länge unterteilt, und diese Abschnitte dem Netz getrennt präsentiert. Die Bewertungen des Netzes für all diese Teilabschnitte des Musikstückes werden dann aufsummiert und führen so zu einer Gesamtentscheidung. Welche Voraussetzungen sind für dieses Vorgehen erforderlich? Es muß möglich sein, ein Mustikstück nur aufgrund eines solchen kurzen Abschnittes der korrekten Stilrichtung zuzuordnen. Das Problem darf nur innerhalb dieses Rahmens *zeitvariant* sein. Zu diskutieren ist die Dauer dieser Abschnitte. Erinneren wir uns dazu der Resultate der Experimente zur Wahrnehmung von Musik. Bei der Präsentation von 1sec langen Ausschnitten konnten 82,9% der Musikstücke richtig zugeordnet werden. Wurden von den gleichen Musikstücken 3sec lang dauernde Ausschnitte dageboten, so lag die Quote bei 84,9%. Andererseits lagen die Ergebnisse bei den abschnittsvertauschten Stücken (Vertauschung 50ms langer Abschnitte) bei 69,6%. Diese Ergebnisse liefern also schon erste Anhaltspunkte für eine geeignete Länge der Eingabe.

Nun läßt sich allerdings auch die Frage stellen, weshalb man denn nicht die Eingabelänge hinreichend groß wählt, und so das Problem der Zeitinvarianz umgeht. Ein zu großer Eingaberaum wirkt sich nachteilig auf das Lernverhalten Neuronaler Netze aus. Während des Lernvorganges spezialisieren sich die Neuronen der verdeckten Schicht auf relevante Gebiete des Eingaberaumes. Ein überdimensionierter Eingaberaum erschwert natürlich das Auffinden relevanter Teilgebiete. Die Dimension des Eingaberaumes hängt also von mehreren Punkten ab und eine geeignete Eingabedimension muß experimentell bestimmt werden.

Ablauf der Experimente Nach diesen Vorüberlegungen können wir nun zu den Experimenten kommen. Um aussagenkräftige Vergleiche zu ermöglichen, sind jeweils mehrere Netzsimulationen durchgeführt wurden. Variiert wurden neben den Lernparametern auch die Anzahl verdeckter Neuronen. Zum Trainieren der Netze wurde die Datenbasis, welche im zweiten Kapitel beschrieben ist, verwendet. Der Ablauf der Netzsimulationen sieht folgendermaßen aus: Trainiert werden die Netze mit den Trainingsdaten. Während des Trainings wird gleichzeitig auf einer disjunkten Kreuzvalidierungsmenge getestet. Ein Abbruch des Trainings erfolgt, wenn der Klassifikationsfehler auf der Kreuzvalidierungsmenge wieder zunimmt. Die Bewertung der Leistungsfähigkeit des so trainierten Netzes basiert auf eine dritte separate Testmenge. Sofern nichts Gegenteiliges angegeben ist, sind alle Erkennungsleistungen auf dieser Testmenge gemessen wurden.

6.1.1 Rhythmus

Wie in Kapitel 5 beschrieben, ist ein wesentliches Merkmal eines Musikstückes sicherlich der Rhythmus. Die Autokorrelationsfunktion der Kurzzeitenegie erfaßt die Regelmäßigkeiten der Betonung. In Abbildung 5.1 aus dem Kapitel 5 sind solche Autokorrelationen verschiedener Musikstücke dargestellt.

Die Neuronen der Eingabeschicht werden also mit den Koeffizienten der Autokorrelationsfunktion der Kurzzeitenergie aktiviert. Im ersten Experiment erhält das Netz Eingaben von jeweils 1 Sekunde Länge. Das Eingabefeld wird jeweils nur um $0,6$ Sekunden verschoben, so daß sich die Eingaben um $0,4$ Sekunden überlappen. Da die Kurzeitenergie auf einem Fenster von 50ms und einer Fensterverschiebung von 40ms berechnet wird, entsprechen einer Sekunde 25 Koeffizienten der Autokorrelationsfunktion.

Ergebnisse Die Gesamtentscheidung des Netzes entsteht durch Mittelung von 50 Einzelbewertungen. Die 50 Einzelbewertungen kommen durch Zerlegung der Musikstücke in 1 Sekunden lange Abschnitte zustande. Die Abschnitte werden jeweils um $0,4$ Sekunden verschoben. Ein 30 Sekunden langes Musikstück wird so in 50 Teile zerlegt. Hinweise auf das gelernte Verhalten des Netzes geben die Erkennungs- und Dominanztabellen in Abbildung 6.2. Die Erkennungsleistung beträgt $73,6\%$.

Probleme bereiten wie erwartet die Stile, die bezüglich des Rhythmus ähnlich sind - Rock, Pop und Techno. Die langsameren, balladenartigen Stücke

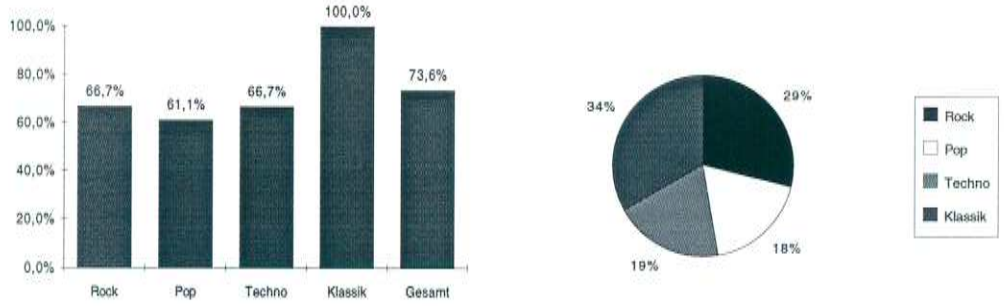

Abbildung 6.2: Erkennung und Dominanz des Rhythmus-Netzes

wie zum Beispiel etwa *Sade* werden der klassischen Stilrichtung zugeordnet. Ebenso verhält es sich bei den Verwechselungen von Techno mit Klassik. Dort werden die eher *trance*-artigen Stücke fehlklassifiziert. Daraus resultiert die Dominanz des klassischen Stils. Der Anteil der als Klassik klassifizierten Stücke ist 34%. Die klassischen Stücke werden wie erwartet richtig erkannt, auch die etwas moderneren wie die *Bilder einer Ausstellung* von Mussorgsky.

Beachtung sollte man auch der zeitlichen Entwicklung der Klassifikation schenken. Dies kann sichtbar gemacht werden, indem die Gesamtentscheidung nur durch Mittelung auf einen Teil der Einzelbewertungen berechnet wird. Den zeitlichen Verlauf der Erkennungsleistung kann man Abbildung 6.3 entnehmen.

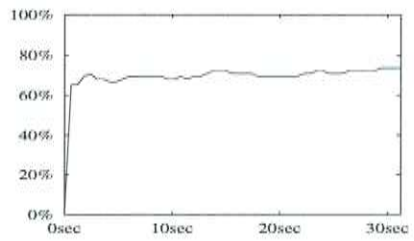

Abbildung 6.3: zeitlicher Verlauf der Erkennungsleistung des Rhythmus-Netzes

Analyse erlernten Verhaltens Erkennung und Dominanz einzelner Musikstile in der Tabelle 6.2 zeigen das erlernte Verhalten des Netzes. Wie kommt dieses Verhalten aber zustande? Um diese Frage zu beantworten,

müssen wir die Netzwerkverbindungen betrachten. Dabei gehen wir in zwei Schritten vor. In dem ersten Schritt wird untersucht, welche verdeckte Neuronen sich auf welche Eingaberegionen spezialisiert haben. Haben wir dies getan, können wir den Zusammenhang zwischen den verdeckten Neuronen und den Musikstilen untersuchen.

- Spezialisierung auf Eingaberegionen
- Wirkung der Spezialisten auf die Musikstile

Resultat dieser Analyse sind Regeln, die den Zusammenhang zwischen Rhythmus und Musikstil wiedergeben. Die Verbindungen der Eingabeschicht werden dabei als die Prämissen der Regeln aufgefaßt. Die Konklusionen ergeben sich aus den Verbindungen der verdeckten Schicht zu der Ausgabeschicht. Diese Interpretation soll keineswegs eine Nähe zu klassischen regelbasierten Ansätzen aufzeigen. Nein, die Analyse dient vielmehr zur Erklärung des erlernten Verhaltens. Dadurch werden Erkennungsleistung, Dominanz einzelner Stile und Konfusionen plausibel.

Spezialisierung auf Eingaberegionen Klären wir nun, wie sich die Neuronen der verdeckten Schicht auf den Rhythmus-Eingaberaum spezialisiert haben. Aufschluß darüber geben die Gewichtsvektoren $\vec{w}$ der Eingabeneuronen zu den verdeckten Neuronen. Ist der Eingabevektor $\vec{x}$ ähnlich zu dem Gewichtsvektor $\vec{w_i}$ (bezüglich des Skalarprodukts), dann hat das zu dem Gewichtsvektor korrespondierende Neuron i eine hohe Aktivierung $s_i = 1/(1 + e^{-(\vec{w_i} * \vec{x} - \Theta_i)})$, wobei Θ_i die Schwelle des Neurons i definiert.

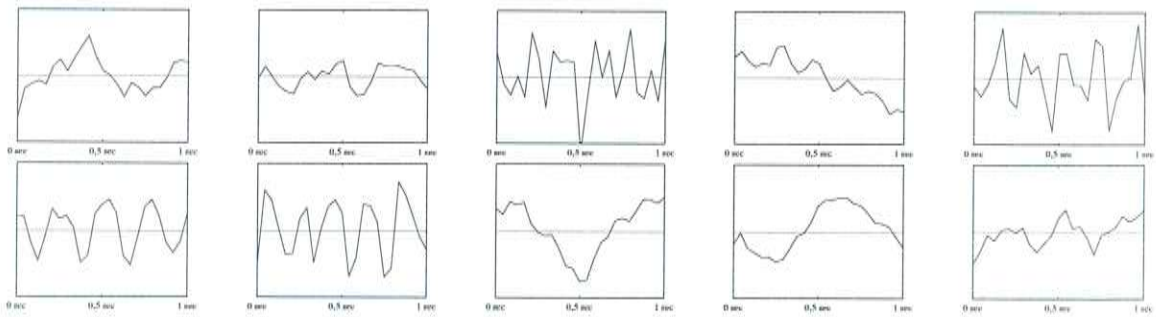

Abbildung 6.4: verdeckte Neuronen des Rhythmus-Netzes

Die Numerierung der Gewichtsvektoren der verdeckten Neuronen in Abbildung 6.4 ist zeilenweise von 1 bis 10. Die zehn verdeckten Neuronen des Rhythmus-Netzes haben sich auf unterschiedliche Teile des Eingaberaumes spezialisiert (zur Erinnerung: Der Eingaberaum wird durch die 25 Autokorrelationskoeffizenten der 1sec langen Musikabschnitte gebildet). Das dritte und fünfte Neuron sowie das siebente Neuron in Abbildung 6.4 haben sich auf die schnell varierende Rhythmusverläufe spezialisiert, die eher für Techno typisch sind. Die langsam varierende Rhythmusverläufe werden dagegen durch die Neuronen 1,2,4,8,9 und 10 repräsentiert. Bei Pop-Rhythmen wird das sechste Neuron aktiviert.

Wirkung der Spezialisten auf die Musikstile Die soeben analysierten Spezialisierungen der verdeckten Neuronen auf unterschiedliche Eingaberegionen machen sich auch in den Verbindungen der verdeckten Neuronen zu den Ausgabeneuronen bemerkbar. Diese Verbindungen sind in Tabelle 6.1 ersichtlich. Die Schwellwerte der Ausgabeneuronen sind in der ersten Zeile separat angegeben. Darauf folgen die Verbindungsgewichte der verdeckten Neuronen zu den Ausgabeneuronen Rock, Pop, Techno und Klassik. Die Schwellwerte sind so angegeben, daß hohe Werte das Neuron eher hemmen und für eine geringere Aktivierung sorgen. Klar erkennbar ist, daß die Neuronen, die sich auf die langsam varierenden Rhythmusverläufe spezialisiert haben, starke Verbindungsgewichte zu dem Klassik-Ausgabeneuron aufweisen. Stark negativ gewichtet sind die Verbindungen zu den verdeckten Neuronen, die sich auf die schnell varierende Verläufe spezialisiert haben. Analog verhält es sich mit den Verbindungen zu dem Techno-Ausgabeneuron. Auffallend ist, daß sich kein verdecktes Neuron für Rock-typische Rhythmusverläufe interessiert. In Verbindung mit dem negativen Schwellwert des Rock-Ausgabeneurons kann das Verhalten des Netzes interpretiert werden, daß wenn der eingegebene Rhythmus-Verlauf unähnlich zu den gelernten Verläufen ist, daß sich für Rock entschieden wird. Daraus ergeben sich exemplarisch folgende Regeln.

- *Ist der Rhythmus ähnlich langsam wie die der Muster 1,2,4,8,9 oder 10, dann ist der Musikstil eher Klassik*
- *Ist der Rhythmus ähnlich schnell wie die der Muster 3,5,6 oder 7, dann ist der Musikstil eher kein Klassik*

- *Ist der Rhythmus ähnlich zu dem Muster 6, dann ist der Musikstil eher Pop*
- *Paßt der Rhythmus zu keinem Muster, dann ist der Musikstil eher Rock.*

Verbindung	Rock	Pop	Techno	Klassik
Schwellw.	-1.30	0.27	1.62	0.26
1	-1.53	1.16	-12.32	**3.87**
2	-3.08	-5.27	**13.39**	**1.81**
3	-2.21	-1.74	**8.86**	-11.77
4	-1.61	-1.52	-1.52	**2.15**
5	-0.93	-0.24	**6.34**	-9.78
6	0.46	**3.56**	-5.30	-10.01
7	-0.32	0.82	**4.36**	-10.73
8	-1.57	-0.15	-14.86	**4.47**
9	-2.93	-0.03	-13.99	**3.80**
10	-4.98	-4.83	**9.77**	**4.16**

Tabelle 6.1: Ausgabeneuronen des Rhythmus-Netzes

Bemerkenswert ist, daß die Verbindungen des zehnten verdeckten Neurons sowohl zu dem Klassik-Ausgabeneuron als auch zu dem Techno-Ausgabeneuron stark positiv korreliert sind. Auch der langsame Rhythmusverlauf des zweiten Neuron sollte nicht erwarten lassen, daß ein stark positive Verbindung dieses Neurons zu dem Techno-Ausgabeneuron existiert. Verantwortlich für diese gelernten Verbindungen können Beispiele des Technostils sein, die eher *trance*-artig sind. Dies kann zum Teil die Verwechselungen von Techno mit Klassik erklären. Auch die Dominanz der klassischen Musik kann erklärt werden. Der Anteil der Stücke, die dem klassischen Musikstil zugeordnet werden, ist überproportional mit 34%. Von den 10 verdeckten Neuronen haben sich 5 Neuronen auf klassische Rhythmusverläufe bei Betrachtung der Verbindungen zu dem Klassik-Ausgabeneuron spezialisiert. In der verdeckten Schicht ist demnach Klassik überrepräsentiert, genauer formuliert sind es die langsamen Rhythmusverkäufe, die überrepräsentiert sind.

Anpassung des Eingaberaums Die bisher referierten Ergebnisse fokussieren auf das gelernte Verhalten des Neuronalen Netzes. Aber es gibt sicherlich noch mehr Gesichtspunkte, die es verdienen, beachtet zu werden.

Bei der Bewertung des gelernten Verhaltens eines Neuronalen Netzes muß auch der Eingaberaum berücksichtigt werden. Eine Möglichkeit, den Eingabegraum zu vergrößern, besteht darin, das Fenster, in der die Autokorrelationskoeffizienten berechnet werden, zu erweitern. In Kapitel 5 ist die Frage eines geeigneten Fensters bereits angesprochen worden - mit dem Hinweis, daß dazu ein Vergleich der Erkennungsleistungen vonnöten ist. Das bisher verwendete Fenster umfaßt die Zeit von einer Sekunde. Analog zu den Wahrnehmungsexperimenten kann dem Netz auch ein 3 Sekunden großes Fenster als Eingabe dargeboten werden. Die Klassifikation eines 30 Sekunden langes Musikstückes wird dann aufgrund der Einzelbewertungen der 10 Rhythmus-Eingabevektoren vorgenommen.

Fenstergröße	Überlappung	Erkennungsleistung
1 sec	0 sec	72,2%
1 sec	0,4 sec	**73,6%**
3 sec	0 sec	70,8%
3 sec	1 sec	65,3%

Tabelle 6.2: Vergleich der Erkennungsleistungen des Rhythmus-Netzes bei unterschiedlichen Eingaberäumen

Zudem kann auch erwogen werden, die Fenster nicht überlappend anzuordnen. Die Auswirkungen dieser Eingabeveränderungen sind in Tabelle 6.2 festgehalten. Anhand dieser Ergebnisse können wir den Einfluß der Fenstergröße und der Fensterverschiebung analysieren.

1. Größe des Rhythmusfensters
   Die etwa 30 Sekunden langen Musikstücke werden bei der 3sec-Variante in 10 Rhythmusfenster a 3 Sekunden zerlegt. Die Gesamtentscheidung ensteht dann durch Mittelung der 10 Einzelbewertungen. Eine feinere Zerlegung der Musikstücke erreicht man bei der 1sec-Variante. Es entstehen dann 30 Einzelbewertungen. Werden die Fenster zu klein gewählt, dann entstehen Fehler durch ungenaue Berechnungen des Rhythmus. Bei zu großen Fenstern können Fehler entstehen, wenn sich der Rhythmus innerhalb des Fensters geändert hat. Die Berechnung des Rhythmus basiert ja auf der Annahme, daß innerhalb des Fensters die Regelmäßigkeiten der Betonung gleich sind.

Mit einem Netz auf Basis nichtüberlagernder 3sec-Fenster werden 70, 8% aller Musikstücke richtig erkannt. Die Vergrößerung des Eingaberaumes bewirkt also eine Verschlechterung im Vergleich zu den 72, 2% bei der 1sec-Variante. Weshalb ist dies nun so? Zunächst ist zu prüfen, ob diese Unterschiede überhaupt von Bedeutung sind. In dem Test ist jeder Musikstil durch 18 Stücke repräsentiert, die gesamte Testmenge umfaßt also 72 Musikstücke. Die prozentuale Differenz von $72,2\% - 70,7\%$ entspricht einer zusätzlichen Verwechselung. Gleichwohl sind die Zuordnungen beider Klassifikatoren unterschiedlich. Vergleicht man Klassifikatoren auf Rhythmus-Basis so muß man beachten, daß die Zuordnungen der Rockstücke eher aufgrund von Regeln der Art 'Im Zweifelsfall gilt Rock' zustande kommen. Auf Rhythmen der Popstücke hat sich nur ein verdecktes Neuron spezialisiert. Konsequenterweise sollten also im Blickpunkt des Vergleichs die Klassik- und Technostücke sein. Und da sehen die Zuordnungen der 1sec-Variante weitaus besser aus. Bezogen auf die Klassik- und Technostücke erkennt die 1sec-Variante 3 Musikstücke mehr. Die 1sec-Variante hat also im Vergleich zu der 3sec-Variante zwei Rock/Pop Stücke mehr falsch klassifiziert. Nachdem dieser Punkt geklärt ist, können wir uns nun der Frage nach dem Warum zuwenden. Weshalb funktioniert die 1sec-Variante besser? Betrachten wir eine Folge von Abschnitten eines Musikstückes. Jeder Abschnitt ist nun mehr oder weniger für den Musikstil charakteristisch. Werden nun durch ein zu großes Fenster charakteristischen Abschnitte durch nicht-charakteristische Abschnitte überlagert, so werden diese charakteristische Abschnitte bei der Auswertung nur unzureichend berücksichtigt. Damit können die Unterschiede der Erkennungsleistungen der 1sec- und 3sec-Variante erklärt werden.

2. Fensterüberlappung
   Kommen wir nun zun den anderen Ergebnissen. Werden die 1sec-Fenster überlappend angeordnet, so ergibt sich einer Verbesserung von 72, 2% auf 73, 6%. Bei Betrachtung der 3sec-Fenster sieht das Bild umgekehrt aus. Diese Resultate scheinen widersprüchlich zu sein. Bevor wir versuchen, eine Lösung für dieses Problem zu finden, sollten wir prüfen, ob die Unterschiede vom Bedeutung sind. Der interessierte Leser wird wissen wollen, bei welchen Musikstücken die Zuordnungen unterschiedlich ausfallen. Und in der Tat, die Unterschiede sind in dem Rock/Pop-

Bereich zu finden. Unter der Annahme, daß der Rhythmus nicht das entscheidene Merkmal für die Unterscheidung von Rock und Pop ist, ist eine Beurteilung der Fensterüberlappung aufgrund dieser Ergebnisse nicht zulässig.

Zusammenfassend kann also festgestellt werden, daß Neuronale Netze auf Rhythmus-Basis in der Lage sind, Klassik- und Technostücke zu erkennen. Zur Berechnung des Rhythmus genügen 1sec-Fenster. Die Trennung von Rock und Pop ist eher unzureichend. Der Rhythmus ist allerdings auch kein signifikantes Merkmal von Rock- und Popmusik.

6.1.2 Klangfarbe

Im vergangenen Abschnitt hat sich also gezeigt, daß der Rhythmus einen Beitrag zur Unterscheidung der Musikstile leisten kann. Dies gilt insbesondere für den Anteil der klassischen Musik. Nicht zu übersehen waren allerdings auch die Schwierigkeiten bei der Unterscheidung von Rock- und Popmusik. Welche Eigenschaften eines Stückes bestimmen nun, ob ein Lied als Rockmusik wahrgenommen wird? Hinweise geben die gespielten Instrumente. Eine E-Gitarre läßt eher auf Rockmusik schließen, ein Synthesizer könnte ein Merkmal für Popmusik oder auch für Techno sein. Mit der Klangfarbe werden diese Merkmale erfaßt. Im Kapitel *Extraktion relevanter Merkmale* wurden bereits zwei Methoden zur Extraktion der Klangfarbe diskutiert. An dieser Stelle können nun die dort offen gelassenen Fragen beantwortet werden, nämlich ob Cepstren oder eher Melscales besser geeignet sind und an welcher Stelle die Lifterung einsetzen sollte. Ebenso kann geklärt werden, wieviele Melscale-Koeffizienten zur Beschreibung der Klangfarbe berechnet werden sollen. Abhängig von der Anzahl der Koeffizienten muß die Größe des Fensters bestimmt werden. Um die Güte des Merkmals Klangfarbe zu evaluieren, müssen also folgende drei Punkte berücksichtigt werden, als da sind: Art (Melscales,Cepstren), Anzahl der Koeffizienten und Fenstergröße.

Ergebnisse Mit 81,9% richtig erkannten Musikstücken (Abbildung 6.5) ist es evident, daß die Klangfarbe ein signifikantes Merkmal für die Unterscheidung von Musikstilen ist. Die Auswertung der referierten Ergebnisse gliedert sich in zwei Teile. Im ersten Teil beschäftigen wir uns mit dem gelernten Verhalten des besten Netzes. Diskutiert werden Konfusionen und der zeitliche

Verlauf der Erkennung. Eine genauere Analyse der unterschiedlichen Erkennungsleistungen der Netze auf Cepstren- und Melscale-Basis ist Bestandteil des zweiten Teils der Auswertung. Richten wir nun also unser Blick auf das Netz mit der 81,9% Erkennungsleistung.

Der Eingaberaum dieses Netzes wird durch Cepstral-Koeffizienten gebildet. Wie im Kapitel *Extraktion relevanter Merkmale* berichtet, werden die Cepstren auf einem 50ms-Fenster berechnet, welches jeweils um 40ms verschoben wird. Von den so erhaltenen 513 Koeffizienten werden jeweils nur die ersten 5 Koeffizienten genommen. Dies geschieht aufgrund der Annahme, daß die Grundtondauer größer 0,48 msec (bzw. die Tonhöhe kleiner 2052 Hz) ist. Die Eingabe des Netzes besteht nun aus Cepstren aus 10 aufeinanderfolgenden Fenstern. Damit stehen dem Netz zeitvariante Informationen in Umfang von $10 * 40ms = 0,4$ Sekunden zur Verfügung. Mit dem so gebildeten Eingaberaum kann das Neuronale Netz 81,9% aller Musikstücke richtig erkennen.

Der Abbildung 6.5 können die Erkennungsleistungen bei den einzelnen Musikstilen entnommen werden. Klassik wird mit Abstand am besten erkannt, es werden 94,4% aller klassischen Stücke richtig erkannt. Dies muß auch in Zusammenhang mit der Dominanz gesehen werden. Der Anteil der Musikstücke, die für Klassik gehalten wurden, ist 1/4, dies entspricht gerade dem Anteil der vorhandenen klassischen Stücke (Gleichverteilung). Die auf den ersten Blick vermeintlich gute Erkennung der Popstücke (88,9%) muß dagegen relativiert werden. So entscheidet sich das Netz in 36,1% aller Fälle für Popmusik. Von den Technostücken werden 83,3% richtig erkannt.

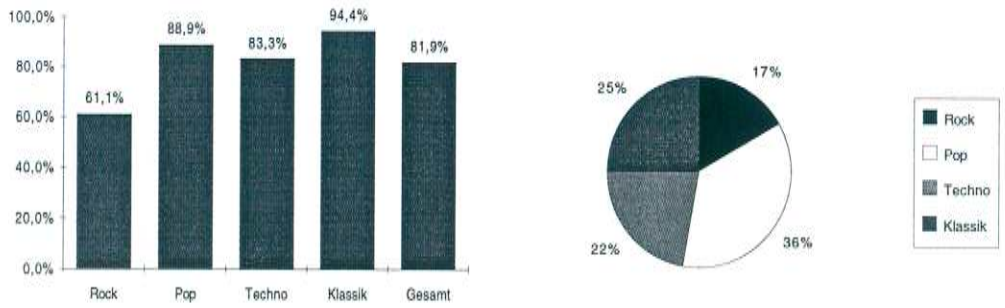

Abbildung 6.5: Erkennung und Dominanz des Klang-Netzes

Wie auch bei den klassischen Stücken ist der Anteil der für Technomusik

gehaltener Stücke ausgeglichen. Daß der Technostil recht gut erkannt wird, ist ein Indiz, daß sich die harten schnellen Grundschläge auch in der Klangfarbe widerspiegeln. Schauen wir uns dazu ein Ausschnitt aus dem Spektrum eines Technostückes an.

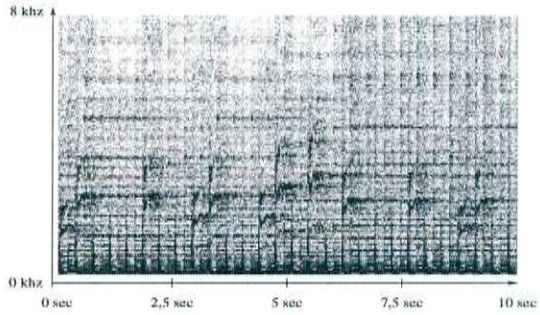

Abbildung 6.6: Spektrogramm eines Technostückes

Der abgebildete Ausschnitt des Spektrums ist von etwa 10 Sekunden Dauer. Deutlich fallen die Spitzen in den unteren Frequenzanteilen auf. Diese Spitzen charakterisieren die harten Grundschläge. Zählt man die Anzahl der Spitzen, so kann man auf die Anzahl der Grundschläge (beats per minute) schließen. Dies wären hier zirka 170 bpm. Mit diesen Informationen kann das Netz die technotypischen Merkmale lernen. Dies erklärt die gute Erkennung des Technostils. Zusammenfassend liegen die Schwierigkeiten also eher bei Rock und Pop, während Klassik und Techno von Netz recht gut erkannt werden.

Ein weiterer wichtiger Aspekt ist der zeitliche Verlauf der Erkennung. Um es deutlicher zu betonen: Gemeint ist hier nicht der zeitliche Rahmen an Informationen, die das Netz gleichzeitig erhält, sondern die serielle Ordnung. Der zeitliche Rahmen bei diesem Netz ist $0,4$ Sekunden. Wieviel solcher Zeitfenster eines Musikstückes benötigt also das Netz, um den Musikstil herauszufinden. In der Abbildung 6.7 ist die Entwicklung der Erkennungsleistung im Verlauf von 30 Sekunden erkennbar. Es zeigt sich ein deutlicher Aufwärtstrend von etwa 60% nach den ersten $0,4$ Sekunden ausgehend. Bemerkenswert sind die kurzen Unterbrechungen des Anstiegs. Diese Unterbrechungen signalisieren nicht-repräsentative Abschnitte der Musikstücke.

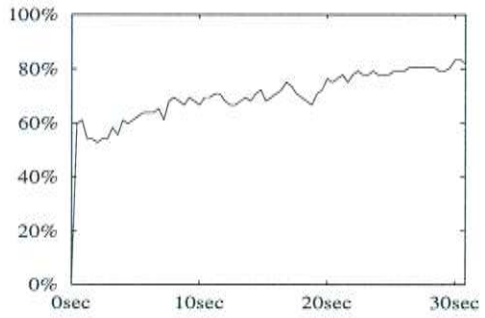

Abbildung 6.7: zeitlicher Verlauf der Erkennungsleistung des Klang-Netzes

Anpassung des Eingaberaums Nachdem wir nun anhand einiger Grafiken und Diagramme das gelernte Verhalten des Klangfarbennetzes analysiert haben, kommen wir nun zum zweiten Teil der Auswertung. Untersuchen wir also den Einfluß des Eingaberaumes auf die Erkennungsleistung. Bei der folgenden Analyse können wir uns auf die Tabellen 6.3 und 6.4 stützen.

Cepstren	krit. Grundtonhöhe	Fenstergröße	Erkennungsleistung
3	3420 Hz	0,6 sec	62,5%
5	2052 Hz	0,04 sec	68,1%
5	2052 Hz	0,4 sec	**81,9%**
5	2052 Hz	1,0 sec	77,8%
10	1026 Hz	0,2 sec	77,8%

Tabelle 6.3: Vergleich der Erkennungsleistung des Klang-Netzes bei erzeugungsbasierter Modellierung der Klangfarbe

1. Grundtondauer der Cepstralkoeffizienten
   Den Cepstralkoeffizienten liegt die Idee zugrunde, Anregung und Resonanz zu trennen. Von Interesse ist nicht der Grundton als solcher, sondern das Verhältnis der Energien der Partialtöne. Eine Trennung des Signals erfordert die Kenntnis der Grundtondauer, zumindest eine untere Grenze. Wird als Grenze für die Tonhöhe etwa 3400 Hz angenommen, dann verbleiben nur die ersten 3 Koeffizienten. Ganz offensichtlich ist der so erzeugte Eingabraum ungünstig. Es werden lediglich 62,5% der Musikstücke richtig erkannt. Dies hier sichtbar gewordene

Problem klang schon in dem Kapitel *Extraktion relevanter Merkmale* an. Mit drei Koeffizienten kann die Klangfarbe nicht hinreichend detailliert beschrieben werden. Also müssen mehr Koeffizienten berücksichtigt werden. Werden nun aber mehr Koeffizenten berücksichtigt, so nimmt auch die Anzahl zu schätzender Modellparameter zu. Zum Ausgleich sollte also die Fenstergröße verringert werden. Einen etwa gleich großen Eingaberaum erhält man, wenn das Fenster von 0,6 Sekunden auf 0,4 Sekunden verkleinert werden. Mit diesem Eingaberaum bekommt man dann die 81,9% Erkennungsleistung, also deutlich besser. Nun kann dieser Weg weiter verfolgt werden und die Anzahl der Cepstralkoeffizienten auf 10 erhöht werden. Bei einer weiteren Verringerung des Eingabefensters auf 0,2 Sekunden, werden 77,8% der Musikstücke richtig erkannt. Diese Verschlechterung der Erkennungsleistung kann nun mit einer unsauberen Trennung von Anregung und Resonanz erklärt werden. Bei den verwendeten 10 Koeffizienten erfolgt die Trennung nur noch dann korrekt, wenn die Grundtonhöhe kleiner als 1026 Hz ist. Der Anteil an unsauberen Trennungen nimmt also zu. Insgesamt gesehen, scheint die Berücksichtigung von 5 Koeffizienten am günstigsten. In dem Fall liegt die kritische Tonhöhe bei 2052 Hz.

2. Fenstergröße
   Bei der Untersuchung des benötigten Kontextes beschränken wir uns auf Eingaberäume, die durch 5 Cepstralkoeffizienten gebildet wurden, da es sich ja bereits gezeigt hat, daß diese Eingaberäume günstig sind. Wird das Fenster sehr klein gewählt (0,04 Sekunden), dann werden nur 68,1% der Musikstücke richtig erkannt. Die Erkennung von Musikstilen ist also offenbar kein zeitinvariantes Problem, sondern erfordert zeitlichen Kontext. Wird das Fenster auf eine Breite von einer Sekunde vergrößert, so hat das Netz weitaus mehr Kontextinformationen zur Verfügung. Im Vergleich zu dem 0,4 Sekunden Fenster ist aber auch hier die Erkennung schlechter, die Erkennungsleistung beträgt 77,8%. Dies kann aufgrund der durch die Eingaberaumvergrößerung entstehenden Probleme bei der Schätzung der Modellparameter verursacht werden. Insgesamt gesehen, ist die Wahl der Fenstergröße von notwendigen Kompromissen geprägt.

3. Melscale-Koeffizienten
   Die zweite Möglichkeit, die Klangfarbe zu repräsentieren, sind Melscale-Koeffizienten. Während bei Cepstren von der Erzeugung des Signals ausgegangen wird, wird bei den Melscales von der Wahrnehmung ausgegangen. Das Spektrum wird ausgehend von psychoakustischen Erkenntnissen geglättet. Die Melscale-Koeffizienten entstehen durch die Zusammenfassung von Fourier-Koeffizienten eines Frequenzbandes. Die Größe des Eingaberaumes bestimmt sich durch die Anzahl der verwendeten Filterbänke. Je mehr Filterbänke verwendet werden, desto genauer ist die Frequenzauflösung. Zum einen möchte man eine zu genaue Frequenzauflösung vermeiden, um die Grundfrequenzanteile zu verringern, andererseits muß die Frequenzauflösung hoch genug sein, um das Verhältnis der Energien der Partialtöne zu repräsentieren. Zudem muß wie auch bei den Cepstralkoeffizienten die Anzahl der Koeffizienten in Relation zu der Fenstergröße gesehen werden. Wird der Eingaberaum durch 16 Filterbankkoeffizienten und einem Fenster von 0, 2 Sekunden gebildet, dann werden 72, 2% Erkennungsleistung erzielt. Verringert man die Anzahl der Filterbänke auf 8 bei gleichzeitiger Vergrößerung des Fensters auf 0, 4 Sekunden, dann erhält man 73, 6% Erkennungsleistung. Diese Ergebnisse zeigen ganz klar, daß den Cepstralkoeffizienten den Vorzug zu geben ist. Die Modellannahmen von der Signalerzeugung bei den Cepstralkoeffizienten erweisen sich als unkritisch im Vergleich zu dem Wahrnehmungsmodell bei den Melscale-Koeffizienten.

Melscales	Fenstergröße	Erkennungsleistung
16	0,2 sec	72, 2%
8	0,4 sec	73, 6%

Tabelle 6.4: Vergleich der Erkennungsleistung des Klang-Netzes bei wahrnehmungsbasierter Modellierung der Klangfarbe

Zusammenfassend kann festgestellt werden, daß die Klangfarbe relevante Informationen zur Erkennung von Musikstilen enthält. Wichtige, in dem Kapitel *Extraktion relevanter Merkmale* offen gelassene, Fragen konnten geklärt werden. So eignen sich Cepstren besser als Melscales zur Klangfarbenrepräsentation. Bei der Trennung von Anregung und Resonanz kann von einer Grundtonhöhe von etwa 2000 Hz ausgegangen werden.

6.1.3 Kombination

Nachdem wir nun die Relevanz einzelner Merkmale für die Musikstilklassifikation herausgearbeitet haben, ist die Zeit gekommen, die Merkmale zu integrieren. Untersuchen wir also die Frage, ob ein Klassifikator, der über Informationen von Rhythmus und Klangfarbe gleichzeitig verfügt, besser Musikstile erkennen kann, als wenn die Merkmale getrennt verarbeitet werden. Zur Beantwortung dieser Frage erzeugen wir einen neuen Eingaberaum, der Informationen zu Rhythmus und Klangfarbe enthält.

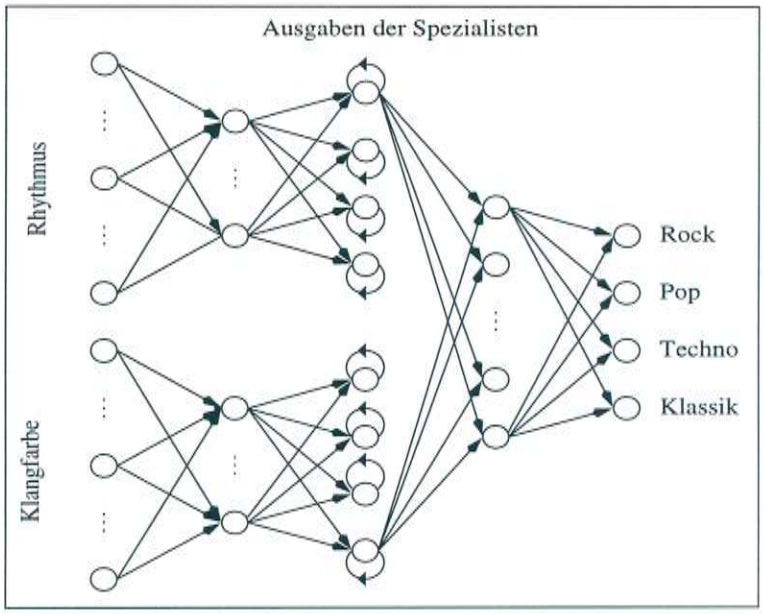

Abbildung 6.8: modulare Netzstruktur

modulare Netzstruktur Naheliegend ist es, die Ausgaben des Rhythmus- und des Klangfarbennetzes als Basis für die Gesamtentscheidung zu verwenden. Im Idealfall sollten die Aktivierungen der Ausgabeneuronen die a-posteriori-Wahrscheinlichkeiten approximieren. So wird der Eingaberaum des Kombinationsnetzes durch die Ausgaberäume der Spezialnetze gebildet. Vorteil einer modularen Bauweise ist die Möglichkeit, die Spezialnetze getrennt voneinander zu trainieren. Abbildung 6.8 zeigt die Netzstruktur des Gesamtsystems. Ausgehend von den angelegten Rhythmus- und Klangfarbenkoef-

fizienten berechnen die Spezialnetze ihre Ausgaben. Das Kombinationsnetz hat die Aufgabe, die Spezialisten hinsichtlich ihrer Bedeutung für die Problemlösung zu gewichten und eine Gesamtentscheidung zu formulieren.

Kombination der Spezialnetze Die Eingabeneuronen des Kombinationsnetzes werden mit den zeitlich gemittelten Aktivierungen der Ausgabeneuronen des Spezialnetze aktiviert. Durch die Zerlegung des Musikstückes in kleine zeitlich konstante Teile setzen sich die Ausgaben der Spezialnetze aus einer Reihe von Einzelbewertungen zusammen.

Seien die Ausgabeaktivierungen der Spezialnetze mit $s_i(t)$ bezeichnet. Der Index i läuft dabei von 1 bis 8, da das Rhythmusnetz und das Klangnetz jeweils 4 Ausgabeneuronen haben. Die Eingabeschicht des Kombinationsnetzes wird dann mit $k_i(t+1) = k_i(t) + s_i(t+1)$ aktiviert. Zum Zeitpunkt t wird also das Kombinationsnetz mit den aufsummierten Entscheidungen der Spezialnetze vom Zeitpunkt 0 bis einschließlich t aktiviert. Der Raum der $k_i(t)$ definiert den Eingaberaum des Kombinationsnetzes.

Zur Kombination wurde auf Rhythmusseite die 1sec-Variante mit einem Überlapp von $0,4$ Sekunden und auf Klangfarbenseite die 5 Cepstren Variante mit $0,4$ Sekunden Kontextfenster verwendet.

Ergebnisse Mit einem solchen Kombinationsnetz, daß auf die Ergebnisse der trainierten Rhythmus- und Klangfarbennetze aufsetzt, können $87,5\%$ aller Musikstücke dem richtigen Stil zugeordnet werden. Das Ergebnis ist beachtlich. Die Erkennungsleistungen der Teilnetze (Rhyhmus: $73,6\%$ und Klangfarbe: $81,9\%$) können also bei gemeinsamer Betrachtung der Merkmale deutlich übertroffen werden.

Die $87,5\%$ Erkennungsleistung sollten allerdings nicht mit den $84,9\%$ Erkennungsleistung bei den Wahrnehmungsexperimenten verglichen werden. Die Vp hatten nur 3 Sekunden lange Ausschnitte zur Verfügung, während die $87,5\%$ bei wesentlich längeren Ausschnitten erreicht worden sind. Gibt man dem Netz nur die 3 Sekunden zur Verfügung, dann ist die Erkennungsleistung bei etwa 78%. Nach weiteren 2 Sekunden liegt dann die Erkennungsleistung bei 80%. Der zeitliche Verlauf der Erkennungsleistung ist in Abbildung 6.10 dargestellt.

Die größte Verbesserung in Vergleich zu der Ergebnissen der Spezialnetze ist das ausgewogene Verhältnis von Rock und Pop (siehe Abbildung 6.9).

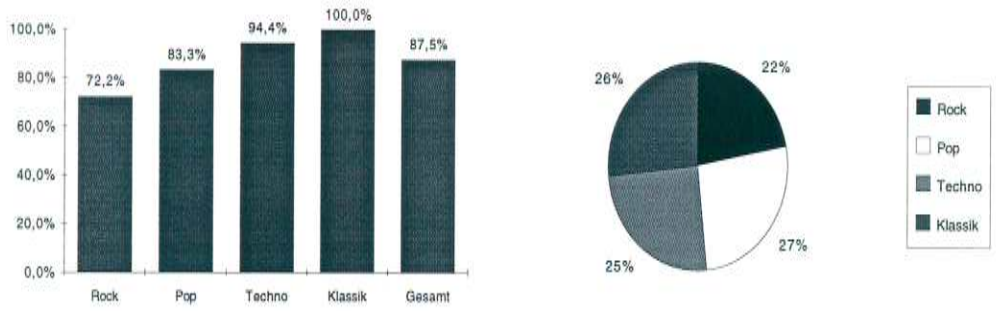

Abbildung 6.9: Erkennung und Dominanz des Kombinationsnetzes

War bei dem Klangfarbennetz der Anteil der Popmusik noch über 36%, so ist hier der Popanteil auf etwa 27% gesunken. Die erkannten Musikstile sind annähernd gleich verteilt. Weiterhin hat die Erkennung von Rock von 61, 1% auf 72, 2% deutlich zugenommen. Hier bringt also das gemeinsame Wissen um Rhyhmus und Klangfarbe dem Netz erhebliche Vorteile. Auch die Erkennung von Techno hat sich stark verbessert. Das Klangfarbennetz kann Techno zu 83, 3% erkennen, daß Rhythmusnetz 66, 7%, zusammen werden jetzt 94, 4% der Technostücke richtig erkannt.

Ein weiterer Vorteil dieser Art der Integration ist die Möglichkeit, eine Klassifikation vorzunehmen, ohne das Ende des Musikstückes abzuwarten. Eine Gesamtentscheidung des Kombinationsnetzes kann auch dann abgefragt werden, wenn noch nicht alle Einzelbewertungen der Spezialnetze vorliegen. Es werden dann nur die Einzelbewertungen, die bis zum jetzigen Zeitpunkt vorliegen, aufsummiert und dem Kombinationsnetz als Eingabe übergeben. Somit ist der zeitliche Verlauf der Erkennungsleistung berechenbar, wie er in Abbildung 6.10 wiedergegeben ist.

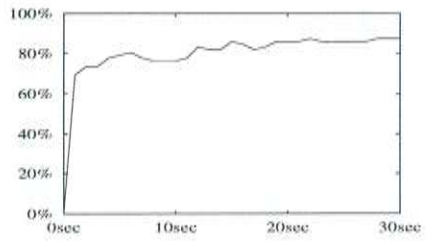

Abbildung 6.10: zeitlicher Verlauf der Erkennungsleistung des Kombinationsnetzes

Analyse erlernten Verhaltens Von Interessse ist ebenfalls, wie stark die einzelne Gewichtung der Spezialnetze an der Gesamtentscheidung ist. Dies kann in ähnliche Weise analysiert werden, wie das gelernte Verhalten des Rhythmusnetzes analysiert worden ist. Wir visualiseren zunächst die Verbindungsgewichte der Eingabeschicht zu der verdeckten Schicht und betrachten dann die Verbindungen der verdeckten Schicht zu der Ausgabeschicht. Der 8-dimensionale Eingabevektor besteht aus den Aktivierungen der 4 Ausgabeneuronen des Rhythmusnetzes und denen des Klangfarbennetzes. Die Reihenfolge ist jeweils Rock, Pop, Techno, Klassik. So enthält der 5.te Koeffizient des Eingabevektors die Aktivierung der Rockunit des Klangfarbennetzes. Wenn sich zum Beispiel das Klangfarbennetz für Rock entscheidet und der 5.te Koeffizient des Gewichtsvektors eines verdeckten Neuron sehr groß ist, dann wird dieses verdeckte Neuron aktiviert.

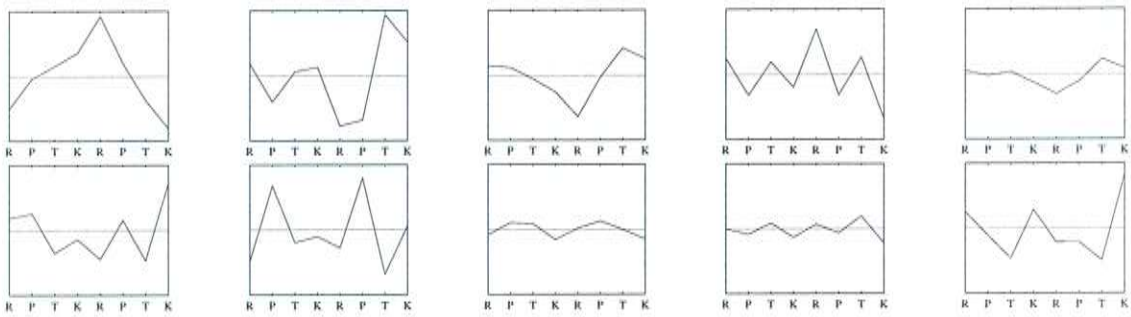

Abbildung 6.11: verdeckte Neuronen des Kombinationsnetzes

An den Verbindungsgewichten zu den verdeckten Neuronen kann also gewissermaßen die Zuständigkeit der Neuronen abgelesen werden. Die Verbindungsgewichte der verdeckten Neuronen zu den 4 Ausgabeneuronen zeigen dann an, welche verdeckten Neuronen für welchen Musikstil relevant sind. Um es mit Begriffen regelbasierter Systeme zu formulieren: Die Verbindungen der Eingabeschicht zur verdeckten Schicht stellen die Prämissen der Regeln dar, die Konklusionen der Regeln werden durch die Verbindungen der verdeckten Schicht zur Ausgabeschicht repräsentiert.

Die Numerierung der verdeckten Neuronen ist zeilenweise von 1 bis 10. Die Schwellwerte der Ausgabeneuronen sind jeweils in der ersten Zeile des entsprechenden Abschnitts der Tabelle 6.5 angegeben. Diskutieren wir nun die gelernten Netzwerkverbindungen des Kombinationsnetzes.

Verbindung	Rock	Pop	Techno	Klassik
Schwellw.	-1.04	-1.66	-1.80	0.69
1	**3.84**	0.48	-7.93	-9.31
2	-4.45	-10.18	**7.86**	1.11
3	-3.89	0.81	**6.51**	1.49
4	0.82	-6.56	-2.77	-3.20
5	-2.94	-1.03	2.87	-0.03
6	-3.68	**1.40**	-4.55	**6.73**
7	-4.05	**1.22**	-5.53	-5.80
8	-1.44	-0.64	-1.43	-2.67
9	-0.51	-2.23	0.42	-2.94
10	-0.63	-0.52	-10.82	**7.17**

Tabelle 6.5: Ausgabeneuronen des Kombinationsnetzes

1. Verbindungen zu dem Rock-Ausgabeneuron
   Betrachten wir die Verbindungen von der verdeckten Schicht zu dem Rock-Ausgabeneuron. Nur das erste verdeckte Neuron besitzt eine stark ausgeprägte postive Korrelation zu dem Rock-Neuron. Und das verdeckte Neuron 1 wird dann aktiviert, wenn der 5.te Koeffizient des EIngabevektors sehr hoch ist. Mit anderen Worten: Aktiviert das Klangfarbennetz sein Rock-Ausgabeneuron, dann wird das verdeckte Neuron 1 des Kombinationsnetzes aktiviert und damit auch das Rock-Ausgabeneuron des Kombinationsnetzes. Bemerkenswerterweise scheint das Rock-Ausgabeneuron des Rhythmusnetzes keinen Einfluß auf das Kombinationsnetz zu haben. Aber dies ist auch klar, wenn wir uns an die Analyse des Rhythmusnetzes erinnern. Dort hat sich kein verdecktes Neuron auf Rock-Rhythmen spezialisert. Das Rhythmusnetz hat eher eine Regel 'im Zweifelsfall Rock' gelernt. Somit sollte das Rhythmusnetz im Rockbereich auch keinen Einfluß haben.

2. Verbindungen zu dem Pop-Ausgabeneuron
   Die beiden verdeckten Neuronen 6 und 7 haben eine positive Verbindung zu dem Pop-Ausgabeneuron. Interessant ist der Gewichtsvektor des 7. ten Neuron. Dieses Neuron wird aktiviert, wenn jeweils die Pop-Ausgabeneuronen des Rhythmus- und des Klangfarbennetzes aktiviert. Im Bereich Pop haben sowohl das Klangfarben- als auch das

Rhythmusnetz Einfluß. Im übrigen ist Neuron 7 negativ zu dem Rock-Ausgabeneuron korreliert. Wenn sich die Spezialnetze für Pop entscheiden, dann soll sich das Kombinationsnetz für Pop und gegen Rock entscheiden. Etwas verwirrend ist das 6.te verdeckte Neuron. Es wird aktiviert, wenn das Klangfarbennetz sein Pop -und sein Klassikneuron aktiviert. Eine Erklärung wäre, daß das Klangfarbennetz sein Klassik-Ausgabeneuron auch bei Pop aktiviert. Das Klangfarbennetz ordnet aber keine Popstücke dem klassischen Stil zu.

Die Analyse der Verbindungen zu den übrigen Ausgabeneuronen kann in analoger Weise fortgesetzt werden. Es zeigt sich, daß die Ausgabeaktivierungen des Klangfarbennetzes von größerer Bedeutung sind, als die des Rhyhmusnetzes. Dies ist aufgrund der Erkennungsleistungen unmittelbar einsichtig. Der Rhythmus trägt aber auf jeden Fall auch seinen Anteil am Erfolg. Ohne Rhythmusnetz ist die Erkennungsleistung 81, 4%, mit dagegen 87, 5%. Die Hinzunahme des Rhythmus bewirkt hauptsächlich Verbesserungen bei Techno und Klassik.

Zusammenfassend konnte die Analyse der Netwerkverbindungen den Nachweis erbringen, daß das erlernte Verhalten der Neuronalen Netze plausibel und nachvollziehbar ist. Damit haben wir ein wichtiges Argument für die Anwendung konnektionistischer Verfahren gewonnen. Gegner konnektionistischer Ansätze behaupten immer wieder, die Ansätze hätten mit 'Magie' zu tun. Diese Behauptung kann zumindest für das Problem der Musikstilklassifikation als widerlegt angesehen werden.

6.2 Versteckte Markov Modelle

Das erfolgreiche Lösen eines Klassifikationsproblems erfordet die Auseinandersetzung mit zwei Problemfeldern. Die Extraktion relevanter Merkmale und das damit verbundene Auffinden eines geeigneten Eingabraumes hat in dem Abschnitt *Neuronale Netze* aufgrund der Vorüberlegungen in Kapitel *Extraktion relevanter Merkmale* weitesgehend Beachtung gefunden. Untersuchungen zur Wahl einer geeigneten Architektur des Klassifikators waren dagegen nicht Gegenstand des letzten Abschnittes, zum Einsatz gelangten ausschließlich Neuronale Netze. Ein Vergleich von Architekturen fehlt bisher. Dieser Abschnitt ist der Untersuchung alternativer Architekturen gewidmet.

Bayes-Entscheidung Auf dem Gebiet der Spracherkennung werden mit Erfolg Klassifikatoren eingesetzt, die auf einer *Bayes*-Entscheidung aufsetzen. Eine Entscheidung zugunsten der Klasse ω_i wird vorgenommen, wenn die Klasse ω_i die größte a-posteriori-Wahrscheinlichkeit $p(\omega_i|x)$ aller Klassen hat. Dabei bezeichnet x das beobachtete akustische Ereignis, dies ist in unserem Fall das Musikstück, genauer die relevanten Merkmale des Musikstückes, Rhythmus und Klangfarbe. Klassifikatoren basierend auf der Bayes-Entscheidung minimieren den Klassifikationsfehler. Der Bau eines solchen Klassifikators erfordert die Kenntnis der a-posteriori-Wahrscheinlichkeiten. Wegen $p(\omega_i|x) * p(x) = p(x|\omega_i) * p(\omega_i)$ können wir das Problem auf die Bestimmung der klassenbedingten Wahrscheinlichkeiten $p(x|\omega_i)$ und a-priori-Wahrscheinlichkeiten $p(\omega_i)$ reduzieren. Dem Leser ist letztere Größe schon bekannt. Die Musikstücke in der Datenbasis sind gleichverteilt. Es gilt: $p(rock) = p(pop) = p(techno) = p(klassik) = 0,25$. Es verbleibt die Aufgabe, die klassenbedingten Wahrscheinlichkeiten zu schätzen. Dieses Problem ist schon in zahlreichen Arbeiten bearbeitet wurden, so daß wir uns auf deren Erkenntnisse stützen können.

Gaußklassifikator Wir betrachten im folgenden zwei Möglichkeiten zur Schätzung der Wahrscheinlichkeitsverteilungen. Der ersten Variante liegt die Annahme zugrunde, die beobachteten Daten gehorchen einer Gaußverteilung. Damit reduziert sich das Problem auf die Bestimmung von Mittelwert und Varianz. Lockert man die Voraussetzung, indem man als Verteilung eine Mischung von Gaußverteilungen ansetzt, dann erfordert dies die Bestimmung der Mischungsgewichte. Dies läßt sich durch den EM-Algorithmus oder dessen EM* Variante bewerkstelligen. In einem iterativen Verfahren werden die Merkmalsvektoren x den einzelnen Gaußverteilungen zugeordnet und aufgrund dieser Zuordnung Mischungsgewichte, Mittelwerte und Varianzen bestimmt. Die EM* Variante vereinfacht den Zuordnungsprozeß, indem von einem winner-takes-all Prinzip ausgegangen wird. Einzelheiten zur Herleitung des Algorithmus sind in in dem Buch von Schukat-Talamazzini [20] beschrieben. Aufgrund der Verteilungsannahme werden diese Klassifikatoren auch *Gaußklassifikatoren* genannt.

HMM Variante 2, die Schätzung der klassenbedingten Wahrscheinlichkeitsverteilungen, geht von einem zweistufigen Zufallsprozeß aus. Modelliert

wird ein System, welches die akustischen Ereignisse x emittiert. Dieses System wird als ein endlicher Automat begriffen. Die Wahrscheinlichkeit, ein akustisches Ereignis x zu beobachten, ist von dem augenblicklichen Systemzustand abhängig. Jedem Zustand wird also eine Emissionswahrscheinlichkeitsverteilung zugeordnet. Diesem Zufallsprozeß überlagert ist ein zweiter stochastischer Prozeß. Dieser dient zur Modellierung der Zustandswechsel des Systems. Dabei wird vorausgesetzt, daß der augenblickliche Systemzustand lediglich von dem vorigen Systemzustand abhängig ist. Solche Modelle zur Schätzung der klassenbedingten Wahrscheinlichkeitsverteilungen tragen den Namen *Versteckte Markov Modelle*. Die Anwendung des EM bzw. EM* Algorithmus nennt man bei diesen Modellen Baum-Welch- bzw. Viterbi-Training.

Vorgehensweise Wichtig für iterative Verfahren ist ein geeigneter Ausgangspunkt. Es ist eine Initialisierung der zu lernenden Parameter erforderlich. Mittelwerte und Varianzen lassen sich initial durch den Einsatz eines Ballungsverfahren bestimmen. Sollen etwa die initialen Parameter einer Mischung von 10 Gaußverteilungen bestimmt werden, so werden mit dem *k-means*-Algorithmus $k = 10$ Klassenzentroide gefunden, mit der die Mittelwerte initialisiert werden. Sind die Mittelwerte gefunden, können dann initiale Werte der Varianzen und Mischungsgewichte berechnet werden. Hinweise dazu finden sich in jedem Grundlagenwerk, etwa in [20].

Die Vorgehensweise bei dem Training dieser Klassifikatoren geschieht in gleicher Weise wie bei den Neuronalen Netzen. Trainiert wird auf der Trainingsmenge. Abgebrochen wird das Training bei schlechter werdender Erkennungsleistung der Kreuzvalidierungsmenge. Zum Testen kommt die Testmenge zum Einsatz.

Bevor wir uns den Ergebnissen der Experimente mit diesem Klassifikationsansatz zuwenden, sollten wir diskutieren, inwieweit uns die Erkenntnisse zur Extraktion relevanter Merkmale von Nutzen sind, die in den Experimenten mit den Neuronalen Netzen gewonnen wurden. So hat sich gezeigt, daß sich Cepstren besser als Melscale-Koeffizienten für die Repräsentation der Klangfarbe eignen. Auch wissen wir, daß die kritische Grundtonhöhe bei etwa 2000 Hz liegen sollte. Dies sind Erkenntnisse, die unabhängig von der Architektur des Klassifikators seien sollten. Zu berücksichtigen ist allerdings die Aussagekraft. So sind die Unterschiede zwischen Melscales und Cepstren sehr groß (73,6% gegenüber 81,9%), dagegen haben die Unterschiede bei

Fenstergröße	Überlappung	Größe=5	Größe=10	Größe=15	Größe=20
1 sec	/	62,5%	63,9%	61,1%	/
3 sec	/	56,9%	59,7%	61,1%	59,7%
1 sec	0,4 sec	55,6%	61,1%	62,5%	62,5%
3 sec	1 sec	55,6%	**65,3%**	62,5%	/

Tabelle 6.6: Vergleich der Erkennungsleistungen des Gaußklassifikators auf Rhythmus-Basis bei unterschiedlichen Eingaberäumen

Verkleinerung der Grundtonhöhe auf 1024 Hz (= 10 Cepstral-Koeffizienten) nicht solche Signifikanz (81,9% und 77,8%). Zur Validierung dieser Ergebnisse sind Experimente zur Bestimmung der kritischen Grundtonhöhe bei diesem Klassifikationsansatz angebracht. Zu beachten ist auch die Abhängigkeit der Größe des Eingaberaumes von dem Lernverfahren des Klassifikators. Es ist also erforderlich, die Größe des Eingabefensters an den Klassifikator anzupassen. Dies gilt für den Rhythmus als auch die Klangfarbe. Nach diesen Bemerkungen können wir uns nun den Experimenten zuwenden.

6.2.1 Rhythmus

In Abschnitt *Neuronale Netze, Rhythmus* sind die Details und die Bedeutung des Merkmals Rhythmus bereits herausgearbeitet worden. Daher kann an dieser Stelle auf allgemeine Bemerkungen verzichtet werden, und wir konzentrieren uns stattdessen auf die Besonderheiten dieses Klassifikationsansatzes. Wir nehmen also an, daß die beobachteten Rhythmusmerkmale sich gemäß einer Mischung von Gaußverteilungen verhalten. Experimentell zu bestimmen sind Eingabefenster und eine eventuelle Überlappung der Fenster. Zudem ist es natürlich erforderlich, eine geeignete Modellgröße zu finden, d.h. die Anzahl der Gaußverteilungen.

Ergebnisse Tabelle 6.6 enthält zusammengefaßt die Ergebnisse des Gaußklassifikators. Die erste Spalte gibt die Länge des Eingabefensters an, innerhalb dessen der Rhythmus berechnet wird. Die Überlappung der Fenster ist zum einem 0,4 Sekunden bei Berechnung des Rhythmus in 1sec Fenster und 1 Sekunde Überlappung bei den 3sec-Fenster, Die weiteren Spalten beinhalten die Erkennungsleistungen bei unterschiedlichen Modellgrößen. Damit ist die

Anzahl der Gaußverteilungen gemeint, die linear kombiniert werden. Es ist augenscheinlich, daß die Ergebnisse des Gaußklassifikators hinter den Ergebnissen des Neuronalen Netzes weit zurückbleiben. Bevor wir die Klassifikatoren ausführlicher vergleichen, analysieren wir den Einfluß der Fenstergröße und der Fensterverschiebung.

1. Fenstergröße
   Beginnen wir mit dem Vergleich der Varianten ohne Überlappung. Die 1sec-Variante ist mit 63,9% besser als die 3sec-Variante mit 61,1%. Die Vorteile der 1sec-Variante sind hauptsächlich im Techno - und Klassikbereich zu finden, in der die Ergebnisse der Klassifikatoren auf Rhythmus-Basis von Bedeutung sind. Bei den Varianten mit Überlappung sieht das Bild umgekehrt aus. Dort erreicht die 1sec-Variante 62,5%, die 3sec-Variante dagegen 65,3%. Wie kommt dieser Widerspruch zustande? Aufschluß gibt eine detaillierte Auflistung der Klassifikation der Musikstücke. Die 3sec-Variante kann besser Rock- und Popstücke klassifizieren, während bei Techno- und Klassikstücken die 1sec-Variante Vorteile hat.

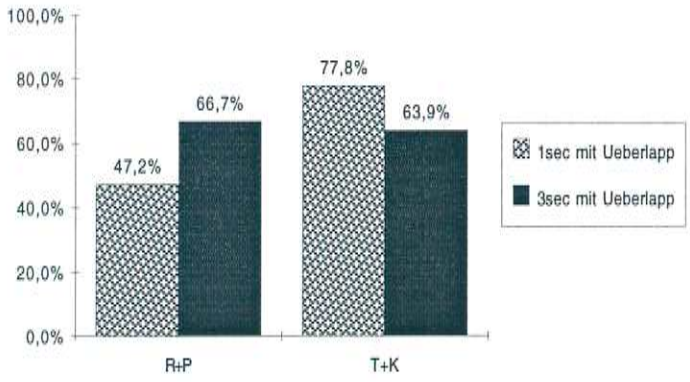

Abbildung 6.12: Vergleich der Erkennungsleistungen von Rock/Pop und Techno/Klassik bei unterschiedlichen Fenstergrößen

In Abbildung 6.12 werden die Ergebnisse der 1sec-Variante mit Überlappung mit 62,5% Erkennungsleistung mit der 3sec-Variante mit Überlappung und 65,3% Erkennungsleistung verglichen. Die Vorteile der 3sec-Variante liegen eindeutig im Rock- und Popbereich, während im

Techno- und Klassikbereich die 1sec-Variante besser ist. Wird nun davon ausgegangen, daß der Rhythmus lediglich ein relevantes Merkmal für den Klassik- und Technobereich ist, aber kein entscheidenes Merkmal zur Trennung von Rock und Pop ist, dann sind die Ergebnisse der 3sec-Variante zu relativieren. Zudem leidet die Signifikanz der 65,3% aufgrund der Tatsache, daß bei geringer Änderung der Modellgröße sich nicht annähernd gleich gute Ergebnisse reproduzieren lassen. Insofern sollten den 65,3% keine Bedeutung zugemessen werden, die ihnen nicht zusteht.

2. Überlappung
   Aufgrund der Ergebnisse in Tabelle 6.6 lassen sich keine eindeutigen Aussagen über die Auswirkungen einer überlappender Fensterverschiebung treffen. Während bei der 1sec-Variante die Überlappung von Nachteil ist (62,5% statt 63,9%), scheint bei der 3sec-Variante die Überlappung von Vorteil zu sein (65,3% statt 62,5%). Diese Ergebnisse kommen aufgrund einer Verschiebung der Trennlinie zwischen Rock und Pop zustande. Ein ähnlich widersprüchliches Verhalten war auch bei den Neuronalen Netzen zu beobachten. Es bleibt festzuhalten, daß der Rhythmus ein relevantes Merkmal für die Musikstile Techno und Klassik ist, sich aber nicht jedoch für die Trennung von Rock und Pop eignet.

Vergleich der Klassifikatoren Das Verhalten des Gaußklassifikators stimmt bezüglich des Nicht-Könnens im Rock- und Popbereich mit dem Neuronalen Netz überein. Die Rock- und Popprobleme sind also weniger ein durch den Klassifikator verursachtes Problem, sondern lassen sich dahingehend deuten, daß der Rhythmus kein signifikantes Merkmal dieser Musikstile ist, was im übrigen konform zur menschlichen Wahrnehmung ist. Eine eindeutige Aussage zur Fenstergröße in der der Rhythmus berechnet wird, ist nicht möglich. Bei Neuronalen Netzen konnten mit den kleineren Fenster 73,6% der Musikstücke erkannt werden, bei Verwendung größeren Fenster waren es 70,2%. Mit der 1sec-Variante des Gaußklassifikators sind 63,9% richtig erkannt wurden, 65,3% bei der 3sec-Variante. Sofern lediglich der Klassik- und Technobereich betrachtet wird, kann dennoch eine Aussage zugunsten der 1sec-Variante getroffen werden, da in diesem Sektor die 1sec-Varianten beider Klassifikatoren besser sind. Insgesamt gesehen, können Neuronale Netze

die Rhythmus-Merkmale besser verarbeiten als die Gaußklassifikatoren.

6.2.2 Klangfarbe

Wie in den Vorbemerkungen am Anfang dieses Abschnittes betont, ist der Zweck dieser Untersuchung der Vergleich des Klassifikationsansatzes. Zur Modellierung der klassenbedingten Wahrscheinlichkeiten werden Normalverteilungen als auch versteckte Markov Modelle in Betracht gezogen. Auch bei diesem Klassifikationsansatz muß eine geeignete Fenstergröße gefunden werden. Zudem scheint eine Validierung der kritischen Grundtonhöhe angebracht. Dazu können die Ergebnisse des Gaußklassifikators herangezogen werden. Nachdem diese Größen auf Basis der Ergebnisse des Gaußklassifikators bestimmt wurden sind, können darauf aufbauend Experimente mit den HMM's durchgeführt werden. Die Gefahr einer unzulässigen Generalisierung ist gering, da sich eine Mischung von Gaußverteilungen als ein Ein-Zustands-HMM, dessen Emissionswahrscheinlichkeiten durch solche Gaußverteilungen parameterisiert werden, interpretieren läßt.

Ergebnisse der Gaußklassifikatoren Beginnen wir also mit der Modellierung durch Gaußverteilungen. Die zu bestimmenden Werte für die Grundtonhöhe und Fenstergröße definieren gemeinsam die Größe des Eingaberaumes und müssen deshalb als voneinander abhängig gesehen werden. Wie auch bei den Gaußklassifikatoren auf Rhythmus-Basis ist es erforderlich, die Anzahl der zu mischenden Gaußverteilungen zu variieren. Zusammengefaßt sind die Ergebnisse in Tabelle 6.7 dargestellt.

Grundtonhöhe	Fenstergröße	Größe=5	Größe=10	Größe=15	Größe=20
2052 Hz	0,12 sec	72,0%	76,4%	70,1%	/
2052 Hz	0,20 sec	**77,8%**	75,0%	73,0%	/
2052 Hz	0,28 sec	75,0%	73,6%	73,6%	/
1026 Hz	0,12 sec	65,3%	66,7%	73,6%	72,2%
1026 Hz	0,20 sec	77,8%	69,4%	/	/
1026 Hz	0,28 sec	76,4%	69,4%	/	/

Tabelle 6.7: Vergleich der Erkennungsleistungen des Gaußklassifikators auf Klang-Basis bei unterschiedlichen Eingaberäumen

Bevor wir den Einfluß der Grundtonhöhe berücksichtigen, diskutieren wir die Ergebnisse der oberen 3 Spalten bei einer Grundtonhöhe von 2052 Hz (dies sind 5 Cepstral-Koeffizienten). Das beste Resultat wird bei einer Fenstergröße von $0,2$ Sekunden erreicht. Es werden $77,8\%$ der Musikstücke richtig erkannt. Eine weitere Vergrößerung der Fenstergröße auf $0,28$ Sekunden bewirkt ein Absinken der Erkennungsleistung. Die Größe des Eingaberaumes beginnt kritisch zu werden. Auch läßt sich die Anzahl der zu mischenden Gaußverteilungen nicht mit Erfolg erhöhen. Werden statt 5 Gaußverteilungen 10 Gaußverteilungen miteinander kombiniert, dann sinkt die Erkennungsleistung auf $75,0\%$. Das Lernverhalten ist also weitaus ungünstiger als bei Neuronalen Netzen. Mit Neuronalen Netzen war es möglich, $81,9\%$ der Musikstücke richtig zu klassifizieren.

Werden 10 Cepstral-Koeffizienten zur Bildung des Eingaberaumes verwendet, d.h. die kritische Grundtonhöhe liegt bei 1026 Hz, dann werden auch $77,8\%$ Erkennungsleistung erreicht. Aufgrund der Erkennungsleistung dieses Klassifikationsansatzes kann demnach keine exakte Aussage zur Wahl der Grundtonhöhe getroffen werden. Bei gleicher Erkennungsleistung ist der kompakteren Repräsentation der Vorzug zu geben. Zudem sollten die Ergebnisse der Neuronalen Netze aufgrund ihrer besseren Erkennungsleistung stärker gewichtet werden. Es kann also bei der Berücksichtigung von nur 5 Cepstralkoeffizienten verblieben werden.

Zustandübergangsstrukturen von HMM Aus diesen Ergebnissen lassen sich einige Restriktionen für die Experimente mit den HMM's ableiten. Die Fenstergröße kann auf $0,2$ Sekunden festgelegt werden und die Anzahl der Cepstal-Koeffizienten auf 5 beschränkt werden. Es verbleibt die Aufgabe, eine geeignete Zustandsübergangsstruktur zu suchen. Dazu müssen wir uns überlegen, was die Zustände repräsentieren sollen. Machen wir deshalb einen kleinen Ausflug in die Spracherkennung. Ein Phonem kann in ein Anglitt, Kern und Abglitt gegliedert werden. Es besteht also eine zeitliche Reihenfolge. Konsequenz für die Modellierung kann ein Modell bestehend aus 3 Zuständen sein, Zustandsübergänge sind nur jeweils in den Nachfolgezustand erlaubt. Ist nun ein solches Modell für die Klangfarbe eines Musikstückes adäquat? Sicherlich nicht. Wir können uns vorstellen, daß die Zustände für spezielle Klangereignisse zuständig sind. Dann ist eine strikte zeitliche Ordnung unsinnig. Vielmehr sollte es möglich sein, 'Klangereignisse' auszulassen.

Auch Wiederholungen im Ablauf der Folge der Klangereignisse sind denkbar. Diese Überlegungen legen es nahe, eine Modellierung vorzunehmen, wie sie in der folgenden Abbildung gezeigt wird.

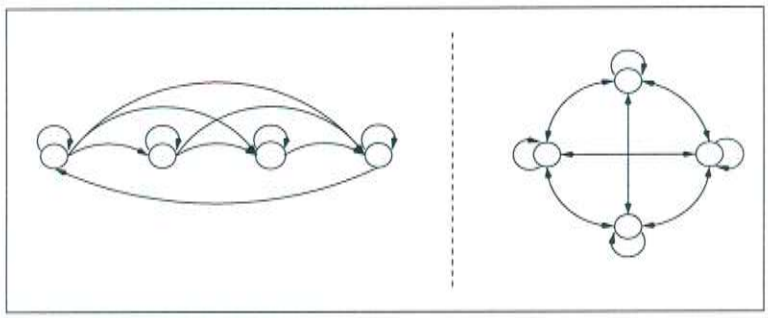

Abbildung 6.13: ergodische Zustandsübergangstrukturen von versteckten Markov Modellen

Initialisierung von HMM Ein solche Modellierung kann nur erfolgreich sein, wenn es gelingt die Emissionswahrscheinlichkeitsverteilungen vernünftig zu initialisieren. Werden für diese Verteilungen einzelne Gaußverteilungen zugrunde gelegt, dann kann ein Ballungsalgorithmus angewendet werden. Werden für das Rock-Modell etwa 4 Zustände vorgesehen, dann werden 4 Ballungen der Rock-Klangfarbenvektoren berechnet und man erhält so Startwerte für die Mittelwerte und Varianzen. Eine ähnliche Vorgehensweise ist möglich, wenn die Emissionswahrscheinlichkeitsverteilungen durch mehr als eine Gaußverteilung parameterisiert werden. Bevor der Ballungsalgorithmus angeworfen wird, müssen die Klangfarbenvektoren den einzelnen Zuständen zugeordnet werden, um so die Datenbasis zu berechnen, auf der der Ballungsalgorithmus aufsetzt. Eine solche Zuordnung der Vektoren läßt sich durch ein bereits trainiertes gleich strukturiertes Modell definieren. Ein Merkmalsvektor wird demjenigen Zustand zugeordnet, dessen Zustand die höchste Emissionswahrscheinlichkeit hat. Ein solches gleich strukturiertes Modell können wir aber schon trainieren. Bevor wir das größere Modell mit den Mischverteilungen trainieren, trainieren wir das Modell mit den einzelnen Gaußverteilungen.

Ergebnisse der HMM-Klassifikatoren Nachdem dieser Punkt geklärt ist, analysieren wir die Ergebnisse der Tabelle 6.8. Ist die Zustandsübergangsstruktur fast ergodisch, dann werden 79,2% aller Musikstücke richtig erkannt. Werden die Restriktionen an Rücksprüngen weiter gelockert, dann erhält man mit der voll ergodischen Zustandsübergangsstruktur 76,4%.

Modell	Größe=1	Größe=2	Größe=3
fast ergodisch	68,1%	**79,2%**	73,6%
voll ergodisch	68,1%	76,4%	72,2%

Tabelle 6.8: Vergleich der Erkennungsleistungen bei ergodisch strukturierten HMM auf Klang-Basis

Durch die Hinzunahme weiterer möglicher Zustandsübergänge erhöht sich die Gefahr, daß die Neuabschätzung der Übergangswahrscheinlichkeiten unzuverlässiger wird. So hat der Übergang des 2.ten Zustands zu dem 3.ten Zustand der ergodischen Struktur bei dem Rock-Modell die Wahrscheinlichkeit Null. Ein deutliches Anzeichen, daß die Parameterschätzung Probleme bereitet. Auffallend ist der große Sprung bei Erhöhung der Anzahl der Gaußverteilungen von 1 auf 2. Dies ist bei beiden Zustandsübergangsstrukturen zu sehen. Dies liegt nicht zuletzt an der besseren Initialisierung. Der Ballungsalgorithmus zur Initialisierung der kleinen Modelle basiert auf der gesamten Trainingsmenge. Dagegen kann bei der Initialisierung der größeren Modelle die Zuordnung der Merkmalsvektoren zu den Zuständen des kleinen Modells verwendet werden.

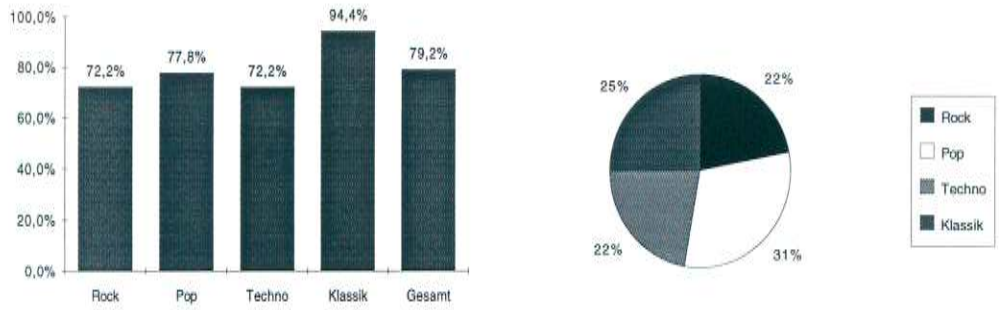

Abbildung 6.14: Erkennung und Dominanz bei HMM auf Klangbasis

Vergleich der Klassifikatoren Vergleichen wir nun die Ergebnisse mit denen der Gaußklassifikatoren. Mit den Gaußklassifikatoren konnten 77,8% erreicht werden, mit den HMM 79,2%. Bei Modellierung durch eine Mischung von Gaußverteilungen repräsentieren einzelne Gaußverteilungen bestimmte Klangereignisse. Durch die Mischungsgewichte kann die Wichtigkeit und Relevanz der Klangereignisse modelliert werden. Die HMM's haben eine ähnliche Möglichkeit. Sie modellieren etwas exakter. Durch die Übergangswahrscheinlichkeiten kann die Reihenfolge von Klangereignissen ausgedrückt werden. Auf ein Ereignis des Typs A folgt ein Ereignis des Typs B. Mit Gaußklassifikatoren kann ausgedrückt werden, daß das Ereignis des Typs A für den Musikstil X typisch ist und das der Musikstil Y typischerweise Klangereignisse des Typs B hat. Die Ergebnisse lassen vermuten, daß erstere Modellierung dem Problem eher gerecht wird.

Zusammenfassung Wir haben nun einige Ergebnisse mit diesem Klassifikationsansatz diskutiert und können nun die beiden Klassifikationsansätze vergleichen. Sowohl bei den Rhythmus- als auch bei den Klangfarbeneingaberräumen erweisen sich Neuronale Netze als besser. Neuronale Netze auf Rhythmus-Basis können 73,6% der Musikstile richtig klassifizieren, Gaußklassifikatoren nur 65,3%. Auf Klangfarben-Basis erreichen Neuronale Netze 81,9%, mit dem anderen Klassifikationsansatz sind es 79,2%. Das Lernverhalten Neuronaler Netze zeigt seine Vorteile. Sie sind in der Lage diskriminierend zu lernen. Ihnen ist es möglich, Sachverhalte der Art 'Ein Ereignis des Typs A ist untypisch für Musikstil Y' auszudrücken. Über dieses Potential verfügen Gaußklassifikatoren und HMM's nicht. Eine Modellierung mit den versteckten Markov Modellen ist offenbar für das Problem der Musikstilklassifikation nicht angemessen.

6.3 Modellierung zeitabhängiger Phänomene

Die Informationsverarbeitung zeitabhängiger Phänomene ist von besonderen Schwierigkeiten gekennzeichnet. Zur Verarbeitung eines Ereignisses zum Zeitpunkt t ist es erforderlich, auch vergangene Ereignisse zu berücksichtigen. Relevante Ereignisse zur Musikstilerkennung sind dabei Rhythmus- und Klangereignisse. In diesem Abschnitt konzentrieren wir uns auf die Klangereignisse. Um diese Ereignisse hinsichtlich ihrer zeitabhängigen Bedeutung

geeignet zu verarbeiten, müssen eine Reihe von Fragen geklärt werden.

6.3.1 Umfang der Zeitabhängigkeit

Es ist sicherlich nicht in jedem Fall erforderlich, *alle* bisherigen Ereignisse zu berücksichtigen. Eine zentrale Frage lautet deshalb, in welchem Umfang vergangene Ereignisse berücksichtigt werden müssen. Diese Frage ist in den vergangenen Kapiteln bereits mehrfach angesprochen worden. Fassen wir die Ergebnisse kurz zusammen.

1. Experimente zur Wahrnehmung von Musik
   Bei der Darbietung unverfremdeter Ausschnitte mit 1 Sekunde Dauer sind 82,9% der Stücke richtig erkannt worden. Sind die gleichen Stücke 3 Sekunden lang gespielt worden, dann war die Erkennungsleistung 84,9%. Mit einem Signifikanztest konnte nicht nachgewiesen werden, daß diese Unterschiede nicht zufällig sind. Die Nullhypothese konnte nicht abgelehnt werden. Die Schlußfolgerung war, daß die Musikstilerkennung innerhalb eines zeitlich lokalen Fensters möglich ist. Von Bedeutung ist dieses Ergebnis für die maschinelle Informationsverarbeitung aber nur dann, wenn sich die Ergebnisse der menschlichen Musikwahrnehmung übertragen lassen. Dies ist als Arbeitshypothese vorausgesetzt worden.

2. Experimente zur Anpassung des Eingaberaums
   In dem Abschnitt *Neuronale Netze, Klangfarbe* diskutierten wir die Ergebnisse von Netzsimulationen mit unterschiedlichen Fenstergößen. Bei sehr kleinen Fenstergrößen von 0,04 Sekunden arbeitete das Netz häufig fehlerhaft. Das Netz erkannte lediglich 68,1% der Musikstücke korrekt. Eine Vergrößerung der Fenster auf 0,4 Sekunden brachte einen erheblichen Fortschritt. Die Erkennungsleistung lag dann bei 81,9%. Eine weitere Fenstervergößerung auf 1 Sekunde ließ die Erkennungsleistung jedoch wieder auf 77,9% sinken. Eine Fenstervergößerung bewirkt aber automatisch eine Erhöhung der zu trainierenden Modellparameter. Dies hat zur Folge, daß die Schwierigkeiten bei der Spezialisierung auf relevante Eingaberegionen zunehmen - mit der Konsequenz, daß die Generalisierungsfähigkeit abnehmen kann.

Die Experimente zur Wahrnehmung von Musik und zur Anpassung des Eingaberaums geben Anhaltspunkte, in welchem Umfang Kontextinformationen

erforderlich sind. Aus diesen Experimenten können aber keineswegs eindeutige Aussagen zur Zeitabhängigkeit abgeleitet werden. Sowohl bei den Experimenten zur Wahrnehmung von Musik als auch bei den Experimenten zur Anpassung des Eingaberaums sind Sachverhalte zu berücksichtigen, die die Aussagekraft beeinträchtigen.

Es gibt also Gründe genug, sich bei der Informationsverarbeitung nicht darauf zu verlassen, daß alle notwendigen Informationen in dem Zeitfenster enthalten sind. Und es ist auch keine Lösung, daß Zeitfenster hinreichend groß zu wählen, so daß in jedem Fall alle zeitabhängigen Ereignisse enthalten sind. Wenn man das Zeitfenster also nicht beliebig groß machen kann, dann muß man die Informationsverarbeitung anpassen, um die Unsicherheiten bei der Zeitinvarianz auszugleichen.

6.3.2 Partiell rekurrente Netzwerke

Es gibt Neuronale Netze, die zeitliche Abhängigkeiten berücksichtigen. Bei Jordan-Netzwerken werden die Aktivierungen der Ausgabeschicht zum Zeitpunkt t als Eingabe zum Zeitpunkt $t + 1$ mitverwendet. Elman-Netzwerke sind analog aufgebaut - bis auf den Unterschied, daß hier die Aktivierungen der verdeckten Schicht statt der Ausgabeschicht genommen werden.

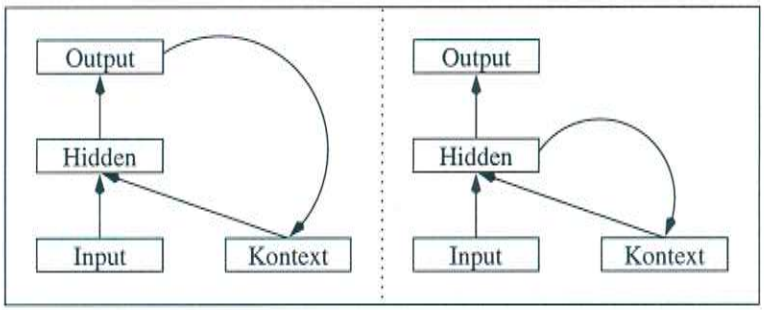

Abbildung 6.15: Jordan- und Elman-Netzwerke

Zeitabhängigkeit wird bei diesen Netzwerken also in Form von Kontext berücksichtigt. Dies entspricht im wesentlichen der Situation bei einer Vergößerung der Eingabefenster. Eine Vergößerung der Eingabefenster läßt sich als das Hinzufügen einer Kontextschicht zu der Eingabeschicht auffassen, wenn man die Kontextschicht mit den vergangenen Aktivierungen der Eingabeschicht füllt.

Zeitabhängige Phänomene werden bei diesen Ansätzen nur durch Kontextinformationen repräsentiert. Das Netz ist gezwungen, die Art der Zeitabhängigkeit aus den Kontextinformationen zu lernen. Besser wäre eine direktere Modellierung der Zeitabhängigkeit.

Ein weiterer Nachteil dieser Netzwerke ist konzeptioneller Art. Es werden auf einer Stufe Informationen unterschiedlicher Abstraktionsniveaus verarbeitet. Dies gilt insbesondere bei den Jordan-Netzwerke. Cepstral-Koeffizienten, die in der Eingabeschicht anliegen und approximierte a-posteriori-Wahrscheinlichkeiten der Kontextschicht sind keine gleichartigen Informationen. Die Eingabe der verdeckten Neuronen besteht dann aus einer Kombination aus 'Äpfeln' und 'Birnen'. Das Gruppenprinzip fordert jedoch, daß gemeinsam verarbeitete Informationen auch gleichartig sein sollen.

6.3.3 zeitabhängige Informationsverarbeitung

Eine angemessene Modellierung zeitabhängiger Phänomene berücksichtigt neben den Unsicherheiten der Zeitinvarianzannahme auch generelle Prinzipien, wie das Einfachheits- und Gruppenprinzip. Zudem sollten alle sinnvollen Möglichkeiten genutzt werden, Problemwissen bei der Modellierung zu berücksichtigen. Dies betrifft in diesem Fall die *Art* der Zeitabhängigkeit. Wird dem Netz lediglich Kontext zur Verfügung gestellt, dann muß das Netz allein die Art der Zeitabhängigkeit aus den Kontextinformationen lernen. Aus den bisherigen Experimenten haben wir aber Hinweise zu der Art der Zeitabhängigkeit. Fassen wir deshalb die Ergebnisse der bisherigen Versuche in Hinblick auf die Zeitabhängigkeit zusammen.

Neuronale Netze gehen von der vollen Zeitinvarianzannahme aus. Alle relevanten Informationen sind in dem Eingabefenster enthalten. Das Verhalten des Netzes kann durch Regeln folgender Art interpretiert werden: Ist das Klangereignis zum Zeitpunkt t vom Typ A, dann ist der Musikstil eher vom Typ X und nicht vom Typ Y. Die Konklusionen sind diskriminativer Art. Die Prämissen erfassen nur das aktuelle Klangereignis. Die Gesamtentscheidung entsteht durch Summation aller Einzelbewertungen.

Gaußklassifikatoren sind bezüglich der Prämisse ähnlich zu den Neuronalen Netzen. Auch sie berücksichtigen nur das aktuelle Klangereignis. Die Konklusionen sind aber natürlich nicht diskriminativer Art. Es gibt keine Abgrenzung zwischen den einzelnen Musikstilen. Jedes Modell eines Musikstils wird getrennt trainiert.

Kommen wir nun zu den versteckten Markov Modellen. Die Modellierung mit HMM's ist auch nicht diskriminativ. Die Parameter der Gaußklassifikatoren und der HMM's werden mit dem gleichen Lernverfahren eingestellt. Aber die Art der Prämisse ist unterschiedlich. Die Emissionswahrscheinlichkeit hängt von dem aktuellen Zustand ab. Die Wahrscheinlichkeit für den aktuellen Zustand hängt wiederum vom letzten Zustand und dessen Klangereignis ab. Dies kann man als Prämissen der Art 'Auf Klangereignis A folgt Klangereignis 'B' auffassen.

Wenn wir die Erkennungsleistungen der Architekturen vergleichen, dann stellen wir zunächst fest, daß diskriminative Lernverfahren auf jeden Fall von Vorteil sind. Auf dem zweiten Blick sehen wir aber auch, daß die Erkennungsleistung des HMM besser als die des Gaußklassifikators ist. Damit kann folgende Idee motiviert werden. Die Modellierung zeitabhängiger Phänomene soll zum einen Abhängigkeiten der Art 'Auf Klangereignis A folgt Klangereignis B' gerecht werden und zum anderen auch diskriminativ sein.

Modell	Zeitabhängigkeit	Lernverfahren	Erkennung
Neuronales Netz	aktuelles Ereignis	diskriminativ	$81,9\%$
Gaußklassifikator	aktuelles Ereignis	nicht diskriminativ	$77,8\%$
HMM	auf Ereignis A folgt B	nicht diskriminativ	$79,2\%$
?	auf Ereignis A folgt B	diskriminativ	?

Tabelle 6.9: Modellierung zeitabhängiger Phänomene

Modellierung zeitlich abhängiger Klangereignisse Nachdem nun die Idee da ist, müssen wir uns um folgende Punkte kümmern. Zunächst muß untersucht werden, auf welche Art und Weise eine Spezialisierung auf bestimmte Klangereignisse erreicht werden kann. Wenn wir die Klangereignisse kennen, dann können wir überlegen, wie die Abhängigkeiten der Ereignisse modelliert werden. Im dritten Schritt kommt dann das diskriminative Lernen zur Musikstilerkennung unter Berücksichtigung der zeitlichen Abhängigkeiten.

Fangen wir mit der Spezialiserung auf Klangereignisse an. Dies ist nicht weiter schwer. Als wir die gelernten Netzwerkverbindungen des Rhythmusnetzes in Abschnitt *Neuronale Netze, Rhythmus* analysierten, haben wir gesehen, daß die verdeckten Neuronen sich auf bestimmte Eingaberegionen spe-

zialisiert haben. Bei dem Klangnetz ist es nun genauso. Dazu betrachten wir die zeitlichen Verläufe der verdeckten Neuronen des Klangnetzes.

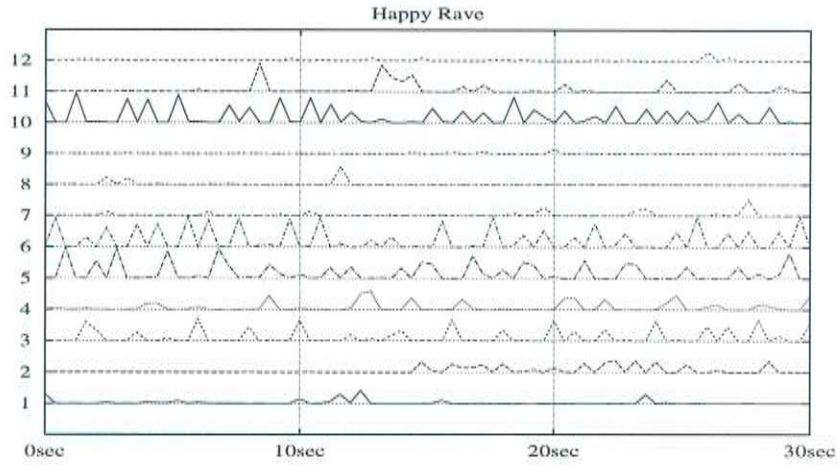

Abbildung 6.16: zeitlicher Verlauf der Aktivierungen verdeckter Neuronen bei einem Techno-Stück

In Abbildung 6.16 sind die Aktivierungen bei der Eingabe eines Technostückes zu sehen. Die verdeckte Schicht des Klangnetzes besteht aus 12 Neuronen. Das 9.te und 12.te Neuron sind während der ganzen Zeit kaum aktiv. Dagegen sind die für den Techno-Stil so typischen Regelmäßigkeiten deutlich in dem zeitlichen Verlauf der Aktivierungen des 5.ten, 6.ten und 10.ten Neurons ersichtlich. Die hohen Aktivierungen dieser Neuronen passen auch zu den stark positiven Verbindungen zu dem Techno-Ausgabeneuron. Die verdeckten Neuronen haben sich also auf bestimmte Klangereignisse spezialisiert. Dies ist genau das erwünschte Resultat.

Extraktion zeitlicher Abhängigkeiten Modellieren wir nun die zeitlichen Abhängigkeiten der Klangereignisse. Mit dem winner-takes-all Prinzip legen wir zu jedem Zeitpunkt das dominierende Klangereignis fest. Dann kann man dann ganz einfach Sachverhalte 'Auf Ereignis A folgt Ereignis B' modellieren. Durch Berechnen der Häufigkeit dieses Ereignisfolgen erhält man Bigramm-Wahrscheinlichkeiten $P(EreignisB|EreignisA)$. Die zeitlichen Abhängigkeiten können natürlich auch komplizierter sein, so lassen sich in gleicher Weise Trigramm-Wahrscheinlichkeiten berechnen. Bei dem Zählen

der Ereignisübergänge wird zunächst der gesamte zeitliche Verlauf berücksichtigt. Prinzipiell kann man jedoch auch nur zeitlich lokale Ausschnitte berücksichtigen.

Neben den Ereignisübergängen selbst kann auch die Dauer der Ereignisse interessant sein. Dies kann modelliert werden, indem man zählt, wie lange ein Spezialist die höchste Aktivierung unter allen anderen Spezialisten hat. Damit kann man zum Beispiel die Breite der Spitzen in Abbildung 6.16 erfassen.

Auch die Aktivierungen selbst sind für die Musikstilerkennung wichtig. Es kann die durchschnittliche Aktivierung berechnet werden. Ebenso die maximale Aktivierung der Spezialisten und die Varianz der Aktivierung. Zusammenfassend können folgende Merkmale relevant sein:

1. Unigramm-Wahrscheinlichkeiten: $P(EreignisA)$
2. Bigramm-Wahrscheinlichkeiten: $P(EreignisB|EreignisA)$
3. Trigramm-Wahrscheinlichkeiten: $P(EreignisC|EreignisA, EreignisB)$
4. Dauer der Ereignisse
5. Mittelwert, Maximum und Varianz der Aktivierung

Selektion zeitlicher Abhängigkeiten Für die Musikstilerkennung ist sicherlich nicht die Berücksichtigung aller möglichen Ereignisübergänge erforderlich und möglich. So gibt es $12*11*11 = 1452$ Trigramm-Übergänge. Viele davon kommen in den Musikstücken nicht vor. Null-Wahrscheinlichkeiten sind auch ein Problem bei statistischen Sprachmodellen. Eingesetzt werden dann Rückfallstrategien auf Bigramm- und Unigramm-Wahrscheinlichkeiten in Kombination mit Interpolationsmechanismen. Dies ist hier nicht nötig.

Ereignisübergänge, die nur selten vorkommen, werden einfach ignoriert. Es findet eine Auswahl relevanter Merkmale statt. Ein Merkmal ist dabei relevant, wenn der Wert des Merkmals bei unterschiedlichen Musikstilen deutlich abweicht.

$$Relevanz(MerkmalA) = |\sum_{StilX} value(x, A) - \sum_{StilY} value(y, A)|$$

Eine genauere Bewertung der Merkmale kann erreicht werden, wenn auch die Varianz berücksichtigt wird. Bei der Merkmalsselektion werden derzeit

aber nur die Mittelwerte verwendet. Die Selektion bezieht sich dabei nicht nur die Trigramm-Wahrscheinlichkeiten, sondern auch auf die übrigen Merkmale. Damit ist es auch möglich, ungenutzte Spezialisten aufzuspüren. Enthält die verdeckte Schicht des Klangnetzes zu viele Neuronen, dann haben sich einige Neuronen auf keine Klangereignisse spezialisieren können und bringen nur Schwierigkeiten bei der Generalisierung. Durch die Selektion werden solche 'toten' Spezialisten eliminiert.

Die Selektion erfolgt für jede Merkmalsgattung getrennt, da die Relevanz-Bewertungen unterschiedlicher Merkmalsgattungen nicht vergleichbar sind. Verbesserungen durch Normieren des Relevanz-Maßes sind sicherlich möglich, werden aber derzeit nicht weiter verfolgt. Nach der Bewertung der Merkmale auf Relevanz werden dann die n besten Merkmale weiter verarbeitet. Die übrigen Merkmale werden ignoriert.

Systemstruktur Nun ist es an der Zeit, die gefundenen zeitlichen Abhängigkeiten in einen diskriminativen Lernprozeß einzubinden. Dies geschieht am einfachsten durch ein Neuronales Netz. Eingabe des Netzes sind die relevanten Merkmale der zeitlichen Abhängigkeiten. Gelernt wird die Erkennung des Musikstils. In Abbildung 6.17 ist die Systemstruktur zur Modellierung zeitlich abhängiger Phänomene schematisch wiedergegeben.

Der gesamte Lernprozeß gliedert sich also in folgende Schritte: Zunächst wird ein einfaches Klangnetz zur Musikstilerkennung trainiert. Die Ausgaben des Klangnetzes werden nicht weiter berücksichtigt. Die verdeckte Schicht des Netzes hat sich auf Klangereignisse spezialisiert. Aus den Aktivierungen der verdeckten Schicht werden dann zeitabhängige Merkmale berechnet. In einem Auswahlprozeß werden probleminvariante Merkmale entfernt. Die übrig gebliebenen Merkmale bilden die Eingabe für ein weiteres Neuronales Netz. Mit diesem Netz wird dann die Erkennung der Musikstile gelernt.

1. Spezialisierung auf Klangereignisse durch Lernen des Klangnetzes
2. Extraktion zeitabhängiger Merkmale
3. Selektion der relevanten Merkmale
4. Neuronales Netz zur Musikstilerkennung

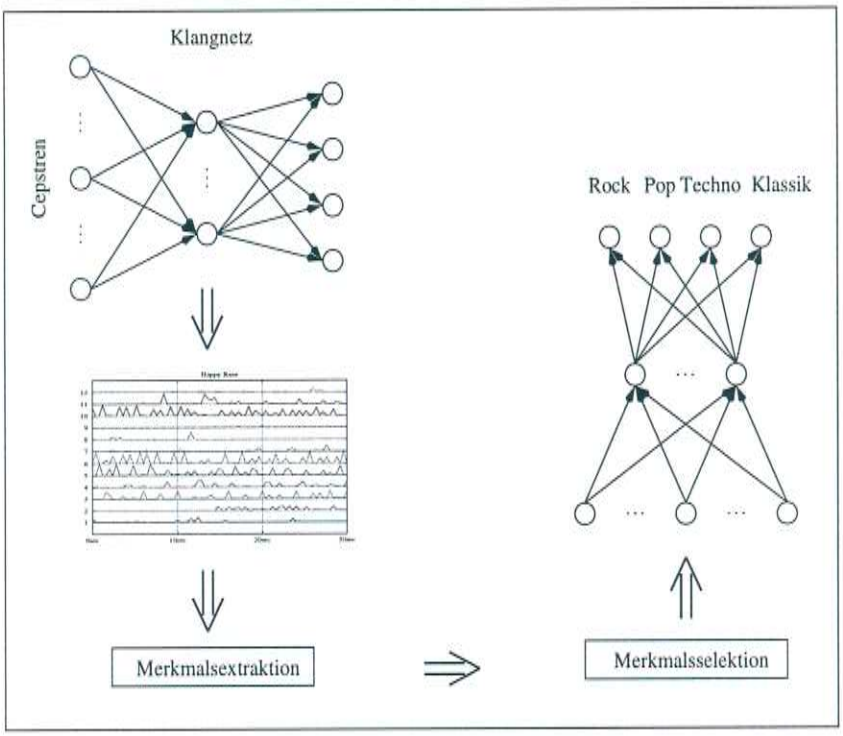

Abbildung 6.17: Systemstruktur zur Modellierung zeitlich abhängiger Phänomene bei der Musikstilerkennung

Diese Systemstruktur hat einige Vorteile. Es wird das Wisssen über die Art der Zeitabhängigkeit berücksichtigt. Das Netz muß nicht allein aus Kontextinformationen die Zeitabhängigkeiten lernen. Modelliert wird die Dauer von Ereignissen und Ereignisübergänge. Die Ereignisübergänge können auch mehrstufig sein. Die Berechnung von n-gramm Übergangswahrscheinlichkeiten macht dies möglich. Die Verarbeitung der Informationen ist so ausgelegt, daß jeweils Informationen gleichen Abstraktionsniveaus auf gleicher Stufe verarbeitet werden. Die Informationsverarbeitung ist konform zu dem Gruppenprinzip.

Experimente Nach der Einführung der Systemstruktur kommen wir nun zu den Experimenten. Welche der vorgeschlagenen Zeitabhängigkeiten sind für die Musikstilerkennung relevant? Können so die Unsicherheiten der Zeitinvarianzannahme ausgeglichen werden?

Übergänge			Dauer	Aktivierung			Anzahl der	Erkennungs-
Uni	Bi	Tri		Mittel	Maximum	Varianz	Merkmale	leistung
Merkmalsprüfung: mittlere Aktivierung								
12	26	0	12	0	7	0	57	84,7%
12	26	0	12	6	7	0	63	81,9%
12	26	0	12	12	7	0	69	80,5%
Merkmalsprüfung: maximale Aktivierung								
12	26	0	12	0	0	0	50	83,3%
12	26	0	12	0	12	0	62	84,7%
Merkmalsprüfung: Bigramm-Übergänge								
12	14	0	12	0	7	0	45	79,2%
12	29	0	12	0	7	0	60	83,3%
Merkmalsprüfung: Trigramm-Übergänge								
12	26	3	12	0	7	0	60	83,3%
12	26	7	12	0	7	0	64	**86,1%**
12	26	11	12	0	7	0	68	80,5%

Tabelle 6.10: Erkennungsleistung bei zeitabhängiger Modellierung der Klangereignisse

Die Ergebnisse der Untersuchungen sind in der Tabelle 6.10 zusammengefaßt dargestellt. Die Sprünge in der Parameteranzahl sind aufgrund des Selektionsprozesses diskontinuierlich. Bei den eindimensionalen Merkmalen (Unigramm, Dauer, Aktivierung) ist die Merkmalsanzahl durch die Anzahl der verdeckten Neuronen des Klangnetzes begrenzt. Es kann also maximal jeweils 12 dieser Merkmale geben. Die Gesamtanzahl der zur Klassifikaton verwendeten Merkmale ist in der vorletzten Spalte angegeben. In der letzten Spalte steht die resultierende Erkennungsleistung.

Diskutieren wir die Ergebnisse. Die Merkmale der mittleren Aktivierung sind anscheinend für die Musikstilerkennung nicht relevant. Ohne diese Merkmale ist die Erkennungsleistung 84,7%. Benutzt der Klassifikator alle mittleren Aktivierungen, dann sinkt die Erkennungsleistung auf 80,5%. Dagegen sind die maximalen Aktivierungen für die Musikstilerkennung von Vorteil. Es werden aber nicht die maximalen Aktivierungen aller verdeckten Neuronen gebraucht. Es reicht, die 7 besten dieser Merkmale zu nehmen.

Die nächsten Versuche galten den Übergangswahrscheinlichkeiten. Auch

dort werden nicht alle Bigramm-Wahrscheinlichkeiten gebraucht. Werden nur 14 Bigramme zugelassen, dann ist die Erkennungsleistung 79, 2%. Eine Erhöhung der Anzahl der Bigramme auf 26 führt zu einer Steigerung der Erkennungsleistung auf 84, 7%. Eine weitere Vergrößerung auf 29 Bigramme bewirkt dann eine Abnahme auf 83, 3%.

Auch die Verwendung von Trigrammen kann sinnvoll sein. Werden nur die ersten 3 Trigramme zusätzlich genommen, dann sinkt die Leistung jedoch auf 83, 3%. Durch die Berüchsichtigung der 7 relevantesten Trigramme kann die Erkennungsleistung auf 86, 1% gesteigert werden. Aber eine weitere Erhöhung der Trigramm-Anzahl auf 11 läßt die Erkennungsleistung überraschend stark auf 80, 5% sinken. Die hinzugenommen Trigramm-Merkmale verändern den Eingaberaum offenbar so, daß das Netz Schwierigkeiten bei der Spezialisierung auf relevante Eingaberegionen bekommt. Dies läßt darauf schließen, daß die Relevanz-Berechnung zu stark vereinfacht worden ist. Mit einer linearen Diskrimanzanalyse könnte man versuchen, die relevanten Merkmale besser herauszufiltern.

Vergleich zeitabhängige - zeitunabhängige Modellierung Das diskriminative Lernen zeitabhängiger Phänomene bringt eine Verbesserung der Erkennungsleistung auf 86, 1%. Ohne die Berücksichtigung der zeitabhängiger Phänomene ist die Erkennungsleistung des Neuronalen Netzes 81, 9%. Die vorgeschlagene Systemstruktur ist also in der Lage, die Unsicherheiten der Zeitinvarianzannahme auszugleichen. Zudem bleiben die Vorteile des diskriminativen Lernens erhalten. Die Erkennungsleistung der HMM's, die ja auch Zeitabhängigkeiten berücksichtigen, ist nur 79, 2%. Die zu Beginn des Abschnitts besprochene Tabelle 6.9 zur Motivation des diskriminativen Lernens zeitabhängiger Phänomene kann nun vervollständigt werden.

Modell	Zeitabhängigkeit	Lernverfahren	Erkennung
Neuronales Netz	aktuelles Ereignis	diskriminativ	81, 9%
Gaußklassifikator	aktuelles Ereignis	nicht diskriminativ	77, 8%
HMM	auf Ereignis A folgt B	nicht diskriminativ	79, 2%
ETM-NN[1]	auf Ereignis A folgt B	diskriminativ	**86, 1%**

Tabelle 6.11: Modellierung zeitabhängiger Phänomene, Teil 2

[1] Explict Time Modelling with Neural Networks

Auch der Rechenaufwand bleibt durch die hierarchische Informationsverarbeitung begrenzt. Das Verarbeiten jeweils gleichartiger Informationen auf einer Stufe zeigt seine Vorteile. Durch die Extraktion und Selektion zeitabhängiger Merkmale kann die Art der Zeitabhängigkeit auf explizite Weise berücksichtigt werden. Das Netz muß nicht allein aufgrund von Kontextinformationen die Art der Zeitabhängigkeit lernen.

6.4 Gesamtsystem

Das Gesamtsystem verknüpft die in den letzten Abschnitten untersuchten Einzelelemente. Berücksichtigt werden dabei die relevanten Merkmale Rhythmus und Klangfarbe. Die im Kapitel *Extraktion relevanter Merkmale* angesprochene Nulldurchgangsrate zur Modellierung der Melodie wird nicht genutzt. Der potentielle Nutzen dieses Merkmals liegt weniger in der Relevanz zur Musikstilerkennung sondern eher als Hilfsmerkmal zur besseren Verarbeitung der Merkmale Rhythmus und Klangfarbe. Änderungen in der Nulldurchgangsrate könnten zur Segmentierung des Signals dienen. Die untersuchten Ansätze zur Informationsverarbeitung benötigen dieses Wissen jedoch weniger.

Die Kombination von Rhythmus und Klangfarbe geschieht auf möglichst hohem Abstraktionsniveau. Ein Kombinationsnetz integriert die Teilentscheidungen der beiden Spezialnetze für Rhythmus und Klangfarbe zu einer Gesamtentscheidung. Ein solche Kombination ist bereits in dem Abschnitt *Neuronale Netze, Kombination* beschrieben worden. Das Klangnetz wird nun aber durch das *ETM-NN* ersetzt. Das diskriminative Lernen zeitabhängiger Phänomene berücksichtigt die Unsicherheiten der Zeitinvarianzannahme bei den Klangereignissen. Das Spezialnetz auf Rhythmusseite bleibt so, wie es im Abschnitt *Neuronale Netze, Rhythmus* beschrieben ist. Die Unsicherheiten bei der Zeitinvarianzannahme auf Rhythmusseite sind weniger kritisch.

Zu den Ergebnissen. Die Vorteile der Integration von Rhythmus und Klangfarbe sind beachtlich. Die Erkennungsleistung des Gesamtsystems liegt bei $88,9\%$. Die Vorteile der Integration von Rhythmus und Klangfarbe haben sich auch bereits bei Verwendung des einfacheren Klangnetzes ohne Modellierung expliziter Zeitabhängigkeiten gezeigt. Das im Abschnitt *Neuronale Netze, Kombination* beschriebene System kann $87,5\%$ der Stücke richtig erkennen, obwohl die Erkennungsleistung des zur Kombination verwendeten

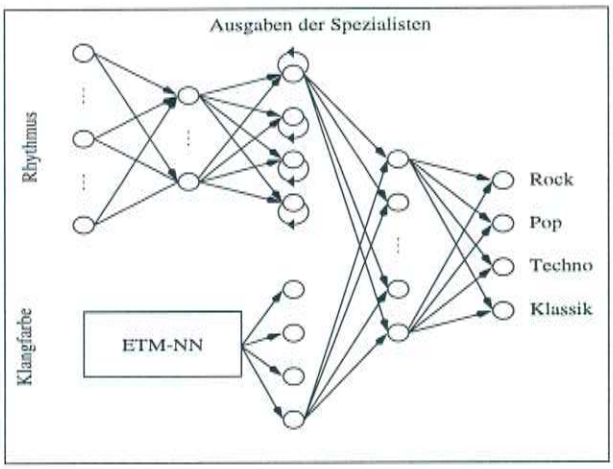

Abbildung 6.18: Gesamtsystem zur Musikstilerkennung

Klangnetzes bei 81,9% liegt. Tabelle 6.12 enthält einen Vergleich der Ergebnisse bei der Kombination mit unterschiedlichen Klangspezialisten. Die erste Zeile enthält die Angaben zu dem ursprünglichen System, welches im Abschnitt *Neuronale Netze, Kombination* vorgestellt wurde und keine explizite Modellierung zeitlicher Abhängigkeiten der Klangereignisse vornimmt. Die Systeme in den weiteren zwei Zeilen verwenden das *ETM-NN* zur Modellierung dieser Abhängigkeiten.

Rhythmusspezialist	Klangspezialist	Gesamtsystem
NN 73,6%	NN 81,9%	87,5%
NN 73,6%	ETM-NN 84,7%	**88,9%**
NN 73,6%	ETM-NN 86,1%	86,1%

Tabelle 6.12: Erkennungsleistung des Gesamtsystems

Es zeigt sich, daß bei zunehmender Verbesserung des Klangspezialisten die Integrationseffekte abnehmen. Ist die Erkennungsleistung des Klangspezialisten bereits allein 86,1%, so wird der Rhythmusanteil schlicht ignoriert. Die Netzwerkverbindungen des Rhythmusnetzes zu dem Kombinationsnetz sind dann kaum ausgeprägt. Die Erkennungsleistung des Gesamtsystems verbleibt dann bei 86,1%. Wird dagegen der etwas schlechtere Klangspezialist mit 84,7% genommen (ebenfalls *ETM-NN*), dann wird auch das Rhythmus-

netz zur Gesamtentscheidung berücksichtigt. Die Erkennungsleistung des Gesamtsystems ist dann 88,9%. Zur Kombination der Spezialisten ist es also erforderlich, daß die Leistungsfähigkeit beider Teilsysteme nicht allzu unterschiedlich ist.

In Abbildung 6.19 sind abschließend die Erkennungsleistungen einzelner Musikstile detailliert wiedergegeben. Techno und Klassik sind die Musikstile, die am besten erkannt werden. Aber auch die Erkennung von Rock und Pop ist gut. Kein Musikstil dominiert das Gesamtverhalten. Das Verhalten ist ausgewogen.

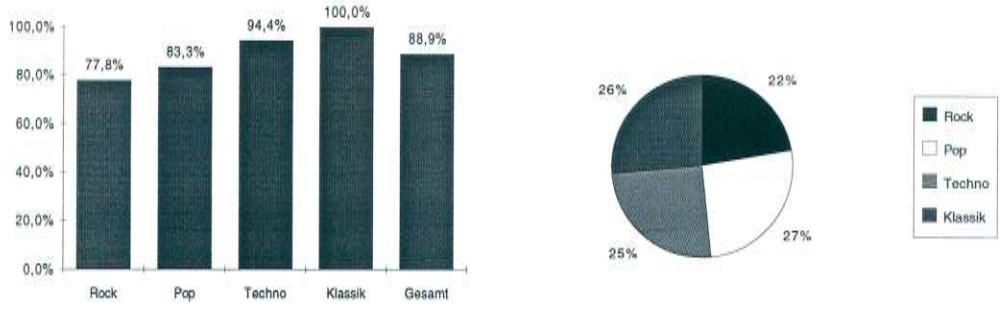

Abbildung 6.19: Erkennung und Dominanz des Gesamtsystems

Kapitel 7

Fazit

Was war das Ziel der Diplomarbeit? Zu Untersuchen war ein Teilaspekt der Wahrnehmung von Musik - die Erkennung der Stilrichtung. Ein Teilziel ist dabei das Finden einer angemessenen Repräsentation der Musik gewesen. Das zweite Teilziel bestand in der Entwicklung eines Systems, welches automatisch die Stilrichtung des Musikstückes erkennt. Voraussetzung dafür war eine Darstellung der Musikstücke, die nur relevante Eigenschaften enthält und in der die probleminvarianten Eigenschaften eliminiert sind. Zudem soll das System auch praktisch einsetzbar sein. Die maschinelle Erkennung soll in möglichst kurzer Zeit erfolgen und zuverlässig sein.

Beide Teilziele sind erreicht worden. Es gelang, eine kompakte Darstellung der Musikstücke zu finden. In den Experimenten konnte gezeigt werden, daß Rhythmus und Klangfarbe relevante Merkmale zur Erkennung des Musikstils sind. Zur Berechnung dieser Eigenschaften genügen zeitlich lokale Fenster. Die Informationsverarbeitung berücksichtigt dabei Unsicherheiten der Zeitinvarianzannahme durch die explizte Modellierung zeitlicher Abhängigkeiten. Das Gesamtsystem zur Erkennung der Musikstile arbeitet in $88,9\%$ aller Fälle (bezogen auf die verwendete Datenbasis) korrekt. Erkennung und Dominanz der Musikstile sind ausgeglichen. Es wird keine einzelne Musikrichtung bevorzugt. Die einzelnen Komponenten des Systems sind so ausgelegt, daß eine Entscheidung in Echtzeit möglich ist.

Zudem zeichnet sich das System durch seine Flexibilität aus. Ohne Änderungen der Gesamtstruktur sind einige Versuche zur Trennung von Musik und Sprache durchgeführt wurden. Mit dem Rhythmusnetz konnten in $94,4\%$ aller Fälle die korrekte Klasse gefunden werden. Das Klangnetz trennte die

Klassen Musik und Sprache in 97, 2% aller Fälle korrekt. Die Sprecher der Trainingsmenge und der Testmenge waren dabei disjunkt.

Das Einfachheits- und das Gruppenprinzip prägten die Vorgehensweise bei den Untersuchungen. Beide Prinzipien tragen zu einer hierarchischen Informationsverarbeitung bei und können nicht unabhängig voneinander betrachtet werden. Wichtig für den Erfolg ist auch eine angemessene Berücksichtigung des Problemwissens. Es darf sich nicht auf nur einzelne Formen des Wissens beschränkt werden. Bei der Suche nach einer geeigneten Musikrepräsentation ist die Beachtung von Expertenwissen unumgänglich. Gezeigt hat sich ebenfalls der Nutzen experimenteller Studien zur Wahrnehmung von Musik. Die Datenbasis mit einer Vielzahl von Beispielen hat es dann erst möglich gemacht, das richtige Verhalten der konnektionistischen Klassifikatoren zu erlernen.

Vertreter der klassischen KI argumentieren gelegentlich gegen die Informationsverarbeitung mit Neuronalen Netzen mit dem Hinweis, daß die erlernten Netzwerkverbindungen nicht verstanden werden und dies eher ein 'Black Box' Verhalten ist. Bezogen auf die Anwendung Neuronaler Netze in dieser Arbeit konnte das Gegenteil nachgewiesen werden. Exemplarisch wurde dies bei dem Rhythmusnetz durchgeführt. Resultat dieser Analyse ist eine Übersetzung der Netzwerkverbindungen in eine Gruppe von Regeln.

Was bleibt als Ausblick? Gelöst wurde ein Teilaspekt bei der Wahrnehmung von Musik. Die Erkennung akustischer Ereignisse umfaßt aber ein weitaus größeres Spektrum. Dort wären interessante Nachfolgearbeiten denkbar. Das vorgestellte diskriminative Lernverfahren *ETM-NN* zur Modellierung zeitlich abhängiger Phänomene ist dabei sicherlich ein geeigneter Ansatzpunkt. Eine wettbewerbsorientierte Spezialisierung auf relevante Ereignisse könnte die Eigenschaften des Lernverfahrens dabei noch verbessern.

Denkbar wäre auch die Entwicklung eines adaptiven Systems, welches die von Hörern bevorzugten Musikrichtungen erkennt. Dazu kann auf dieses System aufgesetzt werden. Die Aktivierungen der Neuronalen Netze für die Musikstilerkennung können als eine höhere Repräsentation der Stücke aufgefaßt werden. Dies gilt insbesondere für die Ausgabeschicht des Kombinationsnetzes, dessen Aktivierungen a-posteriori-Wahrscheinlichkeiten approximieren. Diese Neuronenaktivierungen können die Eingabe für ein induktives Lernverfahren bilden.

Literaturverzeichnis

[1] K. Bös: *Statistikkurs 1, Einführung in die statistischen Auswertungsmethoden für Sportstudenten, Sportlehrer und Trainer*, Czwalina 1986

[2] G. Bamberg, F. Baur: *Statistik*, Oldenbourg 1996

[3] J.C. Brown: *Calculation of a constant Q spectral transform*, Journal of the Acoustical Society of America, Vol 89, No. 1

[4] J.C. Brown: *Musical fundamental frequency tracking using a pattern recognition method*, Journal of the Acoustical Society of America, Vol 92, No. 3

[5] P. Coli, G. de Poli, G. Lauzzana: *Auditory modelling and self-organizing neural networks for timbre classification*, Journal of New Music Research, Vol. 23, 1994

[6] S. Dixon: *Multiphonic Note Identification*, Proccedings of the 19th Australasian Computer Science Conference, Melborne

[7] Hecht-Nielsen: *Neurocomputing*, Addison-Wesley 1990

[8] J. Hertz, A. Krough, R. Palmer: *Introduction to the theory of neural computation*, Addison-Wesley 1991

[9] G. Klemm: *Untersuchungen über den Zusammenhang musikalischer und sprachlicher Wahrnehmungsfähigkeiten*, Lang 1996

[10] P. Lindsay, D. Norman: *Einführung in die Psychologie, Informationsaufnahme und -verarbeitung beim Menschen*, Springer 1981

[11] T. Lee, A.Bell, R. Lambert: *Blind Separation of delayed and convolved sources*, NIPS 1996

[12] M. Leman: *Music and Schema Theory, Cognitive Foundations of systematic musicology*, Springer 1995

[13] M. Minsky, S. Papert: *Perceptrons*, MIT Press 1988

[14] H. de la Motte-Haber: *Handbuch der Musikpsychologie*, Laaber-Verlag 1996

[15] John R. Pierce: *Klang, Musik mit den Ohren der Physik*, Spektrum der Wiss., 1985

[16] Philip Quinlan: *Connectionism and psychology, a psychological perspective on new connectionistic research*, Harvester Wheatsheaf 1991

[17] L. Rabiner, R. Schafer: *Digital processing of speech signals*, Prentice-Hall, 1978

[18] S. Russell, P. Norvig: *Artificial intelligence: a modern approach*, Prentice Hall 1995

[19] J. Saunders: *Real-Time Discrimination of Broadcast Speech-Music*, ICASSP 1996

[20] Schukat-Talamazzini: *Theorie der Spracherkennung* Vieweg 1995

[21] K. Spence, G. Swayne: *Das grosse Buch der Musik*, Herder, 1984

[22] Schoenebeck, Reiss, Noll: *Musiklexikon für junge Erwachsene*, Cornelsen, 1994

[23] A. Tanguiane *Artificial perception and music recognition*, Springer 1993

[24] P. Toiviainen, M. Kaipainen, J. Louhivuori: *Musical timbre: Similarity ratings correlate with computational feature space distance*, Journal Of New Music Research, Vol 24, 1995

[25] V. Vapnik *The nature of statistical learning theory*, Springer 1995

[26] V. Vapnik *Estimation of Dependences based on emprical data*, Springer 1982

[27] Alex Waibel et al. *JANUS-II — Translation of Spontaneous Conversational Speech* in Proc. ICSLP-96 pp. 409 ff, Atlanta, Georgia 1996

[28] van der Waerden: *Mathematische Statistik*, Springer 1971

[29] Yost, Popper, Fay: *Springer Handbook of auditory Research, Psychophysics*, Springer 1993