

Johann-Markus Batke

Untersuchung von Melodiesuchsystemen
sowie von Verfahren zu ihrer
Funktionsprüfung

Untersuchung von Melodiesuchsystemen sowie von Verfahren zu ihrer Funktionsprüfung

vorgelegt von
Diplom-Ingenieur
Johann-Markus Batke
aus Mülheim an der Ruhr

Von der Fakultät IV – Elektrotechnik und Informatik
der Technischen Universität Berlin
zur Erlangung des akademischen Grades
Doktor der Ingenieurwissenschaften
– Dr.-Ing. –
genehmigte Dissertation

Berlin 2006
D 83

Bibliografische Information der Deutschen Bibliothek

Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.ddb.de> abrufbar.

1. Aufl. – Göttingen : Cuvillier, 2006

Zugl.: Berlin, Technische Universität Berlin, Diss., 2006

ISBN 10: 3-86727-085-6

ISBN 13: 978-3-86727-085-4

© CUVILLIER VERLAG, Göttingen 2006

Nonnenstieg 8, 37075 Göttingen

Telefon: (05 51) 5 47 24-0

Telefax: (05 51) 5 47 24-21

www.cuvillier.de

Alle Rechte vorbehalten. Ohne ausdrückliche Genehmigung des Verlags ist es nicht gestattet, das Buch oder Teile daraus auf fotomechanischem Weg (Fotokopie, Mikrokopie) zu vervielfältigen.

1. Auflage, 2006

Gedruckt auf säurefreiem Papier

ISBN 10: 3-86727-085-6

ISBN 13: 978-3-86727-085-4

Promotionsausschuss

Vorsitzender: Prof. Dr.-Ing. Reinhold Orglmeister

1. Bericht: Prof. Dr.-Ing. Thomas Sikora

2. Bericht: Prof. Dr.-Ing. Peter Noll

3. Bericht: Prof. Dr.-Ing. Wolfgang Hess

Tag der Einreichung: 7.4.2006

Tag der wissenschaftlichen Aussprache: 26.9.2006

Inhaltsverzeichnis

1	Einleitung	1
1.1	Aufgabenstellung	1
1.2	Fachgebiete	4
1.3	Wirtschaftliche Bedeutung	5
1.4	Kapitelübersicht	5
2	Musik und Melodie	9
2.1	Der Melodiebegriff	9
2.2	Begriffe der Musiktheorie	11
2.2.1	Akustik	11
2.2.2	Notenschrift	12
2.2.3	Klavierwalze	16
2.2.4	Akkordsymbolschrift	17
2.2.5	Kammerton a	18
2.2.6	Intervall	19
2.2.7	Tonsystem und Skalen	20
2.2.8	Tonart	21
2.2.9	Cent-System	21
2.2.10	Temperaturen	22
2.2.11	Intonation	23
2.2.12	Monophonie und Polyphonie	23
2.3	Musikrepräsentation	24
2.3.1	Noten	24
2.3.2	MIDI	25
2.3.3	Melodiekontur	27
2.4	Singen von Melodien	29
2.5	Zusammenfassung	31
3	Musiksuchmaschinen	33
3.1	Beispiele für Musiksuchsysteme	34
3.1.1	Musicline	34

3.1.2	notify!	34
3.1.3	Musipedia	35
3.1.4	Vodafone-MusicFinder	35
3.1.5	Weitere Systeme	37
3.1.6	Merkmale	37
3.2	Zielbestimmung	39
3.2.1	Musskriterien	40
3.2.2	Wunschkriterien	41
3.2.3	Abgrenzungskriterien	42
3.3	Einsatz von QBH-Systemen	42
3.3.1	Anwendungsbereich	42
3.3.2	Zielgruppen	42
3.3.3	Betriebsbedingungen	43
3.4	Umgebung eines QBH-Systems	44
3.4.1	Software	44
3.4.2	Hardware	44
3.4.3	Orgware	44
3.4.4	Schnittstellen	44
3.5	Funktionen	45
3.5.1	Benutzeroberfläche	46
3.6	Zusammenfassung	46
4	Multimedia-Standards	49
4.1	MPEG-7	49
4.1.1	Anwendungsbereiche	50
4.1.2	Gliederung des Standards	53
4.1.3	Auditive Inhaltsbeschreibung (Part 4)	56
4.2	MPEG-21	62
4.3	SMIL	63
4.4	Zusammenfassung	63
5	Monophone Transkription	65
5.1	Die Transkriptionsaufgabe	66
5.2	Tonhöherkennung	67
5.2.1	Verfahren der Kurzzeitanalyse	69
5.2.2	Zeitbereichsverfahren	74
5.2.3	Diskussion	75
5.3	Rhythmuserkennung	78

5.4	Eigene Untersuchungen	79
5.4.1	Tonhöhererkennung	79
5.4.2	Rhythmuserkennung	87
5.4.3	Praktische Versuche	89
5.5	Zusammenfassung	95
6	Polyphone Transkription	97
6.1	Melodie-Transkription für Query-by-Humming-Systeme	99
6.1.1	MIDI-Transkription	99
6.1.2	Manuelle Transkription	100
6.1.3	Transkription aus Audiosignalen	100
6.2	Automatische Transkription von Musiksignalen	102
6.2.1	Tonhöhererkennung	102
6.2.2	Rhythmuserkennung	105
6.2.3	Melodieerkennung	106
6.3	Eigene Untersuchungen	108
6.3.1	Übersicht	108
6.3.2	Filterbank	109
6.3.3	Momentanfrequenzen	112
6.3.4	Trennung von Melodie- und Bassbereich	115
6.3.5	Harmonische Analyse	115
6.3.6	Melodieagenten	118
6.3.7	Praktische Versuche und Evaluierung	119
6.4	Zusammenfassung	123
7	Melodievergleich	125
7.1	Datenbanken	126
7.1.1	Grundlagen und Begriffe	126
7.1.2	Suchmethoden	127
7.2	Zeichenkettensuche	129
7.2.1	Längste gemeinsame Teilsequenz (LCE)	130
7.2.2	Globaler Abgleich	133
7.2.3	Lokaler Abgleich (LAL)	133
7.2.4	Längste gemeinsame Zeichenkette (LCT)	135
7.2.5	Ähnlichkeitsberechnung	135
7.3	Indizierung	135
7.3.1	Koordinatenvergleich (CM)	138
7.3.2	Summe der Häufigkeiten (SF)	140

7.3.3	Ukkonen-Messung (UK)	140
7.4	Spezielle Verfahren	140
7.4.1	TPBM I	140
7.4.2	Direkte Messung	142
7.5	Anwendung in Melodiesuchsystemen	143
7.5.1	Diskussion der Verfahren	143
7.5.2	Einfluss der Melodielänge	144
7.5.3	Implementierung	144
7.6	Zusammenfassung	145
8	Melodiedatenbanken	147
8.1	Bewertung von Musiksuchsystemen	147
8.1.1	Relevanz	148
8.1.2	Vollständigkeit und Präzision	149
8.2	Statistik der Melodiedatenbank	150
8.2.1	Referenzdateien	150
8.2.2	Parameter der Melodiedatenbank	153
8.2.3	Die Bedeutung statistischer Parameter der Melodiedatenbank für die Suche	157
8.3	Melodievergleich	168
8.3.1	Indizierung	168
8.3.2	Zeichenkettensuche	174
8.4	Anfragefehler	176
8.4.1	Anfragelänge	176
8.4.2	Melodiefehler	177
8.4.3	Feldversuch	184
8.5	Zusammenfassung	186
9	Zusammenfassung und Ausblick	189
9.1	Zusammenfassung	189
9.2	Ausblick	194
A	Titel der Melodiedatenbank	197
	Lebenslauf Johann-Markus Batke	219
	Nachwort	221

Abbildungsverzeichnis

1.1	Anwendungsszenario eines QBH-Systems	2
2.1	Beispiel Notenschrift – BACH: „Partita in A-Moll“	13
2.2	Klavierwalze und MIDI-Notendarstellung.	19
2.3	Übersicht Repräsentationsformen für Musik	25
2.4	Noten von „As Time Goes By“ mit verschiedenen Konturdarstellungen	28
3.1	Benutzerschnittstelle der Internetseite Musipedia	36
3.2	Schematische Darstellung eines QBH-Systems	40
3.3	Anwendungsfall-Diagramm für die Nutzung eines QBH-System	45
3.4	Sequenz-Diagramm für die Prozesse eines QBH-Systems	46
3.5	Die Nutzerschnittstelle des QBH-Systems <i>Queryhammer</i>	47
4.1	Fokus des Standards MPEG-7	50
4.2	Abstrakte Darstellung möglicher Anwendungen von MPEG-7	51
4.3	Die Kernelemente des MPEG-7-Standards: D, DS und DDL	52
4.4	Übersicht über den Konformitätstest von Deskriptoren	54
4.5	Übersicht über die Konformitätsprüfung von Anwendungen	55
4.6	Datenstruktur des MPEG-7 <i>MelodyType</i>	56
4.7	Datenstruktur des MPEG-7 <i>MeterType</i>	57
4.8	Datenstruktur des MPEG-7 <i>ScaleType</i>	58
4.9	Datenstruktur des MPEG-7 <i>KeyType</i>	59
4.10	Datenstruktur des MPEG-7 <i>MelodyContourType</i>	60
5.1	Monophone Transkriptionsstufe in <i>Queryhammer</i>	66
5.2	Verarbeitungsschritte der Grundfrequenzerkennung	68
5.3	Zeitsignal eines gesungenen „na“	70
5.4	Übersicht von Grundfrequenzanalyseverfahren	71
5.5	AKF des Signals aus Abbildung 5.3	72
5.6	AMDF des Signals aus Abbildung 5.3	73

5.7	LDS des Signals aus Abbildung 5.3	73
5.8	Frequenzauflösung der AKF-Methode vs. chromatische Tonleiter	77
5.9	Detaillierte Darstellung der monophonen Transkriptionstufe in <i>Queryhammer</i>	80
5.10	Fensterung und AKF-Berechnung im Verfahren nach BOERSMA .	82
5.11	Noten, Zeit- und Grundfrequenzverlauf einer gesungenen An- frage	84
5.12	Nachverarbeitung des Grundfrequenzverlaufs	86
5.13	Beispiel zur Rhythmuserkennung	90
5.14	Testszenarien zur Bewertung von monophonen Transkriptions- systemen	92
5.15	Fehlerquoten bei monophoner Transkription	94
6.1	Polyphone Transkriptionsstufe in <i>Queryhammer</i>	98
6.2	Blockschaltbild des Transkriptionsverfahrens <i>PreFEst</i>	109
6.3	<i>PreFEst</i> Filterbank	111
6.4	<i>PreFEst</i> Bandpassfilter	115
6.5	<i>PreFEst</i> Melodieagenten	118
6.6	Wahrscheinlichkeiten der Grundfrequenz einer Melodie	120
7.1	Vergleichsstufe in <i>Queryhammer</i>	125
7.2	Übersicht verschiedener Distanzmaße	129
7.3	Noten und MPEG-7-MelodyContour zum Kinderlied „Hänschen klein“	131
8.1	Noten der untersuchten „Top-10-Single-Charts“	151
8.1	Noten der untersuchten „Top-10-Single-Charts (Fts.)“	152
8.2	Statistische Kenngrößen der MIDI-Datenbank	154
8.3	Die Zustandsübergangsmatrix für die Konturwerte der MIDI- Datenbank.	156
8.4	Vollständigkeit vs. Datenbankumfang (MIDI)	160
8.5	Vollständigkeit vs. Datenbankumfang (Zufallsmelodien)	162
8.6	Vollständigkeit vs. Datenbankumfang (Markov-Melodien)	163
8.7	Vollständigkeit V_1 vs. Datenbankumfang (MIDI, Zufall, Markov)	165
8.8	Vollständigkeit V_3 vs. Datenbankumfang (MIDI, Zufall, Markov)	166
8.9	Vollständigkeit V_{10} vs. Datenbankumfang (MIDI, Zufall, Markov)	167
8.10	Vollständigkeit vs. N-Gramm-Länge	170
8.11	Vollständigkeit vs. N-Gramm-Länge, gemittelt	171

8.12	Vollständigkeit vs. Anfragelänge für verschiedene Melodiefehler	172
8.13	Vollständigkeit vs. Anfragelänge für verschiedene Melodiefehler, gemittelt	173
8.14	Vollständigkeit vs. Edierkosten, Zeichenkettensuche	175
8.15	Vollständigkeit vs. Anfragelänge, Zeichenkettensuche, mit Melodiefehlern	175
8.16	Vollständigkeit vs. Anfragelänge, Zeichenkettensuche	177
8.17	Vollständigkeit vs. Auslassungsfehler	179
8.18	Vollständigkeit vs. Einfügingsfehler	180
8.19	Vollständigkeit vs. Edierfehler	181
8.20	Vollständigkeit vs. Fehler, Mittelwert über Distanzmaße	182
8.21	Vollständigkeit vs. Fehler, Mittelwert über Fehlerkategorien . . .	183
8.22	Vollständigkeit vs. Distanzmaß, für MIDI-Anfragen und Probanden	185
8.23	Vollständigkeit vs. Distanzmaß, Probanden einzeln	187

Tabellenverzeichnis

2.1	Bezeichnung der Oktaven in der Musik	14
2.2	Tempobezeichnungen in verschiedenen Sprachen.	17
2.3	Tempobezeichnungen und entsprechende M.M.-Werte	18
2.4	MIDI Notenummern	26
2.5	Konturwerte des PARSONS-Codes	27
2.6	Konturwerte der MPEG-7 <i>MelodyContour</i>	29
4.1	Intervallzuordnung der MPEG-7- <i>MelodyContour</i> in Cent	61
5.1	Einfügungsfehler bei gesummten Einzelnoten	95
6.1	Ergebnisse des „Melody Transcription Contest“ ISMIR 2004 . . .	122
7.1	Beispiel Kostenmatrix zur Ermittlung der längsten gemeinsa- men Teilsequenz	132
7.2	Beispiel Kostenmatrix zur Ermittlung des lokalen Abgleichs . . .	134
7.3	Beispiel Kostenmatrix zur Ermittlung der längsten gemeinsa- men Zeichenkette	136
7.4	Trigramme zum Beispiel „Hänschen klein“	139
8.1	Titel und Interpreten der deutschen „Top-10-Single-Charts“ . . .	152
8.2	Edierkosten für die Distanzmaße der dynamischen Program- mierung	174

Symbolverzeichnis

2rt	Index Normierung 2. Wurzel	144
9rt	Index Normierung 9. Wurzel	144
AKF	Autokorrelationsfunktion	71
AMDF	average magnitude difference function, Betragsdifferenzfunktion	72
AV	audiovisuell	50
BPM	Beats per Minute, Schläge pro Minute	15
CM	coordinate matching	138
D	Deskriptoren	52
DDL	Description Definition Language	52
DFT	diskrete Fouriertransformation	76
DI	Digital Item	62
DM	direct measure	142
DP	dynamische Programmierung	129
DS	Description Scheme	52
FFT	fast Fourier transform	76
GFA	Grundfrequenzanalyse	67
HTML	Hypertext-Markup-Language	49
HTTP	Hypertext-Transfer-Protokoll	49
LAL	local alignment	133
LCE	<i>longest common subsequence</i>	130
LCT	<i>longest common substring</i>	135
LDS	Leistungsdichtespektrum	72
len	Index Normierung logarithmisch	144
len	Normierung Länge	144

M.M.	Metronom Mälzel, Schläge pro Minute	15
MDS	Multimedia Description Schemes	54
MPEG	Moving Pictures Experts Group	4
non	Index ‘keine Normierung’	144
QBH	Query by Humming	1
SF	sum of frequencies	140
SMIL	Synchronized Multimedia Integration Language	63
TPBM	time pitch beat matching	140
UK	Ukkonen-Messung	140
WWW	World-Wide-Web	49
XM	eXperimentation Model	54
ZCR	zero crossing rate	74
$\lambda(t)$	Momentanfrequenz	112
A	Kostenmatrix	130
p	Vektor mit Konturwerten einer Melodie	130
q	Vektor mit Konturwerten einer Anfrage	130
$\mathbf{x}_N(k)$	diskretes Zeitsignal, Block der Länge N	70
$A(d)$	Betragsdifferenzfunktion	72
f	Frequenz	22
F_0	Grundfrequenz	71
f_s	Abtastrate	76
k	Zeitindex	70
m	Frequenzabweichung in Cent	22
N	Blocklänge	70
$P(q, D)$	Präzision	149
$R(d)$	Autokorrelationsfunktion	71
$S(k)$	Leistungsdichtespektrum	73
T_0	Grundperiodendauer	72
$V(q, D)$	Vollständigkeit	149

$w(n)$	Fensterfunktion	81
$X(k, n)$	Kurzzeitfouriertransformierte	111
$x(t)$	reellwertiges Zeitsignal	112
$x_a(t)$	analytische Zeitsignal	112
$\hat{x}(t)$	Hilbert-Transformierte	112

Durch die Flut von digitalen Musikstücken im Internet und auf Datenspeichern gewinnt die Suche nach einzelnen Musikstücken gegenwärtig stark an Bedeutung. So ermöglicht etwa ein mobiles Wiedergabegerät wie der *Apple iPod* ohne weiteres die Speicherung von 5.000 Titeln [4]. Bislang können einzelne Musikstücke aus einer solchen Datenmenge nur durch die Eingabe von Titel oder Interpret ausgewählt werden. Ein typisches Problem ist es aber, dass dem Musiksuchenden diese Informationen nicht gegenwärtig sind, sondern er sich lediglich an die Melodie erinnern kann.

Technisch gesehen handelt es sich beim Auffinden einzelner Musikstücke aus einer großen Menge von Titeln um die Suche in einer Musikdatenbank. Wenn dem Suchenden nur die Melodie bekannt ist, muss folglich die Eingabe der Melodie in das Suchsystem möglich sein, um eine wie oben beschriebene Suchanfrage bedienen zu können. Diese Anforderung wird von einem Query-by-Humming-System (QBH-System) erfüllt: Die Melodie kann dem System vorgesummt werden und eine Anzahl ähnlicher Melodien der Musikdatenbank werden als Suchergebnis präsentiert. In Abbildung 1.1 ist das Benutzungsszenario eines solchen QBH-Systems dargestellt.

1.1 Aufgabenstellung

Für den Erfolg der Suche nach Melodien ist es von entscheidender Bedeutung, wie die Suchanfrage und der Datenbankbestand miteinander verglichen werden. Bei den meisten QBH-Systemen wird eine *symbolische* Melodiedarstellung verwendet, in welcher die Melodien der Musikstücke in der Datenbank gespeichert werden. Eine solche symbolische Darstellung kann zum Beispiel der PARSONS-Code sein, der den Verlauf der Melodie mit nur drei Buchstaben beschreibt: „U“ für aufwärts (up), „D“ für abwärts (down) und „R“ für gleichbleibend (repeat) [154]. Eine Darstellung wie der PARSONS-Code wird auch als *Melodiekontur* bezeichnet. Wird die Suchanfrage an das QBH-System in eine ebensolche Melodiekontur umgewandelt, kann sie mit dem Datenbankbestand verglichen werden. Ergebnis dieses Vergleichs ist schließlich eine Liste

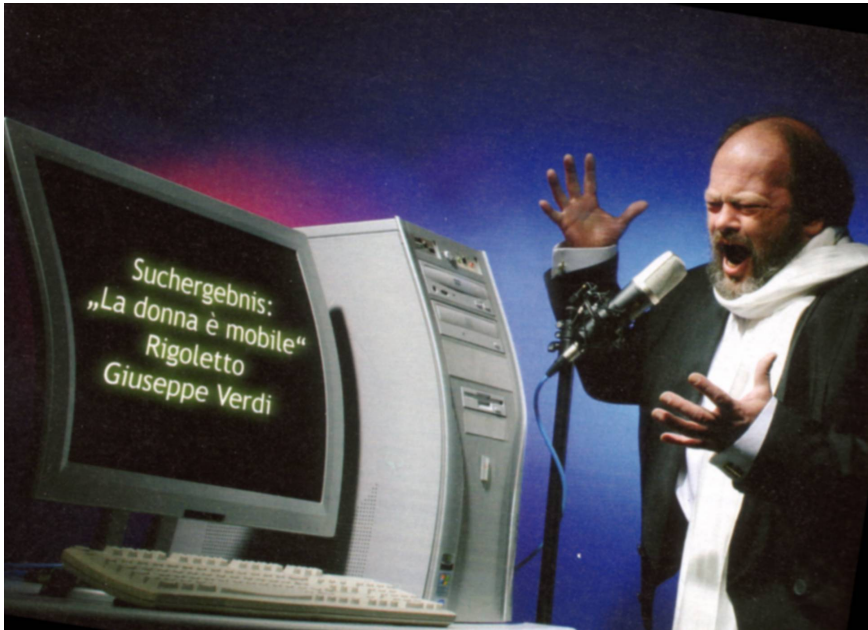


Abbildung 1.1: Ein Query-by-Humming-System in Anwendung: Der musikalische Vortrag des Nutzers wird vom QBH-System verarbeitet und die ähnlichste Melodie als Ergebnis einer Datenbanksuche angegeben. Quelle: [117]

mit den ähnlichsten Musikstücken. Die Musikstücke selbst können über Titel und Interpret, Noten oder als Audiodatei in der Datenbank gespeichert sein und dem Nutzer angeboten werden.

Für den Betrieb eines QBH-Systems sind damit im Wesentlichen zwei Aufgaben zu erfüllen: zuerst muss die gesummte Suchanfrage in eine symbolische Darstellung wie oben beschrieben gebracht werden. Dieser Vorgang ist die *Transkription* eines *monophonen* Gesangssignals in eine symbolische Darstellung. Die zweite Aufgabe besteht im *Vergleich* der erhaltenen Symbole mit dem Inhalt der Melodiedatenbank, um die ähnlichsten Melodien zu finden. Diese Aufgabe macht die Verwendung geeigneter *Ähnlichkeitsmaße* notwendig.

Bei beiden Aufgaben handelt es sich um komplexe Vorgänge. So ist die Transkription von gesummen Signalen an Signalverarbeitungsschritte wie

eine Grundfrequenzanalyse und eine Rhythmuserkennung gebunden. Ähnlich wie zum Beispiel bei der Aufgabenstellung der automatischen Spracherkennung ist vieles, was für den Menschen leicht zu interpretieren ist, technisch nur durch aufwendige Verfahren zu analysieren. Der Vergleich von Melodien wird auf symbolischer Ebene durchgeführt. Die Darstellung der Melodie durch Symbole als Melodiekontur bedingt eine gewisse Ungenauigkeit, aber auch Verallgemeinerung der Repräsentation und führt daher auch zu einer Vergrößerung der Treffermenge. Diese größere Treffermenge wird zu Gunsten einer möglichst großen Freiheit beim Stellen der Suchanfrage bewusst in Kauf genommen. Beide Aufgaben, Melodietranskription und -vergleich, sind Gegenstand zahlreicher Untersuchungen in der Literatur [37, 41, 50, 69, 76, 88, 90, 101, 124, 132, 142, 153, 161, 165, 175, 198]. Bei allen Untersuchungen werden unterschiedliche Datenbestände und Bewertungskriterien herangezogen, damit ist ein Vergleich der Untersuchungen schwierig oder nicht sinnvoll. Ein Ziel dieser Arbeit ist es daher, möglichst flexibel einsetzbare Verfahren zur Bewertung von QBH-Systemen zu erarbeiten.

Um ein QBH-System anzubieten, ist der Aufbau einer Melodiedatenbank eine wichtige zu erfüllende Aufgabe. Audiodateien sind für Musikdatenbanken das am meisten verwendete Medium und liegen in großer Anzahl als Wellenform- oder MP3-Datei vor. So bietet zum Beispiel allein das Internetportal *mp3.de* über 150.000 frei erhältliche Titel an [5]. Beim kommerziellen Anbieter *Musicline* werden online über 1,8 Millionen Titel angeboten [11], über das Internetportal *iTunes* mehr als 2 Millionen Titel [4]. Das Portal *mp3.com* schließlich gibt 6 Millionen Titel an [2]. Um eine Melodiedatenbank aufzubauen, müssen symbolische Beschreibungen für die gespeicherten Musikstücke gefunden werden. Bei der Menge der verfügbaren Titel ist klar, dass dies auf automatischem Wege erfolgen sollte. Da es sich im Gegensatz zu gesummen Anfragen bei Musik meist um polyphone, d. h. mehrstimmige Signale handelt, muss zur Melodietranskription aus solchen Audiosignalen zusätzlich noch die richtige, also melodieführende Stimme ausgewählt werden.

Diese Aufgabe ist sehr schwierig und wird derzeit noch nicht beherrscht. Daher greifen bestehende QBH-Systeme meistens auf MIDI-Datenbanken zurück, bei denen sich die Melodie technisch einfach extrahieren lässt. MIDI steht für „music instrument digital interface“ – dieses Format enthält symbolische Informationen darüber, wie Noten auf einem elektronischen Musikinstrument gespielt werden. Daneben werden auch oft Melodien durch Musiker transkribiert und manuell in Melodiedatenbanken eingetragen. In jüngster Zeit hat es im Bereich der Forschung zur automatischen Transkription von polyphoner

Musik erhebliche Fortschritte gegeben [87,111]. Im Rahmen dieser Arbeit werden daher aktuelle Verfahren auf ihre Eignung für QBH-Systeme untersucht und bewertet.

Melodiedatenbanken und ihre Eigenschaften bestimmen in einem Suchsystem das Ergebnis und sind von zentraler Wichtigkeit für QBH-Systeme. Über die Eigenschaften von Melodiedatenbanken, die sich über statistische Methoden herleiten lassen, sind bislang keine Untersuchungen veröffentlicht worden. In der vorliegenden Arbeit sollen Melodiedatenbanken unter diesem Aspekt untersucht werden.

Der große Anteil existierender QBH-Systeme verwendet den oben beschriebenen PARSONS-Code zur symbolischen Darstellung von Melodien. Jüngere Arbeiten der *Moving Pictures Experts Group* (MPEG) bieten aber weitere, standardisierte Möglichkeiten zur Melodiebeschreibung im Standard MPEG-7. Darüberhinaus werden dort aber auch Definitionen angeboten, die zur Beschreibung von inhaltsbezogenen, abstrakten Daten geeignet sind. Damit lassen sich Systemschnittstellen wie die von QBH-Systemen repräsentieren. Im Rahmen dieser Arbeit wird die Anwendung des Multimedia-Standards MPEG-7 für QBH-Systeme ausführlich untersucht und diskutiert.

1.2 Fachgebiete

Die in einem QBH-System genutzten Technologien berühren ganz unterschiedliche Disziplinen und Fachgebiete. In Hinblick auf die Verarbeitung gesummer Melodien bzw. der Musiksignale für die Melodiedatenbank steht man vor Aufgaben der *Audiosignalverarbeitung*. Der dabei kardinale Begriff der Melodie und was sich damit verbindet, wird in den *Musikwissenschaften* beschrieben. Der Vergleich von symbolischen Melodiedarstellungen wird mit Methoden der *Informatik* vollzogen. Die Gewinnung von Daten aus Musiksignalen kann in diesem Zusammenhang den Disziplinen *Maschinenhören* und *Data-Mining* zugeordnet werden. Die Einbettung eines QBH-Systems in ein Netzwerk gehört gleichermaßen in das Aufgabengebiet der Informatik wie der *Elektrotechnik*. Es handelt es sich somit um eine *Multimedia-Anwendung*, die beispielsweise für *digitale Bibliotheken* genutzt werden kann.

Die Betrachtung von QBH-Systemen bringt also eine Fülle verschiedener Themen und Fachgebiete mit sich. Im Rahmen dieser Arbeit soll eine Darstellung der wichtigsten Grundlagen der genannten Fachgebiete erfolgen, um eine fundierte Bewertung von QBH-Systemen zu ermöglichen.

1.3 Wirtschaftliche Bedeutung

Nach dem Aufkommen von illegalen Musik-Tauschbörsen im Internet wie zum Beispiel „Napster“ um 1999 hat die Musikindustrie mittlerweile reagiert und bietet selbst Musik zum kostenpflichtigen Download im Internet an. Der große Ansturm auf solche Angebote steht Pressestimmen zufolge unmittelbar bevor [148]. Für 2005 prognostizierte der Bundesverband der phonographischen Wirtschaft 20 Millionen legale Downloads in Deutschland, nach acht Millionen im Vorjahr.

Feststellbar ist, dass der klassische Handel mit Tonträgern sich in Richtung Internethandel mit Dateien verschiebt. Die Firma *Apple* gibt für ihr Portal *iTunes* 35.000 verkaufte Titel pro Tag an [3], die Konkurrenz, u. a. auch *Microsoft*, folgt mit eigenen Angeboten. Auch unter diesem Gesichtspunkt kommt QBH-Systemen eine große Bedeutung zu.

1.4 Kapitelübersicht

Im Rahmen dieser Arbeit werden alle Teile eines QBH-Systems und Verfahren zu ihrer Funktionsprüfung untersucht. Zur praktischen Untersuchung wurde am Fachgebiet Nachrichtenübertragung ein eigenes System *Queryhammer* implementiert und ausführlich untersucht [27,28,66,91,180,195,196]. Die Komponenten des Systems werden in den einzelnen Kapiteln detailliert vorgestellt. Aus der Aufgabenstellung ergibt sich thematisch folgende Gliederung:

Musik und Melodie Melodiesuchsysteme sind Gegenstand dieser Arbeit, daher ist es notwendig, zunächst den Begriff der Melodie näher einzugrenzen. Damit verbunden werden in Kapitel 2 Begriffe der Musiktheorie erläutert, die zur Beschreibung von Musik und Melodien verwendet werden können. Ihre Eignung für QBH-Systeme wird im Einzelnen erörtert. Danach werden die technischen und musikalischen Aspekte der Melodiedarstellung erläutert. Weiterhin ist diesem Kapitel ein Abschnitt über das Singen von Melodien zugeordnet. Er enthält eine Übersicht von vorhandenen Untersuchungen in der Literatur, die das Singen von Melodien und speziell das Verhalten von Nutzern von QBH-Systemen behandeln.

Musiksuchmaschinen In Kapitel 3 wird der Stand der Technik anhand einiger Melodie- und Musiksuchsysteme dargestellt. Besondere Merkmale dieser Systeme wie verschiedene Eingabemöglichkeiten und an-

geschlossene Datenbanken werden diskutiert; die verwendeten Techniken werden kurz dargelegt. Die grundsätzlichen Anforderungen an ein QBH-System werden erläutert, darüberhinaus die Möglichkeiten der Anwendung in Netzwerken diskutiert. Abhängig vom Verwendungszweck wird erörtert, welche Anwender und Nutzergruppen ein Interesse an QBH-Systemen haben. Danach wird die technische Umgebung von QBH-Systemen mit Auswirkungen auf Hard- und Software erläutert. Aus diesen Betrachtungen werden schließlich die Zielbestimmungen für das Beispielsystem *Queryhammer* abgeleitet.

Multimedia-Standards Für die Beschreibung der Melodiekontur, aber auch weiterer Schnittstellen eines QBH-Systems eignen sich Multimedia-Standards, die in Kapitel 4 dargestellt werden. Während sich der Standard MPEG-7 vor allem auf die Beschreibung von Inhalten von Multimedia-daten konzentriert, können andere Standards wie MPEG-21 oder SMIL besonders für die Präsentation der Ergebnisse und Einbindung in Netzwerke benutzt werden. Die durch die Verwendung solcher Standards entstehenden Möglichkeiten werden erläutert und der besondere Bezug zu QBH-Systemen wird hergestellt.

Monophone Transkription Die Eingabe von gesummten Melodien und ihre Umwandlung in eine Melodiekontur bedeutet die Transkription eines monophonen Gesangssignals. Kapitel 5 klärt die verschiedenen Schritte der Signalverarbeitung, insbesondere Techniken der Grundfrequenzanalyse und Rhythmuserkennung. In eigenen Untersuchungen zur Transkription von gesummten Anfragen wird untersucht, welche Fehler auftreten und welche Ursachen sie haben. Weiterhin wird ein Verfahren zur möglichst objektiven Beurteilung von Transkriptionssystemen entwickelt.

Polyphone Transkription Kapitel 6 geht der Frage nach, wie Melodien in die Melodiedatenbank eingetragen werden können. Ein Überblick über bislang verwendete Methoden wie die Extraktion der Melodien aus MIDI-Dateien oder die Transkription durch Musiker diskutiert deren Vor- und Nachteile. Anschließend werden die aktuellen Möglichkeiten der Signalverarbeitung zur automatischen Extraktion von Melodien aus Musiksignalen dargelegt. Eigene Untersuchungen zeigen die Einsatzmöglichkeiten geeigneter Verfahren und wie weit damit die Transkription von Melodien aus polyphonen Signalen möglich ist.

Melodievergleich Liegen Anfrage und Melodie der Datenbank symbolisch vor, so kann der Vergleich dieser beiden Melodien erfolgen. Dieser Vergleich macht Ähnlichkeitsmaße notwendig, die eine anschließende Bewertung zulassen. Für QBH-Systeme, die mit Konturdarstellungen arbeiten, kommen Verfahren zur Zeichenkettensuche in Frage. Kapitel 7 erläutert häufig in QBH-Systemen verwendete Verfahren, ein besonderes Augenmerk gilt der Tauglichkeit der Verfahren für den Standard MPEG-7.

Melodiedatenbanken In Kapitel 8 wird untersucht, wie sich Eigenschaften der Melodiedatenbank auf Suchergebnisse in QBH-Systemen auswirken. Ausgehend von einigen statistischen Parametern wird ein Modell für Melodiedatenbanken vorgeschlagen. Danach werden die Ergebnisse aller vorangegangenen Kapitel zusammengefasst und anhand des Beispielsystems *Queryhammer* untersucht. Dazu wird erörtert, auf welchem Wege die Güte des Suchergebnisses eines QBH-Systems beurteilt werden kann. Für das Ergebnis einer Suchanfrage an ein QBH-System sind verschiedenste Parameter wie Größe und Inhalt der Datenbank, Qualität der Transkription und Art und Weise des Melodievergleichs ausschlaggebend. Die Abhängigkeit dieser Parameter voneinander wird untersucht und diskutiert.

Die **Zusammenfassung** der Arbeit gibt die wesentlichen Ergebnisse der Arbeit wieder, formuliert notwendige Konsequenzen für QBH-Systeme und gibt einen Ausblick auf weitere Themen.

*Musik wird oft nicht schön
gefunden,
weil sie stets mit Geräusch
verbunden.*

Wilhelm Busch

Im Vordergrund dieses Kapitels steht die Frage, *wonach* überhaupt gesucht wird, wenn man nach Musik sucht. Der Begriff „Musik“ leitet sich von „*musiké téchne*“ (griechisch: Kunst der Musen) ab und ist eine künstlerische Lebensäußerung des Menschen [199]. Besonderes Kennzeichen eines bestimmten Musikstücks ist oft die Melodie, vor allem in der westlichen Musik. Was unter dem Begriff Melodie allgemein verstanden werden kann und in welchem Sinn er für Query-by-Humming-Systeme (QBH-Systeme) verwendet werden soll, wird im nächsten Abschnitt dargelegt. Um Melodien beschreiben und notieren zu können, ist das Verständnis einiger Begriffe der Musiktheorie notwendig, die im darauf folgenden Abschnitt erklärt werden. Dies führt zur Frage, wie Musik technisch zu erfassen ist – Abschnitt 2.3 beschreibt daher verschiedene Formen der Musikrepräsentation. Schließlich wird das Singen von Melodien in Abschnitt 2.4 diskutiert; dies ist bedeutsam, da ein Mensch dabei natürlicherweise auch Einfluss auf die Melodie nimmt.

2.1 Der Melodiebegriff

Der Begriff der Melodie ist nicht klar definiert und bedarf näherer Erläuterung. In [53] findet man:

Melodie: die in der Zeit sich entfaltende selbständige Tonbewegung, die sich gegenüber weniger selbständigen Tonfolgen (Neben-, Begleit-, Füllstimmen) auszeichnet durch innere Folgerichtigkeit oder Gesanglichkeit oder leichtere Fasslichkeit oder durch Festigkeit und Geschlossenheit ihrer Gestalt und die als konkrete Erscheinung auch das rhythmische Element in sich enthält. [...]

Damit ist eine Fülle von Merkmalen angesprochen, die eine Melodie kennzeichnen können. Zunächst versteht man unter einer Melodie eine zeitliche Tonfolge, nicht das gleichzeitige Erklängen mehrerer Töne in einem Akkord [18]. Diese Tonfolge zeichnet sich meist dadurch aus, dass sie besonders einprägsam und markant ist, je nach Vermögen sollte sie auch gut zu singen sein. Nicht nur die Töne selbst, sondern auch die zeitliche Information als Rhythmus sind wichtig für eine Melodie.

Eine Melodie kann mehrere *Motive* enthalten, die eine „kleinste Einheit musikalischer Ausdrucksbedeutung bildet“ [53, 141]. So besteht der Beginn BEETHOVENS neunter Sinfonie aus einem kurzen und markanten Motiv; mit RICHARD WAGNER wurde das an bestimmte Personen, Ideen oder Orte verknüpfte *Leitmotiv* populär.

Von der Gesangsmelodik (beschränkter Tonumfang, singbare Tonschritte), die auch in einem Großteil der Instrumentaltitel (zum „Mitsingen“) anzutreffen ist, kann die Instrumentalmelodik (mit einem Tonumfang je nach Instrument, häufig auch Extremlagen der Oktave als Effekt, z. T. große Intervallsprünge, Akkordbrechungen, Tonleiterausschnitte) abgegrenzt werden. Die Geschichte der populären Musik belegt, dass in bestimmten Zeitabschnitten die spieltechnischen Möglichkeiten einzelner Instrumente die Melodik entscheidend geprägt haben, z. B. die Violine den Wiener Walzer, die Gitarre die Rockmusik, das Klavier Ragtime und Boogie-Woogie. Auch die Melodik der Jazz- und Rockimprovisationen ist abgesehen von verbreiteten Standardfloskeln weitgehend instrumental bedingt (beim *Scat* werden typische Instrumentallinien gesungen). Der afrikanische bzw. afroamerikanische Einfluss zeigt sich im Blues, im Jazz und im Rock einerseits durch spezielle melodieformbildende Fakten wie z. B. das Ruf-Antwort-Prinzip (Call and Response), Melodiemuster (Pattern), Ostinati und Riffs, andererseits durch individuelle Gestaltungsmittel wie z. B. Tongebung, Phrasierung, Akzentuierung, Verzierung usw.

Was als Melodie bezeichnet wird, hängt stark vom musikalischen Genre ab. Volks- oder auch Rockmusik sind z. B. oft in Strophe und Refrain unterteilt. Diese Teile weisen jeweils eigene Melodien auf, die durch Phrasierung und Liedtext stark variiert werden können. In der klassischen Musik der westlichen Welt treten häufig Eröffnungsmelodien oder Hauptthemen auf, die im Folgenden verändert werden. In der klassischen Musik können Melodien nebeneinander, das heißt zeitgleich in Erscheinung treten. In diesem Fall spricht man von Polyphonie, wie sie beispielsweise für JOHANN SEBASTIAN BACHS Fugen typisch ist.

Für populäre wie auch klassische Musik sind für die Melodie vor allem Tonfolge und Rhythmus der Melodietöne wichtig, während in der zeitgenössischen Musik zunehmend der Qualität der Töne eine besondere Rolle zukommt. *Klangfarbenmelodie* nennt ARNOLD SCHÖNBERG eine Folge von Klangfarben, deren Beziehung untereinander mit einer Art Logik wirkt wie sonst die Tonhöhe [53]. Ein anderes Beispiel ist GÖRGY LIGETIS Komposition „Aventures“ von 1962. Nicht mehr die Rhythmik oder Metrik sind bestimmend, sondern vielmehr die Vorstellung eines Klangkontinuums, das eher Lautgedichten ähnlich ist [137].

Im Jazz ist die notierte Melodie oft nur eine Skizze dessen, was tatsächlich gespielt werden soll und erfährt weitgehende Variationen durch den Musiker. Die Melodie dient hier als Ausgangspunkt für Improvisationen. Es gibt aber auch klar festgelegte Melodien im Jazz, die gemäß des Notentextes gespielt werden. JOBIMS bekannte Komposition „One Note Samba“ ist ein schönes Beispiel für diesen Fall, in dem die Melodie abschnittsweise zwischen purer Rhythmik und Melodik wechselt. In der klassischen indischen Musik besteht ein starker Bezug der Melodie zu Ton und Rhythmus, allerdings nicht zu Harmonien. In nahezu allen Bereichen der populären Musik überwiegt die Bedeutung der Melodie gegenüber anderen Elementen des Musikalischen. Ausnahmen sind die Entwicklungen im Dancefloor-Bereich, wie insbesondere Housemusic und Techno, dort gibt es keine klare Melodie, die Musik lässt sich besser anhand rhythmischer Merkmale charakterisieren.

2.2 Begriffe der Musiktheorie

In diesem Abschnitt werden Begriffe erklärt, die zur syntaktischen (und semantischen) Beschreibung von Musik und Melodien verwendet werden können. Weiterhin wird ihre Bedeutung für QBH-Systeme erläutert.

2.2.1 Akustik

Natürliche Grundlage der Musik ist der Schall [136]. Er ist die Voraussetzung dafür, dass Töne, Klänge und Geräusche entstehen können.

Ton Töne sind die Grundelemente der Musik [13]. Der Ton ist definiert als das Ergebnis einer einfachen sinusförmigen Schwingung, die Tonhöhe ist durch die Frequenz der Sinusschwingung festgelegt [199]. Ein solcher Ton lässt sich nur synthetisch erzeugen; man unterscheidet daher

diese „reinen“ Töne von den „natürlichen“ Tönen, die auch als *Klang* bezeichnet werden.

Klang Ein Instrumentalton, etwa der einer Geige, ist ein natürlicher Ton. Das Gemisch von dem (gespielten und erklingenden) Grundton und den gleichzeitig erklingenden Ober- oder auch Partialtönen ist für das Instrument charakteristisch und wird in der Musik wie in der Elektroakustik als Klang bezeichnet [136, 191]. Die menschliche Stimme ist in dieser Hinsicht auch ein Instrument; ein gesungener Ton besteht ebenfalls aus dem Grundton und mehreren Obertönen. Sogenannte Formanten bestimmen bei der Stimme die Färbung dieses natürlichen Tones. Entsprechend wird mit dem Begriff *Klangfarbe* das Resultat der Gewichtung der einzelnen Partialtöne bezeichnet.

Innerhalb der Psychoakustik wird dem Begriff Klang lediglich das Zusammenklingen zweier reiner oder natürlicher Töne zugeordnet [200] – nennt man einen natürlichen Ton einen Klang, so ist in diesem Fall die Bezeichnung *Klanggemisch* zu verwenden [191].

Geräusch Als Geräusch werden im Unterschied zu Ton und Klang unperiodische Schwingungen bezeichnet, die keine exakt bestimmbare Tonhöhe aufweisen. Frequenz und Stärke ändern sich in der Zeit und unterliegen keiner Gesetzmäßigkeit. Die Teilschwingungen haben kein obertöniges Verhältnis.

Für QBH-Systeme spielen Klänge bzw. natürliche Töne eine große Rolle, da sie die akustischen Informationsträger einer Melodie sind. Um eine Melodie einem QBH-System zugänglich zu machen, muss sie in geeigneter Weise notiert werden. Daher folgt nun die Darstellung der Notenschrift.

2.2.2 Notenschrift

Schon in der griechischen Antike gab es Buchstaben und von Buchstaben abgeleitete Symbole zur Bezeichnung von Tönen. Sie stehen für das, was man messen kann: unterschiedliche Tonabstände, Klassifizierung von Intervallen, Aufbau von Skalen, Ordnung ganzer Tonsysteme. Auch Melodien wurden mit dieser Buchstaben-Notenschrift aufgezeichnet [62].

Das graphische Festhalten von Tonhöhen und Tondauern bezeichnet man als Notation [13]. Die Notation auf fünf Notenlinien, die heute in der abendländischen Musik verwendet wird, wurde von GUIDO VON AREZZO um das



(a) Notenmanuskript eines anonymen Kopisten aus der 1. Hälfte des 18. Jahrhunderts, überschrieben mit „Solo pour la Flute traversiere par J. S. Bach“.



(b) Notendruck der Urtext-Edition des Verlags Zimmermann, Frankfurt

Abbildung 2.1: Beispiel Notenschrift – JOHANN SEBASTIAN BACH: „Partita in A-Moll“, BWV 1012, handschriftlich (Staatsbibliothek zu Berlin, Stiftung Preuß. Kulturbesitz, P968) und gedruckt (entnommen aus [25]).

Jahr 1000 eingeführt [136]. Vor AREZZO wurden Melodien und traditionelle Gesänge üblicherweise mündlich überliefert. Altangesehene Musikschulen sahen sich damals in ihrer Existenz bedroht und leisteten lange und heftig Widerstand gegen die schriftliche Weitergabe der Musik [95].

5-Liniensystem

Das wichtigste Symbol zur Aufzeichnung von Musik ist die **Note** (lateinisch nota = Zeichen) [199]. Noten werden meist in ein Zeilensystem mit fünf Linien eingeordnet, ihre Platzierung lässt die Tonhöhe erkennen. Diese Darstellung wird in der angelsächsischen Literatur auch als „common western music notation“ (CWMN) bezeichnet [187]. Abbildung 2.1 zeigt ein Beispiel.

Zur Angabe der absoluten Tonhöhe dient der **Notenname**. Man benutzt die ersten sieben Buchstaben aus dem Alphabet: „a b c d e f g“. Das „b“ spaltete sich im 10. Jahrhundert in einen tieferen Ton „b rotundum“ und einen höheren Ton „b quadratum“. Letztere Note wurde allmählich mit einem „h“ bezeichnet [199], das als weiterer Notenname hinzukam. Die Note „h“ wird im Angelsächsischen heute allerdings mit „b“ bezeichnet, dagegen ist im deutsch-

Tabelle 2.1: Bezeichnung der Oktaven.

Oktave	enthaltene Noten	alternative Bezeichnung
fünfgestrichene Oktave	nur c^5	c''''
viergestrichene Oktave	$c^4 - h^4$	$c''' - h'''$
dreigestrichene Oktave	$c^3 - h^3$	$c'' - h''$
zweigestrichene Oktave	$c^2 - h^2$	$c' - h'$
eingestrichene Oktave	$c^1 - h^1$	
kleine Oktave	$c - h$	
große Oktave	$C - H$	
Kontra-Oktave	$C_1 - H_1$	
Subkontra-Oktave	nur $A_2 - H_2$	

sprachigem Raum mit „b“ der um einen Halbton tiefere Ton gemeint (im Angelsächsischen „bb“, lies: b flat).

Äquivalent zu den Notennamen aus dem Alphabet werden auch die Silben „do, re, mi, fa, so, la, si“ verwendet. Das Benennen der Töne durch diese Silben bezeichnet man als **Solmisation**, die wie die Notenschrift auf AREZZO zurückgeht [13].

Die Notennamen wiederholen sich zyklisch im Abstand einer Oktave. Zur Unterscheidung der Oktaven sind verschiedene Bezeichnungen üblich. Eine Übersicht gibt Tabelle 2.1. Der Kammerton a als Referenzton liegt in der eingestrichenen Oktave und wird mit a' oder a^1 bezeichnet, vgl. Abschnitt 2.2.5. Im MIDI-Standard sind die Oktaven fortlaufend von $-1, 0, 1, \dots, 9$ durchnummeriert, siehe Abschnitt 2.3.2.

Notenwerte und Pausenzeichen

Die Tondauer einer Note wird durch die *Notenform* bestimmt [136]. Ausgangspunkt für die Einteilung der Notenwerte ist die *Ganze Note*, deren Dauer willkürlich festgelegt wird. Relativ dazu gibt es *Halbe*, *Viertel*, *Achtel*, *Sechzehntel* usw.; jedem Notenwert zeitlich entsprechend gibt es auch ein Pausenzeichen. Darüber hinaus gibt es eine Fülle von weiteren Zeichen, mit denen länger gehaltene Noten, unregelmäßige Unterteilungen usw. dargestellt werden können [136, 199].

Aus den Notenwerten und Pausen ergibt sich der Rhythmus zu einer Melodie, der von den Tonhöhen getrennt untersucht werden kann. Der Rhythmus bezieht sich auf einen Takt, dessen musikalische Bedeutung nun erläutert wird.

Metrum und Takt

Neben der Gliederung der Musik in kurze und lange Tondauern ist die Folge der Betonungen (metrische Akzente) von grundlegender Bedeutung für die Gestaltung einer Melodie [199]. Das Verhältnis von betonten (schweren) und unbetonten (leichten) Zählzeiten oder auch Schlägen nennt man *Metrum* (griechisch *Métron* = Maß).

Der *Takt* eines Stückes fasst bestimmte Gruppen von Zählzeiten unter Beachtung der Betonungsverhältnisse zusammen [199]. Takte werden durch Taktstriche voneinander getrennt. Der Taktstrich kennzeichnet die neue betonte Zählzeit, meist die Eins des folgenden Taktes, und ist gleichzeitig Ausdruck des rhythmischen Ordnungsprinzips.

Die *Taktart* gibt an, in wieviele Zählzeiten ein Takt untergliedert ist; sie wird gewöhnlich nach dem Notenschlüssel in Form eines mathematischen Bruchs angegeben. Der Nenner bezeichnet dabei den rhythmischen Grundwert, also die Länge einer Zählzeit. Der Zähler gibt Auskunft darüber, wieviele Zählzeiten in einem Takt enthalten sind. Üblichste Taktart ist der $\frac{4}{4}$ -Takt, die auch oft mit „C“ notiert wird, so auch im Beispiel der Partita in Abbildung 2.1.

Tempo

Das absolute Tempo, mit dem Noten gespielt werden sollen, lässt sich durch die Anzahl der Zählzeiten festlegen, die pro Minute gespielt werden. Damit handelt es sich im technischen Sinne um eine Frequenz, die in der Einheit *Schläge pro Minute* (englisch: beats per minute, BPM) angegeben wird. Im deutschsprachigen Raum ist auch die Angabe in der Form *M.M.* = 60 üblich, in diesem Beispiel also 60 Schläge pro Minute. *M.M.* steht dabei für *Metronom Mälzel* [53]. Es ist benannt nach JOHANN NEPOMUK MÄLZEL, der 1816 das Metronom auf Anregung LUDWIG VAN BEETHOVENS konstruierte [13]¹. Dieser wünschte sich eine präzisierte Tempodefinition, da bisher die Tempobezeichnung nur in Worten angegeben wurde (siehe unten).

¹Tatsächlich handelt es sich um ein Plagiat: MÄLZEL hatte die Erfindung von DIETRICH NIKOLAUS WINKEL aufgegriffen und mit leichten Änderungen zu seiner gemacht.

Einige Komponisten des 20. Jahrhunderts wie BELA BARTOK oder JOHN CAGE geben die gesamte Aufführungsdauer als Zeitmaß an, aus der ein zu spielendes Tempo nur ungefähr geschätzt werden kann. Den in Abbildung 2.1 gezeigten Noten kann das Tempo sogar nur aus der Bezeichnung „Allemande“ (deutscher Tanz) entnommen werden: es handelt sich um einen Schreittanz und es obliegt dem Musiker ein dafür angemessenes Tempo zu wählen.

Vor der Erfindung des Metronoms wurde das Tempo mit Worten beschrieben, wobei dabei häufig Ausdrücke aus dem Italienischen verwendet wurden [18]. Die Angabe dieser Bezeichnungen ist auch heute noch zusätzlich zu einer Metronomzahl üblich, da sie Informationen über den Charakter des bezeichneten Stückes tragen. So bezeichnen die Ausdrücke „Presto“ wie „Allegro“ ein hohes Tempo, aber mit „Allegro“ wird zusätzlich eine freudige Spielweise angedeutet (italienisch *allegro* = fröhlich, lustig). „Presto“ hingegen ist mit Virtuosität konnotiert. In Tabelle 2.2 ist eine Übersicht dargestellt.

Den angegebenen Tempobezeichnungen lassen sich M.M.-Werte zuordnen, die jedoch nur als Richtwert betrachtet werden können und nicht jeder musikalischen Situation gerecht werden. Tabelle 2.3 zeigt die Zuordnung eines Metronoms der Firma Wittner (Modell MT-50).

Das Tempo ist nicht notwendigerweise konstant, sondern kann beschleunigt oder verlangsamt werden. Mit den Ausdrücken *Accelerando* – *schneller*, *Rallentando*, *Ritardando*, *Ritenuto* – *langsamer* wird das Spieltempo über einen vorgegebenen Zeitraum, meist ein bis mehrere Takte, erhöht oder vermindert. Natürlich sind auch abrupte Tempowechsel mit der Vorgabe einer neuen M.M.-Bezeichnung möglich. Mit *A tempo* wird die augenblickliche Wiederaufnahme eines zuvor verlangsamten Tempos gefordert. Die Bezeichnung *Rubato* erlaubt sogar die freie Anpassung des gewählten Tempos, um einen expressiven Vortrag zu ermöglichen.

Tempobezeichnungen wie Presto oder Largo lassen sich als Text ausdrücken und sind gut geeignet, um eine qualitative Beschreibung eines Tempos anzugeben und daher auch gut für Musiksuchsysteme verwendbar.

2.2.3 Klavierwalze

Mit der Verwendung von elektronischen Musikinstrumenten, die mit MIDI-Signalen arbeiten, hat sich die Klavierwalzendarstellung von Noten etabliert. Sie leitet sich von den zur Steuerung von walzengesteuerten Klavieren benötigten Lochkarten ab, wie sie in Abbildung 2.2a dargestellt ist.

Tabelle 2.2: Tempobezeichnungen in verschiedenen Sprachen.

italienisch	Largo	sehr langsam
	Larghetto	etwas schneller als Largo
	Adagio	langsam
	Andante	schreitend
	Moderato	moderat
	Allegretto	noch nicht Allegro
	Allegro	schnell
	Presto	schnell
	Prestissimo	sehr schnell
	Larghissimo	so langsam wie möglich
	Vivace	lebhaft
	Maestoso	majestätisch (meist langsam)
französisch	Grave	langsam und festlich
	Lent	langsam
	Modéré	moderates Tempo
	Vif	lebhaft
	Vite	schnell
deutsch	Langsam	
	Mäßig	
	Lebhaft	
	Rasch	
	Schnell	

Die Repräsentation von MIDI-Daten als Klavierwalze (piano roll) ist eine Zeit-Tonhöhen-Darstellung einzelner MIDI-Ereignisse, die einzelnen Töne werden als Rechtecke dargestellt. Deren vertikale Position beschreibt die Tonhöhe, die horizontale Position die Einsatzzeit und die Breite die Tondauer. Die Klavierwalze wird im Rahmen dieser Arbeit für die Visualisierung von MIDI-Noten verwendet.

2.2.4 Akkordsymbolschrift

Mehrklänge werden oft nur als Symbol notiert. Die dafür verwendbare Akkordsymbolschrift ist besonders im Bereich des Jazz verbreitet, wenn Harmo-

Tabelle 2.3: Gängige Tempobezeichnungen und entsprechende M.M.-Werte.

Bezeichnung	M.M.
Largo	40–60
Larghetto	60–66
Adagio	66–76
Andante	76–108
Moderato	106–120
Allegro	120–168
Presto	168–208

niefortschreitungen eine besondere Rolle spielen. Eine typische Jazz-Kadenz wird zum Beispiel mit

$$A^{m7} \ D^7 \ G^{maj7}$$

notiert [39]. Die Großbuchstaben geben den Grundton des Akkords an, „m“ bezeichnet die Mollterz, mit der „7“ wird angezeigt, dass zu dem Dreiklang aus Grundton, Terz und Quinte noch die kleine bzw. große (bei „maj“) Septime gespielt wird. Da sich die Akkorde auf die Stufen einer Tonleiter beziehen lassen, ist auch die Schreibweise

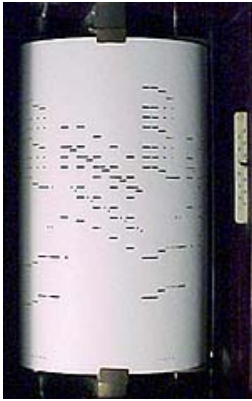
$$II^{m7} \ V^7 \ I^{maj7}$$

möglich; die römischen Ziffern bezeichnen die Stufen der Tonleiter, auf denen der entsprechende Akkord gebildet wird. Im Beispiel oben handelt es sich also um G-Dur.

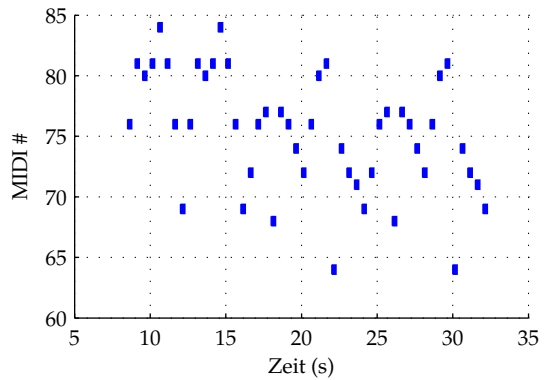
Die Akkordsymbolschrift lässt sich ohne weiteres in Textform angeben und ist daher für die Eingabe in Musiksuchsysteme gut geeignet. Um die Bedeutung der Akkordsymbolschrift zu verstehen, sind gute Kenntnisse der Harmonielehre notwendig, damit beschränkt sich der Anwenderkreis in diesem Fall auf professionelle Nutzer.

2.2.5 Kammerton a

Die relativen Intervallverhältnisse werden an absolute Tonhöhen geknüpft. So wurde der **Kammerton a** auf 440 Hz bei 20 °C festgelegt (2. internationale Stimmtonkonferenz, London, 1939) [136].



(a) Die Steuerkarte einer Klavierwalze. Quelle: [1]



(b) Notendarstellung der MIDI-Fassung zur Partita aus Abbildung 2.1 als Klavierwalze.

Abbildung 2.2: Klavierwalze und MIDI-Notendarstellung.

Im Verlauf der Musikgeschichte wurde der Kammerton mehrfach geändert. Voraussetzung für die Vorgabe einer einheitlichen Tonhöhe überhaupt war die Erfindung der Stimmgabel 1711 durch den Engländer JOHN SHORE. 1788 betrug das a^1 der Pariser Stimmung 409 Hz, dass der Wiener und Berliner Stimmung um 1850 442 Hz [199]. Gegenwärtig sind aber auch höhere Frequenzen als 440 Hz üblich, insbesondere bei Symphonieorchestern mit Streichern (444–448 Hz kommen durchaus vor). Einige Hersteller stimmen ihre Instrumente bereits von vornherein auf 442 Hz [139]. Die Stimmung des Kammertons spielt bei der Transkription von gesungenen Nutzeranfragen eine wichtige Rolle, wie in Kapitel 5 gezeigt wird.

2.2.6 Intervall

Mit Intervall (lateinisch *intervallum* = „Zwischenraum“) bezeichnet man den Abstand beziehungsweise das Verhältnis zweier Töne, die sowohl zusammen, z. B. in einem Akkord, als auch nacheinander, z. B. in einer Melodie, erklingen können [62]. Intervalle werden mit lateinischen Ordnungszahlen benannt. In einer Stammform treten die reinen Intervalle Prime, Quarte, Quinte, und Oktave auf, in zwei Stammformen die großen bzw. kleinen Intervalle Sekunde, Terz, Sexte und Septime. Für die Melodiekontur gemäß MPEG-7 (vgl. Ab-

schnitt 2.3.3) werden die Intervalle Prime, kleine und große Sekunde sowie kleine Terz unterschieden.

2.2.7 Tonsystem und Skalen

Akustische Ereignisse bedürfen einer systematischen Ordnung, um musikalische Informationsträger werden zu können [136]. Es entstehen dabei nach Kulturraum und Epoche unterschiedliche Systeme. Das abendländische **Tonsystem**, das auf die griechische Antike zurückgeht, selektiert Töne (und Klänge, siehe Abschnitt 2.2.1) und scheidet Gleitsequenzen, Geräusche und Knalle aus.

Dem in der westlichen Musik üblichen Tonsystem liegt primär eine *heptatonische* (7-tönige) Tonleiter zugrunde. Sie ergibt sich aus der Unterteilung des Oktavraums. Für die Unterteilung des Oktavraums sind verschiedene Herleitungstheorien bekannt. Eine **Tonleiter** oder **Skala** bezeichnet eine Zusammenstellung diskreter Tonhöhen innerhalb einer Oktave, die so angeordnet sind, dass man eine größtmögliche Anzahl an konsonanten Kombinationen (bzw. kleinstmögliche Anzahl an Dissonanzen) erhält, wenn zwei oder mehr Töne aus dieser Zusammenstellung zusammenklingen [164].

Hier wird nun die **Quintenschichtung** nach dem pythagoräischen System dargestellt [136]. Töne im Abstand einer Quinte sind im 1. Grade verwandt, sie werden auch als „Wiederholung des Grundtons auf höherer Ebene“ beschrieben [179]. Die pythagoräische Quinte ist rein, die Frequenzen der Töne stehen im Verhältnis 3 : 2. Die Schichtung von reinen Quinten ergibt

- die **halbtonlose Pentatonik** bei 5 Quinten: c-g-d-a-e, als Skala in einer Oktave erhält man die Reihenfolge c-d-e-g-a;
- die **diatonische Heptatonik** bei 7 Quinten: ergänzt man die o. g. Pentatonik um eine Quinte nach oben und nach unten, erhält man die Töne f-c-g-d-a-e-h, zurücktransponiert in eine Oktave also alle Stammtöne c-d-e-f-g-a-h;
- die **halbtönige Chromatik** bei 12 Quinten: eine weitere Schichtung über h hinaus ergibt die Töne fis-cis-gis-dis-ais, unter f hinab b-es-as-des-ges. Die Halbtöne haben je nach Ableitung verschiedene Grundfrequenzen; so differieren z. B. ais und b. Das System schließt sich nicht, denn 12 reine Quinten sind größer als 7 Oktaven. Diese Abweichung bezeichnet man als **pythagoreisches Komma**.

Über die Quintenschichtungen ist das Zustandekommen verschiedener Skalen und der Unterteilung des Oktavraums in 12 Halbtöne erklärt; damit verbunden ist die Existenz verschiedener Temperaturen, die im folgenden Abschnitt erläutert werden.

Neben den genannten sind eine Reihe von Skalenbezeichnungen in Gebrauch, u. a. *Dur, Moll, Ganzton, Vermindert, Jüdisch, Arabisch*. Die chromatische Tonleiter ist Grundlage der später im Rahmen dieser Arbeit beschriebenen automatischen Transkription, wenn Töne von gesungenen Melodien in den 12-Tonraum abgebildet werden sollen.

2.2.8 Tonart

Die Tonart wird durch einen Grundton und eine Skala angegeben, zum Beispiel E-Dur oder C $\sharp$ -Moll. Übliche Skalen sind *Dur, Moll, Pentatonik* und *Chromatik*. An der Skalenbezeichnung kann man die Beziehung der Töne untereinander erkennen (Grund-, Schweb-, Spannungs- oder Leitton), während der Name des Grundtons die musikalische Verwandtschaft festlegt [199].

Abhängig von der gewählten musikalischen Temperatur (siehe Abschnitt 2.2.10) können Tonarten einen bestimmten Klangcharakter haben. So schreibt zum Beispiel QUANTZ in [157]:

[...] um so wohl den Affect der Liebe, Zärtlichkeit, Schmeicheley, Traurigkeit, auch wohl, wenn der Componist ein Stück darnach einzurichten weis, eine wütende Gemüthsbewegung, als die Verwegenheit, Raserey und Verzweiflung, desto lebhafter auszudrücken: wozu gewisse Tonarten, als: E moll, C moll, F moll, Es dur, H moll, A dur und E dur, ein Vieles beytragen können.

Die verschiedenen Tonarten können auch im Multimedia-Standard MPEG-7 wiedergegeben werden (siehe Kapitel 4); sie eignen sie sich gut zur Beschreibung musikalischer Charakteristika.

2.2.9 Cent-System

Kleinere Unterteilungen als Halbtöne sind im 12-tönigen Tonsystem nicht möglich. Daher führte der englische Akustiker JOHN ELLIS 1885 das Cent-System ein, indem er einen Halbtonschritt der gleichtemperierten Stimmung (siehe auch nächster Abschnitt) in 100 weitere Teile unterteilte. Damit ist es

möglich, Abweichungen von der gleichstufigen Stimmung numerisch auszudrücken. 100 Cent entsprechen einem gleichtemperierten Halbtonschritt.

Oft ist die Abweichung einer Note bei der Frequenz f von der gewünschten Grundfrequenz f_0 in Cent gesucht; sie ergibt sich aus

$$m = 1200 \log_2 \frac{f}{f_0}. \quad (2.1)$$

2.2.10 Temperaturen

Mit dem Begriff *Temperatur* ist in der Musik die gewollte Verstimmung einzelner Töne einer Tonleiter gemeint, es wird daher auch der Begriff *Stimmung* verwendet.

Bei der *reinen Stimmung* verwendet man die Intervalle, wie sie sich aus den Obertönen ergeben (reine Oktave (1 : 2), reine Quinte (2 : 3) und reine Quarte (3 : 4), reine große Terz (4 : 5), etc.). Die reine Stimmung wird auch *natürliche* oder *harmonische Stimmung* genannt. Nachteil der reinen Stimmung ist, dass die reine Stimmung immer nur für eine Tonart, z. B. C-Dur, zu erreichen ist. Um auf einem Instrument verschiedene Tonarten spielen zu können, sind daher andere Stimmungen gebräuchlich.

Bei der *gleichstufigen* (auch *gleichtemperierten* oder *gleichschwebenden*) *Stimmung* teilt man die Oktave in 12 gleich große Halbtonschritte auf [13]. Ausgehend von einem Ton mit der Grundfrequenz f_0 errechnen sich die weiteren Tonhöhen der chromatischen Tonleiter in diesem Fall über

$$f(n) = f_0 2^{\frac{n}{12}}, \quad (2.2)$$

wobei $n = 1 \cdots 12$ ist und $f(12)$ die Oktave zu f_0 darstellt.

Unter der Sammelbezeichnung „Wohltemperierte Stimmungen“ fasst man eine Reihe musikalischer Temperaturen zusammen. Charakteristisch für diese ist, dass keine Intervalle vorkommen, welche unakzeptabel von denen reiner Stimmung abweichen. Deshalb sind sämtliche Tonarten spielbar. Außerdem bewahren wohltemperierte Stimmungen im Gegensatz zur gleichstufigen Stimmung die Tonartcharaktere, da Dreiklänge in unterschiedlichen Tonarten unterschiedlich stark temperiert sind.

Umstritten ist die verbreitete Annahme (u. a. in [53]), ob der Begriff wohltemperierte Stimmung mit der gleichstufigen Stimmung identisch ist [13]: Das bekannte Werk „Das Wohltemperierte Klavier“ von JOHANN SEBASTIAN BACH diente demnach nicht zur Demonstration der gleichtemperierten Stimmung,

sondern vielmehr zum Hervorheben der Tonartcharaktere sowie zur Unterstreichung der Ungeeignetheit der damals üblichen *mitteltönigen* Stimmungen für Tonarten mit vielen Vorzeichen. Die erste wohltemperierte Stimmung in Europa war die 1691 von ANDREAS WERCKMEISTER eingeführte WERCKMEISTER-Stimmung.

Es gibt zahlreiche weitere Stimmungen, zum Beispiel nach KIRNBERGER, VALLOTTI, YOUNG oder HEINZ BOHLEN [107]. Für QBH-Systeme ist die verwendete Stimmung in Bezug auf die Transkription von Melodien aus Audiosignalen wichtig. Im Rahmen der Arbeit wird die gleichstufige Stimmung verwendet.

2.2.11 Intonation

Der Begriff Intonation hat in der Musik mehrere Bedeutungen:

- Kurze Einleitung eines Musikstückes
Bei Gregorianischen Gesängen etwa eine vom Vorsänger ausgeführte Einleitung, auch ein kurzes Orgelvorspiel vor einem Gemeindelied.
- Vorgang oder auch Ergebnis des Intonierens eines Musikinstrumentes
Die Klangfarbe und die Tonhöhen der verschiedenen Töne werden individuell abgeglichen. Bei Gitarren bezeichnet man den Grad der Stimmigkeit der Tonhöhen auch als Bundreinheit. Ein Klavier wird durch Abfeilen und Stechen der Hammerköpfe intoniert.
- Das exakte Treffen der gewünschten Tonhöhe durch einen Sänger oder Instrumentalisten.

Im Rahmen dieser Arbeit wird der Begriff Intonation ausschließlich im letztgenannten Sinne verwendet.

2.2.12 Monophonie und Polyphonie

Die *Monophonie* bezeichnet die Einstimmigkeit in der Musik [53]. Für die *Polyphonie* gibt es mehrere Bedeutungen. Im einfachsten Fall ist mit Polyphonie die Mehrstimmigkeit, d. h. das gleichzeitige Erklingen mehrerer Töne gemeint. Weniger verbreitet ist allerdings die Unterscheidung zwischen zwei Arten von „Mehrstimmigkeit“:

- Die *Homophonie* bezeichnet mehrere Stimmen, die in einem geschlossenen Satz überwiegend parallel im selben Intervallabstand laufen.
- Die *Polyphonie* bezieht sich in diesem Zusammenhang auf die voneinander unabhängigen Stimmen. Diese Polyphonie wurde gerade in der Generalbaßzeit synonym mit dem Begriff *Kontrapunkt* verwendet.

In der Signalverarbeitung allgemein wie im Rahmen dieser Arbeit wird der Begriff Polyphonie meist im einfachen Sinne der Mehrstimmigkeit verwendet, da dies zu besonderen Anforderungen an die Signalverarbeitung führt. Nur selten wird mit Polyphonie das Vorhandensein mehrerer Stimmen gemeint, wie etwa in [163] (in diesem Beispiel handelt es sich obendrein um monophone Signale).

2.3 Musikrepräsentation

Die Repräsentation von Musik kann technisch gesehen in zwei Kategorien unterschieden werden: In der *Audio-Repräsentation* werden Aufnahmen von Konzerten oder Studioproduktionen festgehalten, die zum Beispiel als Wave- oder MP3-Datei auf einem Rechner digital gespeichert werden können. Die CD (Compact Disc) ist ein digitaler Tonträger, LP (Langspielplatte) oder MC (Musikkassette) speichern Musik analog.

Die *symbolische Repräsentation* von Musik hingegen meint die Notenschrift oder technische Formate für Musiknotation wie *NIFF* (siehe nächster Abschnitt) oder *Guido*, die Noteninformation enthalten. Über das Format *MIDI* (siehe Abschnitt 2.3.2) kann man die Informationen zu Noten, wie sie auf einem elektronischen Musikinstrument gespielt werden, speichern. Speziell für QBH-Systeme können *Melodiekonturen* verwendet werden, die zum Beispiel in Form des *Parsons-Codes* oder als *MPEG-7-MelodyContour* dargestellt werden. Abbildung 2.3 zeigt eine Übersicht der genannten Formate.

Die Überführung der Audio-Repräsentation in eine symbolische Repräsentation wird im Rahmen dieser Arbeit als *Transkription* bezeichnet. Es folgt nun eine Beschreibung der für QBH-Systeme relevanten symbolischen Formate.

2.3.1 Noten

Die wohl bekannteste Darstellungsform für Melodien und Musik überhaupt ist die Notenschrift, die bereits in Abschnitt 2.2.2 beschrieben worden ist. Sie

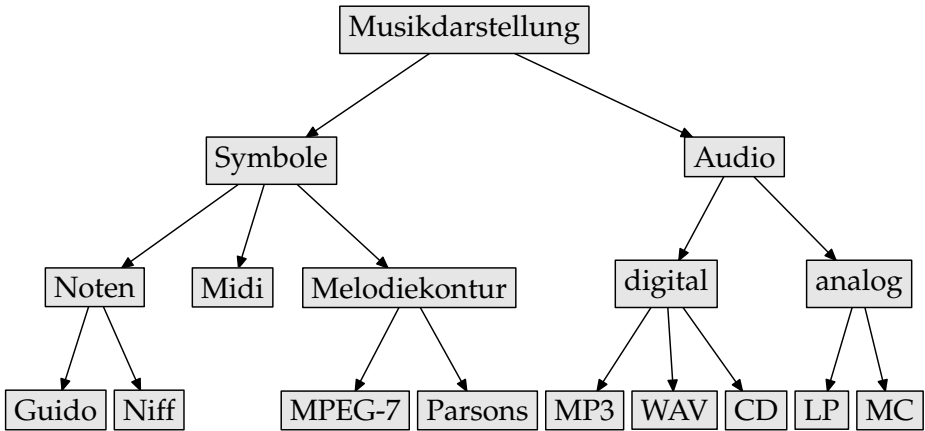


Abbildung 2.3: Übersicht über mögliche Repräsentationsformen für Musik.

ist für den Menschen gut lesbar, während für Maschinen andere, technische Darstellungsformen günstiger sind. Um Noten elektronisch zu speichern, stehen eine Vielzahl von Formaten zur Verfügung. An dieser Stelle soll nur eine exemplarische Auswahl beschreiben werden.

NIFF (notation interchange file format) ist das erste Standardformat für Musiknotation [43]. Es wurde in Zusammenarbeit mehrerer Hersteller von Notationsprogrammen entwickelt und im Herbst 1995 verabschiedet. Obwohl es binär ist, hat es eine hierarchische Struktur, die der in MPEG-7 verwendeten Auszeichnungssprache XML sehr ähnelt. Das Notationsformat *GUIDO* ist eine formale Sprache zur Beschreibung von Noten [7]. Noten können im Textformat gespeichert werden und eignen sich so auch für die Verwendung in Musiksuchsystemen [96]. Formate wie *LilyPond* oder *MusiX_{TEX}* verwenden das Satzprogramm *TEX* zur Ausgabe der Noten [9, 181]. Das MIR-System *Musipedia* benutzt *Lilypond* für die Erzeugung der Notendarstellung der Suchergebnisse [12].

2.3.2 MIDI

Die Abkürzung MIDI steht für *Musical Instrument Digital Interface* und bezeichnet ein Protokoll zur digitalen Datenkommunikation [10]. Ein Konsortium

Tabelle 2.4: Übersicht der MIDI Notennummern über alle Oktaven einer Klaviertastatur [8].

Oktavnr.		Notennummer											
-1	0	1	2	3	4	5	6	7	8	9	10	11	
0	12	13	14	15	16	17	18	19	20	21	22	23	
1	24	25	26	27	28	29	30	31	32	33	34	35	
2	36	37	38	39	40	41	42	43	44	45	46	47	
3	48	49	50	51	52	53	54	55	56	57	58	59	
4	60	61	62	63	64	65	66	67	68	69	70	71	
5	72	73	74	75	76	77	78	79	80	81	82	83	
6	84	85	86	87	88	89	90	91	92	93	94	95	
7	96	97	98	99	100	101	102	103	104	105	106	107	
8	108	109	110	111	112	113	114	115	116	117	118	119	
9	120	121	122	123	124	125	126	127					
	C	C♯	D	D♯	E	F	F♯	G	G♯	A	A♯	B	

von Musikinstrumentenherstellern vereinbarte 1983 die Spezifikation MIDI 1.0, um den Datenaustausch zwischen elektronischen Musikinstrumenten verschiedener Hersteller zu ermöglichen [8]. Durch MIDI können ausschließlich Steuersignale und keine Audiosignale übertragen werden.

Die MIDI-Spezifikation definiert die Beschreibung von Ereignissen wie Note-an (note on) und Note-aus (note off), den Wechsel von Voreinstellungen für Instrumentenklänge (preset change), Betätigung des Haltepedal (sustain), gleitende Tonhöhenänderung (pitch bend) sowie Zeitinformationen (timing information). In Erweiterungen des Standards werden definiert: Klangmuster (sample dumps), Zeitstempel (time code) sowie Dateiformate *Standard MIDI* und *General MIDI* (GM).

MIDI-Daten enthalten Informationen darüber, wie Noten auf einem elektronischen Musikinstrument gespielt werden. In einer MIDI-Datei können aber auch mehrere Spuren niedergelegt werden, die jeweils verschiedene, zusammenspielende Instrumente repräsentieren. Das MIDI-Format kann daher auch zur Melodierepräsentation genutzt werden. Im Rahmen dieser Arbeit wird MIDI als Format für die Melodiedarstellung und für die Klavierwalzendarstellung verwendet.

In Tabelle 2.4 sind die Nummern aller MIDI-Noten dargestellt. Die MIDI-Spezifikation selbst definiert nur Note Nummer 60 als „mittleres C“ (also das

Tabelle 2.5: Die drei Stufen zur Beschreibung einer Melodiekontur. Diese Darstellung wird auch als PARSONS-Code bezeichnet.

Kontur-Wert	Änderung der Tonhöhe
U	up – aufwärts
D	down – abwärts
R	repeat – gleichbleibend
S	same – anstelle von R

c^1 der eingestrichenen Oktave, siehe Tabelle 2.1), alle anderen Noten sind relativ zu dieser Note. Die Nummern der Oktaven sind willkürlich gewählt. Im Rahmen dieser Arbeit werden MIDI-Nummern für die Klavierwalzendarstellung verwendet.

2.3.3 Melodiekontur

Für QBH-Systeme hat sich die sogenannte *Melodiekontur* in vielen Fällen als ausreichend erwiesen. Durch das Weglassen von Information ist diese Melodiedarstellung weniger eindeutig als zum Beispiel eine Notendarstellung, und es lassen sich leichter ungefähr ähnliche Melodien finden. Ein mögliches Format ist der PARSONS-Code.

Parsons-Code

Eine Melodiekontur beschreibt, ob sich die Tonhöhe einer Melodie nach oben oder unten ändert oder gleich bleibt. Damit sind mindestens drei Symbole notwendig, mit der eine Kontur beschrieben werden kann, zum Beispiel „U“ für aufwärts (up), „D“ für abwärts (down) und „R“ für gleichbleibend (repeat) (siehe Tabelle 2.5). Die Darstellung durch U, D und R wird auch als PARSONS-Code bezeichnet [154].

Damit ist eine Beschreibung der Melodiekontur als Zeichenkette möglich, siehe Abbildung 2.4. Die Melodie von „As time goes by“ wird durch die Zeichenkette

* U D D D U U U D D D U U U D D D U

repräsentiert. Weil der ersten Note noch kein Intervall zugeordnet werden kann, notiert man dort einen Stern. Mittels solcher Zeichenketten werden für die Suche nach Ähnlichkeiten zwischen Melodien Zeichensuchverfahren verwendbar [119].



Parsons: *	U	D	D	D	U	U	U	D	D	D	U	U	U	D	D	D	U
Interval: *	1	-1	-2	-2	2	2	3	-2	-1	-2	3	2	5	-1	-2	-2	2
Contour: *	1	-1	-1	-1	1	1	2	-1	-1	-1	2	1	2	-1	-1	-1	1
Beat: 4	5	5	6	6	7	8	9	9	10	10	11	12	13	13	14	14	15

Abbildung 2.4: Notenbeispiel: As Time Goes By - Thema aus dem Film ‘Casablanca’, komponiert von HERMAN HUPFIELD. Drei verschiedene Melodiedarstellungen sind dargestellt: der PARSONS-Code, die *Interval-Methode* und die *MPEG-7-MelodyContour*-Darstellung.

Intervall-Methode

Statt lediglich die Richtung der Melodie anzugeben, ist natürlich auch die exakte Wiedergabe des melodischen Intervalls zwischen den Tönen der Melodie möglich. In Zahlenform bietet sich die Anzahl der Halbtonschritte innerhalb der Skala an, damit wird aus der Melodie in Abbildung 2.4 folgende Intervall-Darstellung möglich:

1 -1 -2 -2 2 2 3 -2 -1 -2 3 2 5 -1 -2 -2 2

Eine Variation der Intervall-Methode ist es, Intervalle größer als eine Oktave modulo 12 zu teilen; damit werden alle Intervalle in eine Oktave abgebildet [187].

MPEG-7 MelodyContour

Im Standard MPEG-7 steht das *Melody Description Scheme* zur Verfügung (siehe auch Kapitel 4) [100]. Die Konturdarstellung der darin enthaltenden *Melody-Contour* erfolgt nuancierter als im PARSONS-Code über 5 Stufen (siehe Tabelle 2.6):

1 -1 -1 -1 1 1 2 -1 -1 -1 2 1 2 -1 -1 -1 1

UITGENBORERD nennt in ihrer Arbeit die *erweiterte Kontur*, die ebenso wie die MPEG-7-MelodyContour in 5 Stufen arbeitet, allerdings in Buchstabendarstellung („UuDdR“) [187].

Tabelle 2.6: Die fünf Stufen der Kontur-Information in dem *MelodyContour Description Scheme* aus MPEG-7 [100].

Kontur-Wert	Änderung als Intervall
-2	eine kleine Terz oder mehr abwärts
-1	einen Halb- oder Ganzton abwärts
0	keine Änderung
1	einen Halb- oder Ganzton aufwärts
2	eine kleine Terz oder mehr aufwärts

Darüber hinaus werden im MPEG-7-*ContourType* aber auch Informationen über den Rhythmus festgehalten (vgl. Abschnitt 4.1.3). Damit können Distanzmaße verwendet werden, die sowohl Melodiekontur wie Rhythmus vergleichen [108]. In Abbildung 2.4 ist ein Beispiel der Anwendung der MPEG-7-*MelodyContour* dargestellt.

2.4 Singen von Melodien

In der Arbeit von LESAFFRE et al. wird das spontane Verhalten von QBH-System-Nutzern untersucht [120]. Im Rahmen des MAMI-Projektes (MAMI: Musical Audio Mining) wurden 30 Musikstücke aus verschiedensten Genres wie Pop-Musik, Chansons, Heavy-Metal, Kinderliedern und klassischer Musik ausgewählt. Am „Query-by-Voice“-Experiment waren 72 Probanden beteiligt. Das Ergebnis der statistischen Auswertung bezieht sich auf die Suchanfragen, die manuell – d. h. von Musikwissenschaftlern annotiert und analysiert – ausgewertet worden sind:

Zeitvorgaben Die durchschnittliche Anfrage dauert ungefähr 14 s und setzt 634 ms nach Start der Aufnahme ein.

Anfrageformen Es wurden sechs Anfrageformen angegeben, die am häufigsten gewählt sind das Singen von Liedtext und von Silben. Pfeifen ist die drittpopulärste Methode, während Summen, Klatschen oder Kommentare nur eine selten gewählte Anfrageform sind. Für 40 % aller Anfragen wurde eine Kombination von wenigstens zwei Anfrageformen gewählt.

Silben Die am häufigsten gesungenen Silben sind: /na/, /n@/, /la/, /t@/, /da/, /di/ und /d@/.

Einfluss von Alter, Geschlecht und musikalischer Erfahrung Der Einfluss von Alter, Geschlecht und musikalischem Hintergrund ist erheblich. Junge Leute tendieren dazu, ihre Anfrage früher zu beginnen und erzielen bessere Suchergebnisse. Männer starten ihre Ergebnisse später als Frauen und verwenden eine größere Anzahl von Silben zum Singen. Musiker stellen längere Anfragen als Nichtmusiker und verwenden weniger Text.

Über das Verhalten der Nutzer hinaus spielen natürlich die Fehler in einer Suchanfrage, die vom Nutzer gemacht werden, eine erhebliche Rolle. Typische Gesangsfehler von Nutzern von QBH-Systemen sind [41,135]:

Einfügungen und Auslassungen Noten werden der gesungenen Melodie zugefügt oder weggelassen. Diese Veränderungen sind häufig, können aber, wie gezeigt werden wird, auch durch den Transkriptionsvorgang verursacht werden.

Transponierung Die Suchanfrage wird in einer anderen Tonart oder einer anderen Oktave gesungen als das Original. Bei der Verwendung von Melodiekonturen ist dieser Fehler irrelevant.

Tempo Die Änderung des Tempos ist unerheblich, sofern es nicht als Suchkriterium verwendet wird.

Modulation Die Transponierung der Suchanfrage kann sich während des Anfrageverlaufs ändern. Für jede Modulation ergibt sich ein Fehler in der Melodiekontur.

Tempowechsel Der Sänger beschleunigt oder verlangsamt das Tempo (Rubato). Abhängig davon, wie die Rhythmuserkennung des Transkriptionssystems arbeitet (Tempoerkennung oder Tempoverfolgung, siehe Abschnitt 5.4.2), kann dieser Fehler abgefangen werden.

Einzelfehler Der Sänger singt eine Note falsch intoniert. Für die Suchanfrage ergibt sich ein Edierungsfehler.

Zusammenfassend kann man sagen, dass als Fehler eines QBH-System-Nutzers Einfügungen und Auslassungen, Modulationen, Tempowechsel und Einzelfehler auftreten. Diese führen bei der Transkription zu Fehlern in der Melodiedarstellung. Transponierungen oder ein falsch gewähltes Tempo hingegen werden durch die Darstellung einer Melodie als Melodiekontur abgefangen.

CARRÉ et al. untersuchen speziell Intonationsprobleme bei gesungenen Anfragen bei einem QBH-System [41], die also in die Kategorie „Einzelfehler“ der Untersuchungen von [135] fallen, bzw. in die Kategorie „Transponierung“, sofern sich die falsche Intonation in eine Richtung (nach oben oder unten) fortsetzt. Mehr als 25 % der falsch intonierten Intervalle liegen im Bereich eines Vierteltons. Es wird gezeigt, dass die Vermeidung der Quantisierung der Tonhöhe in diskrete Stufen der Melodiekontur den Sucherfolg bei QBH-Systemen signifikant verbessert, von der Verwendung von Melodiekonturen wird daher abgeraten. McNAB stellt in seinen Untersuchungen ebenfalls Intonationsprobleme fest, vor allem beim Singen von Intervallen, die größer als zwei Halbtonschritte sind [132]. Eine systematische Untersuchung liegt nicht vor. PAUWS testet Sänger anhand von Beatles-Songs und unterscheidet dabei geübte und ungeübte Sänger [150]. Er stellt dabei in seinen Untersuchungen u. a. fest, dass die Erstellung der Melodiekontur unabhängig davon ist, ob ein Sänger geübt ist oder nicht.

2.5 Zusammenfassung

QBH-Systeme sind Musiksuchsysteme, daher wurden in diesem Kapitel Begriffe der Musiktheorie zur Beschreibung von Musik dargestellt und in Bezug auf Musiksuchsysteme diskutiert. Wie erläutert wurde, ist der Begriff „Melodie“ sehr vielschichtig. Für QBH-Systeme kommen vor allem Melodien in Frage, die sich einfach singen lassen, wie zum Beispiel Volkslieder oder Pop-Musik – Klangfarbenmelodien sind über QBH-Systeme unzugänglich. Weiterhin sollen nur monophone Melodien betrachtet werden.

Die Repräsentation von Melodien kann in die Kategorien Audio-Repräsentation und Symbol-Repräsentation unterteilt werden. Die Überführung einer Melodie in Audio-Repräsentation in eine symbolische Darstellung beschreibt der für diese Arbeit bedeutsame Vorgang der Transkription. Als symbolische Darstellung für Melodien werden besonders Melodiekonturen gemäß des Standards MPEG-7 betrachtet. Darüber hinaus gibt es eine Fülle weiterer Möglichkeiten, um Melodien symbolisch zu beschreiben, z. B. die Notenschrift, MIDI-Informationen oder den PARSONS-Code. Weiterhin wurden die wichtigsten Begriffe aus dem Bereich der Musiktheorie erläutert, die ebenfalls zur Beschreibung von Musik geeignet sind, z. B. Tempobezeichnungen, die Akkordsymbolschrift für Akkordfolgen und Tonartbezeichnungen wie Dur und Moll. Es wurden technische Informationen zu Begriffen der Musiktheorie ge-

geben. Die Definition von musikalischen Intervallen und der musikalischen Temperatur ist notwendig, um die Transkription von Melodien aus Audiosignalen in Noteninformationen vorzunehmen.

Das Singen von Melodien ist ein nicht-technischer Aspekt für Melodiesuchsysteme, der am Ende dieses Kapitels beschrieben wurde. Da die Funktionalität von QBH-Systemen beurteilt werden soll, sind die Fehler, die bereits beim Stellen einer gesungenen Suchanfrage auftreten können, von großer Bedeutung. Der Literatur ist zu entnehmen, dass Fehler als Einfügung, Auslassung, Transponierung, geändertes Tempo und geänderte Tonart und als Einzelfehler enthalten sein können.

*EDV-Systeme verarbeiten,
womit sie gefüttert werden.
Kommt Mist rein, kommt Mist
raus.*

André Kostolany

Suchmaschinen wie *Google*, *Altavista* oder *Fireball* sowie Webverzeichnisse wie *Yahoo* oder *Web.de* sind wichtige Hilfsmittel bei der Suche im Internet [15]. Allen genannten Suchmaschinen und Suchdiensten ist gemeinsam, dass die Suchanfrage ausschließlich durch eine Textbeschreibung formuliert werden kann. Diese Anfrageform ist für Musik nur begrenzt sinnvoll, da in Text zu beschreibende Informationen wie Titel und Interpret nicht immer bekannt sind, Beschreibungen des musikalischen Genre, Liedtext oder andere Informationen nicht eindeutig sein müssen. Andere Beschreibung für Musik, die sich in Text ausdrücken lassen, sind die Akkordsymbolschrift, Tempobezeichnungen u. ä (siehe letztes Kapitel). Häufig ist ihre Verwendung aber nur für fachkundige Nutzer wie Musiker oder Musikwissenschaftler interessant.

Die Melodie als besonderes Kennzeichen der Musik lässt sich nicht ohne weiteres mit Text beschreiben. Stattdessen besteht die unmittelbarste Beschreibung einer Melodie darin, sie in einem musikalischen Vortrag wiederzugeben, im einfachsten Falle also zu singen. Eine solche Melodiebeschreibung kann von Query-by-Humming-Systemen (QBH-Systemen) verarbeitet werden.

Ebenso wie die o. g. Suchmaschinen werden QBH-Systeme häufig als Web-Applikation oder Anwendung mit Internetzugriff gestaltet, so dass es sich um eine auf Melodien spezialisierte *Suchmaschine* handelt. Suchmaschinen, die nicht nur für Melodien, sondern zum Auffinden von Musik überhaupt geeignet sind, bilden die Gruppe der *Music-Information-Retrieval-Systeme* (MIR-Systeme) oder kurz *Musiksuchsysteme*. Sie stellen damit eine QBH-Systemen übergeordnete Gruppe dar [59].

Die vorangegangenen Ausführungen machen bereits klar, dass die Melodie nicht die einzige mögliche Musikbeschreibung ist. Im folgenden Abschnitt dieses Kapitels werden daher verschiedene bereits bestehende MIR- bzw.

QBH-Systeme vorgestellt, und anhand dieser Beispiele werden allgemeine Unterscheidungskriterien und Merkmale solcher Systeme abgeleitet. In Abschnitt 3.2 werden dann die Zielbestimmungen für ein QBH-System festgelegt. Danach werden spezielle Anwendungen und mögliche Nutzergruppen von QBH-Systemen diskutiert. Für die verschiedenen Anwendungsfälle sind unterschiedliche Betriebsbedingungen denkbar, die im darauf folgenden Abschnitt dargestellt werden. Schließlich wird eine genaue Beschreibung der Funktionen gegeben, die für die Einbettung eines QBH-Systems in eine vernetzte Umgebung wie zum Beispiel das Internet wichtig sind.

3.1 Beispiele für Musiksuchsysteme

Dieser Abschnitt gibt eine Übersicht bestehender Musik- und Melodiesuchsysteme und häufig gewählter technischer Lösungen für Musiksuchsysteme. Zuerst werden einige beispielhafte Systeme vorgestellt, danach werden die gemeinsamen Merkmale und Unterscheidungskriterien diskutiert.

3.1.1 Musicline

Das System *Musicline* [11] ist ein kommerzielles Angebot der Phononet GmbH und wurde vom Fraunhofer-Institut IDMT entwickelt [72]. Es bietet ein Java-Applet an, das im Internetbrowser des Anwenders ausgeführt wird. Über das Java-Applet können gesummte Suchanfragen aufgenommen und an einen zentralen Server gesendet werden. Das Ergebnis der Suche erscheint in Listenform als Internetseite mit Kaufhinweisen für die einzelnen Titel.

Es werden etwa 3500 Titel angeboten, hauptsächlich aus dem Genre Pop-Musik. Der Inhalt der Datenbank wird aus MIDI-Dateien extrahiert. *Musicline* verwendet wie das im Rahmen dieser Arbeit implementierte System *Queryhammer* den Multimedia-Standard MPEG-7 [72].

3.1.2 notify!

notify!WhistleOnline ist ein Melodiesuchsystem, bei dem die Suchanfrage durch Pfeifen (query by whistle, QBW) gestellt und danach noch ediert werden kann [118]. Es wurde im Rahmen eines von der DFG geförderten Projektes zum Thema *Digitale Musikbibliotheken* entwickelt [14].

Die Suchmaschine ist als Client/Server-Applikation realisiert, d. h. auf dem Rechner des Nutzers wird ein eigenständiges Programm gestartet (Client), das eine Internetverbindung zur Melodiedatenbank (Server) herstellt. In der Melodiedatenbank sind ca. 2.000 Melodien enthalten, die per manueller Transkription aus Musikstücken extrahiert wurden.

3.1.3 Musipedia

Die offene Musik-Enzyklopädie *Musipedia* ist eine nach Melodien durchsuchbare, edierbare und erweiterbare Sammlung musikalischer Themen [12]. Es stehen mehrere Möglichkeiten zur Verfügung, um nach einer Melodie zu suchen: Textbeschreibungen für die Angabe eines Stichworts oder die direkte Eingabe des PARSONS-Codes. Abbildung 3.1 zeigt eine Bildschirmkopie des Systems.

Der PARSONS-Code der Suchanfrage kann aber auch mittels eines Java-Applets generiert werden, indem man die Melodie in ein an den Rechner angeschlossenes Mikrofon pfeift [154] oder ein alternatives Java-Applet mit Klaviatur verwendet. Die Suchphrasen in Textform können mit dem Parsons-Code verknüpft werden. Weiterhin lässt sich das Genre des gesuchten Titels zur Begrenzung der Trefferzahl spezifizieren. Das Ergebnis der Suche erscheint als Titelliste mit Notendarstellung der Melodie, sofern vorhanden.

Für die akustische Eingabe wird das System *Melodyhound* genutzt, das erstmals in den Arbeiten von RAINER TYPKE in [154] vorgestellt wurde. Die Datenbank enthält etwa 10.000 klassische Titel, 2400 Popmusikstücke und 17.000 Volkslieder [12]. Für registrierte Nutzer der Datenbank besteht die Möglichkeit, fehlende Titel selbst dem Datenbankbestand hinzuzufügen.

3.1.4 Vodafone-MusicFinder

Der Vodafone-MusicFinder ist ein kommerzielles System, das die Identifizierung von Musikstücken per Mobiltelefon ermöglicht [16]. Der Mobiltelefon-Nutzer muss einen Titel, den er gerade hört, für 30 s seinem Telefon vorspielen. Auf Basis der Daten, die durch diese akustische Stichprobe ermittelt werden können, wird in einer Musikdatenbank mit etwa 1,7 Millionen Titeln gesucht, eine Liste der besten Treffer wird als Kurztext-Nachricht (short message service, SMS) zurück zum Mobiltelefon gesendet. Da bei der Untersuchung der akustischen Stichprobe eine Art Fingerabdruck des Musiksignals erstellt wird, wird dieses Verfahren auch als „Audio-Fingerprinting“ bezeichnet.



Abbildung 3.1: Die Benutzerschnittstelle der Internetseite *Musipedia* (Bildschirmkopie). Als Suchanfrage können der Parsons-Code und Suchwörter eingegeben werden. Weiterhin lassen sich Java-Programme für die akustische Eingabe (Query-by-Wistling) oder über eine Klaviatur aufrufen.

3.1.5 Weitere Systeme

Mit den gerade beschriebenen Systemen wurden einige typische Systeme und deren Merkmale vorgestellt. Weitere bestehende QBH-Systeme mit ähnlichen Merkmalen sind *Cubyhum* [149], *MELDEX* [131], *QBH* [76], *Search by Humming* [32], *Sound Compass* [116] oder *Super MBox* [101].

3.1.6 Merkmale

Die vorgestellten MIR-Systeme sind aus verschiedenen Motivationen konzipiert worden und dienen verschiedenen Anwendungen. Sie lassen sich unter folgenden Gesichtspunkten vergleichen (siehe auch [183]):

Zugriffsmethoden Die meisten MIR-Systeme werden dem Nutzer als Internetseite präsentiert; es ist aber auch möglich, ein spezielles Programm auf einem Rechner auszuführen. Weiterhin gibt es die Möglichkeit, einen Suchdienst über Mobil-Telefone anzubieten, die Kurztext-Nachrichten empfangen können.

Anfrageformen Die beschriebenen Systeme lassen akustische, aber auch symbolische Melodie- und Musikbeschreibungen zu. Diese verschiedenen Beschreibungen werden im nächsten Abschnitt genauer dargestellt.

Datenbank Der Inhalt der Datenbank des MIR-Systems ist das wichtigste Merkmal – Umfang, Form und Inhalt bestimmen das mögliche Ergebnis einer Recherche.

Während die Zugriffsmethoden allein von den technischen Gegebenheiten abhängen, sind Anfrageform und Datenbank auch von musikalischen Aspekten abhängig. Daher folgt nun eine eingehende Betrachtung dieser beiden Merkmale.

Anfrageformen

Es lassen sich verschiedene, auf das Ziel der Suche angepasste Anfrageformen unterscheiden:

Akustisch Query-by-Humming (QBH) ermöglicht die Suche nach einer Melodie durch Summen. Der Nutzer des Systems summt die gesuchte Melodie auf Silben wie „na“ oder „da“, was über ein Mikrofon aufgezeichnet

wird. Dieses Signal wird dann für die Suchanfrage weiterverarbeitet. Für einige Systeme wird auch Pfeifen als Vortragsform vorgeschlagen, in diesem speziellen Fall wird auch von Query-by-Wistling (QBW) gesprochen. Noch spezieller ist Anfrageform Query-by-Tapping (QBT) – der Nutzer gibt in diesem Fall nur den Rhythmus des gesuchten Stückes durch Händeklatschen oder rhythmische Tastatureingabe vor.

Beispiel Die Anfrageform Query-by-Example (QBE) bietet die Möglichkeit, dem Suchsystem eine Probe bzw. ein Muster eines vorhandenen Musikstücks darzubieten. Diese Probe kann zum Beispiel ein kurzer Ausschnitt eines Musiktitels sein, der als Audiodatei vorliegt. Da bei der Weiterverarbeitung der Probe eine Art Fingerabdruck ermittelt und in der Datenbank gesucht wird, spricht man auch von *Audio Fingerprinting*.

Text Durch Text lassen sich Titel, Interpret, Genre und weitere Information eines gesuchten Musikstücks beschreiben. Diese Anfrageform wird vor allem bei herkömmlichen Internetsuchsystemen verwendet.

Noten Da Musik über Notenschrift repräsentiert werden kann, ist die Eingabe von Noten selbstverständlich geeignetes Mittel zur Suchanfrage. Allerdings setzt die Eingabe von Noten bzw. einer Beschreibungssprache für Noten musikalische Grundkenntnisse voraus. Sie bietet sich damit einer weniger breiten Anwenderschicht an.

Alle genannten Anfrageformen lassen sich miteinander kombinieren, so ist beispielsweise ein gesummter Vortrag mit dem Texthinweis auf das Genre Popmusik geeignet, um die Trefferliste einer Datenbank Anfrage einzugrenzen. Nutzt man verschiedene Darstellungsformen zur Repräsentation der Suchanfrage, spricht man auch von *multimodalen* Systemen [89].

Datenbanken

Nach der Beschreibung der verschiedenen Anfrageformen werden nun die besonderen Merkmale der Datenbank in einem Musiksuchsystem dargestellt. Inhalt dieser Datenbank ist ein Vorrat von Musikstücken. Die Musikrepräsentation bzw. das Format des Datenbankinhalts bestimmen, auf welcher Abstraktionsebene die Suchanfrage durchgeführt werden kann. Besondere Merkmale von MIR-Datenbanken sind:

Datenbankformat Gemeint sind Angaben zum Format des Datenbankinhalts, in dem gesucht wird. Möglich sind beispielsweise *Audiodateien*, *Audio-Fingerabdrücke*, *Noten*, *MIDI-Dateien* oder *Melodiekonturen*. Häufig werden MIDI-Datenbanken verwendet [75, 76, 90, 116, 124, 130], damit ist die Suche nach einer symbolen Darstellung wie einer exakten Notenfolge oder einer Melodiekontur möglich.

Musikmerkmale Die Merkmale der Musik, die zur Datensuche herangezogen werden, können Tonhöhe, Melodiekontur, Harmoniedarstellung (gemeint sind Akordstufen wie Tonika, Subdominante, Dominante), Tonart, Taktart, Rhythmus oder Tempo sein. Für Melodiekonturen wird oft der PARSONS-Code mit drei Symbolen („UDR“) verwendet [12]. Eine Ausnahme stellt das von Lu et al. vorgestellte QBH-System dar [124], hier wird nur mit zwei Symbolen „UD“ gearbeitet.

Indizierung Die Indizierung des Datenbankinhalts dient der Beschleunigung der Suche. Einige für Melodiedatenbanken geeignete Indizierungstechniken werden in Kapitel 7 näher beschrieben.

Vergleich Mit dem Begriff *Vergleich* sind Verfahren und Methoden zum Ähnlichkeitsbestimmung gemeint. Sie hängen direkt vom Inhalt der Datenbank und von der Form der Anfrage ab; für Melodiekonturen geeignete Techniken werden ebenfalls in Kapitel 7 dargestellt.

Sammlung Angaben zur Art der Sammlung können sich auf das Genre beziehen (zum Beispiel speziell *Volksmusik*, siehe MELDEX [131]), oder auf den Komponisten (etwa das *Bachwerkeverzeichnis* oder das *Köchelverzeichnis* für die Kompositionen MOZARTS).

Größe der Datenbank Der Umfang des Datenbankinhalts, in dem gesucht werden kann, hat Einfluss auf den Sucherfolg einer Anfrage.

3.2 Zielbestimmung

Im Rahmen dieser Arbeit soll ein Beispielsystem implementiert und untersucht werden, das derzeitigen Systemen entspricht und die Bewertung üblicher Techniken für Melodiedarstellung und -vergleich zulässt. Die folgenden Abschnitte beschreiben, welchen Anforderungen ein QBH-System genügen muss und welchen zusätzlichen Anforderungen es genügen sollte, bzw.

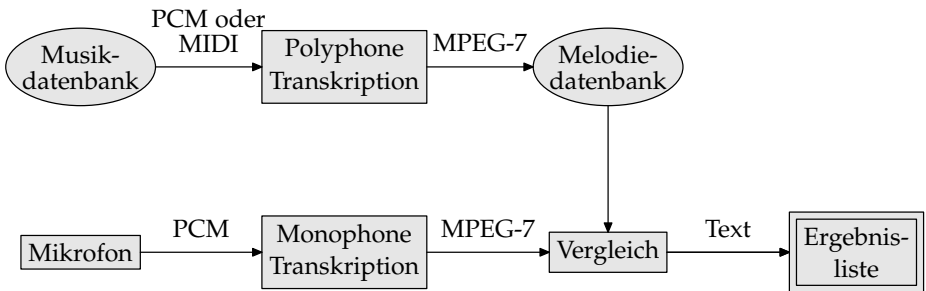


Abbildung 3.2: Schematische Darstellung eines QBH-Systems. Ein Mikrofon für die Anfrageaufzeichnung, die monophone Transkription und die Vergleichsstufe werden von den Musskriterien eines QBH-Systems gefordert. Die polyphone Transkriptionsstufe hingegen gehört zu den Wunschkriterien.

welche Anforderungen nicht unbedingt zu erfüllen sind. Dies führt zur Unterscheidung von Muss-, Wunsch- und Abgrenzungskriterien.

3.2.1 Musskriterien

Query-by-Humming ermöglicht es, eine gesummte Melodie als Suchanfrage für eine Recherche in einer Melodiedatenbank zu verwenden. Für ein QBH-System ergibt sich damit ein schematischer Aufbau wie in Abbildung 3.2 dargestellt. Die vom Nutzer gesummte Anfrage wird über das Mikrofon aufgezeichnet und durch die monophone Transkription weiterverarbeitet. Mittels der gefundenen symbolischen Darstellung wird in der Datenbank nach der ähnlichsten Melodie gesucht. Das Ergebnis der Suche wird in einer Liste von zum Beispiel den zehn besten Treffern dem Nutzer präsentiert.

Folgende Anforderungen sind daher zwangsläufig zu erfüllen:

- Um eine Suchanfrage zu starten, müssen alle notwendigen Parameter aus der Nutzereingabe extrahiert werden. Benötigte Informationen sind Tonhöhe, Anfänge der Noten und ihre Dauer. Die extrahierte Information muss in eine Melodiedarstellung übersetzt werden, die mit der Datenbank verglichen werden kann.

- Die Melodiedarstellung in der Datenbank muss für den Vergleich mit der Nutzeranfrage geeignet sein. Die Datenbank selbst muss ein hinreichend großes Korpus besitzen.
- Der Vergleich von Anfrage und Datenbankbestand soll eine Liste mit den ähnlichsten Melodien ergeben. Die Präsentation der ähnlichsten Melodien muss für den Nutzer die gewünschte Auskunft über Titel und ggf. Interpret darstellen, dabei werden die ähnlichsten Titel am besten platziert.

3.2.2 Wunschkriterien

Folgende Punkte sind für QBH-Systeme eine sinnvolle Ergänzung, müssen aber nicht unbedingt erfüllt werden:

- Transkription der Melodien der Melodiedatenbank

Die Melodiedatenbank enthält die Melodien, in denen die Anfrage gesucht wird. Diese Datenbank wird zweckmäßigerweise aus einem Bestand vorhandener Musikstücke transkribiert, die oft gleichzeitig den Gegenstand des Suchinteresses darstellen. Musikstücke sind häufig polyphon, so dass die Melodie aus mehreren Stimmen ausgewählt werden muss. Die in Abbildung 3.2 dargestellte polyphone Transkription ermöglicht die Generierung der Datenbankschlüssel (vergleiche Kapitel 7) aus den Musikstücken einer Musikdatenbank.

- Speicherung der transkribierten Nutzeranfrage

Mit der Möglichkeit, die Suchanfrage zu speichern, kann der Nutzer seine Recherche zu einem späteren Zeitpunkt wiederholen, zu dem etwa der Datenbankumfang vergrößert worden ist. Auch die Änderung des verwendeten Distanzmaßes oder die Benutzung eines anderen Suchsystems mit der gleichen Anfrage sind dann möglich.

- Visualisierung der Nutzeranfrage

Die Nutzeranfrage wird in eine symbolische Darstellung gebracht, eine sinnvolle grafische Darstellung ermöglicht die Kontrolle der Eingabe, ggf. ist auch die Korrektur der Anfrage auf grafischem Wege möglich.

3.2.3 Abgrenzungskriterien

Ein QBH-System dient dem Auffinden von Melodien durch gesummte Suchanfragen in einer Musikdatenbank. Im Rahmen dieser Arbeit soll ausschließlich diese Anfrageform untersucht werden. Darüberhinaus sind natürlich weitere Beschreibungsmöglichkeiten für Melodien vorhanden, beispielsweise durch die Angabe eines Genre, Namen des Interpreten oder Komponisten. Diese zusätzlichen Angaben führen zu multimodalen MIR-Systemen. Echtzeitaspekte werden in den Untersuchungen dieser Arbeit ebenfalls außer Acht gelassen, da alle wesentlichen Aspekte eines QBH-Systems durch Simulationen beleuchtet werden können.

3.3 Einsatz von QBH-Systemen

Der Einsatz von QBH-Systemen ist abhängig vom Anwendungsbereich, für den die Melodiesuche vorgenommen wird. Abhängig vom Nutzer kann wiederum die Anwendung unterschiedlich ausfallen.

3.3.1 Anwendungsbereich

Die Anwendung eines QBH-Systems ist die Melodiesuche, die durch verschiedene Ziele motiviert sein kann:

- Zur *Suche* bzw. zum Abruf von Melodien, zu denen alle weiteren Informationen wie Titel, Interpret oder Komponist unbekannt sind.
- QBH-Systeme können zum Finden von den einer bekannten Melodie *ähnlichen* Melodien eingesetzt werden.
- Schließlich ist es möglich, die *gleiche* Melodie in verschiedenen Musikstücken zu finden.

Der Einsatz von QBH-Systemen ist sowohl im privaten wie auch im kommerziellen Bereich denkbar und sinnvoll.

3.3.2 Zielgruppen

Je nach Bedürfnissen und Kenntnissen unterteilt sich die Gruppe der Anwender in Laien und professionelle Nutzer, d. h. also Nutzer mit besonderen Vorkenntnissen. Laien profitieren besonders von der Möglichkeit, eine gesuchte

Melodie nur summen zu brauchen und ohne die Angabe einer speziellen Melodiedarstellung danach recherchieren zu können. Für professionelle Nutzer kommen über die gesummte Melodieeingabe hinaus auch die Noteneingabe oder andere Formate in Frage. Folgende Gruppen von QBH-Nutzern können unterschieden werden:

Konsumenten sind Internetbenutzer, die mithilfe eines QBH-Systems nach einem Musiktitel recherchieren wollen, um ihn zum Beispiel auf einem Tonträger oder als Datei zu erwerben. Demgemäß werden QBH-Systeme bereits im kommerziellen Musikvertrieb angeboten, vgl. das Angebot von *Musicline* [11].

Musikwissenschaftler Über den Konsum hinaus ist aber auch die professionelle Verwendung von QBH-Systemen denkbar, die zum Beispiel für Musikwissenschaftler zur Analyse großer Musikbestände nützlich sein können.

Juristen Die Justiz hat bei der Verfolgung von Urheberrechtsverstößen ebenfalls ein Interesse am Überblick über einen großen Musikbestand.

Komponisten können sich durch QBH-Systeme neue Inspirationen verschaffen oder die Verwendbarkeit eigener Melodien durch den Vergleich mit ähnlichen Melodien in bestehenden Kompositionen untersuchen.

Bibliothekare Die Verwaltung großer Musik- und Notenbestände ist bislang auf die Textbeschreibung beschränkt; mit Melodiesuchsystemen ist Ordnung und Suche auf Basis des Datenbestands selbst möglich.

3.3.3 Betriebsbedingungen

Normalerweise wird ein QBH-System als Anwendung implementiert werden, die auf einem Computer ausgeführt oder als Internetanwendung über einen Server angeboten wird. Da zu einem QBH-System eine große Datenbank gehört, wird dieser Teil in den meisten Fällen über ein Netzwerk an das System gebunden sein. Die Vergleichsstufe kann ebenso als Teil der Datenbank betrachtet werden, so dass der Datentransfer sich auf die Übertragung der extrahierten, symbolischen Suchanfrage zur Datenbank hin und der Ergebnisliste zur Nutzeranwendung zurück beschränkt.

3.4 Umgebung eines QBH-Systems

In diesem Abschnitt wird die Umgebung eines QBH-Systems für den Fall beschrieben, dass es in einer vernetzten Umgebung als Computeranwendung betrieben wird.

3.4.1 Software

Viele bestehende QBH-Systeme erwarten eine Java-Umgebung (Java Runtime Environment, JRE) für die Ausführung eines Java-Applets im Browser. Prinzipiell ist jede Programmiersprache zur Implementierung eines QBH-Systems möglich. Für mobile Geräte stehen insbesondere Fragen der Effizienz in Bezug auf Speicher- und Energieverbrauch im Vordergrund. Das im Rahmen dieser Arbeit implementierte QBH-System *Queryhammer* ist ein Versuchssystem und basiert daher ausschließlich auf der Interpretersprache *Matlab*.

3.4.2 Hardware

In den meisten Anwendungsfällen wie auch im Rahmen dieser Arbeit wird ein QBH-System auf einem gewöhnlichen Personal-Computer (PC) laufen. Für die Audioaufzeichnung kann eine gewöhnliche Soundkarte verwendet werden. Darüberhinaus ist natürlich die mobile Anwendung von Interesse, als Hardware-Plattform dafür kommen „Pocket-PCs“ oder Mobiltelefone in Frage.

3.4.3 Orgware

Bei internetbasierten Anwendungen muss ein Internetanschluss zur Verfügung stehen, in lokalen Netzwerken wie etwa innerhalb einer Bibliothek zumindest ein lokaler Netzwerkzugang (local area network, LAN), der den Zugriff zum Datenbanksystem ermöglicht.

3.4.4 Schnittstellen

Die Ein- und Ausgabeschnittstellen eines QBH-Systems bestimmen die Flexibilität der Anwendbarkeit des Systems. Im einfachsten Fall ist nur die akustische Eingabe möglich. Beim System *Queryhammer* können darüberhinaus auch

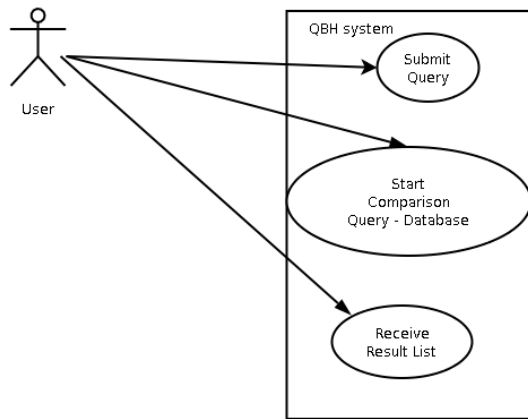


Abbildung 3.3: Ein Anwendungsfall-Diagramm für ein QBH-System. Aktionen des Anwenders sind Freigabe der Anfrage (submit) und Start der Suche.

Audiodateien, Text, MPEG-7-codierte Beschreibung im XML-Format oder Noten im MIDI-Format verarbeitet werden. Prinzipiell eignet sich jede Form der Melodiebeschreibung. Für die Ausgabe wird üblicherweise Text verwendet, der besonders formatiert sein kann und Multimedia-Verknüpfungen enthält. *Queryhammer* erzeugt Textausgaben im HTML-Format. Mögliche Formate und Techniken wie SMIL oder MPEG-21 werden in Kapitel 4 dargestellt.

3.5 Funktionen

Die notwendigen Funktionen eines QBH-Systems werden in diesem Abschnitt anhand von Anwendungsfall- und Sequenzdiagramm dargestellt.

Abbildung 3.3 zeigt ein Anwendungsfalldiagramm für ein QBH-System. Der Anwender summt die Melodieanfrage, die dann zur Weiterverarbeitung freigegeben werden muss (submit). Die daraus extrahierte Melodiekontur wird dann zur Datenbanksuche verwendet (start). Das Ergebnis wird als Liste zurückgegeben und muss dem Anwender zugänglich gemacht werden, üblicherweise durch automatische Anzeige auf dem Bildschirm.

Abbildung 3.4 zeigt das Sequenzdiagramm eines QBH-Systems. Die gesummte Anfrage wird zuerst transkribiert, die gewonnene Melodiekontur wird

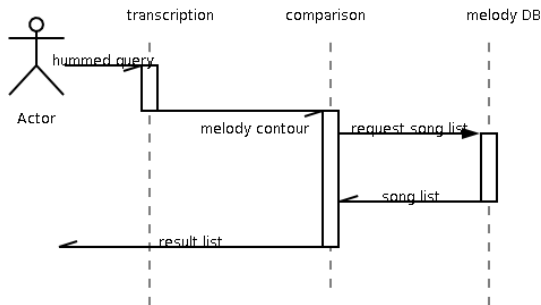


Abbildung 3.4: Sequenz-Diagramm eines QBH-Systems: die ablaufenden Prozesse sind Transkription, Vergleich und Zugriff auf die Melodiedatenbank.

zur Suche an die Datenbank weitergegeben. Das Ergebnis der Suche wird an den Nutzer zurückgegeben. Bei dem Beispielsystem *Queryhammer* lassen sich alle Schritte einzeln aufrufen.

3.5.1 Benutzeroberfläche

Die Benutzeroberfläche eines QBH-Systems ermöglicht die Eingabe der Suchanfrage, ggf. deren Modifikation sowie den Abruf des Suchergebnisses. Die Ausführung der Benutzeroberfläche hängt direkt von den technischen Möglichkeiten des Systems ab.

Queryhammer ist ein Entwicklungssystem, daher sind über die graphische Benutzeroberfläche auch die internen Darstellungen der Melodiekontur zugänglich (vergleiche Abbildung 3.5). Neben der Wellenform der Suchanfrage lassen sich extrahierte Grundfrequenz, MIDI-Ereignisse und Melodiekonturwerte im MPEG-7-Format anzeigen.

3.6 Zusammenfassung

Am Beginn dieses Kapitels stand der Überblick über bestehende Musiksuchmaschinen und die Darstellung ihrer speziellen Merkmale. Zuerst wurden Beispiele einiger bestehender Musiksuchsysteme wie z. B. *Musicline*, *Musipedia* oder *vodafone-Musicfinder* dargestellt. Ausgehend von dieser Übersicht wurden

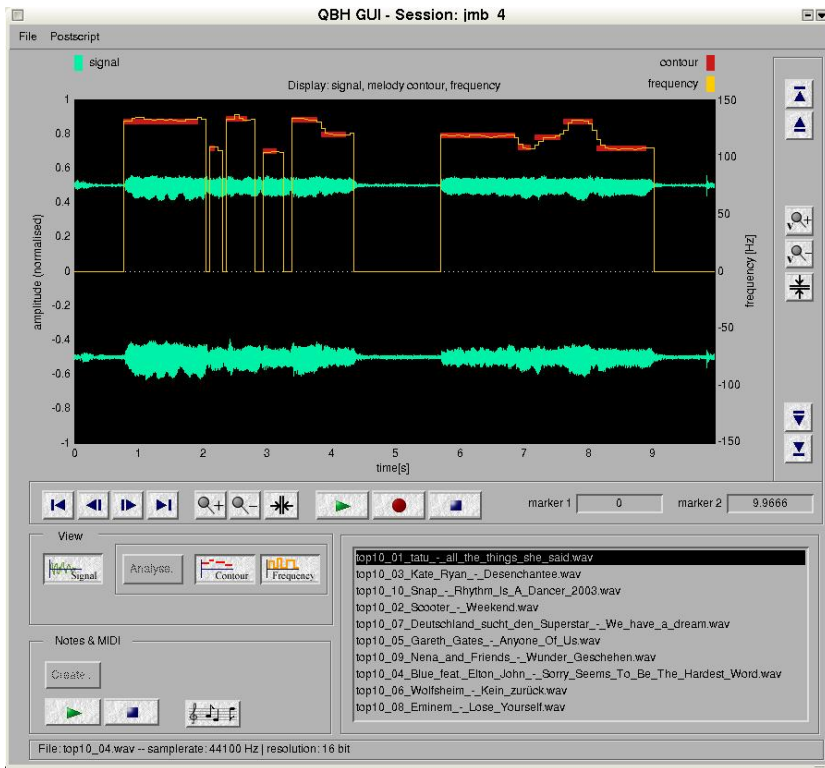


Abbildung 3.5: Die Nutzerschnittstelle des QBH-Systems *Queryhammer* [28].

die wichtigsten Merkmale dargestellt, sie gliedern sich in verschiedene Zugriffsmethoden, Anfrageformen und Datenbankmerkmale. Die am häufigsten gewählte Zugriffsmethode für Musiksuchmaschinen ist die Präsentation des Systems über eine Internetseite. Bei den Anfrageformen kann zwischen der Eingabe von akustischen Signalen, Musikstichproben, Text und Noten unterschieden werden. Query-by-Humming ist demgemäß eine akustische Anfrageform. Die verschiedenen Merkmale der Datenbanken sind das Format des Datenbankinhalts, zur Suche herangezogene Musikmerkmale, Indizierungstechniken, Abgleichmethoden, Art des Datenbankbestands sowie Größe des Datenbankinhalts.

In den Zielbestimmungen für das Beispielsystem *Queryhammer*, an dem im Rahmen dieser Arbeit alle eigenen Untersuchungen durchgeführt werden, wurden Muss- und Wunschkriterien für die Implementierung eines solchen Systems dargestellt. Danach wurden für den Einsatz von QBH-Systemen verschiedene Motivationen der Nutzer diskutiert. Die Unterteilung verschiedener Nutzergruppen von QBH-Systemen wurde grob in professionelle und nicht-professionelle Nutzer vorgenommen. Genauer ist eine Unterteilung in Konsumenten, Musikwissenschaftler, Komponisten, Juristen und Bibliothekare möglich.

Zum Schluss wurden die technische Umgebung eines QBH-Systems in Form von Hard-, Soft- und Orgware beschrieben und die zu implementierenden Funktionen in Form von Anwendungsfall- und Sequenzdiagrammen beschrieben. Das Beispielsystem *Queryhammer* wird als Matlab-Programm implementiert, das Audio-, MIDI- und MPEG-7-Dateien einlesen kann.

*Standards are good – let's have
many of them.*

anonym

Der Austausch elektronischer Daten bedarf einer genau festgelegten Darstellung, wenn Systeme verschiedener Urheber daran beteiligt sind. Auch für Daten, die von Query-by-Humming-Systemen (QBH-Systemen) verarbeitet werden, ist eine standardisierte Darstellung der verarbeiteten Daten wünschenswert. Für QBH-Systeme bietet sich besonders der Standard MPEG-7 an, da dieser Standard speziell auf die Anwendung in Multimedia-Netzwerken abgestimmt ist und die Beschreibung von Metadaten ermöglicht. Während sich die Metadatenbeschreibung in MPEG-7 im Wesentlichen auf die Inhalte generischer Audio- und Videoform konzentriert, gibt es aber auch andere Metadatenstandards für verschiedene Typen von Informationen in diversen Anwendungsbereichen. Beispiele sind *SMIL*, *SMPTE*, *EBU*, *TV-Anytime*, *DIG-35*, *Dublin Core* oder *OCLC/RLG* [125]. Besonders der Standard *SMIL* ist für Anwendungen wie QBH-Systeme geeignet. Darüberhinaus ist bezüglich der Vernetzung und des Zugriffs auf im Netz verteilte Ressourcen auch der Standard MPEG-21 interessant.

In den folgenden Abschnitten dieses Kapitels sollen nun die für QBH-Systeme relevanten Teile der Standards MPEG-7, MPEG-21 und *SMIL* näher dargestellt und diskutiert werden.

4.1 MPEG-7

Das World-Wide-Web (WWW) kann als „gigantische globale Datenbank“ [156] betrachtet werden, deren Daten durchsucht, sortiert und nach bestimmten Kriterien ausgewählt werden können. Durch die Hypertext-Markup-Language (HTML) und über das zugehörige Hypertext-Transfer-Protokoll (HTTP) ist diese Information praktisch jedermann zugänglich.

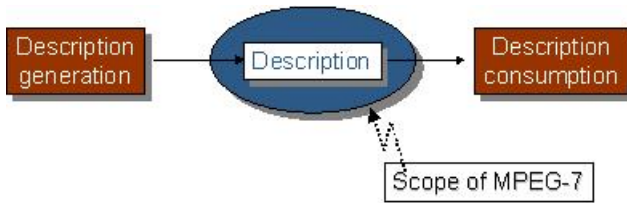


Abbildung 4.1: Der Fokus des Standards MPEG-7 ist die Beschreibung von AV-Inhalten, nicht wie bei den vorangegangenen Standards MPEG-1–4 die Beschreibung bzw. Codierung der Daten selbst. Quelle: [128]

Nicht textbasierte Dokumente wie nach MPEG-1, MPEG-2 oder MPEG-4 komprimierte Audio-Dateien können jedoch nicht bei diesem Verfahren berücksichtigt werden, da sie von Suchmaschinen nicht gelesen werden können. Diese Lücke schließt MPEG-7.

MPEG-7 definiert eine Schnittstelle zur Beschreibung von Multimedia-Inhalten (Multimedia Content Description Interface). Es steht damit in der Folge einer Reihe erfolgreicher ISO/IEC-Standards, die von der MPEG-Gruppe entwickelt wurden [46]. Während die vorangegangenen Standards MPEG-1, -2 und -4 auf die Codierung und Darstellung audiovisueller Informationen (AV-Informationen) abzielen, geht es bei MPEG-7 um die Beschreibung von Multimedia-Inhalten (Abbildung 4.1).

Als relativ junger Standard wird MPEG-7 derzeit noch weiterentwickelt. Übersichtsartikel zu MPEG-7 finden sich in [121, 128, 182] und [125]. Informationen zu Anwendungen, insbesondere zu QBH-Systemen, findet man in [107] oder [65]. Am Fachgebiet Nachrichtenübertragung wurden in Bezug auf MPEG-7 und QBH-Systeme ebenfalls einige Arbeiten durchgeführt [66, 91, 180, 196]. Eigene Arbeiten zu diesem Thema sind [27, 28, 68].

4.1.1 Anwendungsbereiche

Gegenstand des Standards MPEG-7 ist die übergreifende Beschreibung von AV-Inhalten, um das Zusammenwirken verschiedener Systeme und Anwendungen für die Erzeugung, Verwaltung, Verteilung und den Konsum von Multimedia-Inhalten zu ermöglichen (siehe Abbildung 4.2). Es ist eine Fülle von Anwendungen für MPEG-7 möglich:

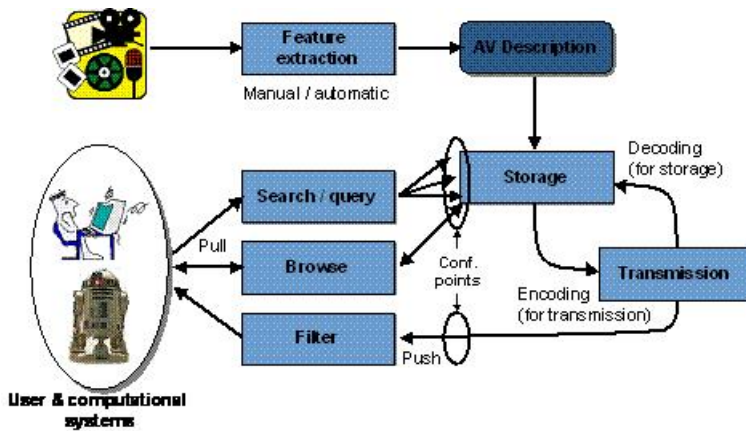


Abbildung 4.2: Eine abstrakte Darstellung möglicher Anwendungen von MPEG-7: Verschiedene Systeme zur Erzeugung, Verwaltung, Verteilung und zum Konsum von Multimedia-Inhalten verwenden standardisierte AV-Beschreibungen. Quelle: [128]

- In **Multimedia-Systemen** wird eine auf den Nutzer zugeschnittene Programmdarstellung angeboten, abhängig von seinen Vorlieben und der bisherigen Anwendung.
- Für Sammlungen oder einzelne Elemente in **Archiven** können Beschreibungen von AV-Inhalten erzeugt und ein nahtloser Austausch zwischen Eigentümern, Verteilern und Verbrauchern ermöglicht werden.
- Über Filter kann die **Anpassung** von Multimedia-Datenströmen an in ihren Ressourcen begrenzte Medien erfolgen, wie etwa bei mobile Netzwerken (UMTS, WLAN).
- Über die Eingabe von Gesang, Musikstücken oder Noten kann im Bereich **Audio/Musik** eine Suchanfrage gestellt werden.
- Die Suche von **Bildern/Grafiken** ist entsprechend mithilfe von Skizzen möglich.
- Bei gegebenen Videoobjekten ist die Beschreibung von **Bewegungen** möglich und damit auch die Suche nach Filmen und Animationen mit zuzuordnenden Bewegungen.

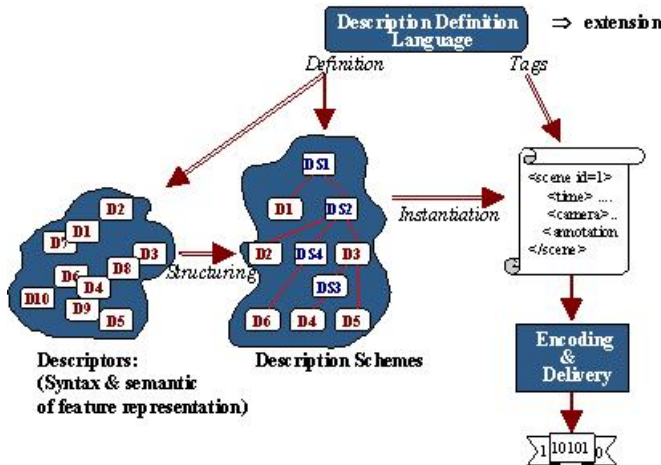


Abbildung 4.3: Die Kernelemente des MPEG-7-Standards: D, DS und DDL.
Quelle: [128]

- **Szenenbeschreibungen** über AV-Inhalte ermöglichen die Suche nach ähnlichen Szenarien.

Die Anwendung von MPEG-7 auf ein QBH-System ist nach Betrachtung dieser Anwendungsfälle unmittelbar einleuchtend. Der Standard beschreibt eine Reihe normativer Elemente, die hier kurz erläutert werden. Mit *Deskriptoren* (D) wird die Syntax und Semantik von AV-Inhalten festgelegt. Auf unterster Ebene werden zum Beispiel für Videosignale Umriss, Farbe, Bewegung und Textur beschrieben, für Audiosignale Merkmale wie Energie, Harmonizität oder Timbre. Allgemein spricht man von „Low-Level“-Deskriptoren. Auf höherer Abstraktionsebene können Merkmale abhängig vom Inhalt beschrieben werden, damit handelt es sich um „High-Level“-Deskriptoren. Generische D beschreiben allgemeine Merkmale. Über *Beschreibungsschemata* (Description Schemes, DS) können komplexere Beschreibungen erstellt werden, indem man Struktur und Semantik der Beziehungen mehrerer D und DS festlegt. Die Definition von D und DS erfolgt mithilfe der *Description Definition Language* (DDL). In Abbildung 4.3 sind die Abhängigkeiten der Kernkomponenten D, DS und DDL graphisch dargestellt.

4.1.2 Gliederung des Standards

MPEG-7 ist ein sehr umfangreicher Standard, daher soll zur besseren Orientierung eine kurze Übersicht über die einzelnen Abschnitte des Standards gegeben werden:

Part 1 – Systems Der Systemteil spezifiziert Funktionalitäten auf Systemebene zur Bereitstellung von Beschreibungen gemäß MPEG-7 und stellt den effizienten Transport und die Synchronisation der Inhalte mit den Beschreibungen sicher [24].

Das Konzept der „Systeme“ hat sich seit der Festlegung der MPEG-1 und MPEG-2 Standards dramatisch entwickelt [24]. Bislang bezogen sich „Systeme“ nur auf Fragen der Architektur, auf Multiplexing und Synchronisation. In MPEG-4 kamen Aspekte der Szenenbeschreibung, Inhaltsbeschreibung und Programmierbarkeit hinzu. In MPEG-7 werden neue Aspekte dem Systembegriff zugeordnet, zu nennen sind hier die Sprachvereinbarung für Beschreibungen, ihre Binärdarstellung und Übertragung.

Part 2 – Description Definition Language Die zur Beschreibung und Definition der Deskriptoren verwendete *Description Definition Language* (DDL) beschreibt Teil 2 des Standards [98].

Die DDL stellt einen der Hauptbestandteile des Standards MPEG-7 dar und bildet die Grundlage aller DS und D. Die DDL definiert syntaktische Vorschriften, um DS und D auszudrücken, zu kombinieren, zu erweitern und verfeinern. Die DDL ist keine Modellierungssprache wie etwa die *Unified Modelling Language* (UML), sondern ein Schema zur Darstellung von AV-Daten. Hervorzuheben ist, dass mit der DDL die Validierung – also die Prüfung auf Gültigkeit und Korrektheit – von MPEG-7-Daten möglich ist.

Part 3 – Visual Die visuellen Deskriptoren und Beschreibungsschemata werden in diesem Teil zusammengefasst. Dabei geht es um Merkmale wie Farbe, Textur, Form und Bewegung [176]. Durch den Vergleich der Deskriptoren wird es möglich, Ähnlichkeiten von Bildern oder Filmen auf Grundlage von visuellen Kriterien zu ermitteln.

Part 4 – Audio Im Audioteil werden Deskriptoren und Beschreibungsschemata für Audioinhalte erläutert [156]. Dieser Teil ist für die vorliegende

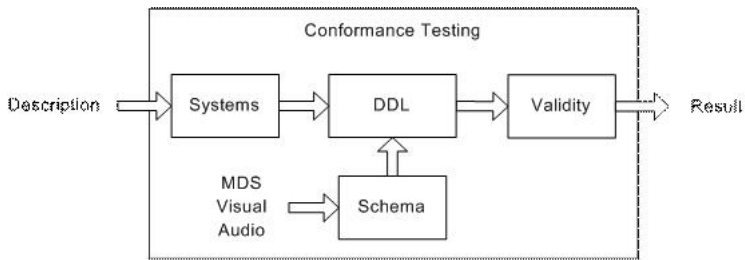


Abbildung 4.4: Übersicht über den Konformitätstest von Deskriptoren.
Quelle: [128]

Arbeit besonders wichtig und wird in Abschnitt 4.1.3 ausführlicher beschrieben.

Part 5 – Multimedia Description Schemes Die AV-Inhalte werden mit Hilfe von *multimedialen Beschreibungsschemata* (Multimedia Description Schemes, MDS) beschrieben, die speziell auf die Kombination von Audio- und Videoinformation zugeschnitten sind. Teil 5 des Standards stellt hierfür das Rahmenwerk dar.

Part 6 – Reference Software MPEG-7 bietet mit dem *eXperimentation Model* (XM) eine Referenzimplementierung an. Die XM-Software ist eine Simulationsplattform für D, DS, CS und die DDL [128]. Neben den normativen werden auch nicht-normative Komponenten benötigt, um vorhandene Datenstrukturen mit Programmen verarbeiten zu können. Datenstrukturen und ausführbare Programme werden als Applikation betrachtet. XM-Applikationen werden unterschieden in Server- und Klientenanwendungen, die für Extraktion und Suche, Filterung oder Umkodierung verwendet werden.

Part 7 – Conformance Der Teil *Conformance* beinhaltet Richtlinien zur Prüfung der Konformität sowohl für Implementierungen von Deskriptoren als auch für Anwendungen [128].

In Abbildung 4.4 wird ein Überblick über die Überprüfung der Konformität von Deskriptoren gegeben. Der Test besteht aus zwei Stufen, dem Systemtest und dem DDL-Test. Der Systemtest beinhaltet die Decodierung der zu überprüfenden Deskriptoren, die in Text- oder Binärform

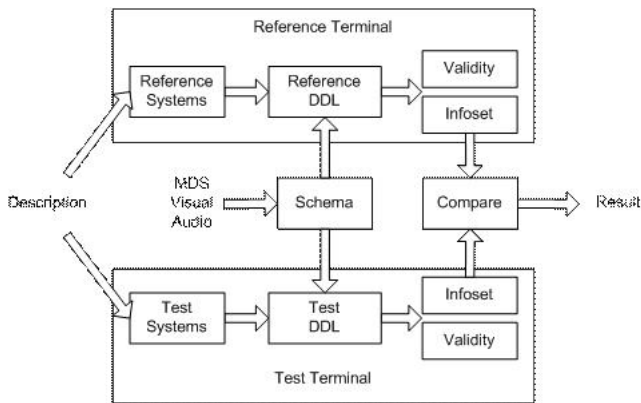


Abbildung 4.5: Übersicht über die Konformitätsprüfung von Anwendungen. Quelle: [128]

vorliegen können. Die Decodierung ergibt ein konformes XML-Textdokument. Der DDL-Test besteht in der Überprüfung des Textes auf Wohlgeformtheit und Validität gemäß des verwendeten Schemas.

In Abbildung 4.5 ist der Konformanztest von Anwendungen dargestellt. Es wird die zu überprüfende Anwendung mit einer Referenzanwendung verglichen. Eine Beschreibung wird in beide Systeme eingespeist, beim Ergebnis wird überprüft:

1. Ist die Antwort des zu überprüfenden Systems gültig im Sinne der Validität?
2. Enthält die Antwort die gleichen Ergebnisse wie die Referenzimplementierung?

Part 8 – Extraction and Use Der Teil *Extraction and Use* enthält Beispiele zur Extraktion und Verwendung von MPEG-7-Beschreibungen, die D und DS verwenden [128].

Part 9 – Profiles Unter diesen Punkt fallen Sicherstellung der Interoperabilität, Kompatibilitätsprüfungen, Tests und Konformitätsprüfungen. Dieser Teil des Standards befindet sich im Entwicklungsstadium.

Part 10 – Schema Definition Dieser Teil bezeichnet einen Datenträger, die MPEG-7 *Schema Definition CD-ROM*.

4.1.3 Auditive Inhaltsbeschreibung (Part 4)

In diesem Abschnitt werden speziell alle D und DS beschrieben, die für die Melodiebeschreibung geeignet sind.

Beschreibungsschema Melodiekontur

Das MPEG-7 *Melody DS* bietet eine Darstellung für Melodieinformationen, die unter den Aspekten Effizienz, Robustheit und Eignung zur Ähnlichkeitssuche entworfen worden ist. Dabei muss eine Eingrenzung des Melodiebegriffs erfolgen; für den Standard MPEG-7 sind Melodien immer monophon und tonal (vergleiche Abschnitt 2.1).

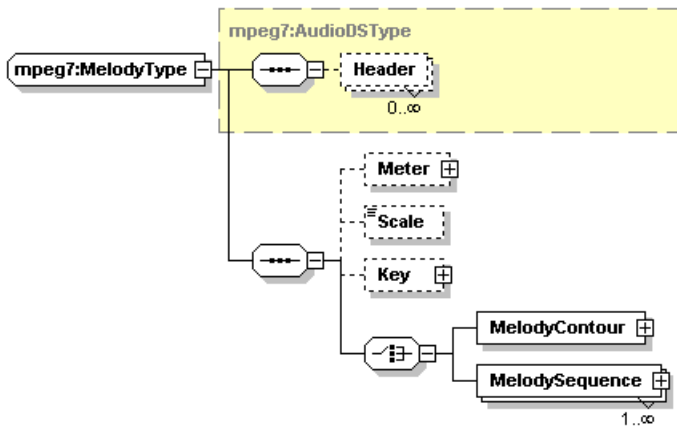


Abbildung 4.6: Datenstruktur des MPEG-7-MelodyType. Quelle: [125]

MelodyType Die Struktur des MPEG-7-MelodyType ist in Abbildung 4.6 dargestellt. Enthalten sind Informationen zu Taktart (meter), Skala (scale) und Tonart (key) der Melodie. Die Darstellung der Melodie selbst findet sich entweder im Feld *MelodyContour* oder im Feld *MelodySequence*. Das Feld *Header* ist optional; die Einträge im Einzelnen sind:

- *Meter*: Die Taktart wird durch den *MeterType* ausgedrückt (Angabe optional).

- *Scale*: Die verwendete Skala wird über einen Vektor mit Halbtonschritten dargestellt (Angabe optional).
- *Key*: Diese Struktur enthält Angaben zur Tonart (Angabe optional).
- *MelodyContour*: eine Struktur mit dem *MelodyContourType* (alternative Auswahl statt *MelodySequence*).
- *MelodySequence* eine Struktur mit dem *MelodySequenceType* (alternative Auswahl statt *MelodyContour*).

Diese einzelnen Einträge und die verwendeten Typen werden nun im Detail erläutert.

Meter Das Feld *Meter* enthält Angaben zur Taktart (siehe Abschnitt 2.2.2). Die Angabe ist ausschließlich als Bruch mit Zähler und Nenner möglich, zum Beispiel 4/4. Eine Abkürzung wie „C“ ist nicht möglich. Die Datenstruktur des *MeterType* ist in Abbildung 4.7 dargestellt. Er besteht aus:

- *Numerator*: Zähler mit Werten 1–128,
- *Denominator*: Nenner, die Werte sind auf Zweierpotenzen festgelegt: $2^0, \dots, 2^7$, d. h. 1, 2, $\dots$, 128.

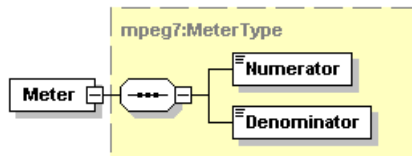


Abbildung 4.7: Datenstruktur des MPEG-7 *MeterType*. Quelle: [125]

Beispiel: Taktarten wie 5/4, 3/2, 19/16 können ohne weiteres mit MPEG-7 dargestellt werden. Komplexe Taktarten wie 3+2+3/8 können lediglich vereinfacht, in diesem Beispiel als 8/8 dargestellt werden:

```

<Meter>
  <Numerator>8</Numerator>
  <Denominator>8</Denominator>
</Meter>
  
```



Abbildung 4.8: Datenstruktur des MPEG-7 *ScaleType*. Es handelt sich um einen einfachen Vektor, mit dem sich die Frequenzen der einzelnen Skalentöne berechnen lassen. Quelle: [125]

Scale Der *Scale* Deskriptor enthält eine Folge von Intervallen, die die Oktave unterteilen. Aus den Intervallen lassen sich dann alle Grundfrequenzen der zur Skala (siehe Abschnitt 2.2.7) gehörigen Töne berechnen. Ausgehend von der Grundfrequenz F_0 des Grundtons der Skala lassen sich die Frequenzen aller weiteren Töne über

$$f(n) = F_0 2^{\frac{n}{12}} \quad (4.1)$$

ausrechnen. Die Information des *Scale*-Deskriptors kann für Referenzzwecke genutzt werden. Die Struktur des *ScaleType* ist einfach ein Vektor, der Fließkommazahlen enthält, so wie in Abbildung 4.8 gezeigt.

- *Scale*: Der Vektor enthält die Parameter n aus Gleichung 4.1. Verwendet man die ganzen Zahlen 1–12, so erhält man die gleichstufig temperierte chromatische Skala, die auch die Voreinstellung des *Scale*-Vektors ist. Um die Werte $s(n)$ aus angegebenen Frequenzen $f(n)$ zu berechnen, verwendet man folgende Gleichung:

$$s(n) = 12 \log_2 \left(\frac{f(n)}{f_0} \right). \quad (4.2)$$

Beispiel: Die Voreinstellung des *Scale* Vektors ist die chromatische Tonleiter unter Verwendung der gleichstufigen Temperatur.

```
<Scale>
  1.0 2.0 3.0 4.0 5.0 6.0 7.0 8.0 9.0 10.0 11.0 12.0
</Scale>
```

Als Beispiel einer Wohltemperierten Stimmung soll die Temperatur *Kirnberger III* angegeben werden [13]:

```
<Scale>
  1.098 2.068 3.059 4.137 5.020 6.098 7.034 8.078 9.103
    10.039 11.117 12.0
</Scale>
```

Die Darstellung mit dem Vektor *Scale* ermöglicht auch die Wiedergabe ungewöhnlicher Stimmungen wie der BOHLEN-PIERCE-Skala. Sie enthält 13 Werte:

```
<Scale>
  1.3324  3.0185  4.3508  5.8251  7.3693  8.8436  10.1760  11.6502
  13.1944  14.6687  16.0011  17.6872  19.0196
</Scale>
```

Key Die Tonart bestimmt das tonale Zentrum eines Musikstücks, siehe Abschnitt 2.2.8. Sie wird mit dem Notennamen des Grundtons und dem Namen einer Skala angegeben. Häufig verwendete Skalen sind Dur und Moll.

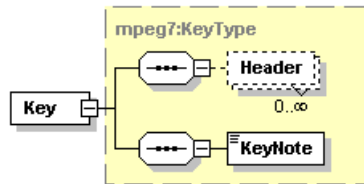


Abbildung 4.9: Datenstruktur des MPEG-7 *KeyType*. Quelle: [125]

Die Struktur des MPEG-7 *KeyType* ist in Abbildung 4.9 gezeigt. Neben dem optionalen *Header* ist das Feld *KeyNote* enthalten, das folgende Datenstruktur hat:

- *degreeNoteType* kann die Zeichen *A, B, C, D, E, F, G* enthalten. Ein optionales Attribut *Display* kann mit einer Zeichenkette belegt werden, die für Anzeigezwecke vorgesehen ist, zum Beispiel 'do' statt 'C'.
- Zwei weitere Attribute können für *KeyNote* gesetzt werden:
 - *accidental* ist ein Aufzählungstyp möglicher Alterationen des Notennamens; mögliche Werte sind *natural* (Voreinstellung), *flat* (b), *sharp* (#), *doubleflat* (bb), *doublsharp* (x).
 - *mode* Der Modus der Tonart, zum Beispiel *major* (Dur) oder *minor* (Moll).

Beispiel: Die Tonart B $\flat$ -Dur („B $\flat$ major“) lautet damit:

```
<Key>
  <KeyNote accidental='flat' mode='major'>B</KeyNote>
</Key>
```

MelodyContourType MPEG-7 bietet zwei speziell für Multimedia-Systeme vorgesehene Melodiedarstellungen an. Der *MelodyContourType* wurde bereits in Kapitel 2 vorgestellt. Neben dem *MelodyContourType* existiert der *MelodySequenceType*, der alternativ gewählt werden kann. Die Besonderheit dieser Melodiedarstellungen ist gegenüber den sonst vorhandenen Melodiebeschreibungen wie PARSONS-Code oder Intervall-Methode die Mitberücksichtigung von rhythmischen Informationen.

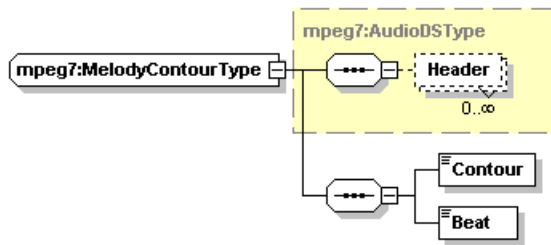


Abbildung 4.10: Datenstruktur des MPEG-7 *MelodyContourType*. Das Feld *Contour* enthält die Intervallwerte der Melodiekontur, das Feld *Beat* enthält die Schläge, auf welche die Konturwechsel fallen. Quelle: [125]

Die Datenstruktur des MPEG-7-MelodyContourType ist in Abbildung 4.10 dargestellt. Er beinhaltet zwei Vektoren, *Contour* und *Beat*.

- *Contour* Dieser Vektor enthält eine 5-stufige Melodiekonturdarstellung wie in Tabelle 4.1 angegeben. Diese Werte werden im MPEG-7-Contour-Type deklariert.
- *Beat* Dieser Vektor enthält die Zählzeiten, zu denen sich die Konturwechsel ereignen, es wird auf ganze Zahlen gerundet. Die Zählzeiten werden über die Taktgrenzen hinweg durchgezählt, d. h. es wird nur im ersten Takt mit Zählzeit eins begonnen, die nächste eins des folgenden Taktes wäre bei einem 4/4-Takt Zählzeit 5.

Tabelle 4.1: Die Intervall-Zuordnung der MPEG-7-MelodyContour mit Angabe des musikalischen Intervalls und Tonhöhenunterschied in Cent.

Contour	Änderung um m Cent	musikalisches Intervall
-2	$m \leq -250$	kleine Terz oder mehr abwärts
-1	$-250 \leq m < -50$	große oder kleine Sekunde abwärts
0	$-50 < m < 50$	reine Prime
1	$50 \leq m < 250$	große oder kleine Sekunde aufwärts
2	$250 \leq m$	kleine Terz oder mehr aufwärts

Die Konturwerte in Tabelle 4.1 werden nach Untersuchung der Melodietöne auf ihre Abweichung in Cent voneinander zugeordnet. Damit können auch noch ungenau intonierte Intervalle zugeordnet werden: die Schwankung um ± 49 Cent wird immer noch als Prime gewertet, eine zu weite große Sekunde mit ± 249 Cent noch richtig zugeordnet.

Beispiel: Ein Beispiel zur Verwendung der MPEG-7-MelodyContour ist bereits in Abschnitt 2.3.3 beschrieben worden. An dieser Stelle soll die entsprechende XML-Darstellung des Deskriptors angegeben werden.

Der *Beat*-Vektor in Abbildung 2.4 startet bei Zählzeit 4, weil die Melodie mit einem Auftakt beginnt. Die nachfolgenden Achtelnoten werden mit 5, 5, 6, 6 gezählt, da ein 4/4 vorgeschrieben ist. Man beachte, dass der *Beat*-Vektor um einen Wert länger ist als der *Contour*-Vektor.

```
<!-- MelodyContour description of "As_time_goes_by" -->
<AudioDescriptionScheme xsi:type="MelodyType">

  <Meter>
    <Numerator>4</Numerator>
    <Denominator>4</Denominator>
  </Meter>

  <MelodyContour>
    <Contour>
      1 -1 -1 -1 1 1 <!-- bar 2 -->
      2 1 -1 -1 2 1 <!-- bar 3 -->
      2 -1 -1 -1 1 <!-- bar 4 -->
    </Contour>
  </MelodyContour>
</AudioDescriptionScheme>
```

```
<Beat>
  4                <!-- bar 1 -->
  5 5 6 6 7 8      <!-- bar 2 -->
  9 9 10 10 11 12<!-- bar 3 -->
  13 13 14 14 15 <!-- bar 4 -->
</Beat>
</MelodyContour>
</AudioDescriptionScheme>
```

MelodySequence Das *MelodyContour DS* ist für QBH-Systeme gut geeignet, kann jedoch für Anwendungen, die genauere Melodiedarstellungen benötigen, unbrauchbar sein. Aus diesem Grund wird das *MelodySequence DS* zur Verfügung gestellt, das eine detailliertere Melodiedarstellung enthält.

Die Melodiebeschreibung erfolgt durch die Intervall-Methode, wobei die Intervallbeziehungen der Melodietöne über die genauen Grundfrequenzen hergestellt werden. Die rhythmischen Merkmale der Melodie werden in ähnlicher Weise beschrieben, indem die Differenzen der Notenlängen angegeben werden. Darüber hinaus ist das Speichern von Liedtext inklusive einer phonetischen Beschreibung möglich.

4.2 MPEG-21

MPEG-21 ist ein Framework für Multimedia-Datenübertragung und Konsum [40]. Ziel ist es, die Nutzung von Multimedia-Ressourcen für eine Vielzahl von Netzwerken und Geräten zu ermöglichen. Zur Zeit gibt es viele technische Plattformen, um die Infrastruktur für die Übertragung von multimedialen Inhalten aufzubauen oder deren Konsum zu ermöglichen. Es gibt jedoch bislang kein Konzept, diese Vielfalt an technischen Elementen zueinander in Beziehung zu setzen [192].

Mit MPEG-21 wird das Konzept des *Digital Item* (DI) eingeführt, das eine Abstraktion für multimedialen Inhalt inklusive verschiedenster Datentypen darstellt. Deskriptoren aus MPEG-7 beschreiben dabei die zugehörigen *Ressourcen*. QBH-Systeme sind eine Anwendung, auf die sich der MPEG-21-Standard sehr gut anwenden lässt. Beispielsweise lassen sich die Ergebnisse einer Suchanfrage, die aus Audiodateien, Lied- und Notentext, Informationen zum Künstler usw. bestehen können, mithilfe eines DI beschreiben. Durch das

plattformübergreifende Konzept in MPEG-21 werden diese verschiedenen Inhalte für jedes verwendete Gerät passend dargestellt.

4.3 SMIL

Die Arbeitsgruppe *Synchronized Multimedia* (SYMM) des *World Wide Web Consortium* (W3C) arbeitet an einer neuen Beschreibungssprache, um Audio, Video, Text und Graphiken für Multimedia-Präsentationen in Echtzeit miteinander verbinden zu können. Diese Beschreibungssprache heißt *Synchronized Multimedia Integration Language* (SMIL, gesprochen wie englisch „smile“) und ist ähnlich wie MPEG-7 als XML-Anwendung konzipiert. Einfach ausgedrückt ermöglicht es SMIL beispielsweise einem Autor einer Multimedia-Präsentation, genau zu spezifizieren, wann ein Bild präsentiert werden soll und dies abhängig davon, wann ein bestimmter Satz gesprochen worden ist.

BLACKBURN untersucht in seinen Arbeiten die Anwendung von SMIL für QBH-Systeme [31]. Speziell eignet sich SMIL dazu, die Sitzung (session) eines QBH-System-Benutzers zu beschreiben.

4.4 Zusammenfassung

In diesem Kapitel wurden verschiedene Konzepte von Multimedia-Standards erläutert, die sich für QBH-Systeme verwenden lassen. Zuerst wurde ein Überblick über MPEG-7 gegeben, der eine Schnittstelle zur Beschreibung von Multimedia-Inhalten definiert. Durch die Beschreibung verschiedener Anwendungen, die für MPEG-7 in Frage kommen, wurde die Eignung dieses Standards für QBH-Systeme klargestellt. Danach wurde das Konzept der in MPEG-7 verwendeten normative Elemente *Deskriptor* (D), *Beschreibungsschemata* (Description Schemes, DS), und der *Description Definition Language* (DDL) umrissen. Aufgrund des Umfangs des Standards wurde eine Übersicht über seine Gliederung gegeben und der für QBH-Systeme besonders wichtige Teil 4 mit den Definitionen der D für die Audiobeschreibung besonders erläutert. Im Mittelpunkt stand dabei die Darstellung des *Melody DS*, welches die Definition des *MelodyContourType* enthält, der im Rahmen der Arbeit verwendet wird.

Im folgenden Abschnitt wurde der Standard MPEG-21 beschrieben, der das Konzept des *Digital Item* (DI) definiert. Das DI verwendet Deskriptoren aus MPEG-7 für den Zugriff auf Multimedia-Daten in einer vernetzten Umgebung.

Einen ebenfalls direkter Bezug zu Multimedia-Daten und Netzwerken stellt der Standard SMIL her, der zum Schluss kurz vorgestellt wurde.

QBH-Systeme sind multimediale Anwendungen, daher ist die Verwendung von Multimedia-Standards möglich und sinnvoll. Im Rahmen dieser Arbeit wird der Standard MPEG-7 ausgewählt, weil sich der Anwendungsfall von QBH-Systemen besonders gut auf die angebotenen Definitionen und Werkzeuge dieses Standards beziehen lässt und von den standardisierten Schnittstellenbeschreibung profitieren kann.

*Das Beste in der Musik steht
nicht in den Noten.*

Gustav Mahler

Die Transkription von gesummen Anfragen in eine symbolische Darstellung ist ein zwingend notwendiger Schritt in einem Query-by-Humming-System (QBH-System). Ergebnis der Transkription ist eine Melodiebeschreibung in dem Format, in dem auch die Titel der Melodiedatenbank vorliegen. Viele Publikationen beschäftigen sich speziell mit der Aufgabenstellung der Transkription, zum Beispiel [50, 90, 133, 194]. Die Transkriptionsstufe wird im Zusammenhang mit QBH-Systemen auch als *akustisches Front-End* (*acoustic front end*) bezeichnet [50]. In bestehenden Systemen wie *Musicline*, *MELDEX* oder *Musipedia* ist dieser Teil als Java-Applet ausgeführt [11, 12, 131]. Java-Applets werden vom Web-Server, der die Internetseite anbietet, zum Browser des Anwenders übertragen und dort ausgeführt.

In diesem Kapitel werden alle Schritte beschrieben, um eine gesumme Anfrage in eine symbolische Darstellung zu transkribieren. Im ersten Abschnitt dieses Kapitels wird die Transkriptionsaufgabe eingegrenzt und die Unterteilung in verschiedene Verarbeitungsböcke motiviert. Abbildung 5.1 zeigt eine Unterteilung der Verarbeitungsböcke in der Transkriptionsstufe. Das Audio-signal der Gesangsanfrage wird zuerst der Tonhöhererkennung zugeführt, anschließend wird eine Rhythmuserkennung vorgenommen. Danach erfolgt die Auswertung aller extrahierten Parameter und die Transkription der Melodie in eine symbolische Darstellung. Diesen Verarbeitungsschritten entsprechend werden in den folgenden Abschnitten die Tonhöhererkennung und die Rhythmuserkennung im Detail erläutert. Eigene Untersuchungen diskutieren Implementierungsaspekte und praktische Probleme, anschließend wird ein Prüfverfahren für Transkriptionssysteme vorgestellt und einige Fehlerquellen für das Ergebnis der Melodietranskription erörtert.

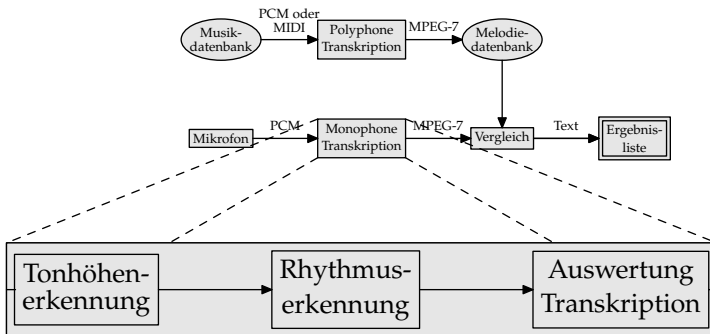


Abbildung 5.1: Die einzelnen Verarbeitungsblöcke der monophonen Transkriptionsstufe in *Queryhammer* sind *Tonhöhen-erkennung* und *Rhythmus-erkennung* zur Analyse aller notwendigen Parameter und *Auswertung/Transkription* zur Erzeugung einer symbolischen Melodiedarstellung.

5.1 Die Transkriptionsaufgabe

Der Vorgang der Transkription ist die Überführung von Musik aus einer Audio-Repräsentation in eine symbolische Repräsentation (vgl. Abschnitt 2.3). Allgemein lässt sich diese Aufgabe auf jedes Musiksignal beziehen. Dabei kann zunächst abhängig vom Inhalt des Audiosignals zwischen der Transkription von monophoner und polyphoner Musik unterschieden werden. Bei monophoner Musik werden alle transkribierten Informationen der Melodie zugeordnet. Bei polyphoner Musik muss zusätzlich entschieden werden, welche Stimme die Melodie führt und transkribiert werden soll; dieses Problem wird getrennt betrachtet und ist Gegenstand von Kapitel 6. Weiter lässt sich unterscheiden, welchen akustischen Ursprungs das zu analysierende Signal ist. Für bestimmte Musikinstrumente kann man besondere Lösungsstrategien verfolgen, beispielsweise die Transkription von Klaviersignalen [127, 162]. Bei QBH-Systemen kommen speziell Transkriptionssysteme für Gesangssignale in Frage; hier kann weiterhin Gesang mit und ohne Text unterschieden werden. Eine gute Übersicht bestehender Transkriptionssysteme, die für die Transkription von Gesang geeignet sind, findet sich in [50].

Zunächst stellt sich die Frage, welche Informationen zur Transkription in eine symbolische Melodie-Darstellung überhaupt notwendig sind. Zielformat im Rahmen dieser Arbeit ist die *MPEG-7-MelodyContour*, die bereits in Kapi-

tel 4 ausführlich beschrieben worden ist. Es müssen alle Informationen und Parameter ermittelt werden, die für die Transkription von Noten erforderlich sind. Es sind von besonderer Bedeutung [29]:

Tonhöhe: sie entspricht der Grundfrequenz eines natürlichen Tons (vergleiche Abschnitt 2.2.1).

Anschlag: Der Anschlag der Note (englisch: onset) wird mit dem Zeitpunkt des Beginns eines Notenereignisses beschrieben.

Tondauer: Die Zeit der Tondauer bezeichnet die Dauer vom Anschlag bis zum Verstummen des Tons (gelegentlich: offset).

Die Ermittlung der Tonhöhe erfolgt durch eine Grundfrequenzanalyse (GFA) des Instrumentenklangs oder der Singstimme. Die Notenanfänge lassen sich im einfachsten Fall aus dem Verlauf der GFA, aber auch aus anderen Daten wie etwa dem Energieverlauf des untersuchten Signals ermitteln. Ebenfalls ergibt sich aus diesen Informationen der Zeitpunkt des Notenendes. Die Tondauer lässt sich aus Start- und Endpunkt der Noten ermitteln. Von zentraler Wichtigkeit für die Melodietranskription sind daher zum einen die Verfahren der GFA zur Tonhöhenerkennung, zum anderen die Methoden der Rhythmuserkennung.

5.2 Tonhöhenerkennung

Um die Tonhöhe eines natürlichen Tons (vergleiche Abschnitt 2.2.1) zu ermitteln, ist eine Grundfrequenzanalyse (GFA) notwendig. Die Aufgabenstellung der GFA ist im Rahmen der Signalverarbeitung ein gut erschlossenes Gebiet, insbesondere im Bereich der Sprachsignalverarbeitung. Schwieriger gestaltet sich die GFA für Musikinstrumentensignale. Ein wesentlicher Grund dafür ist u. a., dass für Sprache ein Signalmodell existiert, das für Musiksignale im Allgemeinen nicht vorhanden ist [93].

Generell lässt sich jedes GFA-Verfahren in die in Abbildung 5.2 gezeigten Stufen untergliedern: *Vorverarbeitungsstufe*, *Extraktionsstufe* und *Nachverarbeitungsstufe* [92]. Die Vorverarbeitung sorgt für die Störbefreiung des untersuchten Signals und ggf. für eine Datenreduktion, um die Arbeit der Extraktionsstufe zu erleichtern. Die Extraktionsstufe selbst nimmt dann die eigentliche

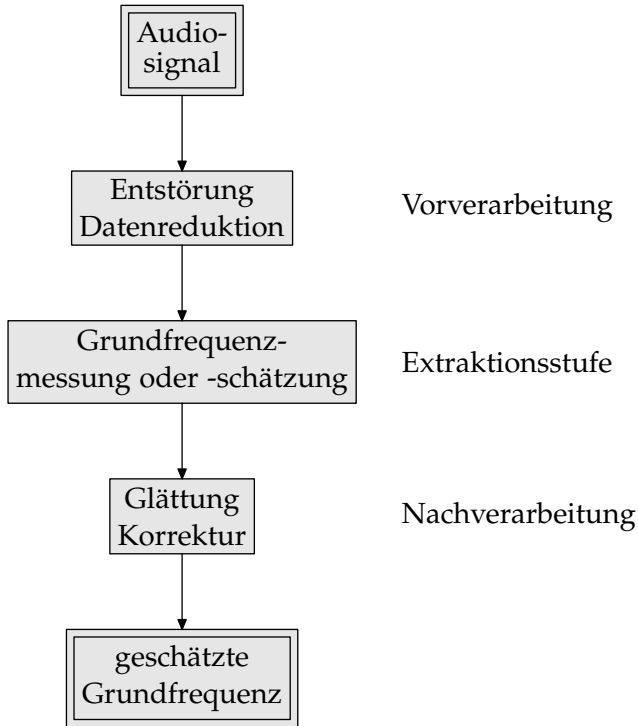


Abbildung 5.2: Verarbeitungsschritte zur Erkennung der Grundfrequenz in Audiosignalen. Die Vorverarbeitungsstufe sorgt für eine Störfreiung und Minderung der zu verarbeitenden Datenmenge. Die Extraktionsstufe führt die eigentliche Grundfrequenzanalyse durch, die Nachverarbeitungsstufe bereitet die gewonnenen Daten für die Weiterverarbeitung auf.

Messung bzw. Schätzung möglicher Grundfrequenzen vor. In der Nachverarbeitungsstufe werden Aufgaben wie Fehlererkennung und -korrektur oder die Glättung der gewonnenen Daten vorgenommen.

Eine Reihe von Schwierigkeiten treten bei der GFA für QBH-Systeme auf. Da es sich um gesungene oder gesummte Signale handelt, ist das untersuchte Signal nicht perfekt periodisch, sondern muss genauso wie Sprache als nicht-stationärer Prozess betrachtet werden [93]. Infolge der Vielfalt der möglichen sinnvollen Artikulationsstellungen des menschlichen Vokaltraktes und infolge der Vielfalt menschlicher Stimmen existiert eine große Anzahl möglicher Zeitstrukturen. Nicht alle Abschnitte eines Gesangssignals sind stimmhaft und weisen eine Grundfrequenz auf, so dass zwischen stimmlosen und stimmhaften Abschnitten unterschieden werden muss [159]. Weiterhin können bei der Aufnahme von gesungenen Anfragen Störgeräusche aus der Umgebung mit aufgenommen werden, welche die GFA erschweren.

Bei gesungenen Anfragen kann zwischen Eingaben unterschieden werden, in denen Liedtext vorgetragen wird und solchen, bei denen auf bedeutungslosen Silben gesungen wird. Häufig verwendete Silben sind /da/, /na/, /ta/, /du/ und ähnliche, siehe auch Abschnitt 2.4. Liedtext ist aus verschiedenen Gründen wesentlich schwerer auszuwerten als einzelne Silben – die Vokale, die die Tonhöheninformation tragen, fallen in der Regel kürzer aus als bei gesungenen Einzelsilben, der Rhythmus ist nicht so klar durch einen bestimmten Konsonanten markiert. Die meisten QBH-Systemen fordern aus diesem Grund vom Nutzer, auf /na/ oder /da/ beim Stellen einer Suchanfrage zu singen.

Die Anzahl existierender und für die Transkriptionsaufgabe verwendbarer GFA-Verfahren ist unüberschaubar groß [92, 93]. Sie lassen sich in zwei Kategorien einteilen, wenn man das Eingangssignal der Extraktionsstufe als Unterscheidungskriterium heranzieht. Falls dieses Signal die gleiche Zeitbasis besitzt wie das ursprüngliche Signal, arbeitet das Verfahren im Zeitbereich. In allen anderen Fällen ist der Zeitbereich innerhalb der Vorverarbeitungsstufe verlassen worden. Da das Eingangssignal zeitveränderlich ist, kann dies nicht anders erfolgen als mit Hilfe einer Kurzzeittransformation.

5.2.1 Verfahren der Kurzzeitanalyse

Bei jedem Algorithmus, der sich der Kurzzeitanalyse bedient, wird in der Vorverarbeitungsstufe eine Kurzzeittransformation durchgeführt [93]. Zu diesem

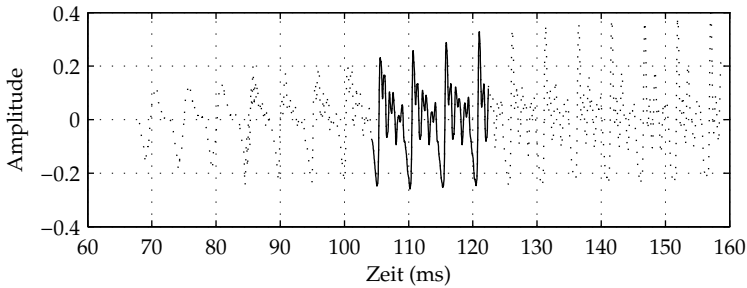


Abbildung 5.3: Ausschnitt aus dem Zeitsignal eines gesungenen „na“: Man erkennt die sich verändernde Signalstruktur. Für die hervorgehobenen drei Perioden kann das Signal als quasistationär betrachtet werden, die Periodendauer beträgt etwa 5 ms.

Zweck wird das Eingangssignal $x(n)$ in eine Folge von Signalblöcken (blocks, frames) $\mathbf{x}_N(k)$ eingeteilt:

$$\mathbf{x}_N(k) = x(n + kM) \text{ mit } n = 0 \dots N-1, k = 0, 1, \dots \quad (5.1)$$

Der Parameter N bezeichnet die Länge des Blocks, die Schrittweite (hopsi-ze) M bestimmt, wie weit der Block k im Signal weiterverschoben wird. N muss hinreichend klein sein, damit der zu bestimmende Parameter als annähernd konstant angenommen werden kann. Andererseits muss N hinreichend groß sein, damit der Parameter überhaupt messbar ist. Für Kurzzeitanalyse-Verfahren zur GFA wird N daher meist so gewählt, dass das Messintervall mindestens 2 bis 3 vollständige Grundperioden enthält [34,93]. Abbildung 5.3 zeigt ein Beispiel.

Die GFA-Verfahren mittels der Kurzzeitanalyse lassen sich weiter unterteilen in *Korrelationsmethoden* unter Verwendung der Autokorrelationsfunktion oder der Betragsdifferenzfunktion, *Frequenzbereichsverfahren* wie die Cepstralanalyse oder die „harmonische Analyse“, *mathematisch motivierte Verfahren* wie die Kleinste-Quadrate-Methode oder die *aktive Modellierung*, die ein Spracherzeugungsmodell zugrunde legt. Abbildung 5.4 zeigt eine Übersicht der genannten Verfahren. Die am häufigsten verwendeten Verfahren werden nun kurz vorgestellt.

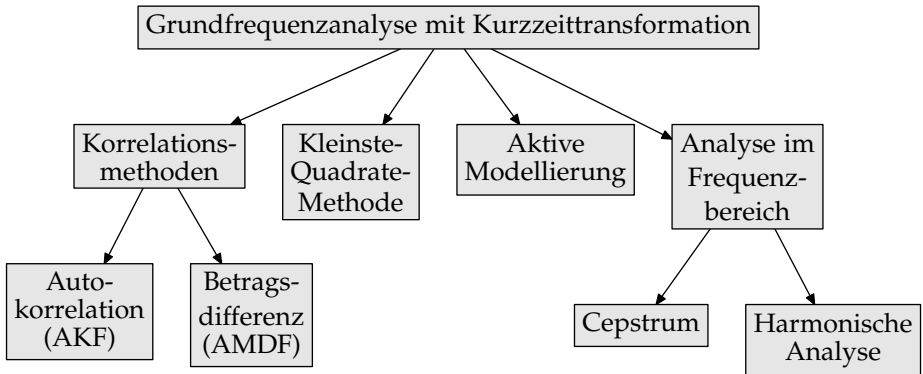


Abbildung 5.4: Verfahren zur GFA in der Übersicht (nach [93]).

Autokorrelationsmethode

Die Autokorrelationsfunktion (AKF) erlaubt eine Aussage darüber, wie ähnlich ein Signal $x(n)$ seiner um die Zeit d verschobenen Version ist [146]:

$$R(d) = \sum_n x(n) x(n + d). \quad (5.2)$$

Für die GFA wird nun das Beobachtungsintervall $x_N(n)$ zur Berechnung der AKF herangezogen.

Für den Signalabschnitt aus Abbildung 5.3 ergibt sich eine AKF wie in Abbildung 5.2.1 dargestellt. Ist die Verzögerung d gleich der Periodendauer T_0 , ergibt sich ein starkes Maximum. Die Grundfrequenz F_0 errechnet sich aus dem Kehrwert der Verzögerung $T_{\max}$, also

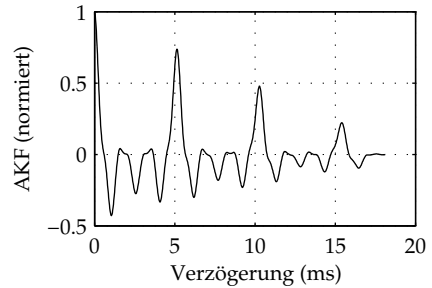
$$F_0 = \frac{1}{T_{\max}}. \quad (5.3)$$

Die AKF-Methode zur GFA ist aufgrund der blockweisen Verarbeitung unempfindlich gegen Phasenverzerrungen und Qualitätsminderungen des untersuchten Signals [92]. Sie ist robust, einfach zu implementieren und wird daher häufig verwendet. Allerdings weist sie eine gewisse Empfindlichkeit gegenüber starken Formanten bei Sprachsignalen auf, die zu Oktavfehlern führt. In Abbildung 5.2.1 ist das zur Grundfrequenz zugehörige lokale Maximum

bei $T_{\max} = 5$ ms am größten, die Grundfrequenz kann in diesem Fall sicher erkannt werden.

Beispiele für die Verwendung der AKF-Methode sind das Transkriptionssystem von BELLO [29] oder das Sprachanalysewerkzeug PRAAT [34]. Die AKF-Methode ist neben der Analyse von Sprach- und Gesangssignalen ebenfalls für die Analyse von Musikinstrumentenklängen geeignet [38].

Abbildung 5.5: Die AKF des Signals aus Abbildung 5.3. Die Maxima liegen bei ca. 5, 10 und 15 ms, die Grundfrequenz kann leicht erkannt werden.



Betragsdifferenzfunktion

Die Betragsdifferenzfunktion (average magnitude difference function, AMDF) kann als Gegenstück zur AKF betrachtet werden [93]. Sie ist definiert als

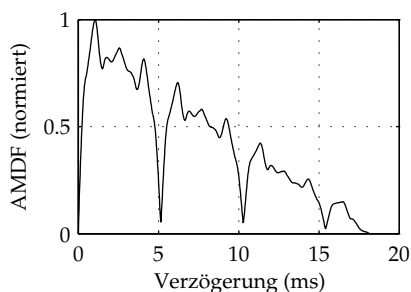
$$A(d) = \sum_n |x(n) - x(n - d)|. \quad (5.4)$$

Ist die Verzögerung d gleich der Grundperiodendauer T_0 , so ergibt sich nun im Gegensatz zur AKF ein starkes Minimum. Weil die Berechnung der Betragsdifferenzfunktion keine Multiplikationen erfordert, lässt sie sich mit relativ geringem Rechenaufwand durchführen. Sie ist daher vor allem für Systeme mit beschränkten Ressourcen empfehlenswert [171]. Außerdem folgt die AMDF dem nichtstationären Prinzip der Kurzzeitanalyse; sie lässt sich bereits unter Verwendung von Messintervallen bestimmen, die noch kürzer sind als die für die AKF-Methode benötigten [93]. Die AMDF wird zum Beispiel im Verfahren YIN [48] verwendet.

Harmonische Analyse

Bei der harmonischen Analyse wird das Frequenzspektrum des untersuchten Signals betrachtet, vorzugsweise das Leistungsdichtespektrum (LDS). Die di-

Abbildung 5.6: Die AMDF des Signals aus Abbildung 5.3. Das erste Minimum bei 5 ms ergibt die Periodendauer der Grundfrequenz.



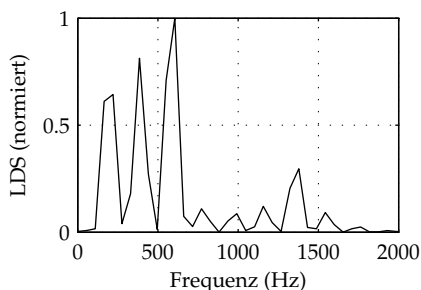
rekte Bestimmung der Grundfrequenz aus dem ersten Maximum des LDS ist jedoch häufig unzuverlässig, da sich etwa durch Störungen ebenfalls spektrale Maxima ausbilden können. Daher wird nach dem harmonischen Spektrum eines natürlichen Tons gesucht, das aus einer Grundschiwingung und harmonischen Teilschwingungen besteht.

In Abbildung 5.2.1 ist das LDS des Signals aus Abbildung 5.3 dargestellt. Es berechnet sich aus

$$S(k) = |\mathcal{F}_N\{\mathbf{x}_N(n)\}|^2, \quad (5.5)$$

wobei der Operator $\mathcal{F}_N$ die diskrete Fouriertransformation mit N Punkten bezeichnet [102, 147]. Man erkennt deutlich ein harmonisches Spektrum mit einer Grundfrequenz von ungefähr 200 Hz.

Abbildung 5.7: Das Leistungsdichtespektrum des Signals aus Abbildung 5.3. Die stark ausgeprägten Maxima liegen bei 200, 400 und 600 Hz, das zur Grundfrequenz gehörige Maximum ist jedoch *nicht* am stärksten.



Cepstralanalyse

Wird das gerade beschriebene LDS logarithmiert und danach in den Zeitbereich zurücktransformiert, so ergibt sich das sogenannte „Cepstrum“ [93,

102, 145, 147]. Die Grundperiodendauer T_0 lässt sich aus dem Maximum des Cepstrums ableiten. Von der Cepstralanalyse ist bekannt, dass sie gegenüber dominanten Formanten unempfindlich ist [93], während jedoch eine gewisse Empfindlichkeit gegenüber verrauschten Signalen besteht. Für Musiksignalverarbeitung wird sie daher üblicherweise nicht verwendet.

5.2.2 Zeitbereichsverfahren

Verfahren zur GFA, die im Zeitbereich arbeiten, sind vor allem dem Bereich der Sprachverarbeitung zuzuordnen und daher auch für die Verwendung in QBH-Systemen interessant. Das zu analysierende Signal wird bei der Zeitbereichsverarbeitung prinzipiell von Periode zu Periode untersucht; daraus ergibt sich auch eine höhere Empfindlichkeit gegenüber lokalen Störungen des Signals [93]. Wichtige Verfahren zur Bestimmung der Grundfrequenz sind die Messung der Nulldurchgangsrate und das GFA-Verfahren nach GOLD und RABINER.

Nulldurchgangsrate

Die Grundfrequenz eines Signals lässt sich durch Abzählen der Nulldurchgangsrate (zero crossing rate, ZCR) bestimmen [106, 159]. Um die tiefen Frequenzen des untersuchten Signals zu bevorzugen und Störungen durch rauschhafte Signalanteile zu unterdrücken, ist eine starke Tiefpassfilterung notwendig (18 dB/Oktave, [93]). Die dadurch bedingte Dynamikänderung des Signals sowie die Notwendigkeit einer ausgeprägten Grundschiwingung (wie sie bei Telefonsignalen zum Beispiel fehlt) sind klare Nachteile des Verfahrens. Vorteil ist die Einfachheit der Methode. Die ZCR wird oft auch zur Klassifizierung der Stimmhaftigkeit von Signalen eingesetzt [83, 106].

GFA nach Gold und Rabiner

Die GFA nach GOLD und RABINER wurde bereits 1969 vorgestellt und ist daher sehr bekannt [77]. Es handelt sich um ein Mehrkanalsystem, d. h. das zu untersuchende Signal wird mit mehreren Algorithmen gleichzeitig untersucht, danach muss eine Entscheidung über das „richtige“ Ausgangssignal herbeigeführt werden [93]. Bei dem Verfahren von GOLD und RABINER werden nach einer Tiefpassfilterung sechs verschiedene Impulsfolgen aus dem Signal abgeleitet, die anschließend parallel untersucht werden. Die redundante Verarbei-

tung des Signals ist insofern besonders interessant, als ihr auch von Musikwissenschaftlern für gehörbezogene Phänomene eine große Bedeutung beigemessen wird [35]. Der Algorithmus ist wenig rechenaufwendig und ergibt daran gemessen gute Ergebnisse. Er wird zur Melodietranskription von McNAB verwendet [132].

5.2.3 Diskussion

In den vorangegangenen Abschnitten sind verschiedene GFA-Verfahren vorgestellt worden, deren Eignung für QBH-Systeme nun diskutiert werden soll. Die für QBH-Systeme wesentlichen Aspekte sind Robustheit, Genauigkeit und Komplexität. Mit Robustheit wird die Unempfindlichkeit gegenüber Signalstörungen bezeichnet, die sich aus dem Verfahren ergibt. Die Genauigkeit bezeichnet die Güte der Frequenzmesswerte, die Komplexität eines Verfahrens macht eine Aussage über den benötigten Berechnungs- und Implementierungsaufwand. Genauigkeit und Komplexität hängen voneinander ab.

Das GFA-Verfahren als Teil des Transkriptionsverfahrens wird, wie zu Beginn dieses Kapitels bereits diskutiert worden ist, häufig innerhalb eines Java-Applets untergebracht. Damit ergibt sich die Anforderung einer kostensparenden Implementierung zugunsten einer schnellen Übertragung des Java-Applets. Da durch die Ausführung des Java-Applets jeder beliebige Rechnerarbeitsplatz in Frage kommt und damit auch jeder beliebige Aufnahmeraum, ist eine gute Robustheit gegen akustische Störungen erforderlich.

Robustheit

Die Zeitbereichsverfahren ZCR und GOLD/RABINER sind zwar wenig rechenintensiv und einfach zu implementieren, aber auch weniger robust als Verfahren der Kurzzeitanalyse. Da für QBH-Systeme eine gewisse Robustheit gegenüber Störungen aus der akustischen Umgebung notwendig ist, sind letztere in jedem Fall vorzuziehen.

Genauigkeit und Komplexität

Bezüglich der Genauigkeit der gemessenen Frequenzwerte ergeben sich für die Transkription von Noten besonders im unteren Frequenzbereich hohe Anforderungen, da die Frequenzeinteilung der Notenskala logarithmisch ist (siehe Kapitel 2). Diskussionen der benötigten Frequenzauflösung findet man nur

selten in der Literatur, etwa bei KLAPURI [111]. Daher wird dieser Aspekt im Folgenden am Beispiel der harmonischen Analyse und der AKF-Methode kurz vorgestellt.

Harmonische Analyse Für die harmonische Analyse wird zur spektralen Zerlegung des Signals meist die diskrete Fouriertransformation (DFT) herangezogen. Die Frequenzauflösung der DFT ist für alle Frequenzen gleich, es gilt

$$\Delta f_{\text{DFT}} = \frac{f_s}{N}, \quad (5.6)$$

wobei f_s die Abtastrate ist und N die Ordnung der DFT angibt. Um eine hinreichend hohe Genauigkeit zu erzielen, ist eine Frequenzauflösung von $\Delta f_{\text{min}} = 3,0869 \text{ Hz}$ notwendig, wenn der tiefste aufzulösende Ton A_1 bei 55 Hz ist und er noch von $G_{\sharp 1}$ unterschieden werden können soll. Damit wäre bei einer Abtastrate von $f_s = 8 \text{ kHz}$ die Ordnung $N = 2592$ für die DFT bzw. $N = 4096$ für die schnelle Fouriertransformation (fast Fourier transform, FFT) notwendig; dies bedeutet neben einer starken zeitlichen Verschmierung (ein Analysefenster hat dann die Länge von 324 ms) einen hohen Berechnungsaufwand.

Häufig wird in der Literatur zur Erhöhung der spektralen Auflösung der DFT bei gleichzeitiger Verwendung kurzer Fensterlängen das Auffüllen mit Nullen (zero padding) vorgeschlagen, wie z. B. in [64]. Diese Vorgehensweise ist jedoch nur für Signale mit einfacher harmonischer Struktur geeignet, da sich durch die Nullenergänzung zwar die Anzahl der berechneten Stützstellen im Spektrum erhöht, keinesfalls aber die Frequenzauflösung zur Untersuchung des Signals [147]. Treten Maxima nicht deutlich aus dem Spektrum hervor, besteht eine starke Empfindlichkeit gegen Störungen.

Autokorrelationsfunktion Die Frequenzauflösung der AKF-Methode ist frequenzabhängig, es gilt

$$\Delta f_{\text{AKF}}(n) = \frac{f_s}{n^2 - n}, \quad (5.7)$$

wobei $\Delta f(n)$ den Frequenzunterschied zwischen zwei AKF-Stützstellen bei n und $n + 1$ darstellt.

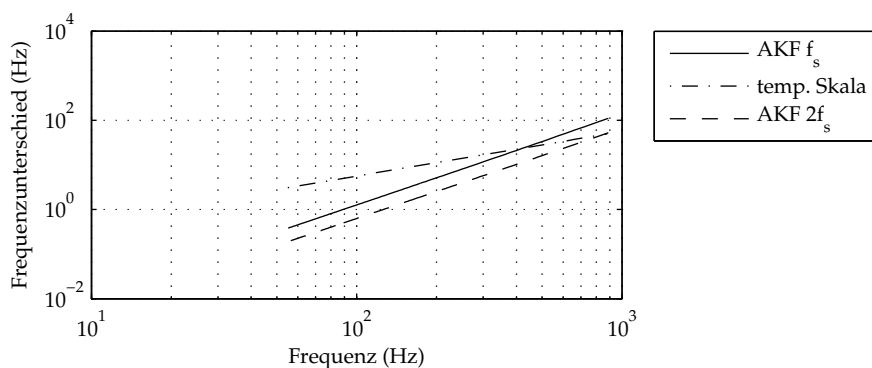


Abbildung 5.8: Die Frequenzauflösung der AKF-Methode im Vergleich mit der gleichstufigen Skala. Die AKF-Methode mit doppelter Abtastrate (AKF2) liefert eine hinreichend hohe Frequenzauflösung für den Frequenzbereich von 80–800 Hz.

In Abbildung 5.8 ist Gleichung (5.7) für $f_s = 8 \text{ kHz}$ über die Frequenz dargestellt, dazu zum Vergleich die Frequenzabstände der Töne einer chromatischen Tonleiter, gestimmt nach der gleichstufigen Skala. Bis etwas über 400 Hz ist der Frequenzunterschied zwischen zwei Frequenzstützstellen der AKF-Methode kleiner als es die gleichstufige Skala erfordert, darüber sind die Frequenzabstände der chromatischen Tonleiter kleiner als die der AKF. Daher ist eine Erhöhung der Frequenzauflösung der AKF-Methode notwendig. Eine einfache Möglichkeit ist es, die Abtastrate zu verdoppeln. Die resultierende Frequenzauflösung ist nun auch im gesamten Frequenzbereich ausreichend, wie in Abbildung 5.8 zu erkennen ist.

Für die monophone Transkriptionsstufe des Beispielsystems *Queryhammer* wird aus oben ausgeführten Gründen die AKF-Methode zur GFA ausgewählt. Ebenso wäre die Verwendung der AMDF möglich. Sie ist robust gegen Störungen, ausreichend genau und mit angemessenem Aufwand zu implementieren. Genaue Informationen zur Implementierung werden in Abschnitt 5.4 gegeben.

5.3 Rhythmuserkennung

Mit dem Begriff „Rhythmuserkennung“ soll im Rahmen dieser Arbeit der Vorgang bezeichnet werden, aus einem Musiksinal rhythmische Merkmale zu extrahieren und in symbolischer Form zugänglich zu machen. Allgemein können hierbei einerseits von der Musiknotation unabhängige Merkmale gemeint sein, andererseits aber auch speziell von der Notenschrift abhängige Elemente (siehe auch Kapitel 2) [66]. Von der Notation unabhängig sind

- die Anschläge,
- die Hauptzählzeiten (beats) und
- das Tempo.

Als Elemente der Notenschrift können extrahiert werden:

- die Anfänge und Längen der Takte,
- Taktart und Taktwechsel,
- die Länge des Auftakts.

Die Extraktion dieser einzelnen Komponenten ist unterschiedlich kompliziert, alle Merkmale sind voneinander abhängig. An dieser Stelle wichtig ist besonders die Abhängigkeit von Tempo und Hauptzählzeiten: Das Tempo ist immer umgekehrt proportional zur Länge einer Zählzeit. Bei der Rhythmuserkennung besonders bedeutsam ist die Erkennung der Anschläge, da über sie die einzelnen Notenereignisse segmentiert werden können. Wenn wie beim MPEG-7-MelodyContour-Deskriptor rhythmische Informationen festgehalten werden, sind auch die Hauptzählzeiten und das Tempo von Interesse.

Eine besondere Aufgabenstellung tritt bei der Rhythmustranskription im Rahmen eines QBH-Systems auf: Bei der Gesangseingabe hält der Sänger gewöhnlich kein festes Tempo ein, d. h. die M.M.-Zahl (vergleiche Abschnitt 2.2.2) ändert sich während des Gesangsvortrags in gewissen Grenzen. In der Literatur wird daher zwischen der *Tempoerkennung* (beat induction) und *Tempoverfolgung* (beat tracking) unterschieden [82]; im ersten Fall wird das Tempo der Zählzeiten geschätzt und im Folgenden als konstant angenommen, im zweiten Fall wird das Tempo regelmäßig neu geschätzt. Um Temposchwankungen für die Eingabe von gesummen Anfragen an QBH-Systeme zu vermeiden, wäre die Vorgabe eines Taktes bzw. des Tempos durch ein Metronom notwendig.

Das aber würde den Nutzerkreis von QBH-Systemen auf musikalisch geübte Anwender einschränken und wird daher üblicherweise nicht vorgenommen.

Ein von SCHEIRER in [170] vorgestelltes Verfahren wird häufig zur Tempo- und damit Zählzeiterkennung verwendet. Es ist in Arbeiten am Fachgebiet Nachrichtenübertragung ausführlich untersucht worden [66] und liefert als Verfahren zur Tempoverfolgung Informationen über das Tempo eines Musiksignals in M.M. in Abhängigkeit von der Zeit. Damit das Verfahren erfolgreich ist, muss allerdings ein relativ starker Puls im Signal vorhanden sein, d.h. Musiksignale, bei denen der Rhythmus durch Schlaginstrumente (mit-)gespielt wird, liefern gute Ergebnisse. Gesummte Signale hingegen weisen kaum perkussive Signalanteile auf und können daher nur schlecht verarbeitet werden. In *Queryhammer* werden die benötigten rhythmischen Informationen daher über den Grundfrequenzverlauf extrahiert.

5.4 Eigene Untersuchungen

In diesem Abschnitt wird die im Rahmen der vorliegenden Arbeit entwickelte Transkriptionsstufe des Systems *Queryhammer* dargestellt. Abbildung 5.9 zeigt eine Übersicht der Verarbeitungsstufen, die zur Transkription der Nutzeranfrage verwendet werden. *Queryhammer* verwendet für die GFA der Tonhöhenerkennung die AKF-Methode mit Erweiterungen, wie sie BOERSMA in [34] vorgeschlagen hat. An die Tonhöhenerkennung schließt sich die Rhythmuserkennung an. Der Weg, aus dem Grundfrequenzverlauf die rhythmische Information zu gewinnen, ist typisch für viele akustische Front-Ends von QBH-Systemen [76, 90, 132, 133] und wird daher auch für das Beispielsystem *Queryhammer* gewählt. Die dabei auftretenden Probleme werden genauer untersucht, um Fehlerquellen für diese typische Vorgehensweise zu identifizieren.

5.4.1 Tonhöhenerkennung

Die Tonhöhenerkennung der monophonen Transkriptionsstufe wird nun gemäß den in Abbildung 5.2 dargestellten Blöcken Vorverarbeitung, GFA und Nachverarbeitung erläutert. Anhand von Signalbeispielen wird die Wirkungsweise dieser Verarbeitungsstufen dargestellt.

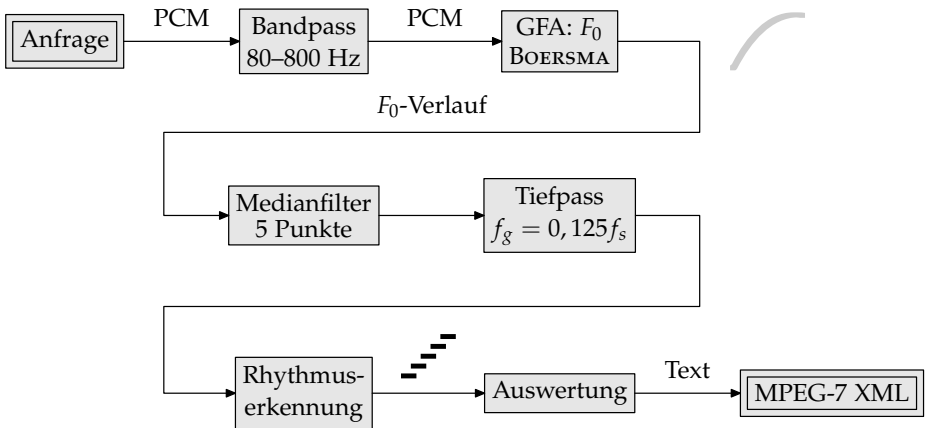


Abbildung 5.9: Das Blockschaltbild der monophonen Transkriptionstufe im Beispielsystem *Queryhammer*: Die Vorverarbeitung erfolgt durch den Bandpass. Nach der GFA wird der F_0 -Verlauf in der Nachverarbeitung median- und tiefpassgefiltert.

Vorverarbeitung

Nachdem die gesummte Anfrage mit einem Mikrofon aufgenommen worden ist, wird zunächst eine Bandpassfilterung durchgeführt, um den zu analysierenden Frequenzbereich zu begrenzen und Umgebungsgeräusche außerhalb dieses Bandes zu unterdrücken. Das Signal wird bei einer Abtastrate von $f_s = 16 \text{ kHz}$ verarbeitet, die in Hinblick auf die Frequenzauflösung ausreichend ist. Die Bandbegrenzung hat einen Durchlassbereich von $80\text{--}800 \text{ Hz}$, was für gesungene Signale in aller Regel ausreichend ist [132]. Dieser Frequenzbereich entspricht einem Notenumfang von $D\sharp\text{--}g^2$. An diese Vorverarbeitung schließt sich die GFA an.

GFA nach Boersma

Das GFA-Verfahren nach BOERSMA stammt aus dem Bereich der Sprachanalyse und ist im Programm PRAAT enthalten [34]. Das Verfahren ist sehr robust gegen Störungen und als AKF-Methode bei entsprechender Abtastrate auch hinreichend genau, wie in Abschnitt 5.2.3 bereits dargelegt worden ist. Am

Fachgebiet Nachrichtenübertragung durchgeführte Untersuchungen zur Verwendung dieses Verfahrens für ein QBH-System finden sich in [196].

Ein Nachteil der AKF-Methode ist, dass sie für Oktavfehler anfällig ist, wie in Abschnitt 5.2.1 schon erläutert worden ist. Das Verfahren nach BOERSMA enthält mehrere Ergänzungen zur AKF-Methode, um dieses Problem zu mindern. Die einzelnen Schritte werden nun anhand eines Beispiels erläutert. Abbildung 5.10a zeigt ein sinusförmiges Zeitsignal $x(n)$ mit einer Grundfrequenz von $F_0 = 100$ Hz und einer ausgeprägten Teilschwingung bei 200 Hz. Diese Signalform ist ein gutes Beispiel für ein Sprachsignal mit stark ausgeprägtem Formanten [34], für das die Gefahr groß ist, dass statt der Grundfrequenz eine Teilschwingung detektiert wird und damit ein Oktavfehler entsteht. Zuerst wird ein Signalabschnitt $\mathbf{x}_N$ mit N Abtastwerten mittelwertbefreit und gefensert:

$$a(n) = (x(n) - m_x(n)) w(n) \quad \forall \quad x(n) \in \mathbf{x}_N. \quad (5.8)$$

Die Fensterfunktion $w(n)$ wird als HANN-Fenster gewählt, siehe Abbildung 5.10b. Das resultierende Signal zeigt Abbildung 5.10c.

Zur GFA wird die AKF des Signalausschnitts berechnet und auf die Energie des untersuchten Signalabschnitts normiert:

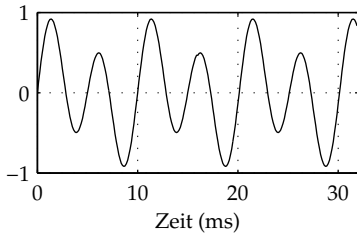
$$R_a(k) = \frac{\sum a(n)a(n+k)}{\sum a^2(n)}, \quad (5.9)$$

vergleiche Abbildung 5.10d. Das erste Maximum der AKF liegt bei 5 ms und bezieht sich damit auf die erste Harmonische – man erkennt die starke Empfindlichkeit der AKF-Methode gegenüber Formanten. Weiterhin bewirkt der abfallende Verlauf von R_a , dass der Zeitpunkt des Maximums fehlerhaft geschätzt wird und damit die geschätzte Grundfrequenz zu hoch ist [34]. Um die höheren Formanten zu unterdrücken, wird daher üblicherweise eine Tiefpassfilterung vorgenommen.

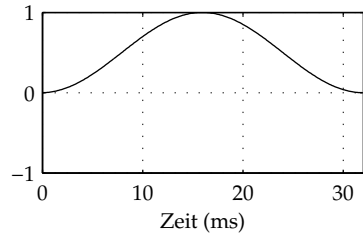
Der nun für das BOERSMA-Verfahren wesentliche Schritt besteht darin, R_a durch die normierte AKF der Fensterfunktion, R_w (vgl. Abbildung 5.10e) zu teilen. Damit ergibt sich eine entzerrte AKF R_x mit

$$R_x(k) \approx \frac{R_a(k)}{R_w(k)}, \quad (5.10)$$

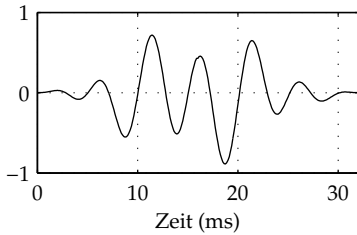
und das Maximum dieser Funktion führt nun zur richtigen Grundfrequenz (Abbildung 5.10f). Gleichung (5.10) gilt exakt für das Signal $x(n) = 1$ (für



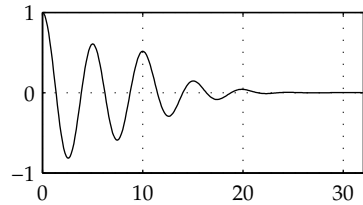
(a) Zeitverlauf eines Sprachsignalauschnitts $x(n)$ mit $T_0 = 10$ ms.



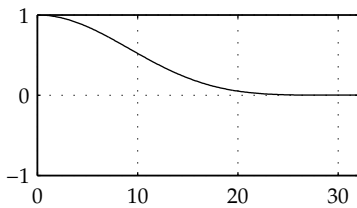
(b) HANN-Fenster $w(n)$.



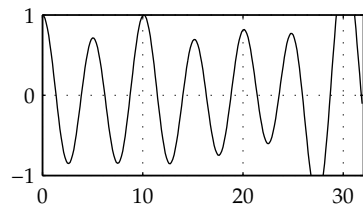
(c) Das gefensterte und mittelwertbefreite Signal $a(n)$.



(d) Die AKF $R_a(d)$ des Signals $a(n)$. Das Maximum liegt bei $T_{\max} = 5$ ms (Oktavfehler).



(e) Die AKF $R_w(d)$ der Fensterfunktion $w(n)$.



(f) Die entzerrte AKF $R_x(d)$ des Signals $a(n)$ mit Maximum bei $T_{\max} = 10$ ms.

Abbildung 5.10: Zur Fensterung und AKF-Berechnung im Verfahren nach BOERSMA.

diesen Fall ohne Subtraktion des Mittelwertes). Für periodische Signale sind die Spitzenwerte nahe dem Wert 1, wie Abbildung 5.10f zeigt.

Durch Störungen des Signals wie Hintergrundrauschen kann es dazu kommen, dass um eine Oktave zu tiefe Frequenzen gefunden werden. Auch perfekt periodische Signalabschnitte sind für diesen Fehler anfällig, da alle Maxima der Funktion $R_x(k)$ ungefähr gleich groß sind. Um dies zu vermeiden, wird von BOERSMA eine Kostenfunktion R_{Okt} vorgeschlagen, die den Betrag der AKF abhängig von der gefundenen Verzögerungszeit τ_{max} macht:

$$R_{\text{Okt}} = R_x(\tau_{\text{max}}) - O_{\text{Okt}}^2 \log(f_{\text{min}} \tau_{\text{max}}). \quad (5.11)$$

Der Parameter O_{Okt} bevorzugt mit wachsenden Werten höhere Grundfrequenzen. Als praktisch haben sich Werte im Bereich $0,01 \dots 0,2$ erwiesen.

Bei rauschhaften Eingangssignalen, wie sie zum Beispiel bei Konsonanten auftreten, bilden sich für die Funktion $R_x(k)$ nur sehr schwache Maxima, da kaum periodische Signalanteile enthalten sind. Es ist zweckmäßig, diese Signalabschnitte durch einen festen Schwellwert R_{min} aus der Grundfrequenzberechnung herauszunehmen. Für Gesangssignale hat sich ein konstanter Wert von 0,3 als günstig erwiesen. Die Verwendung fester Schwellwerte ist aufgrund der blockweisen Normierung des Signals auf die Energie möglich.

Um die Genauigkeit der AKF-Methode weiter zu erhöhen, wird nun noch eine Interpolation der AKF-Werte vorgenommen, wodurch die Lage des Maximums genauer bestimmen werden kann. Es wird eine parabolische Interpolation durchgeführt [155], für die interpolierte Stelle des Maximums gilt dann:

$$k_{\text{max iplt.}} \approx k + \frac{0,5 (R(k+1) - R(k-1))}{2R(k) - R(k-1) - R(k+1)}. \quad (5.12)$$

Der zugehörige Maximalwert der AKF berechnet sich zu

$$R_{\text{max iplt.}} \approx R(k) + \frac{(R(k+1) - R(k-1))^2}{8(2R(k) - R(k-1) - R(k+1))}. \quad (5.13)$$

Durch eine hinreichend hohe Abtastrate von 16 kHz und die Interpolation ist nun eine genaue Berechnung der Grundfrequenz gewährleistet.

Beispiel: Abbildung 5.11a zeigt die Noten einer Benutzeranfrage, die alle MPEG-7-MelodyContour-Werte in aufsteigender Reihenfolge ergibt. Ein QBH-Nutzer singt nun diese Notenfolge auf der Silbe /da/.

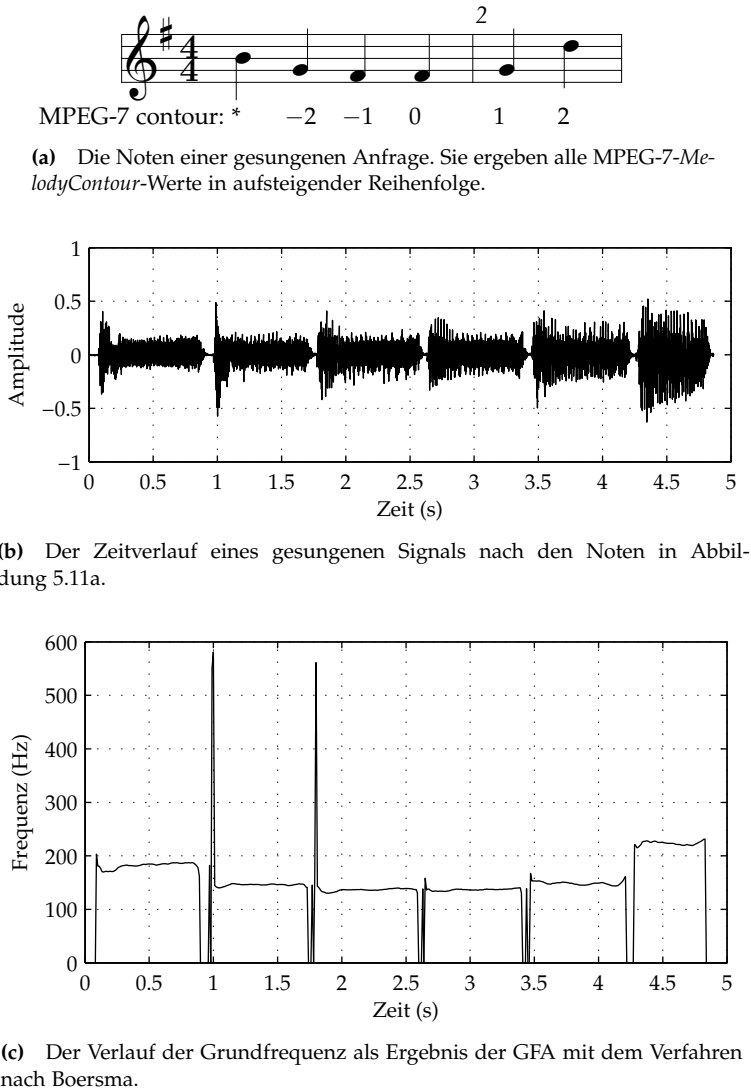


Abbildung 5.11: Noten, Zeit- und Grundfrequenzverlauf einer gesungenen Anfrage an ein QBH-System.

Der Zeitverlauf des aufgenommenen Signals ist in Abbildung 5.11b dargestellt, die Noteneinsätze sind im Zeitverlauf gut zu erkennen. Das Ergebnis der GFA in Abbildung 5.11c zeigt deutliche Ausreißer beim zweiten und dritten Ton. Grund ist Konsonant „d“, bei dem keine Grundfrequenz erkannt werden kann.

Nachverarbeitung

Das Ergebnis der GFA ist für die eigentliche Transkription nur bedingt geeignet, da durch Ausreißer im Grundfrequenzverlauf wie im dargestellten Beispiel leicht Fehler entstehen können. Daher wird eine Nachverarbeitung bzw. Aufbereitung des Grundfrequenzverlaufs vorgenommen.

Die Hauptaufgabe der Nachverarbeitungsstufe ist die Beseitigung von Messfehlern. Zwei Verfahren haben sich bei dieser Aufgabenstellung durchgesetzt: die *Listenkorrektur* und die *Glättung* [93]. Listenkorrekturverfahren benötigen Informationen aus der Extraktionsstufe, um mit deren Hilfe das Ergebnis der GFA korrigieren zu können. Sie können die Messfehler um bis zu eine Zehnerpotenz senken, sind allerdings auch relativ aufwendig in der Implementierung.

Die Glättung des Grundfrequenzverlaufs wird durch die Filterung der ermittelten Grundfrequenzwerte mit Median- und Tiefpassfiltern vorgenommen. Die Glättung mit einem Tiefpassfilter ist geeignet zur Reduktion von Messungenauigkeiten [93], ungeeignet hingegen zur Beseitigung von Grobfehlern. So wird beispielsweise ein Oktavfehler ($f_{\text{Mess}} = 2F_0$, richtig F_0) in einen nichtharmonischen Grobfehler (z.B. $0,8F_0$) verwandelt, der subjektiv sogar als unangenehmer als der ursprüngliche Fehler empfunden werden kann. Aus diesem Grund wurde von RABINER et al. in [158] die Glättung der Grundfrequenzverläufe mit Hilfe der Medianfilterung vorgeschlagen. Damit lassen sich Ausreißer wie zum Beispiel Oktavfehler gut korrigieren.

Beispiel: In Abbildung 5.12 sind die Modifikationen des Grundfrequenzverlaufs dargestellt. Der unbearbeitete Grundfrequenzverlauf zeigt besonders bei den Toneinsätzen starke Ausreißer, wie bereits im vorangegangenen Beispiel erläutert worden ist. Diese können durch eine 5-Punkte-Medianfilterung wirkungsvoll unterdrückt werden. Abbildung 5.12a zeigt den unbearbeiteten Grundfrequenzverlauf aus Abbildung 5.11 im Vergleich mit dem mediangefilterten Verlauf in Abbildung 5.12b.

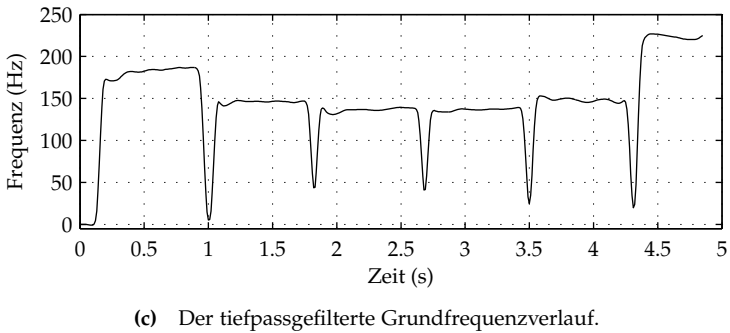
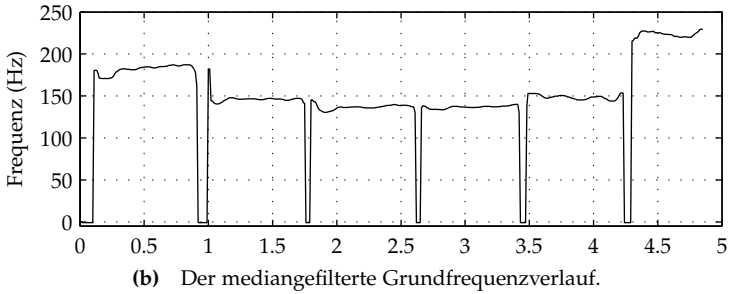
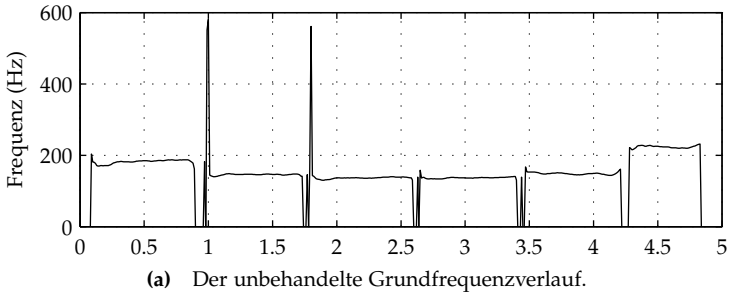


Abbildung 5.12: Nachverarbeitung des Grundfrequenzverlaufs des Signals aus Abbildung 5.11.

Anschließend wird durch eine Tiefpassfilterung eine weitere Glättung des Signals erreicht, die der anschließenden Rhythmuserkennung zuträglich ist, wie im Folgenden gezeigt wird. Den resultierenden Grundfrequenzverlauf zeigt Abbildung 5.12c.

5.4.2 Rhythmuserkennung

Die Rhythmuserkennung dient dem Auffinden der Anschläge, der Bestimmung der Dauer der einzelnen Töne sowie der Ermittlung des Tempos. Durch die Bestimmung der Anschläge und der Tondauer werden die einzelnen Notenereignisse im Signal ermittelt. Für die Rhythmuserkennung werden in bestehenden Verfahren Parameter wie der Grundfrequenzverlauf, die Einhüllende des Zeitverlaufs, Informationen über die Stimmhaftigkeit oder Tonalität des Signals (voicing) und das geschätzte Tempo verwendet [61, 132, 167].

Das System *Queryhammer* orientiert sich an bestehenden Verfahren aus QBH-Systemen, indem für die Erkennung der Anschläge ausschließlich der Grundfrequenzverlauf verwendet wird. Die Ermittlung des Tempos aus gesungenen Nutzeranfragen ist nicht befriedigend zu handhaben, wie in Abschnitt 5.3 erläutert worden ist. Daher wird auf die Tempobestimmung zugunsten einer ausführlichen Untersuchung der Melodiekontur allein verzichtet. Die Notendauer wird für die Konturdarstellung nicht benötigt und braucht daher nicht ermittelt zu werden.

Aufgabe ist es, den Grundfrequenzverlauf so zu unterteilen, dass gleichbleibende Grundfrequenzen einem Notenereignis zugeordnet werden können. Aus den resultierenden Notenereignissen kann schließlich die Melodiekontur transkribiert werden. Dazu wird jeder Wert des Grundfrequenzverlaufs einzeln untersucht. Wenn die untersuchte Frequenz im festgelegten Such- bzw. Gesangsbereich liegt, wird geprüft, welche musikalische Tonhöhe sich aus der Frequenz ergibt und um wieviel Cent sie von der gleichstufigen Stimmung abweicht. Man erhält damit einen Notennamen, die zugehörige Frequenz der musikalischen Tonhöhe in Hertz und eine Abweichung von dieser Frequenz in Cent. Gleichen sich die Notennamen, die zwei aufeinanderfolgenden Frequenzwerten des Grundfrequenzverlaufs zugeordnet werden, so ist ein Notenereignis entdeckt worden, dem alle weiteren Werte des Grundfrequenzverlaufs, die zum gleichen Notennamen ± 50 Cent führen, zugeordnet werden.

Problematisch bei dieser Methode ist, dass sie zu einer starken Segmentierung einzelner Notenereignisse führen kann, wenn die gemessene Frequenz einer gesummen oder gesungenen Note nicht hinreichend konstant ist. Die-

ser Fall liegt vor, wenn der Sänger sich nicht sicher ist und stark in der Tonhöhe schwankt. Dies führt zu einer Minderung der Güte der Transkription von Melodiekonturen, siehe hierzu auch die Untersuchungen in [41]. Aber auch absichtliche Frequenzschwankungen sind zu beobachten, etwa durch die Phrasierung der Noten durch das sogenannte „Anschleifen“, d. h. den Beginn eines Tons mit einer bewusst zu tief gewählten Frequenz zu singen und dann zu korrigieren, oder durch besondere Gestaltungen wie das Vibrato. Singt nun ein Sänger über die Frequenzbereichsgrenzen eines Tons hinweg, so werden zwei oder mehr Notenereignisse erkannt, die aus nur einer tatsächlichen Note hervorgehen.

Dieses Problem hat zwei Ursachen: Zum einen ist der Frequenzbereich, den die Frequenz einer gesungenen Note überstreichen kann, auf 100 Cent beschränkt. Zum anderen werden die Grenzen des zu erkennenden Tones eher überschritten, wenn die für die Segmentierung zugrunde gelegte Grundstimmung nicht der Stimmung des Sängers entspricht, der also „zu tief“ oder „zu hoch“ singt. Die Lösung, dem Benutzer eines QBH-Systems einen Stimmtun anzubieten, ist unbefriedigend. Es ist naheliegend, umgekehrt den Kammerton der Gesangseingabe anzupassen. Dies ist vor allem deshalb sinnvoll, weil damit die zu starke Segmentierung der Gesangseingabe wirkungsvoll unterbunden wird. Die Anpassung des Kammertons ist Gegenstand verschiedener Untersuchungen [90, 122, 133, 194]. Im Folgenden wird ein Algorithmus angegeben, der einfach und effizient ist.

Die Grundfrequenz F_0 wird gemäß Gleichung 4.1 in eine Tonhöhe der gleichstufigen Stimmung umgerechnet, und ein Notenereignis $E(n)$ wird initialisiert. Dabei werden Abweichungen von ± 50 Cent akzeptiert, um den Ton der gleichstufigen Skala zuzuordnen. Alle weiteren Werte von F_0 werden $E(n)$ zugeordnet, solange sie die zulässige Abweichung in Cent nicht überschreiten. Wird sie überschritten, wird ein neues Notenereignis $E(n+1)$ initialisiert und für $E(n)$ der Medianwert aller enthaltenden Grundfrequenzen $F_{m0}(n)$ ermittelt, der die Tonhöhe des Ereignisses wiedergibt.

Für das neue Ereignis $E(n+1)$ wird untersucht, ob $E(n)$ eine fest vorgegebene Minstdauer von $T_{\min} = 200$ ms überschreitet. Ist das der Fall, so wird aus $F_{m0}(n)$ eine neue Stimmung des Kammertons a berechnet. Weicht also $F_{m0}(n)$ um c_{m0} Cent von der bisherigen Stimmung ab, so werden alle nachfolgenden Töne bzgl. eines um c_{m0} Cent korrigierten Kammertons a berechnet. Eine ähnliche Vorgehensweise findet sich bei Mc NAB [133]. Die Beobachtung zeigt, dass Sänger in der Intonation oft abfallen, d. h. die Stimmung des Kammertons a sinkt.

Da ein Mensch nur begrenzt schnell singen kann, werden nur Ereignisse oberhalb einer festen Mindestdauer für die Transkription berücksichtigt [29]. Alle Ereignisse mit $t(n) < 100$ ms werden verworfen, dies entspricht Sechzehntelnoten bei M.M. = 150.

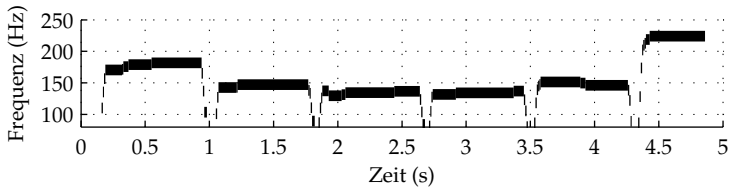
Beispiel: In Abbildung 5.13 ist die Auswertung des in Abbildung 5.12c dargestellten Grundfrequenzverlaufs zu sehen. Zuerst sind die Abschnitte konstanter Frequenz in Bezug auf die gleichstufige Stimmung *ohne* jede weitere Korrektur dargestellt (Abbildung 5.13a), es werden nur Schwankungen der Tonhöhe von ± 50 Cent um konstante Frequenzwerte zugelassen. Man erkennt Gruppen von Ereignissen, die in der Frequenz schwanken, was zu einer starken Segmentierung der gesungenen Noten führt (siehe zum Beispiel die Notenanfänge bei Sekunde 0 und 1).

In Abbildung 5.13b sind die Abschnitte konstanter Grundfrequenz *mit* oben beschriebener Korrektur dargestellt, die Ereignisse entsprechen deutlich besser den gesungenen Noten. Der Verlauf der zugehörigen Stimmung des Kammertons *a* in Abbildung 5.13d lässt ein deutliches Abfallen der Stimmung Note für Note erkennen. Das letzte Intervall ist zu weit aufwärts gesungen, daher steigt die Stimmung des Kammertons zum Ende wieder.

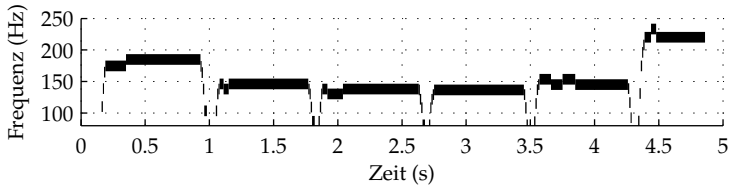
Schließlich werden die für die Transkription der Melodiekontur relevanten Notenereignisse ausgewählt, indem alle konstanten Grundfrequenzabschnitte, die kürzer als $T_{\min} = 100$ ms andauern, verworfen werden. Das Ergebnis in Abbildung 5.13c gibt die gesungene Tonfolge als Klavierwalze wieder, enthält aber einen Fehler: die erste Note ist in zwei Noten segmentiert worden, da ein zu weiter Frequenzbereich beim Singen überstrichen worden ist (der Ton wurde „angeschliffen“). Das Problem der Einfügung einer zusätzlichen Note wird im folgenden Abschnitt genauer untersucht.

5.4.3 Praktische Versuche

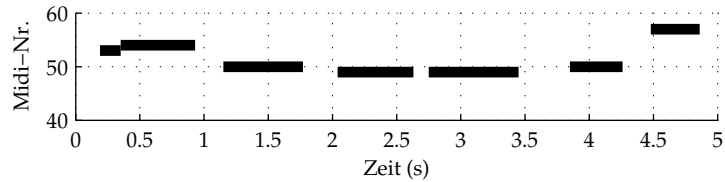
In Abschnitt 2.4 ist bereits diskutiert worden, welche Fehler häufig beim Singen einer Melodieanfrage gemacht werden. In diesem Abschnitt wird nun betrachtet, welche Fehler das monophone Transkriptionssystem beim Ermitteln der Melodiekontur verursachen kann. Zur Untersuchung und Bewertung



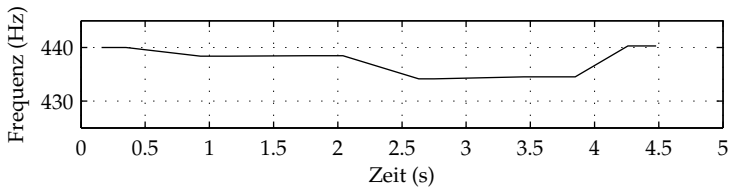
(a) Die Notenereignisse aus dem Frequenzverlauf ohne Korrektur der Stimmung; dargestellt sind die Abschnitte konstanter Frequenz.



(b) Wie Abbildung 5.13a, aber mit Korrektur der Stimmung; die nun erkannten Frequenzen sind über weitere Abschnitte konstant.



(c) Die ausgewählten Notenereignisse in Klavierwalzendarstellung. Der erste Ton wird zu Beginn zu tief intoniert und als weitere Note erkannt.



(d) der Verlauf der zugrunde gelegten Stimmung des Kammertons a. Man erkennt, dass die Stimmung des Sängers mit den Tönen kontinuierlich fällt. Der letzte Ton ist relativ zum Vorgängerton zu hoch intoniert.

Abbildung 5.13: Beispiel für die Rhythmuserkennung: Aus dem Grundfrequenzverlauf in Abbildung 5.12c werden die Notenereignisse ermittelt.

solcher Fehler finden sich in der Literatur verschiedene Ansätze: vergleichende Hörtests sorgen für eine eher subjektive Beurteilung [29], das Abzählen der Notenfehler insgesamt [90, 153] oder Ordnung in verschiedene Kategorien wie Auslassung, Einfügung, Oktavfehler usw. führt zu einer objektiveren Bewertung [50, 130, 142]. Wünschenswert ist ein Testverfahren, das unabhängig vom Vergleich und der Bewertung von Menschen ist, sondern sich allein auf messbare Daten stützt. Daher wird im Rahmen dieser Arbeit ein vergleichendes Bewertungsverfahren entwickelt.

Zur Untersuchung werden nun alle für die Transkription einer Melodiekontur wesentlichen Fehlerkategorien untersucht. Diese sind:

Auslassung Eine gesungene Note wird nicht erkannt, stattdessen wird eine Pause ausgegeben.

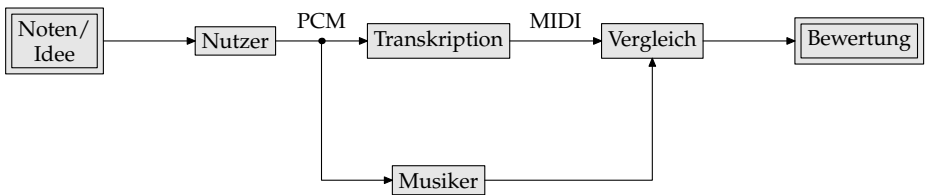
Zusammenfassung Zwei aufeinanderfolgende, gleiche Noten werden zu einer Note zusammengefasst.

Einfügung Eine zusätzliche Note wird in das Transkriptionsergebnis eingefügt.

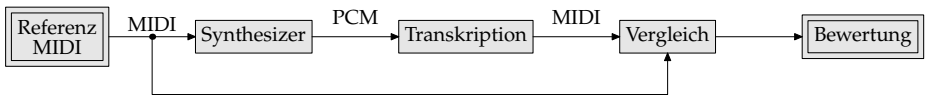
Die Auswirkung dieser Fehler auf die Werte der Melodiekontur selbst und Konsequenzen für das Suchergebnis eines QBH-Systems werden in Kapitel 8 ausführlich diskutiert.

Testverfahren

Das zu untersuchende monophone Transkriptionssystem übersetzt ein PCM-Signal in eine symbolische Darstellung von Noten, beispielsweise MIDI-Daten. Um das Transkriptionssystem testen zu können, muss für die zu extrahierenden Symbole des PCM-Signals eine Referenz vorliegen. Die Referenz wird mit dem Transkriptionsergebnis verglichen, aus dem Vergleich folgt die Beurteilung (siehe Abbildung 5.14). Für die Untersuchung eines Transkriptionssystems für Gesangsanfragen bietet sich die Verwendung gesungener Melodien an. Lässt man einen Sänger eine vorgegebene Notenfolge singen, so kann das transkribierte Ergebnis mit den vorgegebenen Noten (Symbolen) verglichen werden. Bei dieser Vorgehensweise tritt unweigerlich das Problem auf, dass man nicht nur das Transkriptionssystem, sondern auch den Sänger bzw. seine Notenkenntnisse testet. Es ist darum nicht ohne weiteres möglich zu beurteilen, ob alle Noten richtig gesungen wurden.



(a) In dieser Anordnung erfolgt die Bewertung des Transkriptionssystems durch einen Musiker. Ein direkter Vergleich der Transkription mit den ursprünglichen Noten oder der Idee des Nutzers ist nicht möglich.



(b) Erfolgt die Bewertung des Transkriptionssystems unter Zuhilfenahme eines Synthesizers, kann ein direkter Symbolvergleich vorgenommen werden.

Abbildung 5.14: Verschiedene Testszenarien zur Bewertung von monophonen Transkriptionssystemen.

Das Problem mangelnder Notenkenntnis lässt sich umgehen, indem man einen Probanden eine bekannte Melodie freier Wahl singen lässt. Um die Fehler des Transkriptionssystems ermitteln zu können, muss für die zu transkribierende Anfrage eine Referenz generiert werden. Eine Möglichkeit ist es, einen Musiker diese Referenz durch Anhören und manuelle Transkription erstellen zu lassen [50] (Abbildung 5.14a). Dieser Schritt ist wiederum nicht unkritisch, da ein Musiker aus musikalischen Erwägungen viele Gesangsfehler ausgleichen kann, insbesondere dann, wenn er die Melodie schon kennt, ihm Bekanntem annähert und Fehler unwillkürlich ausgleicht.

In der vorliegenden Arbeit wird daher der Weg gewählt, ausgehend von einer MIDI-Datei die Gesangseingabe zu synthetisieren (Abbildung 5.14b). Um ein möglichst gesangsähnliches Signal zu erhalten, wird der General-MIDI-Klang *Doo Voice* (MIDI-Programmnummer 56) verwendet. Zur Synthese des Signals wird ein Synthesizer des Herstellers *Korg* Modell 05 RW benutzt. Der Vorteil dieser Vorgehensweise ist, dass das Testsignal exakt der MIDI-Datei entspricht und die Referenz eindeutig richtig ist. Weiterhin lässt sich die gleiche Eingabe in verschiedenen Tempi generieren und der Einfluss des Tempos auf die Güte der Transkription untersuchen. Nachteil der Vorgehensweise ist, dass ein Synthesizer ohne weitere Maßnahmen keine für Menschen typischen

Abweichungen der Intonation produzieren kann. Daher sollen auch gesungene Nutzersignale untersucht werden. Im Folgenden beschreibt Versuch 1 die Untersuchung synthetischer und Versuch 2 die Untersuchung von Menschen erzeugter Signale.

Versuch 1

Für den Versuch werden die Melodien der untersuchten Top-10 in ihrer MIDI-Version verwendet (siehe Kapitel 8). Dabei wird der Refrain jedes Titels in vier Tempi mit M.M. = 100, 120, 140 und 160 generiert. Die mit dem Synthesizer erzeugte Audio-Datei wird mit dem monophonen Transkriptionsteil des *Queryhammer* wieder in eine MIDI-Datei zurück übersetzt. Durch Auswertung der Klavierwalzendarstellung von Original-MIDI-Datei und transkribierter MIDI-Datei ergeben sich die Fehlerquoten für die einzelnen Fehlerklassen.

Die Ergebnisse des Versuchs sind in Abbildung 5.15 dargestellt. Die ermittelte Fehlerquote in Prozent ist das Verhältnis der fehlerhaften Noten zu allen Noten der jeweils transkribierten Melodie.

- Die Fehlerquote für Auslassungen liegt bei knapp 8 % und steigt mit wachsendem Tempo auf gut 14 %. Diese Fehler entstehen besonders dann, wenn eine Note zu kurz gehalten wird und vom Transkriptionssystem keine Tonhöhe zugeordnet werden kann.
- Am häufigsten sind die Fehler durch Zusammenfassungen im Bereich von 9–18 %. Dieser Fehler tritt bei Tonwiederholungen auf, insbesondere dann, wenn die einzelnen Töne nicht deutlich phrasiert sind und als neues Notenereignis erkannt werden können. Ein Titel wie „All the things she said“ von Tatu (siehe Abschnitt 8.2.1) enthält viele Tonwiederholungen und kann nur bei langsamem Tempo fehlerfrei transkribiert werden.
- Bei keinem der untersuchten Melodien sind Einfügingsfehler aufgetreten. Die Ursache dafür ist die perfekte Intonation des Synthesizers, die bei Menschen ohne besondere musikalische Ausbildung in der Regel nicht gegeben ist. Zusätzliche Noten werden durch Sänger dadurch verursacht, dass eine Note mit so starken und zeitlich ausgedehnten Intonationsschwankungen gesungen wird, dass aufgrund der sich verändernden Tonhöhe zwei Notenereignisse zugeordnet werden. Ein solcher Fehler tritt etwa im Beispiel zu Abschnitt 5.4.2 auf.

Im Umkehrschluss kann festgestellt werden, dass das synthetische Signal nicht in der Praxis auftretenden Gesangssignalen entspricht. Um die Problematik der Einfügsfehler zu illustrieren, soll ergänzend dargestellt werden, wie sich das Transkriptionsmodul bei durch Menschen erzeugten Signalen verhält.

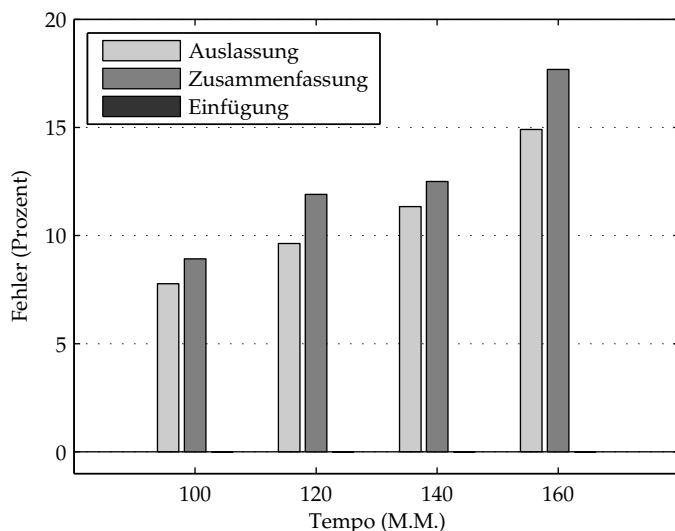


Abbildung 5.15: Die Fehlerquoten der monophonen Transkriptionsstufe in Prozent. Transkribiert wurden per Synthesizer gespielte Melodien (Top-10) in verschiedenen Tempi.

Versuch 2

Da wie oben erläutert eine eindeutige Referenz bei gesungenen Eingangssignalen fehlt, werden nun Abschnitte aus Signalen ausgewählt, die sich eindeutig durch Bewertung des Wellenformverlaufs einer gesungenen Note zuordnen lassen. Diese Note wird transkribiert, ist die Anzahl der erkannten Noten größer eins, so handelt es sich um eine Einfügung.

Insgesamt wurden die Signale von vier Personen untersucht, es wurden je Teilnehmer 10 Signalabschnitte ausgewählt. In Tabelle 5.1 sind die Fehlerquoten der Einfügsfehler je Teilnehmer und im Durchschnitt dargestellt. Es zeigt sich, dass die Zahl der Einfügsfehler sehr stark von der Fähigkeit

des Teilnehmers abhängt, so haben Teilnehmer 1 und 2 wesentlich weniger Einfügingsfehler zu verzeichnen als Teilnehmer 3 und 4. Mit dieser Untersuchung ergibt sich somit eine Aussage darüber, wie intonationssicher, d. h. „wie gerade“ die untersuchten Teilnehmern einen Ton halten können. Für das Transkriptionsverfahren ergibt sich im Gegenzug, dass für die richtige Erkennung der Töne die Bewertung aus dem Frequenzverlauf alleine nicht ausreichend ist. Anders ausgedrückt kann man feststellen, dass das angenommene Modell eines Tones mit einer Frequenz, die maximal ± 50 Cent von der Mittenfrequenz abweicht, in der Realität nicht immer so vorzufinden ist. Um eine höhere Robustheit der Transkription zu erreichen, ist die empirische Ermittlung des Frequenzverlaufs für gesungene Töne denkbar, mit denen ein Transkriptionssystem trainiert werden könnte.

Tabelle 5.1: Einfügingsfehler pro Versuchsteilnehmer bei gesummtten Einzelnoten, die als Signalabschnitt vorselektiert wurden.

Tn	1	2	3	4	Durchschnitt
Fehlerquote (%)	10	10	40	60	30

Es kann abschließend festgestellt werden, dass QBH-Systeme mit einem wie bei *Queryhammer* gewählten akustischen Front-End auf relativ sicher gesungene Anfragen angewiesen sind, um die Anfrage richtig zu transkribieren.

5.5 Zusammenfassung

In diesem Kapitel wurden die Aufgaben der monophonen Transkription in einem QBH-System vorgestellt. Wesentliche Aufgabe der Transkription ist es, Tonhöhe, Anschlag und Tondauer der gesummtten Eingabenoten richtig zu erkennen. Da die Informationen über Anschlag und Tondauer aus dem Grundfrequenzverlauf gewonnen werden können, kommt den Verfahren der Grundfrequenzanalyse (GFA) besondere Bedeutung zu. Daher wurden verschiedene Verfahren zur GFA ausführlich vorgestellt und ihre Eignung für die Transkriptionsaufgabe diskutiert. Hinsichtlich ihrer Genauigkeit und Robustheit ist die Autokorrelationsmethode gut geeignet, sie wird darum im Beispielsystem *Queryhammer* verwendet. Details der Implementierung wurden im Rahmen des Abschnitts über die eigenen Untersuchungen genau beschrieben.

An die GFA schließt sich die Rhythmuserkennung zur Ermittlung der einzelnen Notenereignisse an. Da die einfache Zusammenfassung von Abschnitten gleicher Grundfrequenz zu fehlerhaften Ergebnissen führen kann, wurde ein eigener Algorithmus vorgestellt, mit dem Tonhöhenschwankungen durch den Sänger, beabsichtigt als Phrasierung oder nicht beabsichtigt durch Intonationsunsicherheit, in gewissen Grenzen aufgefangen werden können. Das Problem der Tonhöhenschwankungen innerhalb einer Note führt zu Einfügungen von Noten im transkribierten Ergebnis.

In den abschließenden Untersuchungen wurde ein neues Prüfverfahren zur Untersuchung des monophonen Transkriptionssystems vorgestellt, in dem verschiedene Melodien durch ein gesangsähnliches Synthesizersignal in verschiedenen Tempi erzeugt und transkribiert wurden. Durch dieses Verfahren ist eine objektive Untersuchung der Transkription möglich. Es zeigte sich, dass hier aufgrund der fehlenden Tonhöhenschwankung durch den Sänger ausschließlich Auslassungs- und Zusammenfassungsfehler auftraten.

Als Ergebnis dieses Kapitels kann festgestellt werden, dass die Güte der monophonen Transkription maßgeblich von der Phrasierung und Intonationsunsicherheit des Sängers abhängen. Erfolgt eine intonationssichere Gesangseingabe, kann die Melodie fehlerfrei transkribiert werden. Unsaubere Phrasierung in von undeutlichen Ton- bzw. Notenanfängen führt zu Auslassungs- und Zusammenfassungsfehlern, genauso wie ein zu schneller Vortrag. Die Fehlerquelle liegt in diesem Fall im Transkriptionsverfahren. Einfügungsfehler lassen sich allein auf starke Tonhöhenschwankungen des Sängers zurückführen, die nicht mehr abgefangen werden können.

Polyphone Transkription

6

*Es ist nicht schwer, zu
komponieren. Aber es ist
fabelhaft schwer, die
überflüssigen Noten unter den
Tisch fallen zu lassen.*

Johannes Brahms

Die Melodiedatenbank eines Query-by-Humming-Systems (QBH-Systems) benötigt Melodiebeschreibungen der abgespeicherten Musikstücke. Liegen diese in symbolischer Repräsentation wie MIDI-Dateien vor, so ist es vergleichsweise leicht, andere symbolische Melodiebeschreibungen wie den PARSONS-Code oder die MPEG-7-MelodyContour zu extrahieren. Sind die Musikstücke als Audiodatei gespeichert, benötigt man hingegen aufwendige Signalverarbeitungsverfahren. Die Extraktion von symbolischen Informationen wie Melodiekonturen aus Audiodateien ist schwierig. Diese Aufgabe ist sehr ähnlich zu dem bekannten Problem der automatischen Transkription von Noten. Besonders anspruchsvoll ist dieses, wenn Noten von polyphoner Musik, bei der mehrere Töne gleichzeitig erklingen, transkribiert werden sollen. Das wesentliche Signalverarbeitungsproblem der polyphonen Transkription ist die Mehrfachgrundfrequenzanalyse (MGFA). Der Inhalt der Audiodateien einer Musikdatenbank ist meistens polyphon und die Extraktion der Melodiekonturen aus diesen Audiodateien zum Aufbau einer Melodiedatenbank ist kein notwendiger Bestandteil eines QBH-Systems. Jedoch ist die Datenbank Kernbestandteil einer jeden Suchmaschine und Inhalt, Umfang und Qualität der Datenbank sind ausschlaggebend für die Attraktivität der Suchmaschine.

Abbildung 6.1 zeigt die Verarbeitungsböcke der polyphonen Transkriptionsstufe im Beispielsystem *Queryhammer*: Wie bei der monophonen Transkription sind die Verarbeitungsböcke Tonhöhenerkennung, Rhythmuserkennung und Auswertung vorhanden. Neu ist die Melodieerkennung: Sie entscheidet, welche Teile der extrahierten Information die Melodie darstellen. Dies ist im Gegensatz zum monophonen Fall nicht mehr eindeutig.

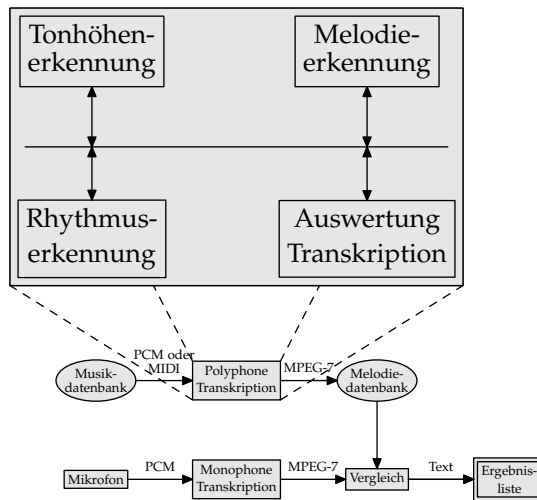


Abbildung 6.1: Die Transkriptionsstufe für polyphone Signale im *Queryhammer*.

Die Arbeitsreihenfolge dieser Stufen ist nicht klar festgelegt. So kann zum Beispiel die Melodieerkennung zuerst bestimmen, welche Teile des Audiosignals transkribiert werden sollen, und danach erfolgt die Tonhöhen- und Rhythmuserkennung. Es ist aber auch möglich, zunächst alle Notenereignisse im untersuchten Musiksignal zu ermitteln, diese nach weiteren Kriterien zu bewerten und dann die Melodie festzulegen.

Abschnitt 6.1 beschreibt die bisher üblichen Ansätze, für QBH-Systeme Melodiedatenbanken zu erstellen. Es wird dargelegt, warum die bisher nicht verwendete polyphone Transkription für diese Aufgabe interessant ist. Danach werden in Abschnitt 6.2 aktuelle Verfahren der automatischen Transkription von Musiksignalen vorgestellt. Abschnitt 6.3 beschreibt das im Rahmen dieser Arbeit implementierte polyphone Transkriptionsverfahren. Weiterhin werden an einigen Beispielen die Möglichkeiten einer solchen Transkriptionsstufe demonstriert.

6.1 Melodie-Transkription für Query-by-Humming-Systeme

Die Melodiebeschreibungen der Datenbanken bestehender QBH-Systeme werden häufig automatisch aus MIDI-Daten (siehe auch Abschnitt 2.3.2) erstellt [11, 187] oder manuell extrahiert und direkt eingegeben [12, 51]. Beide Verfahren werden im Folgenden kurz diskutiert.

6.1.1 MIDI-Transkription

MIDI-Dateien werden gerne verwendet, weil sich die Melodiekontur aus einer bereits symbolischen Darstellung von Musik ohne großen technischen Aufwand herauslesen lässt. Problematisch ist neben der trivialen Voraussetzung, dass eine dem Datenbank-Eintrag entsprechende MIDI-Datei vorhanden sein muss, allerdings die Auswahl der richtigen Spur.

Üblicherweise sind in einer MIDI-Datei mehrere Spuren vorhanden, die verschiedenen Instrumenten zugeordnet sind. Die Melodie selbst kann von jedem auftretenden Instrument wechselweise oder auch gleichzeitig mit anderen Instrumenten zusammen gespielt werden. Die Auswahl der richtigen Spur ist durch Suche nach einer passenden Spurbezeichnung möglich; trägt eine Spur z. B. die Bezeichnung „Melodie“, „melody“, „Gesang“, „voice“ oder „lead“, ist das Auftreten der Melodie in dieser MIDI-Spur zu erwarten [108]. Aus dieser Spur kann dann eine für den untersuchten Titel repräsentative Melodiekontur extrahiert werden.

MIDI-Dateien sind in großer Zahl im Internet zu finden und es wird eine große Anzahl musikalischer Genres abgedeckt. Problematisch bei der Verwendung von MIDI-Dateien ist jedoch:

- Nicht für jedes Musikstück ist eine MIDI-Datei vorhanden. Je nach Genre sind Musikstücke gut oder gar nicht durch MIDI-Dateien darstellbar. Pop-Musik ist sehr gut durch MIDI-Dateien darstellbar, weil diese Musik meist auf Basis von MIDI-gesteuerten Musikinstrumenten produziert wird. Auch klassische Musik des westlichen Kulturkreises lässt sich im MIDI-Format speichern. Ethnische Musik hingegen ist in der Regel sehr schlecht über MIDI darstellbar; in [174] wird dieses Problem speziell für indische Musik diskutiert.

- MIDI-Dateien repräsentieren häufig nicht den genauen Inhalt einer Audiodatei bzw. der Melodie, die man hört. Besonders Gesang als melodieführendes Instrument mit allen Variationen der Stimme, Intonation u. ä. lässt sich unter MIDI nicht direkt wiedergeben und wird oft durch eine abgewandelte instrumentale Fassung ersetzt.
- Wie diskutiert, ist die Melodie in einer MIDI-Datei nicht immer eindeutig gekennzeichnet, eine automatische Auswahl einer bestimmten Spur ist schwierig. Darüber hinaus kann die Melodie auf verschiedene Spuren verteilt sein, was eine weitere Auswertung der MIDI-Informationen erfordern kann.

6.1.2 Manuelle Transkription

Werden Melodien bzw. Melodiekonturen manuell ermittelt, so hängt die Qualität der Transkription von der musikalischen Auffassung des Transkribierenden und dem erstellten Notentext ab. Je nach musikalischem Genre werden Melodien von Musikern beim Spielen leicht von der Notenform abgewandelt (Pop-Musik, einige Sparten der klassischen Musik), in einigen Fällen erhebt die Notation sowieso nur den Anspruch einer musikalischen Skizze und wird stark interpretiert (zum Beispiel im Jazz, siehe auch Kapitel 2). Werden Melodien manuell notiert, so kann dies sehr nuanciert geschehen. In Bezug auf QBH-Systeme bestehen aber auch hier Nachteile:

- Das Verfahren ist für große Datenbankbestände zeitaufwendig und kostenträchtig.
- Die Qualität der Melodiedarstellung hängt stark von der musikalischen Auffassung bzw. den Kenntnissen des Transkribierenden ab und kann dementsprechend schwanken.

6.1.3 Transkription aus Audiosignalen

Den bisher beschriebenen Verfahren steht die Extraktion von Melodiebeschreibungen aus Audiosignalen gegenüber. Diese Aufgabe wird seit Beginn der 70er-Jahre in der Signalverarbeitung als *Automatische Transkription von Musik* erforscht. Folgende Probleme sind besonders bedeutsam:

- Die Transkription polyphoner Musik ist komplex und sehr vom Inhalt des untersuchten Audiomaterials abhängig. Opern-Arien mit langen, im

Vordergrund stehenden Melodietönen sind z.B. wesentlich leichter zu verarbeiten als BACHsche Fugen mit gleichberechtigten Stimmen.

- In den vorhandenen Informationen muss der relevante Teil, die Melodie, erkannt werden. Bewertungskriterien hierfür sind nicht allgemeingültig zu formulieren. In vielen Fällen kann man davon ausgehen, dass eine Melodie dominant bezüglich ihrer Lautstärke ist, damit sie gut zu hören ist. Allerdings ist ein Musikhörer ohne weiteres in der Lage, auch hinsichtlich ihrer Lautstärke im Hintergrund stehende Melodien zu erkennen.

Die Vorteile der Transkription aus Audiodateien gegenüber der Transkription aus MIDI-Daten oder per Hand sind erheblich:

- Audiodateien stellen die Melodie so dar, wie sie vom Musikhörer wahrgenommen wird. Damit werden Abweichungen des Notentextes oder der MIDI-Informationen von der gehörten bzw. gespielten Melodie vermieden.
- Für den häufigen Fall, in dem ein QBH-Nutzer die Audiorepräsentation einer Melodie sucht, erfolgt die Extraktion der Melodiebeschreibung direkt aus dem gesuchten Medium und ermöglicht damit die beste Beschreibung der Melodie.
- Eine kostenträchtige Extraktion per Hand lässt sich vermeiden, wenn eine vollkommen automatische Erzeugung der Melodiebeschreibung möglich ist.

Während bisherige Ansätze die vollständige Erkennung *aller* Notenereignisse anstreben, braucht für die automatische Extraktion von Melodiedarstellungen aus Audiosignalen nur die Information der Melodiestimme extrahiert werden. Weiterhin ist bei QBH-Systemen bei Verwendung der Melodiekontur keine notengenaue Melodiedarstellung erforderlich, wie in Kapitel 5 schon gezeigt worden ist.

Aus den angeführten Gründen wird klar, dass die Transkription der Melodiedarstellung im günstigsten Fall über das Medium erfolgt, das über die Datenbank angeboten werden soll. Für eine Musikdatenbank, die Audiodateien im Wave- oder MP3-Format zur Verfügung stellt, sollten die Melodiekonturen daher aus den angebotenen Audiodateien selbst extrahiert werden.

6.2 Automatische Transkription von Musiksignalen

Die Transkription von Musik aus Audiosignalen wird in der Signalverarbeitung immer mit dem Ziel der möglichst genauen Extraktion der Noteninformation verfolgt. Bislang gibt es keine Transkriptionswerkzeuge, die allgemein für die Erzeugung von Melodiedarstellungen für die Verwendung in QBH-Systemen geeignet sind.

Die Aufgabe der polyphonen Transkription ist eng verbunden mit der Aufgabe der Mehrfach-Grundfrequenzanalyse (MGFA). Die MGFA ist für die Tonhöhenerkennung notwendig, aber auch rhythmische Informationen lassen sich wie im monophonen Fall aus der MGFA ermitteln. Es sollen nun einige aktuelle Ansätze zur MGFA dargestellt werden, die in Hinblick auf die automatische Extraktion von Melodien aus Audiosignalen interessant sind.

6.2.1 Tonhöhenerkennung

In der Literatur sind bereits eine Vielzahl von MGFA-Methoden bekannt [111]. Diese MGFA-Methoden werden zur besseren Übersicht in mehrere Kategorien unterteilt, die im Folgenden kurz beschrieben werden. Die Einteilung ist nicht immer eindeutig möglich, da die Methoden sehr komplex sind und mehreren Verarbeitungsprinzipien gehorchen.

Wahrnehmungsbezogene Gruppierung von Teiltönen

Das menschliche Gehör arbeitet sehr effizient in Bezug auf die Erkennung und Unterscheidung individueller Klänge. Diese kognitive Fähigkeit beschreibt man auch mit dem Begriff der „auditiven Szenenanalyse“ (auditory scene analysis, ASA). Die computergestützte ASA (CASA) wird üblicherweise als zweistufiger Prozess betrachtet, wobei das Eingangssignal zuerst einer Zeit-Frequenz-Analyse unterzogen wird und die erkennbaren Elemente der Audioinformationen dann ihrem Ursprung zugeordnet werden. Vorausgesetzt, dass dieser Vorgang erfolgreich war, kann nun jede einzelne Komponente einer gewöhnlichen GFA unterzogen werden. In der Praxis findet die GFA jedoch schon im Analyseteil der CASA statt. Eine Anwendung der CASA zur polyphonen Transkription findet man in [104].

Gehörmodell-basierte Ansätze

Signalverarbeitungsmodelle des menschlichen Gehörs lassen sich gut für die Aufgaben der MGFA verwenden, insbesondere zur Vorverarbeitung des Signals. Sehr häufig wird das „unitary pitch model“ (UPM) verwendet [134], zum Beispiel in den Algorithmen in [111] oder [174]. Die grundlegenden Signalverarbeitungsschritte sind: eine Analyse mittels einer Bandpassfilterbank zur Modellierung der Frequenzselektivität des Innenohrs, ein Halbwellengleichrichter, der die neuronale Weiterleitung der Hörreize modelliert, anschließend die Berechnung der Autokorrelationsfunktion (AKF) in jedem Bandpass-Kanal und abschließend die Berechnung der Summen-AKF aller Kanäle. Die Berechnung des UPM ist sehr aufwendig, ein Verfahren zur effizienten Berechnung vor dem Hintergrund der Musiktranskription wird von KLAPURI in [114] vorgestellt.

Wandtafel-Architekturen

Wandtafel-Architekturen (blackboard architectures) konzentrieren sich auf die Integration von Wissen. Der Name „Wandtafel“ bezieht sich dabei auf die Metapher, dass eine Runde von Experten um eine Wandtafel versammelt steht und versucht, ein Problem zu lösen [110]. Jeder Experte kann die Entwicklung des Lösungsansatzes verfolgen und bei Bedarf Ergänzungen an die Wandtafel anschreiben.

Eine Wandtafel-Architektur ist aus drei Komponenten aufgebaut:

1. Die Wandtafel als erste Komponente ist ein hierarchisches Netzwerk von Hypothesen. Die Eingabedaten stehen dabei auf unterster Stufe, die Analyseergebnisse auf höherer Ebene. Die Hypothesen sind voneinander abhängig. Die Wandtafel kann auch als Datendarstellungshierarchie betrachtet werden, weil die Hypothesen Daten auf verschiedenen Abstraktionsebenen codieren. Die Intelligenz des Systems wird in Wissensquellen (WQ) abgelegt.
2. Die zweite Komponente besteht aus den Algorithmen, welche die Daten der Wandtafel manipulieren können.
3. Die dritte Komponente, das Steuerprogramm, entscheidet, welche WQ am Zug ist und agieren soll. Da der Zustand der Analyse vollständig in den Wandtafel-Hypothesen enthalten ist, ist es relativ einfach, neue WQ hinzuzufügen und das System zu erweitern.

BELLO und SANDLER beschreiben ein System mit Wandtafel-Architektur zur automatischen Transkription von einfachen polyphonen Audiosignalen [30]. In ihren Untersuchungen werden Aufnahmen transkribiert, die allein Klaviermusik enthalten; es können Akkorde erkannt und dargestellt werden.

Signalmodell-basierter Ansatz

Es ist möglich, die Aufgabe der MGFA als Schätzaufgabe zu beschreiben, wobei die untersuchten Klänge durch ein Signalmodell dargestellt werden und die Grundfrequenz der zu schätzende Parameter des Signals ist.

KLAPURI et al. beschreiben ein automatisches Transkriptionssystem für polyphone Audiosignale, das auf einem solchen MGFA-Verfahren basiert [112]. Es wird nach Klängen gesucht, die ein harmonisches Obertonspektrum besitzen. Frequenzkomponenten, die einem harmonischen Klang zugeordnet werden können, werden sukzessive aus dem Signal entfernt. In mehreren Analysedurchläufen kann das untersuchte Signal auf diese Weise vollständig parametrisiert werden. Das Verfahren funktioniert gut für synthetisch erzeugte Audiosignale, weist jedoch für Signale akustischen Ursprungs Probleme auf.

GOTO beschreibt in [79] das Verfahren *PreFEst*, in dem das Kurzzeitspektrum des Musiksignals modelliert wird. Dazu verwendet er ein Tonmodell, das aus einer Reihe von harmonischen Teiltönen besteht. Über den Estimation-Maximisation-Schätzalgorithmus werden die zugehörigen Grundfrequenzen gefunden. Dieses Verfahren wird in Abschnitt 6.3 näher beschrieben.

Datenadaptive Techniken

In datenadaptiven Systemen sind weder parametrische Modelle noch Kenntnisse über die Quelle vorhanden [111]. Statt dessen werden die Quellensignale aus den Daten geschätzt. Es werden auch keinerlei harmonische Spektren der natürlichen Klänge vorausgesetzt. Für normale Musiksignale ist der Erfolg solcher Verfahren, zum Beispiel die Trennung unabhängiger Komponenten (independent component analysis, ICA) begrenzt. Durch Einführung einiger Beschränkungen für die Quellen lassen sich aber brauchbare Ergebnisse erzielen.

SMARAGDIS und BROWN verwenden eine spektrale Vorverarbeitung in Form der nichtnegativen Matrixfaktorisierung [178]. Das Ergebnis dieser Vorverarbeitung sind Teilspektren mit rhythmisch unabhängigen Komponenten, die sich dann relativ leicht transkribieren lassen. Die Untersuchungen beziehen

sich auf eine polyphone Klavierpassage. Das Verfahren arbeitet jedoch nicht automatisch, die vorgestellten Untersuchungen sind eher prinzipieller Natur.

Weitere Techniken

Ein interessanter Ansatz, der in einem Verfahren in [78] gemacht wird, ist die Verwendung einer Support-Vector-Machine (SVM). Damit wird das Problem der Transkription als Klassifizierungsaufgabe aufgefasst. Weitere Verfahren zur polyphonen Transkription werden in den Dissertationen von KLAPURI [111] und HAINSWORTH beschrieben [87].

6.2.2 Rhythmuserkennung

Die Rhythmuserkennung umfasst, wie bereits in Abschnitt 5.3 beschrieben, die Aufgaben der Anschlags-, Zählzeiten- und Tempoerkennung. Auch im polyphonen Fall können wie im monophonen Fall viele Informationen über Anschläge (Notenanfänge) und Notenlängen aus dem Grundfrequenzverlauf extrahiert werden. Für die Transkription in eine notengerechte Darstellung sind aber weitere Informationen über den Takt und die Taktart notwendig. Ein Verfahren zur Extraktion solcher Informationen wird in [109] vorgestellt. Einen Überblick über Verfahren zur Erkennung der Hauptzählzeiten findet man bei GOUYON [82]. Bei der Transkription in die Notenschrift treten weitere Probleme auf, wenn eine gut lesbare Notation erfolgen soll. So ist eine sinnvolle Unterteilung der Zählzeiten bzw. Quantisierung der Notenlängen bzgl. des Metrums notwendig. CEMGIL et al. [44] beschäftigen sich mit der Quantisierung der Notenlängen unter Berücksichtigung von Noten- und Tempovariationen.

Für polyphone Musikstücke können andere Strategien zur Rhythmuserkennung verfolgt werden als bei monophonen Signalen, insbesondere dann, wenn stark perkussive Klänge vorhanden sind, zum Beispiel die Begleitung durch Schlaginstrumente. ALONSO und RICHARD verwenden in ihrem System für die Anschlagserkennung eine Filterbank, um in polyphonen Signalen das Tempo zu verfolgen [22]. Wie Untersuchungen von ALGHONIEMY und TEWFIK zeigen, lassen sich die Ergebnisse der Rhythmuserkennung auch zur Genre-Klassifikation verwenden [21]. Nicht immer stimmt das Tempo aus Anschlägen exakt mit dem wahrgenommenen Takt überein, Untersuchungen dazu wurden von DIXON und GOEBL durchgeführt [57].

Nahe verwandt mit der Rhythmuserkennung ist die Aufgabe der Schlagzeugtranskription; sie ist möglich durch die Erkennung der Perkussionsinstrumente, für die ein einfaches Signalmodell existiert. Untersuchungen solcher Verfahren findet man in [56,185] oder [73]. Informationen über den Rhythmus fallen in diesen Verfahren als Nebenprodukt an.

6.2.3 Melodieerkennung

Die Frage nach der Melodie eines polyphonen Musikstücks lässt sich nicht immer klar beantworten, letztlich spielen hier vor allem ästhetische Gründe eine Rolle. In Kapitel 2 sind einige Aspekte diskutiert worden. Es lassen sich lediglich wenige technische Parameter festlegen, die das Erkennen einer Melodie aus einem polyphonen Musikstück ermöglichen. Die Melodie ist manchmal die höchste Stimme, aus Gründen der Hörbarkeit meist auch die lauteste. Um die Melodie zu transkribieren, bestehen prinzipiell zwei Möglichkeiten:

- Alle Notenereignisse werden transkribiert und die Noten werden den einzelnen und ebenfalls zu klassifizierenden Stimmen zugeordnet. Die Auswahl der Melodie erfolgt dann unter Verwendung oben genannter Kriterien Tonhöhe und Lautstärke.
- Man wählt zuerst einen Frequenzbereich, in dem die Melodie erwartet wird. Alle Noten dieses Teilbandes werden transkribiert, die lautesten Töne sind per definitionem Melodietöne.

Es handelt sich in anderen Worten entweder um die Auswahl der Melodietöne aus einer *Partitur*, die transkribiert worden ist, oder um die Auswahl bzw. Begrenzung eines Audiosignals zur Melodiesuche.

Auswahl aus der Partitur

Die meisten Untersuchungen zur polyphonen Transkription beziehen sich nicht auf die Melodieauswahl, sondern lediglich auf die Aufgabe der Notenextraktion, z. B. [85,112,129,168]. Häufig wird als polyphones Signal Klaviermusik untersucht [49,126,127,162,178]; damit entfällt die Notwendigkeit, die einzelnen Notenereignisse einem Instrument zuzuordnen. Zudem habe Klaviertöne stark perkussiven Charakter, was die Suche nach Anschlägen im Signal stark erleichtert.

Sind die Noten transkribiert, ist die Auswahl der Melodie aus der erstellten Partiturnotation notwendig. Dazu können die Kriterien Lautstärke, Tonhöhe oder Instrument einschließlich Gesang verwendet werden.

Auswahl aus der Audiodatei

Von SHANDILYA und RAO wird in [174] ein Verfahren zur Extraktion von gesungenen Melodien für ein QBH-System vorgestellt. Die Grundfrequenzerkennung erfolgt dabei über die Berechnung der AKF. Sie wird durch die Vorverarbeitung gemäß eines Gehörmodells ermöglicht; damit wird die Störung der Grundfrequenzschätzung durch die perkussive Begleitung gemindert. Bei den genannten Untersuchungen ist zu beachten, dass sie sich ausschließlich auf Audiosignale des Genres indischer Filmmusik beziehen. Diese zeichnet sich dadurch aus, dass die Abschnitte mit Gesang lediglich durch Perkussionsinstrumente (also keine Instrumente, die hauptsächlich tonal klingen) begleitet werden. Der Ersatz der Audiorepräsentation der Filmmusik durch eine MIDI-Datei zur Melodieextraktion ist gar nicht möglich, da mittels MIDI indische Musik nicht befriedigend repräsentiert werden kann.

GOTO stellt in [79] ein Verfahren zur Erkennung der vorherrschenden Grundfrequenz in Musiksignalen vor. Für die Suche im Signalgemisch wird ein aus Grundton und mitschwingenden Harmonischen bestehendes Signalmodell angenommen. Das untersuchte Signal wird durch eine Filterbank in Melodie- und Bassbereich unterteilt. Für diverse Aufnahmen aus Pop, Jazz und Klassik werden Melodie- und Basslinien gut erkannt. Eine wesentliche Einschränkung dieses Verfahrens ist jedoch, dass die Melodie bezüglich ihrer Lautstärke sehr klar im Vordergrund stehen muss, um als solche erkannt zu werden.

EGGINK und BROWN verwenden einen sehr ähnlichen Ansatz zur Melodieextraktion wie GOTO [63, 64]. Grundlage des Systems ist eine MGFA, die als Vorverarbeitungsstufe lediglich eine Kurzzeitfouriertransformation verwendet. Neben der Information über die Dominanz der erkannten Grundfrequenz im Signalgemisch werden auch Instrumentenklang und Wahrscheinlichkeiten musikalischer Intervallübergänge betrachtet.

Wünschenswert wäre es, für die Melodietranskription das melodieführende Instrument im Audiosignal zu isolieren und dann alleine, d. h. monophon transkribieren zu können. Die CASA (vergleiche Abschnitt 6.2.1) zielt auf die Trennung einzelner Elemente einer auditiven Szene ab. Für diese Aufgabe wird in Arbeiten von CASEY oder SMARAGDIS die Methode der Trennung unabhängiger Quellen (independent component analysis, ICA) vorgeschlagen [42, 177].

Die ICA basiert auf der statistischen Unabhängigkeit der einzelnen Signale in einem Signalgemisch; in diesem Fall ist eine Trennung des Gemisches in die einzelnen Signale der ursprünglichen Quellen möglich [99].

In den Untersuchungen von SMARAGDIS werden einfache synthetische Signale getrennt, nämlich Sinuspulse mit verschiedener Pulsfrequenz und Tonhöhe. Diese sind einfach voneinander zu trennen, da sie sich – statistisch und ebenso spektral betrachtet – leicht voneinander unterscheiden lassen. Bei Musiksignalen sind diese Voraussetzungen keineswegs gegeben, da Musikinstrumente natürliche Töne mit harmonischen Obertonspektren produzieren, die sich (besonders bei im musikalischen Sinne harmonischen Kompositionen) in weiten Teilen überlagern. Wird darüberhinaus ein gemeinsamer Rhythmus gespielt, so ist kaum noch eine statistische Unabhängigkeit der zu beobachtenden Signale gegeben. Die ICA ist daher für die Extraktion von Melodien ungeeignet, da diese in den seltensten Fällen rhythmisch (d. h. statistisch) unabhängig von der begleitenden Musik sind.

6.3 Eigene Untersuchungen

Im Beispielsystem *Queryhammer* wird zur polyphonen Transkription das Verfahren *PreFEst* von GOTO verwendet [79–81]; die Abkürzung steht für *Predominant F_0 Estimation*, womit ausgedrückt werden soll, dass durch dieses Verfahren die im Vordergrund stehenden, also dominanten Grundfrequenzen in einem Musiksignal geschätzt werden.

PreFEst ist von den vorgestellten Transkriptionsverfahren das am besten geeignete Verfahren zur Extraktion von Melodien. Es wird auch in [169] benutzt, die Funktion des Algorithmus wurde u. a. von KLAPURI in [113] geprüft und bestätigt. Ausführliche Untersuchungen des Verfahrens, die am Fachgebiet Nachrichtenübertragung durchgeführt wurden, finden sich in der Diplomarbeit von TAPPERT [180].

6.3.1 Übersicht

In Abbildung 6.2 ist die Übersicht über alle Verarbeitungsblöcke des Verfahrens *PreFEst* dargestellt. Die Vorverarbeitung des Signals erfolgt durch eine Filterbank, Bildung der Kurzzeitfouriertransformierten (short time Fourier transformation, STFT) in den Teilbändern, Berechnung der Momentanfrequenzen und Selektion des zu transkribierenden Frequenzbereichs mittels eines Band-

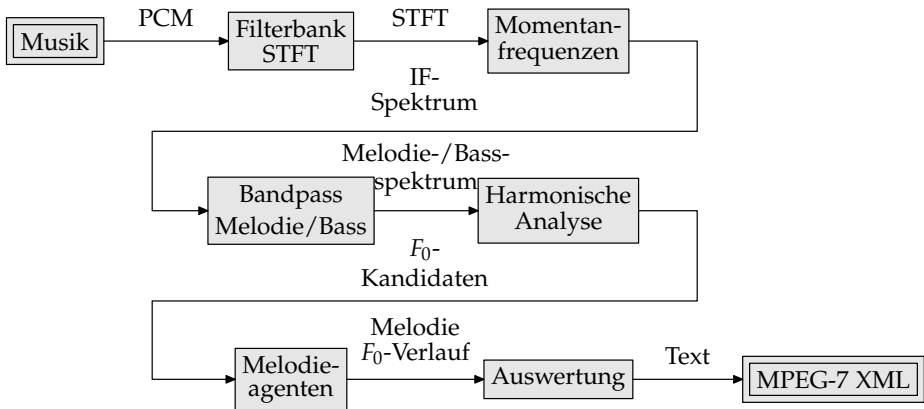


Abbildung 6.2: Das Blockschaltbild des Verfahrens *PreFest* [79] zur Transkription von Melodien aus polyphonen Musiksignalen.

passfilters. Die Schätzung der Grundfrequenzen wird als harmonische Analyse unter Zuhilfenahme eines Tonmodells durchgeführt. Die Analyse der erkannten Grundfrequenzkandidaten geschieht in der Nachverarbeitungsstufe, um die abschließende Auswertung und Erstellung einer Melodiekontur im MPEG-7-MelodyContour-Format zu ermöglichen. Alle Verarbeitungsstufen werden nun einzeln vorgestellt und beschrieben.

6.3.2 Filterbank

Für die harmonische Analyse ist die spektrale Zerlegung des Signals notwendig. In vielen Transkriptionssystemen wird dafür die STFT zur Signalanalyse verwendet [64, 162, 173]. Das in Abschnitt 5.2.3 bereits dargelegte Problem dabei ist, dass die Frequenzauflösung der zugrundeliegenden diskreten Fouriertransformation (DFT) über alle Frequenzen konstant ist, die Tonhöhen in der Musik jedoch einer logarithmisch geteilten Frequenzskala folgen. Um dieses Problem zu umgehen, wird in den gerade genannten Arbeiten eine sehr hohe Ordnung der DFT gewählt, was aber zu einer starken Minderung der Zeitauflösung führt. Die Lösung dieses Problems erfordert eine sog. Constant-Q-Transformation (Q bezieht steht für „quality“, siehe auch [168, 190]): die Güte von Zeit- und Frequenzauflösung bleibt in allen Frequenzbereichen gleich.

Eine andere mögliche Lösung stellt das im Folgenden vorgestellte Verfahren dar.

PreFEst verwendet eine kaskadierte Filterbank, die für die Analyse polyphoner Signale sehr geeignet ist [71,79]. Es handelt sich um ein Multiraten-System, in dem das Eingangssignal in vier Zweigen um jeweils auf die Hälfte der Abtastrate heruntergetastet wird (siehe Abbildung 6.3a) [190,193]. Die Abtastrate des Eingangssignals beträgt 16 kHz, die des untersten Zweigs 1 kHz.

Das Signal $x_l(n)$ jedes Zweigs wird einer STFT *gleicher Ordnung* unterworfen, zum Fenstern wird eine Fensterfunktion $h(n)$ verwendet. Damit erhält man

$$X_l(k, n) = \sum_{-\infty}^{\infty} x_l(n) h(n) e^{-j\omega_k n} \quad (6.1)$$

mit $\omega_k = \frac{2\pi k}{N}$. Da die Ordnung der STFT und damit die Blocklänge des untersuchten Signals in allen Zweigen gleich gewählt ist, die Abtastrate aber verschieden, ist das Analyseintervall in jedem Zweig zeitlich verschieden lang. Um alle Filterbankzweige gemeinsam auswerten zu können, müssen die Blöcke daher zeitlich zentriert werden. Da in [79] und [71] keinerlei Angaben zu dieser Verzögerung gemacht werden, wird kurz die Berechnung der Verzögerungszeiten dargestellt.

Die Filterbank hat L Zweige. Die Abtastrate je Zweig ist $f_{s,l}$ und berechnet sich zu

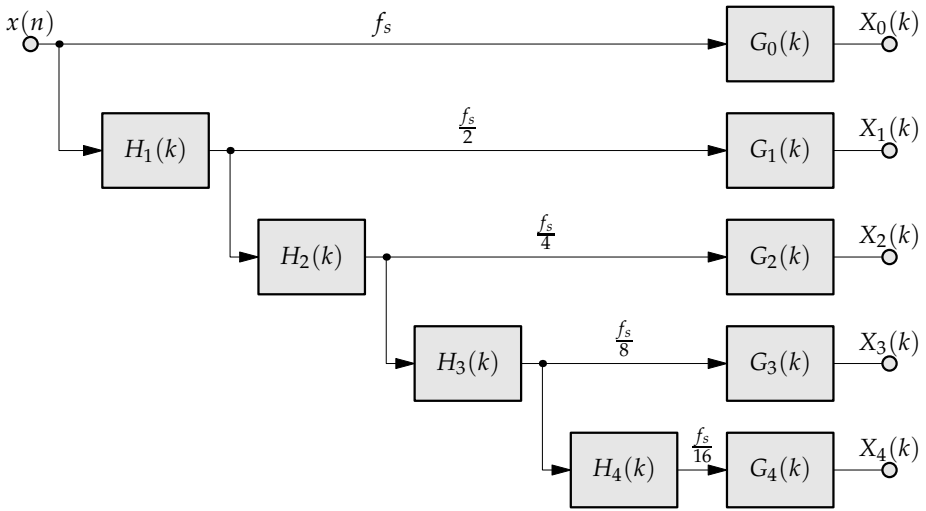
$$f_{s,l} = f_s 2^{-l}, \quad (6.2)$$

wobei $l = 0 \dots L - 1$. Der Eingangsblock habe die Länge N und entspricht der Schrittweite (hopsize), mit der das Eingangssignal untersucht wird. Die STFT habe die Ordnung M . Damit hat enthält der Block für die STFT Werte für die Zeitdauer von

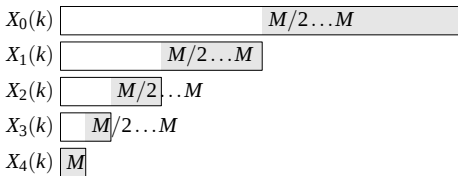
$$n_l = \frac{M}{f_{s,l}}. \quad (6.3)$$

Als gemeinsamer Zeitpunkt der Analyseintervalle wird die Mitte jedes einzelnen Blocks festgelegt. Damit die halbe Zeitdauer jedes Zweigs bei gemeinsamer Betrachtung aller Zweige zeitlich zusammenfällt, müssen die Ausgabeblöcke der Filterbank entsprechend verzögert werden. Es ergibt sich nach kurzer Rechnung eine Verzögerungszeit $t_{d,l}$ von

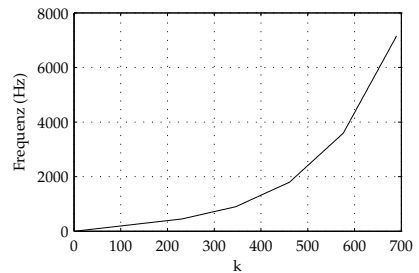
$$n_{d,l} = \frac{M}{2f_s} (2^L - 2^l), \quad (6.4)$$



(a) Das Blockschaltbild der kaskadierten Filterbank.



(b) Das Ausgangsspektrum der Filterbank-Frequenzachse setzt sich aus den farbig unterlegten Teilen der FFT-Spektren zusammen. Nur $X_4(k)$ wird vollständig verwendet.



(c) Zuordnung von Frequenzstützstellen und DFT-Frequenzen; der nichtlineare Verlauf nähert die benötigte logarithmische Frequenzaufteilung an.

Abbildung 6.3: Die kaskadierte Filterbank aus *PreFest*.

die gemäß $N = n f_s$ einer Verzögerung von

$$N_{d,l} = \frac{M}{2^{l+1}} (2^L - 2^l) \quad (6.5)$$

Abtastwerten im Zweig l entspricht. Des weiteren ist der Ausgleich der Verzögerung durch die Tiefpassfilter von je $\frac{N_{TP}}{2}$ notwendig, wobei N_{NP} die Ordnung des Filters darstellt.

Nun können die Spektren der einzelnen Zweige zusammengefasst werden. Es werden von der STFT für die Zweige $1 \dots 4$ der Filterbank nur die Stützstellen $M/2 \dots M$ ausgewertet, wie es in Abbildung 6.3b dargestellt ist. Damit ergibt sich über den gesamten Frequenzbereich eine stückweise lineare Frequenzauflösung, die sich der erforderlichen logarithmischen Frequenzauflösung gut annähert (Abbildung 6.3c). Man erhält ein hybrides Spektrum mit abschnittsweise fallender Frequenzauflösung.

6.3.3 Momentanfrequenzen

Um die spektrale Auflösung der STFT weiter zu verbessern, werden im folgenden Schritt die Momentanfrequenzen (instantaneous frequencies, IF) der Frequenzstützstellen der DFT berechnet [79]. Anwendungsbeispiele für die Momentanfrequenz findet man in der Sprachverarbeitung, zum Beispiel zur GFA [19,20,47,152]; die Reassignment-Methode ist für die Musiktranskription sehr geeignet [84,86], aber auch für die Analyse von Musikinstrumentenklängen [151,166]. Die Definition der Momentanfrequenz erfolgt am anschaulichsten über die Bildung des analytischen Signals [33,105]. Für das analytische Signal $x_a(t)$ des reellwertigen Zeitsignals $x(t)$ gilt:

$$x_a(t) = x(t) + j\hat{x}(t) = A(t)e^{\varphi(t)}, \quad (6.6)$$

wobei $\hat{x}(t)$ die HILBERT-Transformierte von $x(t)$ ist [147]; dann ist die *Momentanfrequenz* die Ableitung der Phase von $x_a(t)$:

$$\lambda(t) = \frac{d}{dt}\varphi(t). \quad (6.7)$$

Die bekannteste Anwendung der Momentanfrequenz ist der Phasenvocoder von FLANAGAN und GOLDEN [74]. Auch dort werden die Zeitsignale $F(\omega_k, t)$

der Frequenzstützstelle ω_k aus einer diskreten STFT (vgl. Gleichung (6.1)) als komplexwertiges Signal einer Filterbank aufgefasst und es gilt:

$$F(\omega_k, t) = a(\omega_k, t) + jb(\omega_k, t) = a + jb. \quad (6.8)$$

Damit kann man über die Definition des Arcustangens in kartesischen Koordinaten die Ableitung der Phasenfunktion berechnen:

$$\lambda(\omega, t) = \frac{a \frac{\partial b}{\partial t} - b \frac{\partial a}{\partial t}}{a^2 + b^2}. \quad (6.9)$$

Problematisch ist nun die Berechnung der partiellen Ableitungen $\frac{\partial a}{\partial t}$ und $\frac{\partial b}{\partial t}$. Im zeitdiskreten Fall ist die einfachste Möglichkeit die näherungsweise Berechnung durch Differenzenbildung zweier benachbarter Zeitwerte des Teilbandsignals. Bessere Ergebnisse sind möglich, wenn man die Reassignment-Methode verwendet, mit der die explizite Berechnung der partiellen Ableitungen vermieden wird [86]. Im Rahmen dieser Arbeit wird daher das Verfahren *PreFEst* entsprechend modifiziert.

Die Methode des „Reassignment“ stammt aus der Signalverarbeitung zur genaueren Zeit-Frequenz-Analyse seismischer Daten [115]. Zur Verbesserung der graphischen Darstellung von Spektrogrammen unter Verwendung der STFT wird für jeden Wert $F(\omega, t)$ die genaue Frequenzstützstelle $\hat{\omega}$ und der genaue Zeitpunkt $\hat{t}$ berechnet. In [23] wird ein einfacher Weg zur Berechnung von $\hat{\omega}$ und $\hat{t}$ angegeben; die explizite Berechnung der partiellen Ableitungen in Gleichung (6.9) kann vermieden werden, indem man zusätzlich zur STFT $F_h(\omega, t)$ mit Fensterfunktion $h(t)$ die STFT $F_f(\omega, t)$ mit $f(t) = \frac{d}{dt}h(t)$ bzw. $F_g(\omega, t)$ mit $g(t) = th(t)$ als Fensterfunktion berechnet (zur Berechnung dieser Fensterfunktionen läuft t von 0 bis zur Länge des Fensters). Damit ist der genaue Zeitpunkt des analysierten Signalsegments

$$\hat{t} = t - \Re \left\{ \frac{F_f(\omega, t)}{F_h(\omega, t)} \right\}. \quad (6.10)$$

Die Frequenzstützstelle ω_k der STFT wird durch die Momentanfrequenz ersetzt, es gilt

$$\hat{\omega} = \lambda(\omega, t) = \omega + \Im \left\{ \frac{F_g(\omega, t)}{F_h(\omega, t)} \right\}. \quad (6.11)$$

Die Momentanfrequenz kann als mittlere Frequenz des untersuchten Signalabschnitts aufgefasst werden [123]. Daher muss sichergestellt sein, dass die

Frequenzauflösung der STFT so hoch ist, dass nicht mehrere Harmonische des Signals in ein Teilband der STFT fallen [87]. Diese Voraussetzung wird durch die Verwendung der Filterbank erfüllt.

Sind die Momentanfrequenzen berechnet, erhält man ein Linienspektrum, bei dem sich die Momentanfrequenz um die Harmonischen im Signal konzentrieren [19,47]. Um die tatsächlichen Frequenzen der Harmonischen zu finden, bildet man die Funktion $\mu(\lambda, \omega)$, welche die Differenz zwischen Momentanfrequenz λ und Frequenz ω angibt:

$$\mu(\lambda, t) = \lambda(\omega, t) - \omega. \quad (6.12)$$

Die Frequenzen der Harmonischen liegen nun dort, wo die Funktion μ Nulldurchgänge hat und ihre Steigung negativ ist [20]. Mit Hilfe dieser beiden Bedingungen lässt sich ein Vektor Ψ_ω von Momentanfrequenzen aus dem Momentanfrequenz-Spektrogramm extrahieren, der im Idealfall nur noch die harmonischen Anteile des Spektrums enthält:

$$\Psi_\omega(\omega, \lambda) = \left\{ \omega \mid \mu(\omega, t) = 0, \frac{\partial}{\partial \omega} \mu(\omega, t) < 0 \right\}. \quad (6.13)$$

Aus dieser Darstellung wird die Amplitudendichteverteilung (ADV) Ψ_A gewonnen:

$$\Psi_A(\omega) = \begin{cases} |X(\omega, t)| & \forall \omega \in \Psi_\omega \\ 0 & \text{sonst.} \end{cases} \quad (6.14)$$

Alle Anteile des IF-Spektrogramms, die nicht den Bedingungen des Gleichungssystems 6.13 genügen, können zu Null gesetzt werden. $\Psi_A(\omega)$ besteht nur noch aus spektralen Linien der Harmonischen.

Nun wird eine Transformation der Frequenzachse von Hertz in Cent durchgeführt, es gilt die Beziehung

$$x = 1200 \log_2 \frac{f}{f_0}; \quad (6.15)$$

als Bezugston wird ein c_0 mit 16,3516 Hz gewählt. Damit erhält man eine Quantisierung der Frequenzachse, die für die Transkription von Musiksignalen geeignet ist. Aus praktischen Gründen werden für die Cent-Achse 10-Cent-Schritte gewählt.

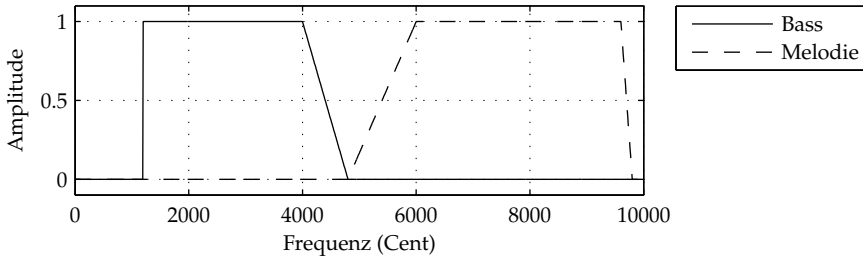


Abbildung 6.4: Der Amplitudengang der Bandpassfilter für Bass- und Melodiebereich.

6.3.4 Trennung von Melodie- und Bassbereich

In dieser Stufe wird die Trennung der Melodie vom übrigen Musiksignal vorgenommen, indem der untersuchte Frequenzbereich durch ein Bandpassfilter begrenzt wird. Goro schlägt zwei Frequenzbereiche für die Suche nach Bass- und Melodielinien vor. Die Übertragungsfunktionen $H_M(x)$ und $H_B(x)$ für Melodie- und Bassbereich sind in Abbildung 6.4 dargestellt, die Bandpassfilterung erfolgt durch Gewichtung der ADV $\Psi_A(x)$. Mit den Filterkurven $H_B(x)$ und $H_M(x)$ ergibt sich für $p_{\Psi_{B/M}}(x)$:

$$p_{\Psi_{B/M}}(x) = \frac{H_{B/M}(x) \Psi_A(x)}{\int_{-\infty}^{\infty} H_{B/M}(x) \Psi_A(x) dx}. \quad (6.16)$$

Die resultierende Wahrscheinlichkeitsdichteverteilung $p_{\Psi_{B/M}}(x)$ hat durch die Normierung auf die Fläche unter dem Produkt $H_{B/M}(x) \cdot \Psi_A(x)$ die Eigenschaft

$$\int_{-\infty}^{+\infty} p_{\Psi_{B/M}}(x) dx = 1. \quad (6.17)$$

Auf eine Unterscheidung zwischen Melodie- und Basslinie wird im Folgenden verzichtet, da die Verarbeitungsschritte für Bass und Melodie identisch sind.

6.3.5 Harmonische Analyse

Die harmonische Analyse folgt der Idee, dass das beobachtete Spektrum das Ergebnis der Überlagerung verschiedener natürlicher Töne ist, für die jeweils

ein Tonmodell angenommen werden kann. Diese einzelnen Töne sind unterschiedlich gewichtet, der Ton mit dem größten Gewicht ist der Ton der zu erkennenden Melodie- bzw. Basslinie.

Aus der bisherigen Vorverarbeitung des Signals steht die Amplitudendichteverteilung (ADV) $p_{\Psi}(x)$ zur Verfügung, die alle spektralen Anteile des untersuchten Frequenzbereichs enthält. Sie wird im Folgenden auch als „gemessene Amplitudendichte“ bezeichnet. Es wird nun angenommen, dass die gemessene ADV $p_{\Psi}(x)$ das Ergebnis einer Überlagerung von Einzeltönen mit jeweils harmonischem Obertonspektrum ist. Die noch genauer zu spezifizierende ADV dieses Spektrums sei $p(x|F)$ und das zugehörige Gewicht $w(F)$. Die Frequenz F bezeichnet die Grundfrequenz eines Einzeltons mit der ADV $p(x|F)$ und stimmt im Idealfall mit der Grundfrequenz F_0 des Melodietons überein. Die Summe der Überlagerung der Einzeltöne bzw. ihrer ADV wird mit $p(x, F; w)$ bezeichnet und sollte im Idealfall mit dem gemessenen $p_{\Psi}(x)$ übereinstimmen:

$$p(x, F; w) = \int_{F_u}^{F_o} \underbrace{w(F)}_{\text{Gewicht}} \underbrace{p(x|F)}_{\text{Tonmodell}} dF. \quad (6.18)$$

F_u und F_o begrenzen den Bereich der möglichen Grundfrequenzen. Die Gewichtungsfunktion $w(F)$ entspricht der zu schätzenden Grundfrequenzverteilung. Für jede Frequenz $F \in [F_u, F_o]$ gibt die Funktion $w(F)$ das Gewicht des Grundtons eines Tonmodells an. Auch für die Mischung $p(x; F, w)$ und die Gewichtungsfunktion $w(F)$ gilt jeweils:

$$\int_{f_u}^{f_o} p(x, F; w) dx = 1 \quad (6.19)$$

und

$$\int_{F_u}^{F_o} w(F) dF = 1. \quad (6.20)$$

Tonmodell

Die Spektrallinien der gerade verwendeten ADV $p(x|F)$ bilden das Tonmodell und werden durch mehrere spektral verteilte Gaußverteilungen nachgebildet.

Die Hüllkurve dieses Obertonspektrums wird ebenfalls durch eine Gaußverteilung beschrieben. Damit erhält man

$$p(x|F) = \sum_{h=1}^N c(h) G(x; F + 1200 \log_2(h), W) \quad h \in \{1, 2, \dots, N\}. \quad (6.21)$$

Diese ADV wird im Folgenden auch einfach als Tonmodell bezeichnet.

Das Tonmodell $p(x|F)$ entsteht also aus der Überlagerung von N Spektrallinien, die mit der Funktion $c(h)$ gewichtet werden. Die Funktion $G(x; \mu, \sigma)$ gibt die Gaußverteilung (Gl. 6.22) mit Mittelwert μ und Varianz σ in Abhängigkeit von der Frequenz x an:

$$G(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (6.22)$$

Die Spektrallinie eines Obertons (Grundton: $h = 1$, Obertöne: $h = 2 \dots N$) wird also durch eine relativ schmale Gaußverteilung mit Standardabweichung σ nachgebildet. Um die kleiner werdende Energie der Obertöne bei wachsender Ordnungszahl h nachzubilden, werden die $c(h)$ wiederum aus einer relativ breiten Gaußverteilung mit Standardabweichung H gewichtet:

$$c(h) = G(h, 1, H) = \frac{1}{\sqrt{2\pi H^2}} e^{-\frac{(h-1)^2}{2H^2}}. \quad (6.23)$$

Schätzung der Grundfrequenz

Die Aufgabe besteht nun darin, die gemessene ADV $p_{\Psi}(x)$ durch Variation der Gewichtungsfunktion $w(F)$ in Gleichung (6.18) nachzubilden. Die Funktion $w(F)$ ist gleichzeitig eine Schätzung der Grundfrequenzverteilung. Die Synthesevorschrift ist durch Gleichung 6.18 gegeben. Der Vektor $w(F)$ muss geschätzt werden, da eine analytische Lösung nicht möglich ist. Dazu wird von Goro der Expectation-Maximisation-Algorithmus (EM-Algorithmus) vorgeschlagen.

Der EM-Algorithmus beschreibt ein Verfahren, bei dem eine Schätzung der Grundfrequenzverteilung $w'(F)$ durch mehrere Iterationen der tatsächlichen Verteilung $w(F)$ angenähert wird. Dabei wird eine neue Verteilung $\bar{w}(F)$ berechnet und für den nächsten Iterationsschritt als momentane Schätzung $w'(F)$ verwendet. Die Iterationsvorschrift lautet

$$\bar{w}(F) = \int_{-\infty}^{+\infty} p_{\Psi}(x) \frac{p(x|F) w'(F)}{\int_{F_u}^{F_o} w'(\eta) p(x|\eta) d\eta} dx \quad (6.24)$$

und wird in der Diplomarbeit von TAPPERT ausführlich beschrieben [180].

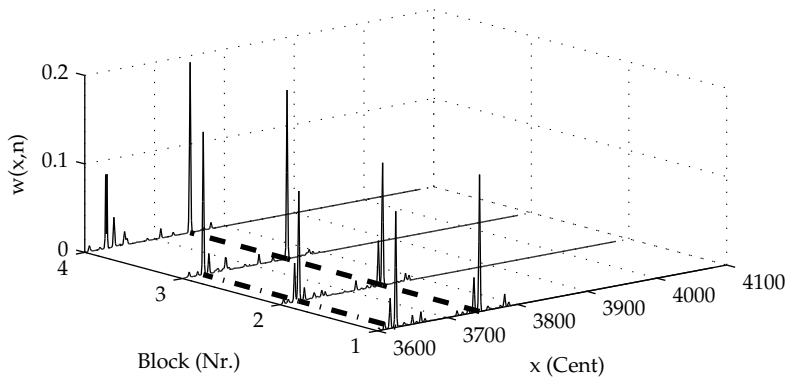


Abbildung 6.5: Die Agenten-Architektur aus dem MGFA-Verfahren *PreFest*. Die gestrichelte Linie und die Strichpunktlinie markieren möglichen Grundfrequenzkandidaten.

6.3.6 Melodieagenten

Um aus den erkannten Grundfrequenzen die Melodie zu erkennen, wird von GOTO ein Analyseverfahren mit einer „Agenten“-Architektur vorgeschlagen. Sie besteht aus einem Spitzendetektor (salience detector) und mehreren sogenannten Agenten, die dynamisch generiert, gestartet und beendet werden. Abbildung 6.5 zeigt eine Visualisierung des Verfahrens. Der Analysevorgang verläuft wie folgt:

- Der Spitzendetektor wählt nach einem dynamischen Schwellwert abhängig vom größten Maximum die zu bewertenden F_0 -Kandidaten aus. Die Agenten teilen sich diese Kandidaten auf, indem jeder Agent sich die nächstliegende Grundfrequenz auswählt. Wird ein Kandidat von mehreren Agenten beansprucht, so gewinnt der Agent mit dem bislang stärksten (lautesten) Grundfrequenzverlauf. Ist der stärkste Kandidat zum Schluss nicht zugeordnet, wird ein neuer Agent generiert.
- Jeder Agent sammelt Strafpunkte. Wird ein Schwellwert überschritten, wird der Agent gelöscht. Strafpunkte werden immer dann vergeben, wenn der Agent keinen F_0 -Kandidaten zugewiesen bekommt. Wird ein Kandidat zugewiesen, wird das Strafpunkte-Konto gelöscht.

- Jeder Agent führt Informationen darüber mit, wie zuverlässig er ist. Die Zuverlässigkeit berechnet sich aus der Stärke des vergangenen und des aktuellen F_0 -Kandidaten. Die Auswahl der erkannten Melodie-GF erfolgt über diese Zuverlässigkeitsbestimmung sowie die Gesamtenergie aller gesammelten F_0 -Kandidaten des Agenten.

Die Agenten verfolgen also Grundfrequenzen, die am dominantesten und möglichst zusammenhängend sind. Im Rahmen der Arbeit von TAPPERT wurde die Agenten-Architektur dahingehend modifiziert, dass die Zuverlässigkeit der Agenten über eine variable Gedächtnislänge eingestellt werden kann [180]. Mit einem Gedächtnis von vier Blöcken lassen sich gegenüber dem von Goto vorgeschlagenen Verfahren verbesserte Ergebnisse erzielen. Ein Beispiel für den zum Schluss bestimmten Grundfrequenzverlauf zeigt Abbildung 6.6.

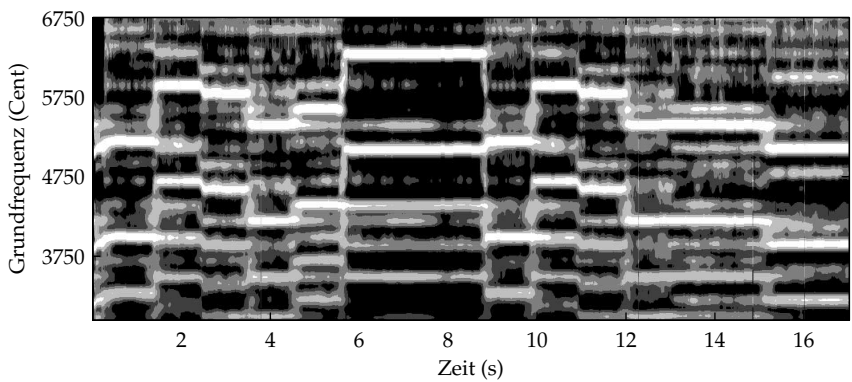
6.3.7 Praktische Versuche und Evaluierung

Die praktischen Versuche und die Evaluierung des polyphonen Transkriptionsteils im Beispielsystem *Queryhammer* werden durch die Ergebnisse der Teilnahme am ISMIR-Melody-Transcription-Contest und die Transkription einiger Melodien aus dem Bereich Pop-Musik dargestellt.

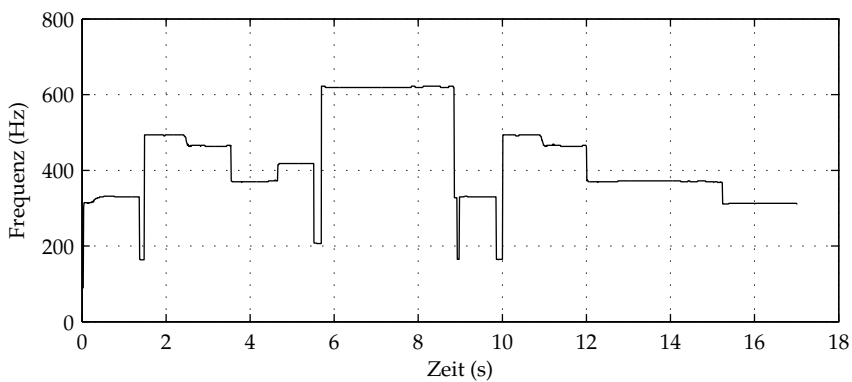
Ergebnisse des ISMIR-Melody-Transcription-Contest

Der „Melody Transcription Contest“ der ISMIR (International Symposium on Music Information Retrieval) vergleicht aktuelle Verfahren zur polyphonen Transkription in einem Wettbewerb. Dieser Wettbewerb wird jährlich durchgeführt, der Transkriptionsteil der vorliegenden Arbeit wurden im Rahmen der Arbeit von TAPPERT 2004 eingereicht [180]. Die Übersicht über alle Teilnehmer sowie eine ausführliche Diskussion der Ergebnisse ist in [78] zu finden. Für den Wettbewerb gibt es öffentlich zugängliches Audiotestmaterial; es handelt sich um eine Sammlung von Audiodateien, deren Länge bei allen Stücken etwa 20 s beträgt. Folgende Genre sind vertreten: polyphone Klänge einer MIDI-Datei mit dominanter Gesangsstimme; Saxophonphrasen mit Begleitung im Hintergrund; synthetischer Gesang mit Begleitung im Hintergrund; Pop-Musik mit Gesang; Operngesang von Frauen- und Männerstimmen mit Orchesterbegleitung.

Es waren insgesamt vier Teilnehmer beim Wettbewerb vertreten, deren Verfahren kurz beschrieben und mit dem vorliegenden Verfahren verglichen werden sollen:



(a) Die Wahrscheinlichkeit der Grundfrequenzen.



(b) Der ausgewählte Grundfrequenzverlauf der Melodie.

Abbildung 6.6: Wahrscheinlichkeiten der Grundfrequenz und extrahierte Frequenz der Melodie für ein polyphones Signal.

Tn 1: Paiva Bei der Vorverarbeitung des Signals wird ein gehörbasierter Ansatz gewählt, die Verwendung von Cochlea-Modell, Korrelogramm und Summen-AKF entspricht dem bereits dargestellten „unitary pitch model“ (vergleiche Abschnitt 6.2.1). Die MGFA erfolgt über die Auswertung der Summen-AKF; die Melodieerkennung wird ähnlich wie im Verfahren *Queryhammer* über die Verfolgung von Maxima (peak tracking) realisiert.

Tn 2: Tappert/Batke (*Queryhammer*) Die Vorverarbeitung des Signals beinhaltet das beschriebene IF-Spektrum, das über eine kaskadierte Filterbank erhalten wird. Durch EM-Schätzung und Agenten-Modell wird die Melodieerkennung vollzogen.

Tn 3: Poliner/Ellis Dieses Verfahren ist mit keinem der anderen Teilnehmer vergleichbar; die Melodietranskription wird als Klassifizierungsaufgabe aufgefasst, bei der eine Support-Vector-Machine (SVM) mit synthetisierten Audiodateien aus MIDI-Daten trainiert wird, um eine Melodienotenerkennung durchzuführen. Für die Signalvorverarbeitung wird eine einfache STFT verwendet, um das Leistungsdichtespektrum zu berechnen. Es werden lediglich Frequenzen unter 2 kHz verarbeitet.

Tn 4: Bello In diesem Verfahren wird wie in *Queryhammer* die Melodieauswahl durch Bandbegrenzung und die Suche nach dominanten Grundfrequenzen vollzogen. Die Suche erfolgt im Zeitbereich, die GFA wird über eine AKF-Methode durchgeführt.

Die Ergebnisse des Vergleichs der Verfahren zum ISMIR Contest 2004 sind in Tabelle 6.1 zusammengefasst [78]. Es werden zur Gütemessung der Melodieerkennung zwei Metriken angegeben:

Metrik 1 vergleicht die ausgegebene Frequenz des untersuchten Verfahrens mit der tatsächlichen (manuell annotierten) Frequenz des zu transkribierenden Musikstücks. Weichen die Frequenzen um mehr als 100 Cent voneinander ab, so wird ein Fehler erkannt. Zum Schluss werden die Fehler über alle Blöcke gemittelt, damit ergibt sich eine mittlere Übereinstimmung in Prozent.

Metrik 2 bildet alle Frequenzen in eine Oktave ab und ist sonst identisch mit Metrik 1. Damit bleiben Oktavfehler unbewertet.

Es wurden insgesamt 20 Stücke der o. g. Genres transkribiert und bewertet. Für das eigene Verfahren sind zwei Werte angegeben (Zeile Tn 2 in Tabelle 6.1): Den auf dem Wettbewerb erzielten Ergebnissen sind Ergebnisse in Klammern nachgestellt, die mit dem endgültig fertiggestellten Verfahren erzielt worden sind. Die Version des *Queryhammer*-Transkriptionswerkzeugs war im Wettbewerb mit einer gegenüber der endgültigen Version stark vereinfachten Agenten-Architektur ausgestattet, mit der lediglich der vierte Platz belegt wurde. Durch die verbesserten Agenten ließ sich eine wesentliche Steigerung erzielen, die zu Platz zwei geführt hätte. Neben der Transkriptionsgüte wurde die Rechengeschwindigkeit der verschiedenen Teilnehmer untersucht. Hier konnte der erste Platz erzielt werden; die durchschnittlichen Verarbeitungszeiten sind ebenfalls in Tabelle 6.1 angegeben (ein nachträglicher Vergleich war hier nicht möglich).

Tabelle 6.1: Die Ergebnisse der Verfahren im „Melody Transcription Contest“ des ISMIR 2004. Die Werte in Klammern bei Tn 2 (mit dem Verfahren aus *Queryhammer*) wurden nachträglich ermittelt.

Tn	Übereinstimmung (%)			Rechenzeit (s)
	Metrik 1	Metrik 2	Schnitt	
1	64,74	65,20	64,97	3346,67
2	57,2 (42,19)	63,18 (55,88)	60,19 (49,03)	60,00 (k.A.)
3	56,14	57,14	56,64	470,00
4	50,85	57,70	54,27	82,50

Ergebnisse der Transkription von Pop-Musik

Da das polyphone Transkriptionssystem zur Erstellung der Melodiedatenbank verwendet werden soll, wurde im Rahmen weiterer Untersuchungen die Transkription der Refrains der Top-10 aus Kapitel 8 untersucht. Resultierend kann man feststellen, dass keine Melodie aus dem vorgegebenen Test-Set in befriedigender Qualität extrahiert werden konnte. Der wesentliche Grund dafür ist darin zu sehen, dass die Melodie nicht stark genug im Vordergrund steht und damit das Verfahren *PreFEst* versagt.

6.4 Zusammenfassung

Gegenstand dieses Kapitels ist die polyphone Transkription, die zur Erstellung von Melodiedatenbanken für QBH-Systeme notwendig ist. Für die polyphone Transkription lassen sich ebenso wie für die monophone Transkription die Teilaufgaben Tonhöhenerkennung und Rhythmuserkennung definieren, zusätzlich muss aber durch eine Melodieerkennung der zu transkribierende Teil des untersuchten Musikstücks bestimmt werden.

Zunächst wurde in diesem Kapitel ein Überblick über derzeit verwendete Methoden zur Erstellung von Melodiedatenbanken für QBH-Systeme gegeben. Zum einen ist die Extraktion von Melodien aus MIDI-Dateien gebräuchlich, aber auch die manuelle Transkription von Melodien durch Musiker wird verwendet. Vor- und Nachteile dieser Vorgehensweisen wurden diskutiert. Besonders für große Datenbestände sind die genannten Verfahren unbefriedigend – dies motiviert die Transkription von Musik aus Audiosignalen. Die automatische Transkription aus Audiosignalen ist ein bislang nicht vollständig beherrschtes Problem, daher wurde eine Übersicht aktueller Lösungsansätze gegeben. In der Übersicht zu bestehenden Methoden der automatischen Transkription konnte das Verfahren *PreFest* als besonders geeignet zur Erstellung von Melodiedatenbanken identifiziert werden. Die Aufgaben von Tonhöhen-, Rhythmus- und Melodieerkennung wurden vor dem Hintergrund der polyphonen Transkription erläutert. Die Umsetzung dieser Teilaufgaben sind im hohen Maße voneinander abhängig, besondere Bedeutung kommt dabei der Mehrfachgrundfrequenzanalyse zu.

Die Implementierung des Verfahrens *PreFest* in *Queryhammer* sowie einiger eigener Modifikationen wurde detailliert beschrieben. Der Vergleich mit anderen Verfahren zur Melodietranskription wurde im Rahmen der Teilnahme am ISMIR-Melody-Transkription-Wettbewerb erfolgreich vollzogen. Weitere Untersuchungen zeigen jedoch, dass der Einsatz polyphoner Transkriptionssysteme generell noch nicht für die Erstellung von Melodiedatenbanken für QBH-Systeme möglich ist.

Melodievergleich

7

*Der Vergleich ist die Wurzel
allen Übels.*

Arthur Schopenhauer

Die Kernaufgabe eines Query-by-Humming-Systems (QBH-Systems) ist der Vergleich von Melodien. In der Vergleichsstufe wird die Ähnlichkeit der gesuchten Anfrage mit den Titeln der Melodiedatenbank bestimmt. In diesem Kapitel werden verschiedene Verfahren betrachtet, die für den Melodievergleich verwendet werden können.

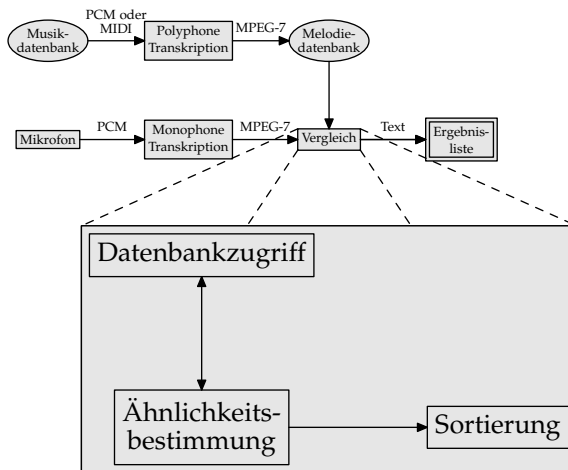


Abbildung 7.1: Die Vergleichsstufe des *Queryhammer* enthält die Blöcke *Indizierung* für den Datenbankzugriff, *Ähnlichkeitsbestimmung* für die Melodiesuche und *Sortierung* für die Erstellung der Ergebnisliste.

In Abbildung 7.1 sind die Verarbeitungsschritte der Vergleichsstufe des Beispielsystems *Queryhammer* dargestellt. Die durch die monophone Transkription ermittelte Melodiekontur wird in das Modul *Ähnlichkeitsbestimmung* ge-

reicht. Um diese Melodiekontur mit dem Bestand der Melodiedatenbank vergleichen zu können, müssen alle Melodiekonturen der Datenbank durchsucht werden. Dies sollte möglichst effizient geschehen und erfolgt über das Modul *Datenbankzugriff*. Schließlich wird das Ergebnis der Ähnlichkeitsbestimmung sortiert und dem Nutzer zugänglich gemacht.

Im folgenden Abschnitt werden zunächst einige Grundlagen zu Datenbanken erläutert und die Begriffe Ähnlichkeitsbestimmung und Datenbankzugriff in diesem Zusammenhang erklärt. Die für die Ähnlichkeitsbestimmung verwendeten Methoden der Zeichenkettensuche und Indizierung werden in den folgenden Abschnitten ausführlich dargestellt. Verfahren der Ähnlichkeitsbestimmung, die speziell für die *MPEG-7-MelodyContour* entwickelt worden sind, werden danach diskutiert. Zum Ende des Kapitels werden besondere Aspekte für die Verwendung von Ähnlichkeitsmaßen in Melodiesuchsystemen erörtert.

7.1 Datenbanken

Eine umfangreiche Sammlung von Informationen, die unter Verwendung eines Computers verarbeitet werden soll, wird *Datenbank* (database) genannt [172]. Der Zugriff auf die Datenbank eines QBH-Systems wird von dem Wunsch nach einem bestimmten Musiktitel gesteuert, nach dem in den Daten gesucht werden muss.

7.1.1 Grundlagen und Begriffe

Das Suchen ist eine grundlegende Operation, die Bestandteil vieler Berechnungsaufgaben ist: das Wiederauffinden eines bestimmten Elements oder bestimmter Informationsteile aus einer großen Menge gespeicherter Information [172]. Die Information liegt normalerweise in Datensätzen vor, denen ein Schlüssel zugeordnet ist. Ziel der Suche ist es, alle Datensätze zu finden, die mit einem bestimmten Suchschlüssel übereinstimmen.

Für ein QBH-System werden die Melodien als Schlüssel gewählt: Die Anfrage des Nutzers wird in eine Darstellung verwandelt, die als Suchschlüssel (auch: Muster) benutzt werden kann. Dieser Suchschlüssel muss in geeigneter Weise mit den Schlüsseln der Datensätze verglichen werden. Die Datensätze enthalten alle weiteren Informationen zur Melodie, also Titel, Komponist, Interpret, Noten, Audio-Dateien usw.

Während der Datenbanksuche wird bei dem Vergleich zweier Melodien respektive Schlüssel die Ähnlichkeit der Schlüssel festgestellt. Ein geeignetes *Ähnlichkeitsmaß* liefert einen skalaren Wert dafür, wie ähnlich sich zwei Schlüssel sind. Als sinnvoller Wertebereich kann z. B. 0 bis 1 gewählt werden: Bei Gleichheit ist das Ähnlichkeitsmaß 1, bei Ungleichheit sind die Werte des Ähnlichkeitsmaßes kleiner 1. Umgekehrt beschreibt ein *Distanzmaß*, wie verschieden zwei Melodien sind, bei 0 liegt z. B. Gleichheit vor, Werte verschieden davon bemessen einen „Abstand“, wobei der Betrag Auskunft über die Größe der Verschiedenheit gibt. Über das skalare Ergebnis des Ähnlichkeits- bzw. Distanzmaßes ist schließlich die Sortierung der Ergebnisliste möglich. Die Begriffe Ähnlichkeits- und Distanzmaß werden im Rahmen dieser Arbeit gleichberechtigt verwendet.

7.1.2 Suchmethoden

Weil der ähnlichste Schlüssel gefunden werden soll, spricht man auch von einer Suchaufgabe. Die einfachste Suchmethode ist die *sequentielle Suche*. Die Datensätze der Datenbank werden in einem Feld gespeichert, das sequentiell auf den Suchschlüssel untersucht wird. Bei großen Datensätzen kann dies aber zu einer beträchtlichen Suchdauer führen. In diesem Fall ist die *binäre Suche* eine mögliche Lösung: Die Menge der Datensätze wird in zwei Teile zerlegt, es wird bestimmt, welchem Teil der Suchschlüssel angehört und die Suche wird in diesem Teil fortgesetzt. Setzt man dieses Verfahren fort, erhält man einen *Suchbaum*.

Die Bestimmung der Ähnlichkeit und die Suche hängen direkt vom Inhalt und Datentyp des Suchschlüssels ab. Wie bereits erwähnt wurde, wird für QBH-Systeme die Melodie als Schlüssel verwendet. Zur Ähnlichkeitsmessung von Melodien sind in der Literatur viele Verfahren vorgeschlagen worden, zahlreiche Untersuchungen beschäftigen sich mit Auswahl und Eigenschaften von Ähnlichkeitsmaßen, siehe [32, 119, 187]. Eine gute Übersicht findet sich in [187]. Eigene Untersuchungen zu diesem Thema werden in [27, 67, 91, 195] dargestellt.

Wird der PARSONS-Code als Melodiedarstellung verwendet, so ergibt sich als Datentyp direkt eine Zeichenkette (string). Ebenso können aber auch die Zahlen in den Vektoren der MPEG-7-MelodyContour als Zeichenketten aufgefasst werden. Das heißt mit anderen Worten, dass die Zeichenkette der Melodieanfrage in der Datenbank des QBH-Systems wiedergefunden werden soll und als Problem der Zeichenkettensuche aufgefasst werden kann [32, 119, 187].

Die Zeichenkettensuche kann exakt oder näherungsweise (approximativ) ausgeführt werden [187].

Exakte Suche

Verfahren der exakten Suche sind für QBH-Systeme unüblich, da die Suchanfrage meistens nicht so gestellt wird, wie die gesuchte Melodie im Datenbestand abgelegt ist [187]. Allerdings werden einige Verfahren der exakten Suche für die approximative Suche verwendet, so zum Beispiel die Bit-Parallel-Methode von BAEZA-YATES und PERLEBERG aus [26] in [119] oder in [76].

Approximative Suche

Bei der exakten Suche ist das Ergebnis der Suche, ob die gesuchte Zeichenkette gefunden wurde oder nicht. Bei der approximativen Suche (auch: „unscharfe Suche“) hingegen können auch ähnliche Zeichenketten gefunden werden. Diese Anforderung besteht für QBH-System im besonderen Maße: Nur selten wird die Suchanfrage fehlerfrei sein, und das Auffinden zur Suchanfrage ähnlicher Zeichenketten bzw. Melodien ist erwünscht.

Die approximative Suche fällt unter den Sammelbegriff der Musteranpassung (pattern matching). Sie ermöglicht es, eine Suche nach einem unvollständig spezifizierten Muster vorzunehmen [172, 189]. Möglich wird dies z. B. durch Berechnung der Editierdistanz [140]. Zur Berechnung der Editierdistanz können Verfahren der dynamischen Programmierung (DP) eingesetzt werden, die in Abschnitt 7.2 erläutert werden. Beispiele für die Anwendung zum Melodievergleich findet man in [97, 118, 140, 149, 154, 160]. Weiterhin ist es möglich, Indizierungstechniken zur Ähnlichkeitsbestimmung von Melodien zu verwenden. Zur Indizierung werden vor allem N-Gramme verwendet, die in Abschnitt 7.3 dargestellt werden. N-Gramme wurden von [58] ausführlich untersucht, sehr häufig werden in der Literatur N-Gramm-Techniken mit denen der DP verglichen [27, 55, 138, 186].

Abhängig von der Melodiedarstellung sind weitere Distanzmaße möglich. Speziell für MPEG-7-MelodyContour wird das Verfahren „TPBM“ vorgeschlagen [108], das Zählzeit für Zählzeit die Kontur zweier Melodiedarstellungen vergleicht. Auch andere Distanzmaße bewerten Zeit- und Tonhöheninformation gemeinsam [75, 184], sind aber für die MPEG-7-MelodyContour ungeeignet, da sie detailliertere Informationen benötigen, als diese Darstellung bietet.

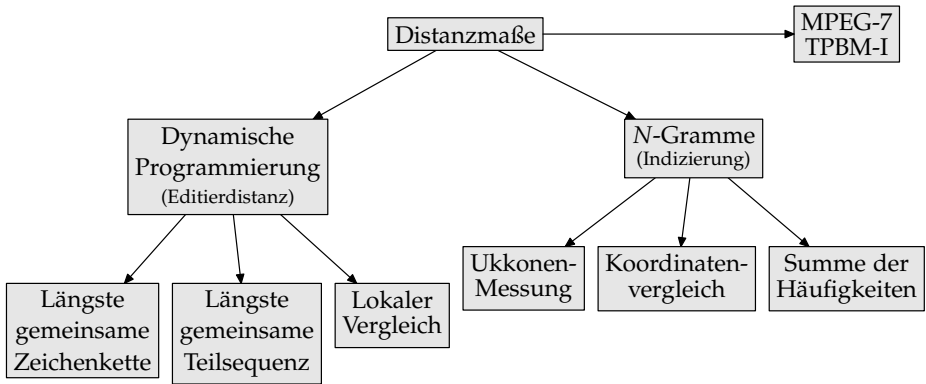


Abbildung 7.2: Übersicht der hier diskutierten Distanzmaße. Die Unterteilung erfolgt anhand der verwendeten Technik. Die dynamische Programmierung dient zur Berechnung von Edierdistanzen, N-Gramme werden zur Indizierung von Datenbanken verwendet. TPBM-I ist ein speziell auf MPEG-7 zugeschnittenes Distanzmaß.

Es werden in diesem Kapitel sowohl Verfahren der Zeichenkettensuche als auch Indizierungsmethoden vorgestellt. Beide Verfahren liefern eine Aussage über die Ähnlichkeit zweier Melodien. Abbildung 7.2 zeigt eine Übersicht der behandelten Verfahren.

7.2 Zeichenkettensuche

Mit Hilfe der Zeichenkettensuche durch Berechnung der Edierdistanz lässt sich ein sehr elementares Verfahren zur Bestimmung von Ähnlichkeiten konstruieren [52]. Es wird ermittelt, wieviele Edieroperationen mindestens benötigt werden, um eine Melodie A in eine Melodie B umzuwandeln. Zulässige Edieroperationen sind dabei das *Löschen*, das *Einfügen* sowie das *Ersetzen* einer Note. Sind die beiden Melodien einander ähnlich, werden nur wenige Edieroperationen benötigt, sind die Melodien dagegen sehr unähnlich, so sind viele Edieroperationen notwendig.

Zur Berechnung der Edierdistanz kann man die Methode der *dynamischen Programmierung* (DP) verwenden. Die DP stützt sich auf das Prinzip „teile und

herrsche“: Um ein umfangreiches Problem zu lösen, zerlegt man es in kleinere Probleme, die unabhängig voneinander gelöst werden können [172]. Die Klasse der Verfahren zur Zeichenkettensuche unter Verwendung der Edierdistanz basieren alle auf einer Kostenmatrix [187]. Diese Kostenmatrix enthält die Ergebnisse des Vergleichs zweier Zeichenketten. Deren Ähnlichkeit kann lokal, d. h. in Bezug auf eine Teilzeichenkette, oder global für die Gesamtübereinstimmung berechnet werden. Es werden nun die am häufigsten verwendeten Edierdistanzen dargestellt.

7.2.1 Längste gemeinsame Teilsequenz (LCE)

Für die Bewertung der „längsten gemeinsamen Teilsequenz“ (longest common subsequence, LCE) wird nach gemeinsamen Zeichen in zwei Zeichenketten in gemeinsamer Reihenfolge gesucht.

Es sei **A** die zu berechnende Kostenmatrix, **q** (wie „query“) und **p** (wie „piece“) die Zeichenketten mit Anfrage und Datenbanktitel; der Index *i* läuft von 0 bis zur Länge der Anfrage und Index *j* von 0 zur Länge des Stücks:

$$A[i, j] = \max \begin{cases} A[i-1, j] & i \geq 1 \\ A[i, j-1] & j \geq 1 \\ A[i-1, j-1] + 1 & q(i) = p(j) \text{ und } i, j \geq 1 \\ 0 & \end{cases} \quad (7.1)$$

Bei diesem Verfahren werde die Elemente $A[i, j]$ heraufgezählt, wenn es sich um eine Übereinstimmung handelt, sonst erhalten sie den Wert der linken oberen Diagonale der Matrix [186]. Das heißt mit anderen Worten, das Einfügen in einer fehlerhaften Anfrage **q** keine Änderung des Ergebnisses bewirken.

Beispiel: Abbildung 7.3 zeigt die Noten und Konturwerte zum Beginn des Kinderlieds „Hänschen klein“. Die vollständige, für das Beispiel verwendete Kontur lautet:

$$\begin{aligned} \mathbf{p} &= \begin{bmatrix} -2 & 0 & 1 & -2 & 0 & -1 & 1 & 1 & 1 & 1 & 0 \dots \\ 0 & 0 & -2 & 0 & 1 & -2 & 0 & -1 & 2 & 2 \dots \\ 0 & -2 \end{bmatrix} \end{aligned} \quad (7.2)$$

Zwischen dieser Kontur und einer fehlerbehafteten Anfrage

$$\mathbf{q} = \begin{bmatrix} -2 & 0 & 1 & -2 & 0 & -1 & 1 & \overset{\text{edt.}}{0} & 1 & 1 & \overset{\text{lös.}}{0} \dots \\ 0 & -2 & 0 & 1 & -2 & \overset{\text{einf.}}{0} & 0 & -1 & 2 & 2 \dots \\ 0 & -2 \end{bmatrix} \quad (7.3)$$

soll nun die Ähnlichkeit mittels LCE festgestellt werden. Die Fehler sind markiert:

- mit „edt.“ ist eine Edierung von 1 im Original nach 0 in der Anfrage gekennzeichnet,
- mit „lös.“ eine Auslöschung (eine 0 fehlt) und
- mit „einf.“ eine Einfügung (eine zusätzliche 0).

Das Maximum der Kostenmatrix $\mathbf{A}$ aus Gleichung (7.1) ist $R_{\text{LCE}} = 21$, wie Tabelle 7.1 zu entnehmen ist. Dies ist leicht nachzuvollziehen: Die Länge der Originalkontur ist 23; die Edierung und Löschung sorgen für je einen Fehler, die Einfügung wirkt sich nicht aus.

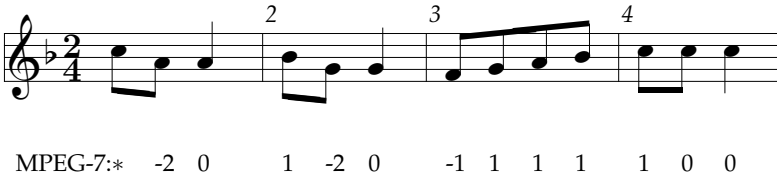


Abbildung 7.3: Der Beginn des Kinderlieds „Hänschen klein“: Notenschrift mit den zugehörigen MPEG-7-MelodyContour-Werten.

Tabelle 7.1: Der Inhalt der Kostenmatrix **A** für die Edierdistanz LCE. Das Maximum der Matrix ist $R_{LCE} = 21$. Gekennzeichnet in der Matrix sind: (a) Edierung, (b) Löschung, (c) Einfügung.

		$\mathbf{p} =$																											
		-2	0	1	-2	0	-1	1	1	1	1	0	0	0	-2	0	1	-2	0	-1	2	2	0	-2					
$\mathbf{q} =$	-2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
	0	0	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	
	1	0	1	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	
	-2	0	1	2	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	
	0	0	1	2	3	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	
	-1	0	1	2	3	4	5	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	
	1	0	1	2	3	4	5	6	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	
	0	0	1	2	3	4	5	6	7	7(a)	7	7	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8
	1	0	1	2	3	4	5	6	7	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8
	1	0	1	2	3	4	5	6	7	8	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	
	0	0	1	2	3	4	5	6	7	8	9	9	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	
	=0	0	1	2	3	4	5	6	7	8	9	9	10	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	
	-2	0	1	2	3	4	5	6	7	8	9	9	10	11	11(b)	12	12	12	12	12	12	12	12	12	12	12	12	12	
	0	0	1	2	3	4	5	6	7	8	9	9	10	11	12	12	13	13	13	13	13	13	13	13	13	13	13	13	
	1	0	1	2	3	4	5	6	7	8	9	10	10	11	12	12	13	14	14	14	14	14	14	14	14	14	14	14	
-2	0	1	2	3	4	5	6	7	8	9	10	10	11	12	13	13	14	15	15	15	15	15	15	15	15	15	15		
0	0	1	2	3	4	5	6	7	8	9	10	11	11	12	13	14	14	15	16	16	16	16	16	16	16	16	16		
0	0	1	2	3	4	5	6	7	8	9	10	11	12	12	13	14	14	15	16	16(c)	16	16	16	17	17	17	17		
-1	0	1	2	3	4	5	6	7	8	9	10	11	12	12	13	14	14	15	16	17	17	17	17	17	17	17	17		
2	0	1	2	3	4	5	6	7	8	9	10	11	12	12	13	14	14	15	16	17	18	18	18	18	18	18	18		
2	0	1	2	3	4	5	6	7	8	9	10	11	12	12	13	14	14	15	16	17	18	19	19	19	19	19	19		
0	0	1	2	3	4	5	6	7	8	9	10	11	12	13	13	14	14	15	16	17	18	19	20	20	20	20	20		
-2	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	14	14	15	16	17	18	19	20	21	21	21	21		

7.2.2 Globaler Abgleich

Bei Bestimmung der LCE wird nicht berücksichtigt, ob beim Vergleich der Zeichenketten Lücken (gaps) gefunden werden. Bei der nächsten betrachteten Variante aus dem Bereich der DP, dem „globalen Abgleich“ oder auch NEEDLEMAN-WUNSCH-Abgleich, werden Treffer positiv bewertet, keine Übereinstimmung zwischen den Zeichen führt zu Strafe, ebenso Einfügungen, Lücken oder Ersetzungen. Für diese Fehler werden verschiedene Kosten gewählt. Die Kostenmatrix berechnet sich durch

$$A[i, j] = \max \begin{cases} A[i-1, j] + d & i \geq 1 \\ A[i, j-1] + d & j \geq 1 \\ A[i-1, j-1] + e & q(i) = p(j) \text{ und } i, j \geq 1 \\ A[i-1, j-1] + m & q(i) \neq p(j) \\ 0 & i, j = 0 \end{cases} \quad (7.4)$$

Dabei sind die Kosten bestimmt durch d für Einfügung oder Löschung, e für Treffer und m , falls keine Übereinstimmung besteht. Üblicherweise werden die Kosten wie folgt gewählt:

$$d = -2, e = 1, m = -1.$$

Damit wird die Edierdistanz zwischen zwei Zeichenketten unter Berücksichtigung verschiedener Edieroperationen bewerkstelligt. Da die Zeichenketten als Ganzes bewertet werden, ist diese Edierdistanz für QBH-Systeme ungeeignet. Eine kleine Modifikation führt jedoch zum wesentlich geeigneteren „lokalen Abgleich“.

7.2.3 Lokaler Abgleich (LAL)

Die Berechnung des lokalen Abgleichs (local alignment, LAL), auch SMITH-WATERMAN-Abgleich, ist sehr ähnlich der des globalen Abgleichs. Auch hier kann das Ergebnis durch die Kostenparameter für Einfügungen, Löschungen und Ersetzungen variiert werden, allerdings kann das Ergebnis nie negativ werden.

$$A[i, j] = \max \begin{cases} A[i-1, j] + d & i \geq 1 \\ A[i, j-1] + d & j \geq 1 \\ A[i-1, j-1] + e & q(i) = p(j) \text{ und } i, j \geq 1 \\ A[i-1, j-1] + m & q(i) \neq p(j) \\ 0 & \end{cases} \quad (7.5)$$

Die Kosten werden üblicherweise wie beim globalen Abgleich angegeben gewählt. Die Matrix **A** zeigt, wie gut sich die beiden Zeichenketten aufeinander abgleichen lassen; weil es sich um den lokalen Abgleich handelt, kann man an der Matrix ablesen, an welchen Stellen der Zeichenketten die beste Übereinstimmung zu erzielen ist.

Beispiel: Tabelle 7.2 zeigt die Kostenmatrix für LAL für die Anfrage aus Beispiel in Abschnitt 7.2.1. Die ersten sieben Zeichen zwischen **q** und **p** stimmen überein, entsprechend wird auf der Diagonalen der Wert 7 erreicht (a). Die Edierung führt zur Strafe -1 , danach stimmen wieder vier Werte überein, der Wert 10 wird erreicht (b). Die Löschung wird mit -2 bestraft, die folgenden fünf passenden Zeichen führen zum Wert 13, siehe (c). Die Einfügung wird wie die Löschung mit -2 bestraft, die letzten fünf übereinstimmenden Werte ergeben schließlich den Wert 16 (d).

Tabelle 7.2: Der Inhalt der Kostenmatrix **A** für LAL. Das Maximum der Matrix ist $R_{LAL} = 16$. Gekennzeichnet in der Matrix sind: (a) Edierung, (b) Löschung, (c) Einfügung, (d) letztes Maximum und beste Übereinstimmung.

		p =																											
		-2	0	1	-2	0	-1	1	1	1	0	0	0	-2	0	1	-2	0	-1	2	2	0	-2						
q =	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
	-2	0	1	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	1	
	0	0	0	2	0	0	2	0	0	0	0	0	1	1	1	0	2	0	0	2	0	0	0	1	0	0	0	0	
	1	0	0	0	3	1	0	1	1	1	1	1	0	0	0	0	3	1	0	1	0	0	0	0	0	0	0	0	
	-2	0	1	0	1	4	2	0	0	0	0	0	0	0	1	0	1	4	2	0	0	0	0	0	0	0	1	0	
	0	0	0	2	0	2	5	3	1	0	0	0	1	1	1	0	2	0	2	5	3	1	0	1	0	0	0	0	
	-1	0	0	0	1	0	3	6	4	2	0	0	0	0	0	0	1	0	3	6	4	2	0	0	0	0	0	0	
	1	0	0	0	1	0	1	4	7(a)	5	3	1	0	0	0	0	1	0	1	4	5	3	1	0	0	0	0	0	
	0	0	0	1	0	0	1	2	5	6	4	2	2	1	1	0	1	0	0	1	2	3	4	4	2	0	0	0	
	1	0	0	0	2	0	0	0	3	6	7	5	3	1	0	0	0	2	0	0	0	1	2	3	3	0	0	0	
	1	0	0	0	1	1	0	0	1	4	7	8	6	4	2	0	0	1	1	0	0	0	0	1	2	0	0	0	
	0	0	0	1	0	0	2	0	0	2	5	6	9	7	5	3	1	0	0	2	0	0	0	1	0	0	0	0	
	q = 0	0	0	1	0	0	1	1	0	0	3	4	7	10(b)	8	6	4	2	0	1	1	0	0	1	0	0	0	0	
	-2	0	1	0	0	1	0	0	0	0	1	2	5	8	9	9	7	5	3	1	0	0	0	0	0	2	0	0	
	0	0	0	2	0	0	2	0	0	0	0	0	3	6	9	8	10	8	6	4	2	0	0	1	0	0	0	0	
	1	0	0	0	3	1	0	1	1	1	1	1	4	7	8	8	11	9	7	5	3	1	0	0	0	0	0	0	
	-2	0	1	0	1	4	2	0	0	0	0	0	0	2	5	8	7	9	12	10	8	6	4	2	1	0	0	0	
	0	0	0	2	0	2	5	3	1	0	0	0	1	1	3	6	9	7	10	13(c)	11	9	7	5	3	0	0	0	
	0	0	0	1	1	0	3	4	2	0	0	0	1	2	2	4	7	8	8	11	12	10	8	8	6	0	0	0	
	-1	0	0	0	0	0	1	4	3	1	0	0	0	0	1	2	5	6	7	9	12	11	9	7	7	0	0	0	
	2	0	0	0	0	0	0	2	3	2	0	0	0	0	0	0	3	4	5	7	10	13	12	10	8	0	0	0	
	2	0	0	0	0	0	0	0	1	2	1	0	0	0	0	0	1	2	3	5	8	11	14	12	10	0	0	0	
	0	0	0	1	0	0	1	0	0	0	1	0	1	1	1	0	1	0	1	4	6	9	12	15	13	0	0	0	
	-2	0	1	0	0	1	0	0	0	0	0	0	0	0	0	2	0	0	1	2	4	7	10	13	16(d)	0	0	0	

7.2.4 Längste gemeinsame Zeichenkette (LCT)

Weiterhin ist es sinnvoll, das Ergebnis in $A(i, j)$ zurückzusetzen, wenn bei der Bewertung die Zeichen nicht übereinstimmen oder eine Ersetzung vorliegt.

$$A[i, j] = \max \begin{cases} A[i-1, j] + d & i \geq 1 \\ A[i, j-1] + d & j \geq 1 \\ A[i-1, j-1] + e & q(i) = p(j) \text{ und } i, j \geq 1 \\ 0 & q(i) \neq p(j) \end{cases} \quad (7.6)$$

Das Maximum in der Matrix repräsentiert die Länge der längsten gemeinsamen Zeichenkette. Daher bezeichnet man dieses Ähnlichkeitsmaß als Suche zur „längsten gemeinsamen Teilzeichenkette“ (*longest common substring*, LCT).

Beispiel: In Tabelle 7.3 ist die Kostenmatrix zum Verfahren LCT dargestellt. Die längsten gemeinsamen Zeichenketten sind 7 Zeichen lang, nämlich der Beginn von **q** und **p** bis zum ersten Fehler $-2 \ 0 \ 1 \ -2 \ 0 \ -1 \ 1$ (a) sowie die Folge $0 \ 0 \ -2 \ 0 \ 1 \ -2 \ 0$ (c).

7.2.5 Ähnlichkeitsberechnung

In den vorangegangenen Abschnitten sind drei unterschiedliche Edierdistanzen vorgestellt worden. Um die Kostenmatrizen zur Ähnlichkeitsberechnung heranzuziehen, verwendet man das Maximum der Kostenmatrix, d. h.

$$R = \max A \quad (7.7)$$

stellt das Ähnlichkeitsmaß dar. Damit stehen abhängig von den Edierdistanzen die Ähnlichkeitsmaße R_{LCE} für die längste gemeinsame Teilsequenz (LCE), R_{LAL} für den lokalen Abgleich zweier Zeichenketten und R_{LCT} für die längste gemeinsame Teilzeichenkette (LCT) zur Verfügung.

7.3 Indizierung

Die Indizierung einer Datenbank dient der Beschleunigung des Zugriffs auf einen bestimmten Datenbankeintrag. Dabei wird der zum Suchschlüssel zugehörige Schlüssel des Datenbankeintrags gesucht, wie in Abschnitt 7.1 schon

Tabelle 7.3: Der Inhalt der Kostenmatrix **A** für LCT. Das Maximum der Matrix ist $R_{LCT} = 7$. Gekennzeichnet in der Matrix sind: (a) Edierung, (b) Löschung, (c) Einfügung, (d) letztes Maximum – das höchste Maximum tritt jedoch bei (a) und (c) auf und zählt als Ergebnis.

		p =																							
		-2	0	1	-2	0	-1	1	1	1	1	0	0	0	-2	0	1	-2	0	-1	2	2	0	-2	
q = 0	-2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
	0	0	1	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	1	
	0	0	0	2	0	0	2	0	0	0	0	0	1	1	1	0	2	0	0	2	0	0	0	1	
	1	0	0	0	3	0	0	0	1	1	1	1	0	0	0	0	3	0	0	0	0	0	0	0	
	-2	0	1	0	0	4	0	0	0	0	0	0	0	0	1	0	0	4	0	0	0	0	0	1	
	0	0	0	2	0	0	5	0	0	0	0	0	1	1	1	0	2	0	0	5	0	0	0	1	
	-1	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	
	1	0	0	0	1	0	0	7(a)	1	1	1	1	0	0	0	0	1	0	0	0	0	0	0	0	
	0	0	0	1	0	0	1	0	0	0	0	0	2	1	1	0	1	0	0	1	0	0	0	1	
	1	0	0	0	2	0	0	0	1	1	1	1	0	0	0	0	2	0	0	0	0	0	0	0	
	1	0	0	0	1	0	0	0	1	2	2	2	0	0	0	0	1	0	0	0	0	0	0	0	
	0	0	0	1	0	0	1	0	0	0	0	0	3	1	1	0	1	0	0	1	0	0	0	1	
	0	0	0	1	0	0	1	0	0	0	0	0	1	4(b)	2	0	1	0	0	1	0	0	0	1	
	-2	0	1	0	0	1	0	0	0	0	0	0	0	0	3	0	0	1	0	0	0	0	0	2	
	0	0	0	2	0	0	2	0	0	0	0	0	1	1	1	0	4	0	0	2	0	0	0	1	
	1	0	0	0	3	0	0	0	1	1	1	1	0	0	0	0	5	0	0	0	0	0	0	0	
	-2	0	1	0	0	4	0	0	0	0	0	0	0	0	0	1	0	0	6	0	0	0	0	0	1
	0	0	0	2	0	0	5	0	0	0	0	0	1	1	1	0	2	0	0	7(c)	0	0	0	1	0
	0	0	0	1	0	0	1	0	0	0	0	0	1	2	2	0	1	0	0	1	0	0	0	1	0
	-1	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0
	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	1	0	0	0
	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	4	0	0
	0	0	0	1	0	0	1	0	0	0	0	0	0	1	1	1	0	1	0	0	1	0	0	5	0
	-2	0	1	0	0	1	0	0	0	0	0	0	0	0	0	2	0	0	1	0	0	0	0	0	6(d)

dargelegt worden ist. Für Melodiedatenbanken sind verschiedene Indizes entwickelt worden, die spezielle Melodiemerkmale als Schlüssel verwenden [31, 51, 116]. Für Melodiekonturen eignen sich prinzipiell alle Techniken, die zur Indizierung von Text verwendet werden.

Für die Indizierung werden Index-Bezeichner und eine Index-Struktur benötigt. In Textdatenbanken werden als Bezeichner zum Beispiel einzelne Wörter verwendet [187]; dies ist in Musikdatenbanken nicht möglich, stattdessen kann man einzelne Noten [51] oder Melodiephrasen indizieren [103]. Die Indizierung einzelner Noten erfordert sehr viele Index-Zugriffe (look ups) für eine einzelne Anfrage. Die Indizierung von Melodiephrasen begrenzt die Möglichkeiten, an welcher Stelle eine Melodie begonnen wird [187].

Eine weitere Technik zur Indizierung ist die Verwendung von N -Grammen (oder auch N -Tupel). Ein N -Gramm ist ein Satz benachbarter Symbole in eine Zeichenkette. Beispielsweise kann man den PARSONS-Code

UDRUDDR

in folgende Trigramme unterteilen:

UDR DRU RUD UDD DDR.

Ein N -Gramm-Index enthält alle N -Gramme, die sich aus dem Bestand einer Melodiedatenbank bilden lassen. Er lässt sich beispielsweise nach der Häufigkeit der N -Gramme ordnen, die in der Datenbank vorkommen. Um einen Suchschlüssel, aus dem ebenfalls N -Gramme gebildet werden, dem passenden Datenbankschlüssel zuzuordnen, werden die Häufigkeiten der entsprechenden N -Gramme verglichen. Somit wird klar, dass durch N -Gramme Ähnlichkeiten festgestellt werden können.

Ausführliche Untersuchungen zu N -Grammen als Ähnlichkeitsmaß und die erfolgreiche Anwendung für Musiksuchsysteme werden von DOWNIE in [58] dargestellt. Bei allen N -Gramm-Techniken werden die gemeinsamen (oder verschiedenen) N -Gramme zweier Zeichenketten gezählt, aus denen sich dann eine Bewertung der Ähnlichkeit ergibt [186]. N -Gramme lassen sich nicht nur mit Buchstaben, sondern auch mit Zahlen bilden und sind daher für die MPEG-7-MelodyContour als Distanzmaß geeignet.

Eine Melodiekontur $\mathbf{p}$, die mit M Intervallwerten beschrieben wird

$$\mathbf{p} = [p(1), p(2), \dots, p(M)], \quad (7.8)$$

wird in N -Gramme

$$\mathbf{p}_N(i) = [p(i), p(i+1), \dots, p(i+N-1)] \quad (7.9)$$

untergliedert, wobei $i = 1 \dots M - N + 1$. Man erhält insgesamt $M - N + 1$ N-Gramme. Mit P_N werde im Folgenden die Menge der gebildeten N-Gramme bezeichnet, also

$$P_N = \{\mathbf{p}_N(1), \dots, \mathbf{p}_N(M - N + 1)\}. \quad (7.10)$$

Die einzelnen N-Gramme $\mathbf{p}_N(i)$ müssen nicht notwendigerweise verschieden voneinander sein. Mit dem Operator $U(\mathbf{p}_N(i), P_N)$ werden die Häufigkeit der N-Gramme $\mathbf{p}_N(i)$ in P_N bezeichnet, d. h.

$$U(\mathbf{p}_N(i), P_N) = \sum_{j=1}^{M-N+1} 1, \text{ falls } \mathbf{p}_N(i) \in P_N. \quad (7.11)$$

Insgesamt lassen sich mit G Symbolen G^N verschiedene N-Gramme bilden.

Beispiel: Das Kinderlied „Hänschen klein“ aus dem Beispiel in Abschnitt 7.2.1 soll nun in einer Trigramm-Darstellung gebracht werden, ebenso die fehlerhafte Anfrage.

In Tabelle 7.4 sind in der Spalte links alle N-Gramme $\mathbf{g}_N(i)$ dargestellt, die sich mit $\mathbf{p}$ und $\mathbf{q}$ bilden lassen. Durch die Fehler in der Anfrage sind die möglichen N-Gramm-Mengen unterschiedlich. Gemeinsam auftretende N-Gramme sind mit einem * gekennzeichnet.

Es werden nun verschiedene Methoden dargestellt, um die Ähnlichkeit zweier Melodiedarstellungen $\mathbf{p}$ und $\mathbf{q}$ mit Hilfe von N-Grammen zu bestimmen.

7.3.1 Koordinatenvergleich (CM)

Beim Koordinatenvergleich (coordinate matching, CM), manchmal auch mit „count distinct measure“ bezeichnet, werden die N-Gramme $\mathbf{g}_N(i)$ gezählt, die sowohl durch $\mathbf{q}$ wie in $\mathbf{p}$ gebildet werden können:

$$R_{\text{CM}} = \sum_{\mathbf{g}_N(i) \in Q_N \cap P_N} 1 \quad (7.12)$$

Für das Beispiel ergibt sich $R_{\text{CM}} = 13$, die Anzahl der mit * gekennzeichneten Zeilen.

Tabelle 7.4: Die Trigramme zum Beispiel „Hänschen klein“.

$\mathbf{p}_3(i) \cup \mathbf{q}_3(i)$			$U(\mathbf{q}_3(i), Q_3)$	$U(\mathbf{p}_3(i), P_3)$	gemeinsam
-2	0	-1	1	2	*
-2	0	0	1	0	
-2	0	1	2	2	*
-1	1	0	1	0	
-1	1	1	0	1	
-1	2	2	1	1	*
0	-2	0	1	1	*
0	-1	1	1	1	*
0	-1	2	1	1	*
0	0	-2	1	1	*
0	0	-1	1	0	
0	0	0	0	1	
0	1	-2	2	2	*
0	1	1	1	0	
1	-2	0	2	2	*
1	0	0	1	1	*
1	0	1	1	0	
1	1	0	1	1	*
1	1	1	0	2	
2	0	-2	1	1	*
2	2	0	1	1	*

7.3.2 Summe der Häufigkeiten (SF)

Bei der Summe der Häufigkeiten (sum of frequencies, SF) wird gezählt, wie oft das N-Gramm $\mathbf{g}_N(i)$, das sowohl in $\mathbf{q}$ wie in $\mathbf{p}$ vorhanden ist, in $\mathbf{p}$ auftritt.

$$R_{\text{SF}} = \sum_{\mathbf{g}_N(i) \in Q_N \cap P_N} U(\mathbf{g}_N(i), P_N) \quad (7.13)$$

Es werden also alle Häufigkeiten $U(\mathbf{g}_N(i), P_N)$ in Tabelle 7.4 addiert, die mit einem * gekennzeichnet sind. Für das Beispiel ergibt sich $R_{\text{SF}} = 17$.

7.3.3 Ukkonen-Messung (UK)

Die Ukkonen-Messung (UK) zählt die N-Gramme, die *nicht* übereinstimmend häufig in beiden Zeichenketten $\mathbf{q}$ und $\mathbf{p}$ vorkommen.

$$R_{\text{UK}} = - \sum_{\mathbf{g}_N(i) \in Q_N \cup P_N} |U(\mathbf{g}_N(i), Q_N) - U(\mathbf{g}_N(i), P_N)| \quad (7.14)$$

Es ist ausreichend, die Vereinigungsmenge $Q_N \cup P_N$ zu untersuchen, da für alle anderen möglichen N-Gramme der Beitrag verschwindet. Da große Werte einen großen Unterschied bedeuten und es sich damit um ein Distanzmaß handelt, während die Methoden Koordinatenvergleich (CM) und Summe der Häufigkeiten (SF) Ähnlichkeitsmaße sind, wird ein Minus-Zeichen vorangestellt. Für das Beispiel ergibt sich $R_{\text{UK}} = -10$.

7.4 Spezielle Verfahren

Neben den Verfahren zur Zeichenkettensuche und Indizierung gibt es Distanzmaße, die speziell auf die Melodiedarstellung eines QBH-Systems zugeschnitten sind. Im Rahmen dieser Arbeit sollen das Verfahren „TPBM I“ sowie die „Direkte Messung“ genannt werden, beide beziehen sich auf die Darstellung einer Melodie im Format MPEG-7-MelodyContour.

7.4.1 TPBM I

Der Algorithmus *TPBM I* (time pitch beat matching) wird in [45] und [108] beschrieben und ist das einzige in der Literatur bekannte Verfahren, das sich

direkt auf die MPEG-7-MelodyContour bezieht. Es werden sowohl der MPEG-7-Melody als auch der MPEG-7-Beat-Vektor ausgewertet. Weiterhin wird auch die Taktart mit einbezogen, die als Information im MPEG-7-MelodyContour-Descriptor enthalten ist; damit ergibt sich ein Triplet $\langle \mathbf{t}, \mathbf{p}, \mathbf{b} \rangle$, das einen Vektor $\mathbf{t}$ für Zähler und Nenner der Taktart, einen Vektor mit Tonhöhen- bzw. Konturwerten $\mathbf{p}$ und einen Vektor mit Taktinformationen $\mathbf{b}$ enthält.

Um das Ähnlichkeitsmaß R zwischen einem Melodiesegment $\Xi = \langle \mathbf{t}_m, \mathbf{p}_m, \mathbf{b}_m \rangle$ und der Anfrage $\Psi = \langle \mathbf{t}_q, \mathbf{p}_q, \mathbf{b}_q \rangle$ zu berechnen, werden folgende Schritte gemacht:

1. Auswertung der Taktart:

Falls die Zähler der Taktart $\mathbf{t}_m$ and $\mathbf{t}_q$ ungleich sind, wird für die Ähnlichkeit $R = 0$ zurückgegeben.

2. Initialisiere den Taktzähler: $n = 1$.

3. Für den Vergleich des Vektors $\mathbf{p}_q$ der Anfrage mit dem Melodiesegment Ξ werden alle Elemente vor dem n -ten Takt aus $\mathbf{p}_m$ gelöscht.

4. Nun erfolgt die Berechnung der Ähnlichkeit der Konturwerte pro Schlag (beat similarity score):

- a) Es werden alle Werte von $\mathbf{p}_m$ und $\mathbf{p}_q$ herangezogen, die in den Schlag fallen. Sie werden im Folgenden als Teilvektor (subset) $\mathbf{s}_q$ und $\mathbf{s}_m$ bezeichnet.
- b) Setze die Zähler $i = 0, j = 0$ und die Ähnlichkeit $R = 0$.
- c) Solange $i \leq |\mathbf{s}_q|$ und $j \leq |\mathbf{s}_m|$ (mit $|\cdot|$ wird die Länge des Vektors bezeichnet)
 - i. Falls die Vektorelemente $s_q(i) = s_m(j)$,
dann $R = R + 1$ und $i = i + 1, j = j + 1$
 - ii. sonst $k = j$ und
falls $s_q(i) \neq 0$ dann $j = j + 1$
falls $s_m(k) \neq 0$ dann $i = i + 1$
- d) Normiere das Ergebnis auf die Länge des Anfrageteilvektors:

$$R = \frac{R}{|\mathbf{s}_q|}.$$

5. Nun wird der Durchschnittswert der Ähnlichkeiten pro Schlag über alle Schläge der Anfrage gebildet. Das Ergebnis ist die Ähnlichkeit ab dem n -ten Takt: R_n .

6. Falls n nicht den letzten Takt in m bezeichnet, dann $n = n + 1$ und wiederhole Schritt 3.
7. Rückgabewert ist $R = \max\{R_n\}$, die höchste Ähnlichkeit bei Start an einem bestimmten Takt.

Die Anfrage Ψ wird also taktweise an die Melodie Ξ „angelegt“ und auf Ähnlichkeit geprüft.

7.4.2 Direkte Messung

Die Darstellung der Rhythmik im *Beat*-Vektor des MPEG-7-MelodyContour-Deskriptors enthält nur ganze Zahlen. Diese Tatsache lässt sich nutzen, um mit geringem Rechenaufwand die Ähnlichkeit zwischen zwei Zahlenfolgen festzustellen. In Arbeiten von EISENBERG et al. wurde dafür die *Direkte Messung* (direct measure, DM) entwickelt [68]. Es werden somit lediglich die rhythmischen Merkmale der Melodien untersucht. Die DM-Messung hat sich als robust gegen Notenaussetzer erwiesen. Der MPEG-7-Beat Vektor weist zwei wesentliche Beschränkungen auf, die die effiziente Berechnung der DM ermöglichen. Alle Vektorelemente sind positive ganze Zahlen und jede Zahl ist gleich groß oder größer als die vorangegangene Zahl.

Gegeben seien die beiden MPEG-7-Beat-Vektoren $\mathbf{u}$ und $\mathbf{v}$. Die iterative Berechnung erfolgt über drei Schritte:

1. Vergleiche die Vektorelemente $u(i)$ mit $v(j)$. Setze $i = j = 1$ für den ersten Vergleich.
2. Falls $u(i) = v(j)$, handelt es sich um einen Treffer. Die Indizes i und j werden um eins erhöht. Weiter mit Schritt 1.
3. Falls $u(i) \neq v(j)$, liegt ein Fehltreffer vor.
 - a) Falls $u(i) < v(j)$, wird nur Index i erhöht; weiter mit Schritt 1.
 - b) Falls $u(i) > v(j)$, wird nur Index j erhöht; weiter mit Schritt 1.

Der Vergleichsprozess wird solange durchgeführt, bis das Ende einer der beiden untersuchten Vektoren als Treffer erkannt worden oder das Ende beider Vektoren erreicht ist. Die Distanz R berechnet sich als Verhältnis aus Fehltreffern M zu Anzahl der Vergleiche V insgesamt:

$$R_{\text{DM}} = \frac{M}{V} \tag{7.15}$$

Die maximale Anzahl der Iterationen für zwei Vektoren der Länge N und M ist gleich der Summe der beiden Längen $N + M$. Damit ist die DM-Messung wesentlich effizienter als die Berechnung mit herkömmlichen Methoden, die wenigstens $N \cdot M$ Operationen benötigen.

7.5 Anwendung in Melodiesuchsystemen

Das Beispielsystem *Queryhammer* enthält zur Untersuchung der verschiedenen Ähnlichkeitsmaße die Implementierung aller dargestellten N-Gramm-Methoden CM, SF und UK sowie aller Zeichenkettensuchverfahren LCE, LCT und LAL. Mit diesen Verfahren werden die üblichsten in der Literatur untersuchten Techniken abgedeckt [138]. In diesem Abschnitt sollen einige Aspekte bei der Anwendung in Melodiesuchsystemen erörtert werden. Die ausführliche praktische Untersuchung der Verfahren erfolgt in Kapitel 8.

7.5.1 Diskussion der Verfahren

Aus der Funktionsweise der vorgestellten Verfahren zur Ähnlichkeitsbestimmung von Melodien ergeben sich verschiedene Konsequenzen. Die Vorteile der DP sind offensichtlich: durch die Berechnung der Edierdistanz lässt sich jedes Distanzmaß dieser Familie gut an die Anforderungen von Musiksuchsystemen und speziell von QBH-Systemen anpassen. Neben der Melodieähnlichkeit liefern Verfahren zur Zeichenkettensuche auch Informationen darüber, wo die Suchanfrage in der Melodie gefunden wird. Nachteile der DP sind der große Speicherbedarf bei längeren Melodien und die langsamere Verarbeitung gegenüber Indextechniken [187].

Der Vorteil der N-Gramm-Methoden ist entsprechend, dass sie schnell arbeiten. Allerdings ist die Verarbeitungsgeschwindigkeit von der Länge der N-Gramme abhängig, so dass N nicht beliebig groß gewählt werden kann. Als Nachteil der N-Gramme ist zu nennen, dass sie bedingt durch die statistische Funktionsweise – es werden N-Gramm-Häufigkeiten gezählt – keine genaue Aussage darüber erlauben, an welcher Stelle einer vorhandenen Melodie die Suchanfrage auftritt.

Das Verfahren TPBM-I erfordert die genaue Einhaltung des Taktes und der Taktart bei der Eingabe, was bei QBH nicht gewährleistet ist, solange dem Nutzer kein Metrum vorgegeben wird. Die Untersuchungen in [108] beziehen sich ausschließlich auf MIDI-extrahierte Anfragen, für ein QBH-System ist die-

ses Verfahren ohne weitere Maßnahmen ungeeignet. Das gleiche gilt für das DM-Verfahren, das speziell für Query-by-Tapping-Systeme entwickelt worden ist [68].

7.5.2 Einfluss der Melodielänge

Die Länge der verglichenen Melodiekonturen steht im unmittelbaren Zusammenhang mit dem Ergebnis der Ähnlichkeitsmessung. Da das Ergebnis ein Skalar ist, sind bei langen Melodiekonturen große, bei kurzen Zeichenketten kleine Werte zu erwarten. Dies bedeutet jedoch nicht, dass die Zeichenketten, die zu großen Zahlen führen, sich ähnlicher sind. Daher soll der Einfluss der Melodielänge durch eine Normierung des Ähnlichkeitsmaßes auf die Länge der Melodie, die untersucht worden ist, ausgeglichen werden.

In der Literatur werden verschiedene Normierungen vorgeschlagen [186]:

- Länge (len)
- Logarithmus der Länge (log)
- Quadratwurzel der Länge (2rt)
- n -te Wurzel der Länge (nrt)
- keine Normierung, als Referenz (non).

Die Bezeichnungen in Klammern geben die im Rahmen dieser Arbeit verwendeten Kurzbezeichnungen an. Es werden alle aufgeführten Normierungen untersucht, für die n -te Wurzel wird wie in [187] $n = 9$ (9rt) gewählt.

7.5.3 Implementierung

Die Berechnung der Ähnlichkeitsmaße SF, CM, UK, LCE, LCT und LAL findet in zwei Schritten statt. Im ersten Schritt werden die Ähnlichkeiten gemäß Gleichung (7.1, 7.5, 7.6, 7.12, 7.13) und (7.14) berechnet. Dieses Ergebnis wird gespeichert und kann für die weitere Auswertung verwendet werden. Im zweiten Schritt werden nun die berechneten Ähnlichkeitsmaße wie in Abschnitt 7.5.2 beschrieben normiert; durch das Zwischenspeichern können die Ähnlichkeitsmaße mehrfach ausgewertet werden, eine zeitaufwendige Neuberechnung entfällt.

7.6 Zusammenfassung

Der Melodievergleich ist die Kernaufgabe eines QBH-Systems. Er erfordert die Bestimmung der Ähnlichkeit zweier Melodiekonturen, wozu verschiedene Ähnlichkeitsmaße herangezogen werden können. Für Melodiekonturen gemäß dem Standard MPEG-7 eignen sich verschiedene Verfahren.

In diesem Kapitel wurden Verfahren der Zeichenkettensuche dargestellt, die mit Hilfe der dynamischen Programmierung berechnet werden können: die Berechnung der längsten gemeinsamen Teilsequenz (LCE), der längsten gemeinsamen Zeichenkette (LCT), und die Methode des lokalen Abgleichs (LAL). Weiterhin können Indizierungstechniken mittels N-Grammen benutzt werden, die ebenfalls eine Aussage über die Ähnlichkeit zweier Melodiekonturen machen. Es wurden die Ähnlichkeitsmaße Koordinatenvergleich (CM), Summe der Häufigkeiten (SF) und die Ukkonen-Messung (UK) vorgestellt. Die speziell für MPEG-7 entwickelten Ähnlichkeitsmaße wie das *Time Pitch Beat Measure I* (TPBM-I) und die *direkte Messung* (DM) eignen sich nur bedingt für QBH-Systeme, da sie das Vorhandensein einer genauen rhythmischen Information in der Anfrage voraussetzen. Sie werden in den vorliegenden Untersuchungen daher nicht berücksichtigt.

Schließlich wurden einige Aspekte für die Verwendung der vorgestellten Ähnlichkeitsmaße in Datenbanken erörtert. Besonders wichtig ist der Aspekt der Normierung des Ähnlichkeitsmaßes auf die Melodielänge, um eine sinnvolle Gewichtung des Ähnlichkeitsmaßes zu erzielen. Im Rahmen dieser Arbeit werden betrachtet: die Normierung auf die Länge (len), auf den Logarithmus der Länge (log), die Quadratwurzel der Länge (2rt), die 9-te Wurzel der Länge (9rt) sowie keine Normierung zu Referenzzwecken (non). Die Implementierung des Beispielsystems *Queryhammer* enthält damit die Ähnlichkeitsmaße LCE, LCT, LAL, CM, SF und UK, die mit den Normierungen len, log, 2rt, 9rt und non kombiniert werden können.

*Was niemand sucht, wird
selten gefunden.*

Johann Heinrich Pestalozzi

In den vorangegangenen Kapiteln dieser Arbeit sind die einzelnen Verarbeitungsstufen eines QBH-Systems dargestellt und ihre Eigenschaften untersucht worden. Die *monophone Transkription* der Nutzeranfrage und die *Vergleichsstufe* ermöglichen die Suche in der *Melodiedatenbank*. Im Mittelpunkt dieses Kapitels steht die Frage, wie der Sucherfolg von den drei genannten Komponenten abhängt. Insbesondere soll betrachtet werden, wie weit der Inhalt der Melodiedatenbank das Ergebnis der Suche beeinflusst. Die monophone Transkription für QBH-Systeme sowie der Vergleich von Melodiekonturen ist in der Literatur untersucht worden, die bisherigen Kapitel bieten eine ausführliche Übersicht. Dagegen sind dem Autor keine systematischen Untersuchungen des Melodiedatenbankinhalts bekannt, der aber wesentlicher Bestandteil eines Musiksuchsystems ist.

8.1 Bewertung von Musiksuchsystemen

Das zentrale Problem bei der Beurteilung von Ergebnissen zu einer Anfrage an ein Informationssuchsystem besteht darin, dass die richtige Antwort bekannt sein muss, um die Antwort des Systems bewerten zu können [70]. Diese Problematik soll anhand der Benutzung eines QBH-Systems kurz erläutert werden.

Ein Nutzer summe eine Melodie in ein QBH-System mit dem Wunsch, Titel und Interpret des Stückes zu erfahren, an dessen Melodie er sich erinnert. Ideal wäre genau *eine* richtige Antwort, die aber tatsächlich nur möglich ist, wenn die Melodie eindeutig ist. Dies ist nicht zwingend der Fall, insbesondere ist die sich ergebende Melodiekontur nicht eindeutig. Trotzdem kann das Suchergebnis für den Nutzer durchaus befriedigend sein. Wird von dem QBH-System eine Liste mit den ähnlichsten Melodien angezeigt, stellt die Aus-

wahl des gesuchten Titels bei einer überschaubaren Anzahl von Möglichkeiten kein Problem dar. Ist der gesuchte Interpret nicht in der Datenbank enthalten, wohl aber eine Bearbeitung der Melodie durch einen anderen Interpreten (sog. „Cover“-Versionen oder „Remixe“ in der Pop-Musik erleben derzeit einen regelrechten Boom), so kann aufgrund der großen Ähnlichkeit möglicherweise zumindest der Titel zur Melodie gefunden werden.

Die Länge der zu durchsuchenden Trefferliste variiert bei bestehenden Systemen. In der Literatur findet man Evaluierungen von QBH-Systemen, in denen untersucht wird, ob der Anfragetitel in den bestplatzierten 5 Titeln enthalten ist [37, 197]. Ebenfalls üblich ist die Betrachtung der bestenplatzierten 10 [45, 69, 97, 124, 197], 20 [197], 40 [154] oder sogar 100 Titel [97]. Die Güte des Suchergebnisses hängt damit nicht ausschließlich davon ab, wie *gut* der gesuchte Titel in der Ergebnis-Liste platziert ist, sondern auch davon, dass er im Anfrageergebnis *leicht* gefunden werden kann. Für die Bewertung von QBH-Systemen kann daher die Untersuchung der Vollständigkeit und Präzision des Ergebnisses verwendet werden [70]. Die Vollständigkeit (recall) gibt Auskunft darüber, ob alle für die Suche relevanten Treffer im Suchergebnis enthalten sind. Die Präzision (precision) ist ein Maß dafür, dass möglichst viele relevante Treffer im Suchergebnis enthalten sind.

In der Literatur zu QBH-Systemen findet man sowohl Messungen der Vollständigkeit [69], als auch der Präzision [60, 173] oder auch beider Maße [36, 144, 149]. Zur Berechnung von Vollständigkeit und Präzision muss bekannt sein, welche Treffer im Ergebnis einer Suchanfrage überhaupt relevant sind. Daher wird nun der Begriff der Relevanz eingeführt.

8.1.1 Relevanz

Um zu überprüfen, ob zu einer Anfrage die richtigen Dokumente gefunden wurden, muss bekannt sein, welche Dokumente in der Datenbank vorhanden sind, die der Anfrage zuzuordnen sind [70]. Dazu verwendet man das Konstrukt der Relevanz, einer Beziehung, die zwischen einer Anfrage und einem Dokument besteht.

Definition Die Relevanz eines Dokuments für eine Anfrage ist eine Relation $r : D \times Q \rightarrow R$, wobei $D = \{d_1, \dots, d_m\}$ die Menge der Dokumente, Q die Menge der Anfragen und R eine Menge von Wahrheitswerten, im Allgemeinen die Menge $\{0, 1\}$, ist.

Die Relation r wird im Allgemeinen durch Befragen von Experten zu konkreten Anfragen und Dokumentenmengen ermittelt und als Tabelle oder in Form von Listen gespeichert. Auch UITDENBOGERD geht so vor, indem sie die Ähnlichkeit einer ausgewählten Menge von Titeln aus ihrer Datenbank von Musikern bewerten lässt [187]. In der vorliegenden Arbeit ist für die getesteten Anfragen jeweils ein relevantes Dokument in der Datenbank vorhanden, das alleine als relevant angenommen wird.

8.1.2 Vollständigkeit und Präzision

Die oben diskutierten Maße Vollständigkeit und Präzision werden nun genau definiert [70].

Definition Sei $D = \{d_1, \dots, d_m\}$ eine Menge von Dokumenten, $q \in Q$ eine Anfrage und D_q die Menge der in D zur Anfrage q gefundenen Dokumente. Sei ferner $r : D \times Q \rightarrow \{0, 1\}$ eine Relevanzrelation und R_q die Menge der Dokumente aus D_q , die zur Anfrage q mit dieser Relevanzfunktion als relevant bewertet werden (d. h. $R_q = r^{-1}q(\{1\})$). Dann heißt

$$P(q, D) := \frac{|D_q \cap R_q|}{|D_q|} \quad (8.1)$$

Precision, Präzision oder **Genauigkeit** der Antwort auf die Anfrage q und

$$V(q, D) := \frac{|D_q \cap R_q|}{|R_q|} \quad (8.2)$$

Recall oder **Vollständigkeit** der Antwort auf die Anfrage q .

Die Präzision gibt also den Anteil der relevanten Dokumente unter den gefundenen Dokumenten an, die Vollständigkeit gibt den Anteil der relevanten Dokumente an, die gefunden wurden. Optimal, nämlich gleich 1, sind die Werte für Präzision und Vollständigkeit natürlich genau dann, wenn $D_q = R_q$ gilt, wenn also genau alle relevanten Dokumente als Antwortmenge zurückgeliefert werden.

Wird die Menge der gefundenen Dokumente festgelegt, ist $|D_q| = N_q = \text{const.}$ Im Folgenden soll diese Bewertungsmethode mit V_N bezeichnet werden, z. B. also V_{10} für die Untersuchung der Vollständigkeit für die 10 besten Treffer in der Ergebnisliste der Suche. Es gelte weiterhin, dass $|R_q| = 1$ ist

(je Suchanfrage nur ein Titel relevant, es sind also keine doppelten Melodien in der Datenbank enthalten), damit kann die Vollständigkeit nur 0 oder 1 betragen.

8.2 Statistik der Melodiedatenbank

In diesem Abschnitt werden statistische Kenngrößen ermittelt, mit deren Hilfe die Eigenschaften einer Melodiedatenbank beschrieben werden können. Ausgehend von vorliegenden Melodien, die aus MIDI-Dateien extrahiert werden, soll eine Melodiedatenbank synthetisiert werden, die die gleichen statistischen Eigenschaften besitzt, aber lediglich zufällig gewählte Melodieinformationen enthält.

8.2.1 Referenzdateien

Im Rahmen dieser Arbeit wird das Genre Pop-Musik untersucht, weil der Internetnutzer als typischer Anwender eines QBH-Systems besonders für dieses Genre Interesse zeigt. Auch das Angebot von *Musicline* ist auf dieses Genre abgestimmt [11].

Um Vergleichbarkeit der Nutzeranfragen untereinander zu wahren, werden die zu summenden Anfragen für Probanden auf eine kleine Menge von Titeln beschränkt. Die Liste der Titel und Interpreten der deutschen „Top-10-Single-Charts“ vom März 2003, im Folgenden kurz „Top-10“, ist in Tabelle 8.1 dargestellt. Es handelt sich um eine Mischung von zwei deutschsprachigen Titeln (Wolfshiem: „Kein zurück“ und Nena: „Wunder geschehen“), einem französischsprachigen (Kate Ryan: „Desenchantee“) und sonst englischsprachigen Titeln. Alle Titel zeigen die für Pop-Musik typische Struktur von Refrain und Strophe. Bei „Rhythm is a Dancer 2003“ von Snap handelt es sich um einen sog. Remix aus dem Jahr 2003, das Original ist aus dem Jahr 1992. Eine vollständige Liste aller Titel findet sich im Anhang A.

Die Titel liegen sowohl im Mp3-Format als auch in einer MIDI-Version vor. Aus der MIDI-Version werden die in der Datenbank abgelegten Melodiekonturen gemäß MPEG-7 extrahiert und gespeichert. Für die Anfragen, die für die weiteren Untersuchungen aus den MIDI-Informationen gewonnen werden, wird ein jeweils signifikanter Teil ausgewählt, entweder Strophe oder Refrain. Die ausgewählten Teile sind in Abbildung 8.1 in Notenform darstellt.



(a) TATU – All the Things She Said



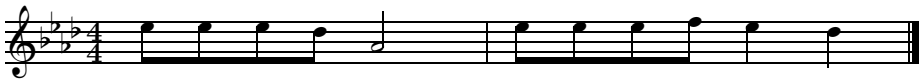
(b) Scooter – Weekend



(c) Kate Ryan – Desenchantee



(d) Blue feat. Elton John – Sorry Seems to Be the Hardest Word



(e) Gareth Gates – Anyone of Us



(f) Wolfsheim – Kein zurück



(g) Deutschland sucht den Superstar – We Have a Dream

Abbildung 8.1: Die Noten zu den Melodien der untersuchten Top-10 aus Tabelle 8.1 (Fortsetzung auf Seite 152).

Tabelle 8.1: Titel und Interpreten der deutschen „Top-10-Single-Charts“ vom März 2003.

Platz	Interpret	Titel
1	TATU	All the Things She Said
2	Scooter	Weekend
3	Kate Ryan	Desenchantee
4	Blue feat. Elton John	Sorry Seems to Be the Hardest Word
5	Gareth Gates	Anyone of Us
6	Wolfsheim	Kein zurück
7	Deutschland sucht den Superstar	We Have a Dream
8	Eminem	Lose Yourself
9	Nena and Friends	Wunder geschehen
10	Snap	Rhythm Is a Dancer 2003



(h) Eminem – Lose Yourself (Begleitstimme)



(i) Nena and Friends – Wunder geschehen



(j) Snap – Rhythm Is a Dancer 2003

Abbildung 8.1: (Fts.) Die Noten zu den Melodien der untersuchten Top-10 aus Tabelle 8.1.

Für die weiteren Melodien der Melodiedatenbank werden die MIDI-Dateien zur Extraktion der Melodiekontur verwendet. Diese wurden bereits von KIM und CHAI für die Entwicklung der MPEG-7-MelodyContour untersucht und stehen im Internet zur Verfügung [45,108]. Es handelt sich überwiegend um Titel aus dem Genre Pop-Musik und einigen Titeln aus dem Bereich der Klassik (10 Titel). Insgesamt hat die Melodiedatenbank für die eigenen Untersuchungen damit einen Umfang von 405 Titeln. Das vollständige Verzeichnis findet sich im Anhang A.

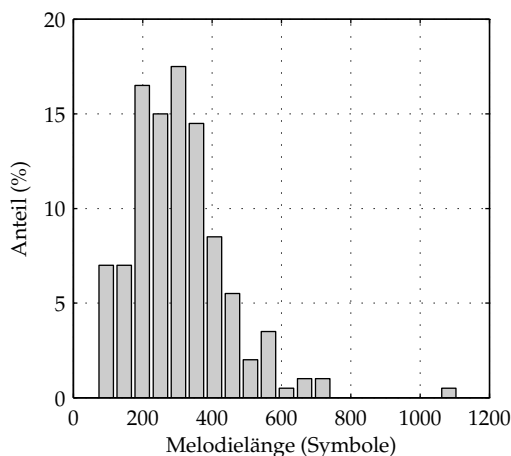
8.2.2 Parameter der Melodiedatenbank

Zur Beschreibung von Melodiedatenbanken sollen nun alle messbaren statistischen Parameter der Melodiedaten betrachtet werden: neben der Größe des Datenbankbestandes kann die Häufigkeitsverteilung der Melodielängen (Noten bzw. Symbole pro Melodie) ermittelt werden. Innerhalb der Melodie kann die Häufigkeitsverteilung der auftretenden Intervalle, speziell bei Melodiekonturen die Häufigkeitsverteilung der verwendeten Kontursymbole betrachtet werden.

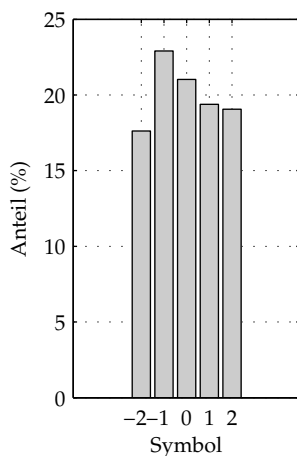
Datenbankumfang

In der Literatur findet man sehr unterschiedliche Datenbankgrößen, die in den verschiedenen Untersuchungen von QBH-Systemen verwendet werden; kleine Datenbanken wie in [108] umfassen nur 50 Titel, in [36] werden 100 Titel und in [76] 183 Titel verwendet. Mittlere Datenbankbestände enthalten etwa 400–1000 Titel [27,45,67,69,124,149]. Darüber hinaus gibt es einige Untersuchungen mit relativ großen Datenbanken wie [197] mit 3000 Titeln, [116,131,154] mit etwa 10.000 Titeln und [37] mit 39.925 Titeln. Die Untersuchungen von DANNENBERG et al. verwenden eine kleine Sammlungen von Titeln (zum Beispiel 258 Titel der Beatles) und extrahieren daraus einzelne Themen, was in einer größeren Datenbank resultiert (2844 Themen) [54].

Betrachtet man die Datenbanken kommerzieller Musikanbieter im Internet, so kommen ganz anderen Datenbankgrößen in Betracht. Etwa 2 Millionen Titel bietet das Portal *iTunes* von Apple an [17], beim Anbieter *mp3.com* sind es sogar 6 Millionen [2]. Da QBH-Systeme gerade bei solchen Größenordnungen hilfreich sein können, ist die Untersuchung von großen Datenbeständen von besonderem Interesse.



(a) Die Häufigkeitsverteilung der Melodielänge.



(b) Die Häufigkeitsverteilung der Melodiekonturwerte.

Abbildung 8.2: Statistische Kenngrößen der untersuchten MIDI-Datenbank.

Melodielänge

Mit der Länge einer Melodie soll im Rahmen dieser Arbeit die Anzahl der Noten bezeichnet werden, nicht die zeitliche Dauer. Weiterhin muss unterschieden werden, ob mit der Melodie eines Titels das musikalische Thema (vergleiche Abschnitt 2.1) gemeint ist oder die melodieführende Stimme. In diesem Fall werden für die Melodiedatenbank die Melodiespuren der MIDI-Dateien verwendet (vergleiche auch Kapitel 6), damit gehen auch alle Wiederholungen des Themas in Melodie ein.

In der Literatur finden sich nur wenige Angaben zur Häufigkeitsverteilung der Melodielänge. KOSUGI et al. verwenden eine Datenbank mit gut 10.000 Titeln, die durchschnittliche Melodielänge beträgt 365,16 Noten. Bei DANNENBERG et al. werden die Themen wie in Abschnitt 8.2.2 beschrieben einzeln ausgewertet und sind im Schnitt 41 Notenwerte lang [54].

Die Häufigkeitsverteilung der Melodielängen der im Rahmen dieser Arbeit verwendeten MIDI-Datenbank ist in Abbildung 8.2a dargestellt. Die Verteilung zeigt eine ähnliche Form wie in den Untersuchungen von [116], der Mittelwert der Melodielänge liegt bei 299,3 Notenwerten pro Melodie.

Symbolverteilung

Es sollen nun Parameter der Melodien selbst ermittelt werden. Für die Darstellung von Melodien als Melodiekontur ist es interessant, die Häufigkeitsverteilung der in der Melodie auftretenden Intervalle zu betrachten. Die Messungen von KIM zeigen, dass bei der Zuordnung der Intervalle zu fünf Konturstufen, wie sie in MPEG-7 vorgenommen wird, alle Kontursymbole gleichverteilt auftreten [45, 108]. Die gemessene Verteilung der Kontursymbole in der eigenen Melodiedatenbank, die aus den in Abschnitt 8.2.1 beschriebenen MIDI-Daten generiert worden ist, entspricht tatsächlich einer Gleichverteilung, wie Abbildung 8.2b zeigt.

Um weitere statistische Parameter auszuwerten, können bedingte Wahrscheinlichkeiten für die Kontursymbole gemessen werden. Dazu betrachtet man die Konturfolgen als wert- und zeitdiskreten Zufallsprozess [94, 143, 146]. Der einfachste Typ eines solchen Prozesses ist eine *MARKOV-Kette erster Ordnung*. Als Zufallsprozess betrachtet kann eine Melodiekontur mit $M = 5$ Symbolen s_m , die ebensoviele *Zustände* $u(k)$ zum Zeitpunkt k einnehmen und nur vom vorangegangenen Zustand $u(k-1)$ abhängen. Die Wahrscheinlichkeit für den neuen Zustand wird mit

$$P_m(k) = P(u(k)|u(k-1)) \quad (8.3)$$

angegeben. Diese werden in einem Vektor der Zustandswahrscheinlichkeiten

$$\mathbf{P}(k) = [P_1(k) \dots P_M(k)]^T \quad (8.4)$$

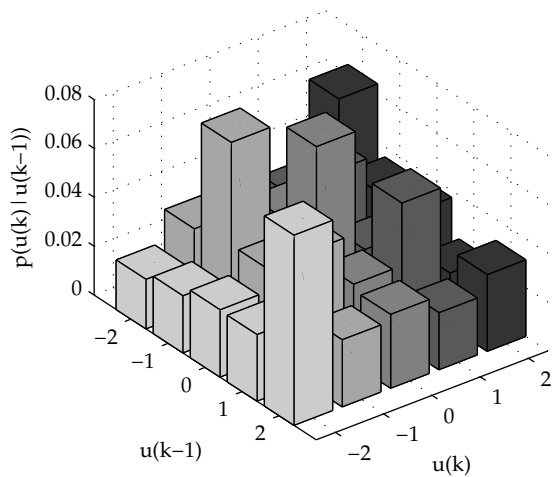
zusammengefasst. Die Wahrscheinlichkeit für den Übergang vom Zustand s_m in s_n soll mit

$$P_{mn}(k) = P(u(k) = s_n | u(k-1) = s_m) \quad (8.5)$$

bezeichnet werden. Es sind insgesamt m^2 Zustandsübergänge möglich, die Zustandsübergangswahrscheinlichkeiten können in einer Matrix $\mathbf{Q}(k)$ zusammengefasst werden, sie besitzt die Elemente

$$\mathbf{Q}(k) = [P_{mn}(k)]. \quad (8.6)$$

Um ein weiteres Beschreibungskriterium für die Konturwerte der MIDI-Dateien zu gewinnen, wird die Zustandsübergangsmatrix aus dem Mittelwert der auftretenden Zustandsübergänge gebildet. In Abbildung 8.3 ist die Matrix



(a) Die grafische Darstellung der Zustandsübergangsmatrix lässt deutlich die häufigen Zustandsübergänge erkennen.

$P_{mn}(k) = P(u(k) = s_n u(k-1) = s_m)$						
		s_n				
		-2	-1	0	1	2
s_m	-2	0.019	0.037	0.027	0.029	0.059
	-1	0.025	0.080	0.043	0.053	0.037
	0	0.026	0.038	0.080	0.030	0.039
	1	0.026	0.053	0.034	0.059	0.022
	2	0.074	0.028	0.029	0.024	0.026

(b) Die einzelnen bedingten Wahrscheinlichkeiten der mittleren Zustandsübergangsmatrix $\bar{Q}$.

Abbildung 8.3: Die Zustandsübergangsmatrix für die Konturwerte der MIDI-Datenbank.

$\bar{Q}$ der untersuchten Melodiedatenbank dargestellt. Es zeigt sich eine Bevorzugung der Zustandsübergangswahrscheinlichkeiten $P(0|0)$, $P(-1|-1)$ und $P(1|1)$ sowie $P(-2|2)$ und $P(2|-2)$. Dies ist musikalisch plausibel, da $P(0|0)$ Tonwiederholungen entsprechen, die besonders rhythmische Informationen wiedergeben, für das Genre Pop-Musik durchaus üblich. Die Wahrscheinlichkeiten $P(-1|-1)$ und $P(1|1)$ treten bei Tonfolgen auf, in denen eine Tonleiter auf- oder abwärts gespielt wird. Mit den stark ausgeprägten Zustandsübergangswahrscheinlichkeiten $P(2|-2)$ und $P(-2|2)$ wird klar, dass Folgen der Symbole „-2 2 -2 2 -2 ...“ häufig auftreten. Damit können Wechsel auf die nächst höhere Akkordnote und zurück beschrieben werden. Die ebenfalls erhöhten Werte für $P(1|-1)$ und $P(-1|1)$ beschreiben diesen Hin- und Herwechsel entsprechend für Töne auf einer Tonleiter.

8.2.3 Die Bedeutung statistischer Parameter der Melodiedatenbank für die Suche

Dieser Abschnitt geht der Frage nach, ob es einen Zusammenhang zwischen den statistischen Parametern der Melodiedatenbank und dem erzielbaren Sucherfolg mit einem QBH-System gibt, das diese Melodiedatenbank verwendet. Offensichtlich ist, dass bei wachsendem Umfang des Melodiebestandes die Suche nach einem in der Datenbank enthaltenen Titel zunehmend schwieriger wird, da zum gesuchten Titel immer mehr ähnliche Titel hinzukommen können. Es ist aber nicht unmittelbar klar, ob Abhängigkeiten des Sucherfolgs vom Genre der Musik bestehen. Da für die Entwicklung und Verwendung von QBH-Systemen eine genaue Beurteilung dieser Zusammenhänge wünschenswert ist, sollen im Folgenden einige Betrachtungen zur Modellierung einer Musikdatenbank ausgeführt werden, die eine weitestgehend neutrale, d. h. von den Testdaten unabhängige Untersuchung von QBH-Systemen ermöglicht.

DANNENBERG et al. untersuchen in [54] den Einfluss der Datenbankgröße in Bezug auf die Güte des Suchergebnisses unter Verwendung verschiedener Suchalgorithmen. Dabei wird festgestellt, dass die Güte der Suchergebnisse mit wachsender Datenbankgröße N_{DB} nur relativ langsam fällt. Zur Modellierung dieses Zusammenhangs wird der Ausdruck

$$R_{MRR} = \frac{1}{\log N_{DB}} \quad (8.7)$$

angegeben. Der Wert R_{MRR} steht für die „mittlere reziproke Platzierung“ (mean reciprocal rank, MRR), also den mittleren Kehrwert des Trefferplatzes in der

Ergebnisliste. Die Untersuchungen an verschiedenen Melodiedatenbanken – eine mit Titeln der Beatles, eine weitere mit Volksliedern – zeigen, dass für die in [54] untersuchten Distanzmaße Gleichung (8.7) den Zusammenhang zwischen Datenbankgröße und Sucherfolg gut beschreibt.

Für die eigenen Untersuchungen soll nun festgestellt werden, ob ein formaler Zusammenhang für die Vollständigkeit des Suchergebnisses gilt. Dabei ist die Betrachtung der verschiedenen Vollständigkeitsmaße V_1 , V_3 und V_{10} ebenso von Interesse wie die Untersuchung der vorgestellten (und im Bereich von Melodiesuchsystemen üblichen) Ähnlichkeitsmaße. Für die Melodiedatenbank soll im Gegensatz zu den Untersuchungen in [54] der allgemein üblichere Fall von vollständigen Melodiespuren statt einzelner Themen betrachtet werden.

Weiterhin wird untersucht, wie weit Abhängigkeiten vom musikalischen Genre der Melodiedatenbank bestehen. Für diesen Zweck wird nun eine Melodiedatenbank mit Zufallsmelodien erzeugt. Die Melodien werden durch Zufallszahlen dargestellt, die gemäß der gemessenen Statistiken der vorhandenen MIDI-Datenbank generiert werden. Als statistische Merkmale werden die Häufigkeitsverteilung der Melodielänge und die Häufigkeitsverteilung der Melodiekonturwerte sowie die Zustandsübergangsmatrix herangezogen. Eine Zufallsdatenbank kann auf diese Weise rein synthetisch in beliebiger Größe generiert werden. Durch Vergleich der erzielbaren Ergebnisse kann die Gültigkeit des Modells überprüft werden.

MIDI-Datenbank

Zunächst werden die Eigenschaften für die Suche in der vorhandenen Melodiedatenbank, im Folgenden kurz als MIDI-Datenbank bezeichnet, untersucht. Die MIDI-Datenbank bietet wie in Abschnitt 8.2.1 beschrieben 405 Melodiekonturen, die aus MIDI-Dateien extrahiert worden sind.

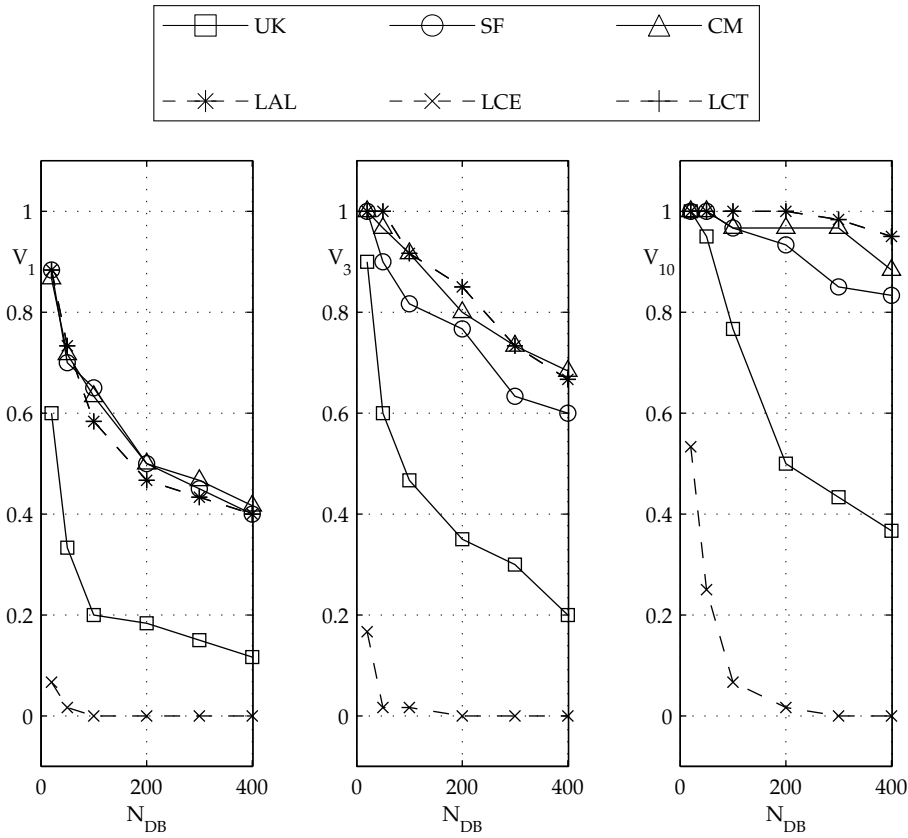
Zur Untersuchung werden als Anfragen die zehn Melodiekonturen aus Tabelle 8.1 sowie 50 weitere, zufällig aus dem MIDI-Datenbestand ausgewählte Melodiekonturen verwendet. Als kleinste Datenbankgröße wird $N_{DB} = 20$ angenommen, damit muss auch die Anzahl der Anfragen pro Untersuchung auf 20 Melodien begrenzt werden, die in diesem kleinsten Datenbanksatz wiedergefunden werden können. Es werden aus diesem Grund drei Messungen mit je 20 Anfragen vorgenommen und dann gemittelt. Die Länge der Anfrage wird auf 8 Noten begrenzt. Dieser Wert hat sich in Voruntersuchungen als kritisch erwiesen – bei kürzeren Anfragen lassen sich kaum noch brauchbare Sucher-

gebnisse erzielen, bei längeren Anfragen zeigen sich evtl. auftretende Unterschiede zwischen den Distanzmaßen nicht mehr so klar. Der Einfluss der Anfragelänge wird detailliert in Abschnitt 8.4.1 untersucht. Als Datenbankgröße wird die Anzahl der verwendeten Melodien zu $N_{DB} = 20, 30, 50, 100, 200, 300$ und 400 gewählt. Es werden die Distanzmaße UK, SF, CM und LCE, LCT und LAL mit jeweils optimaler Normierung verwendet. Die Wahl der Normierung wird in Abschnitt 8.3 ausführlich untersucht.

In Abbildung 8.4 ist die Vollständigkeit der Suchanfrage gemittelt über alle 60 Anfragen dargestellt. In Abbildung 8.4a ist die Vollständigkeit für V_1 dargestellt, d. h. nur wenn der gesuchte Titel erstplatziert ist, ist ein Treffer erzielt. Alle Distanzmaße fallen im Bereich $N_{DB} < 100$ mit wachsender Größe stark ab, danach wird der Verlauf zunehmend flacher. Die besten Ergebnisse erzielen die Verfahren CM, SF und deckungsgleich LCT und LAL (wie zu erwarten war, denn die Anfragen sind fehlerfrei), die schlechtesten Ergebnisse liefern UK und mit Abstand LCE. Nun sollen die Vollständigkeitswerte für größere Ergebnismengen betrachtet werden. Die Ergebnisse für V_3 sind in Abbildung 8.4b dargestellt – alle Ergebnisse außer LCE fallen besser aus als in Abbildung 8.4a, CM, LAL und LCT sind am besten, gefolgt von SF, UK liegt im Mittelfeld, LCE ist unbrauchbar. Auch hier lässt sich ein Abflachen der Kurven für wachsende Datenbankgrößen erkennen. Die Reihenfolge der Distanzmaße wiederholt sich in der Darstellung für V_{10} in Abbildung 8.4c. Die Vollständigkeitswerte sind insgesamt wesentlich höher als bei den Messungen für V_3 und V_1 . Eine vollständige Wiederfindung aller Anfragen ist bis $N_{DB} = 200$ möglich. Um bessere Vollständigkeitswerte bei größeren Datenbanken zu erzielen, sind längere Anfragen notwendig.

Zufallsdatenbank

Es werden nun Melodiedatenbanken untersucht, die lediglich Zufallsmelodien enthalten. Bei der Generierung dieser Zufallsdaten wurde die in Abbildung 8.2 angegebene statistische Verteilung von Melodielängen und Kontursymbolen zugrunde gelegt und mit Zufallszahlen modelliert. Nicht berücksichtigt wird zunächst die Modellierung der Melodien als Markov-Kette, d. h. alle Melodiesymbole treten statistisch unabhängig voneinander auf. Es handelt sich somit um reine „Zufallsmelodien“. Danach wird die gleiche Untersuchung mit Markov-modellierten Zufallsprozessen wiederholt, sie werden in diesem Zusammenhang als „Markov-Melodien“ bezeichnet. Weil die Größe der Datenbank frei wählbar ist, werden die Untersuchungen bis zum Wert



(a) Die besten Ergebnisse für V_1 erzielen die Verfahren CM, SF und deckungsgleich LCT und LAL, die schlechtesten Ergebnisse liefern UK und mit Abstand LCE.

(b) Für V_3 fallen alle Ergebnisse außer LCE besser aus als in Abbildung 8.4a, CM, LAL und LCT sind am besten, gefolgt von SF, UK liegt im Mittelfeld, LCE ist unbrauchbar.

(c) Die Vollständigkeitswerte für V_{10} sind insgesamt wesentlich höher als bei den Messungen für V_3 und V_1 . Eine vollständige Wiederfindung aller Anfragen ist bis $N_{DB} = 200$ möglich.

Abbildung 8.4: Die gemittelte Vollständigkeit von 60 verschiedenen Anfragen in Abhängigkeit vom Umfang der Melodiedatenbank, extrahiert aus MIDI-Dateien.

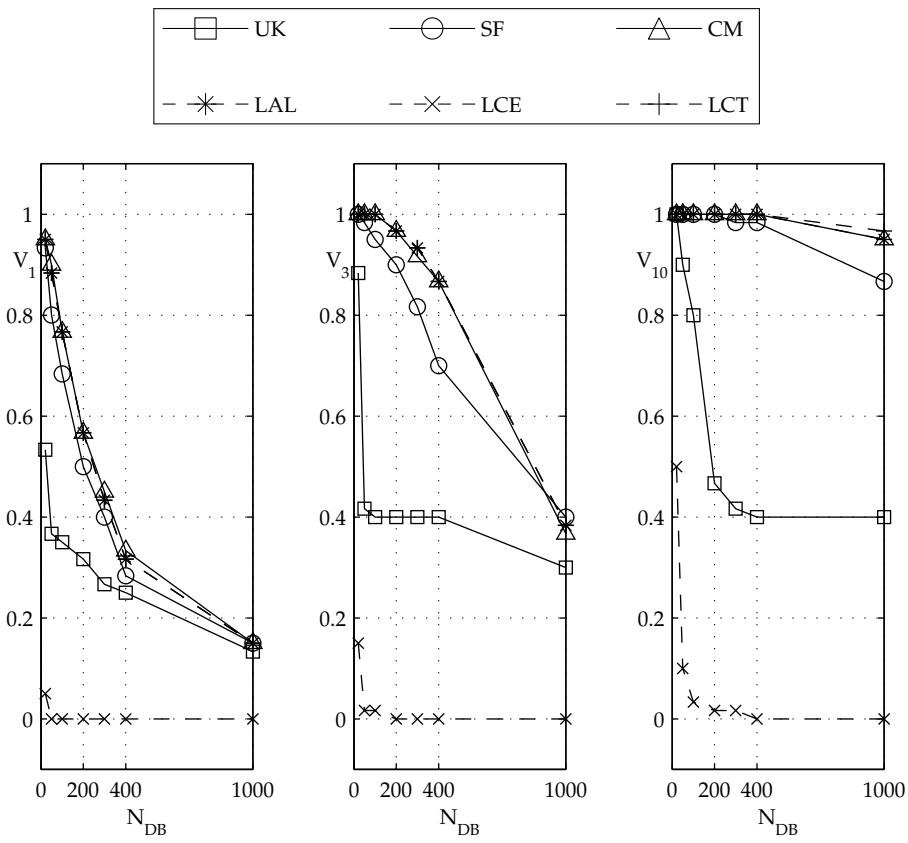
$N_{DB} = 1000$ durchgeführt. Wie in Abschnitt 8.2.2 sind weitaus größere Werte von Interesse, der gewählte Wert dient daher nur einer ersten Prüfung der Gültigkeit der verwendeten Modelle.

Zufallsmelodien Die Datenbank mit den Zufallsmelodien wird ebenso wie die MIDI-Datenbank untersucht; es werden 60 Titel der Datenbank auf eine Länge von 8 Notenwerten gekürzt und als Anfrage verwendet. Untersucht werden wiederum die Vollständigkeitswerte für V_1 , V_3 und V_{10} .

In Abbildung 8.5a sind die Vollständigkeitswerte V_1 dargestellt. Die Reihenfolge der Distanzmaße ist wie bei den MIDI-Daten: Die N-Gramm-Methoden SF und CM sowie die DP-Methoden LAL und LCT sind am besten, UK liegt deutlich darunter, LCE ist unbrauchbar. Bei V_3 fallen die Werte weniger steil, wieder sind CM, SF, LAL und LCT am besten, UK schneidet gegenüber den MIDI-Daten etwas schlechter ab. Bei den Vollständigkeitswerten zu V_{10} wiederholt sich wiederum der Trend der Distanzmaße, mit CM, SF, LAL und LCT lassen sich gute Ergebnisse erzielen, UK liefert für die MIDI-Daten bessere Ergebnisse als bei Zufallszahlen und LCE ist unbrauchbar. Die Suchanfragen bis zu 400 Datenbankeinträgen können für den überwiegenden Teil aller Ähnlichkeitsmaße zu 100 % beantwortet werden.

Markov-Melodien Um den Einfluss des Inhalts der Melodiedatenbank zu untersuchen, soll nun die Modellierung des Zufallsprozesses zur Erzeugung der Zufallsdatenbank gemäß der gemessenen Statistik aus Abschnitt 8.2.2 verwendet werden. Das Ergebnis ist diesmal eine Zufallsdatenbank mit „Markov-Melodien“, die zur Unterscheidung vom vorangegangenen Fall als *Markov-Datenbank* bezeichnet werden soll.

In Abbildung 8.6 sind die Vollständigkeitswerte für die Anfragen wie im bereits vorangegangenen Versuch dargestellt. In Abbildung 8.6a zeigt sich der schnell fallende und dann flacher werdende Verlauf für V_1 , wie er auch in Abbildung 8.5a und für Werte bis 400 in Abbildung 8.4a festgestellt werden kann. Auffällig ist, dass die Ergebnisse aller Ähnlichkeitsmaße außer LCE im gleichen Wertebereich liegen, UK weicht nur für Datenbankengrößen kleiner 200 nach unten ab. Auch in Abbildung 8.6b liegen die Werte enger bei einander als in Abbildung 8.5b oder Abbildung 8.4b. UK ist jedoch über den gesamten betrachteten Wertebereich unterlegen. Das vollständige Wiederfinden der Anfragen in Abbildung 8.6c ist ähnlich wie bei den MIDI-Daten nur bis $N_{DB} = 200$ möglich.

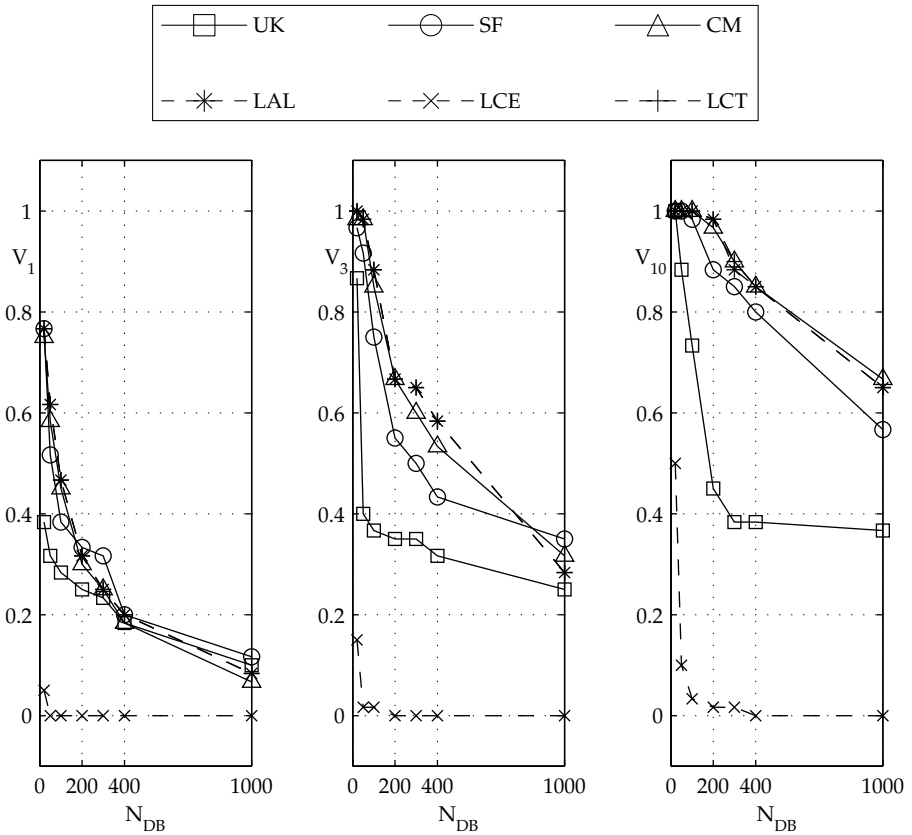


(a) Die N-Gramm-Methoden SF und CM sowie die DP-Methoden LAL und LCT sind für das Maß V_1 am besten, UK liegt deutlich darunter, LCE ist unbrauchbar.

(b) Bei V_3 fallen die Werte weniger steil als bei V_1 , die Reihenfolge der Distanzmaße bleibt.

(c) Bei den Vollständigkeitswerten zu V_{10} wiederholt sich wiederum der Trend der Distanzmaße.

Abbildung 8.5: Die mittlere Vollständigkeit von 60 Anfragen in Abhängigkeit vom Umfang der Melodiedatenbank für Zufallsmelodien.



(a) Für V_1 zeigt sich der schnell fallende und dann flacher werdende Verlauf wie in Abbildung 8.4a.

(b) Auffällig für V_3 ist, dass die Ergebnisse aller Ähnlichkeitsmaße außer LCE im gleichen Wertebereich liegen, UK weicht nur für Datenbankgrößen kleiner 200 nach unten ab.

(c) Das vollständige Wiederfinden der Anfragen für V_{10} ist ähnlich wie bei den MIDI-Daten nur bis $N_{DB} = 200$ möglich.

Abbildung 8.6: Die mittlere Vollständigkeit von 60 Anfragen in Abhängigkeit vom Umfang der Melodiedatenbank für Markov-Melodien.

Vergleich von MIDI- und Zufallsdaten

Um den Vergleich von Zufalls-, Markov- und MIDI-Datenbank zu erleichtern, werden die Vollständigkeitswerte je Distanzmaß und je Vollständigkeitsmaß gemeinsam dargestellt. Da die Ergebnisse der MIDI-Datenbank die Referenz darstellen, wird weiterhin aus den zugehörigen Werten das Ergebnis für den Wert $N_{DB} = 1000$ extrapoliert. Die verwendete Modellfunktion wird in Abschnitt 8.2.3 beschrieben.

Die Ergebnisse sind für das Vollständigkeitsmaß V_1 in Abbildung 8.7 dargestellt. Die gemittelten Kurven für die MIDI-, Zufalls- und die Markovdatenbank haben alle einen qualitativ ähnlichen Verlauf. Die Vollständigkeit der Anfrage fällt immer langsamer für wachsende Datenbankbestände. Die N-Gramm-Methoden SF und CM erzielen bei der Zufallsdatenbank eine bessere Übereinstimmung mit den MIDI-Daten, sonst ist die Übereinstimmung von Markov- und Zufallsdatenbank mit den Referenzdaten etwa gleichwertig. Das extrapolierte Ergebnis der Modellfunktion ist außer bei UK und LCE zu hoch.

In Abbildung 8.8 werden die Vollständigkeitswerte V_3 je Distanzmaß miteinander verglichen. Man sieht, dass die Markov-Melodien das Verhalten der MIDI-Datenbank insgesamt etwas besser annähert als die Zufallsmelodien. Wieder ist das extrapolierte Ergebnis der Modellfunktion für den Wert 1000 mit Ausnahme von UK und LCE etwas zu hoch.

Für die Vollständigkeitswerte V_{10} in Abbildung 8.9 ist die Übereinstimmung von Markov-Melodiedatenbank und MIDI-Datenbank im Vergleich zu den Ergebnissen zu V_3 und V_1 am besten. Der extrapolierte Wert für $N_{DB} = 1000$ liegt nun für SF, CM, LAL und LCT zwischen den gemessenen Werten von Zufalls- und Markov-Melodiedatenbank. Das Markov-Modell eignet sich insgesamt gut zur Modellierung einer Melodiedatenbank, eine weitere Verfeinerung der Modellierung könnte z. B. durch die Wahl einer höheren Ordnung des Markov-Prozesses erfolgen.

Modellbildung

Um die Werte der MIDI-Melodiedatenbank extrapolieren zu können, soll eine Modellfunktion der Vollständigkeitswerte für die gemessenen Daten gebildet werden. Da diese Funktion offensichtlich nichtlinear ist, insbesondere für Werte von V_{10} , wird folgender Ansatz gewählt:

$$V_{\text{Modell}}(n) = \begin{cases} 1 & \text{für } V_{\text{MIDI}} = 1 \\ c_1 n^{c_2} & \text{für } V_{\text{MIDI}} < 1 \end{cases} \quad (8.8)$$

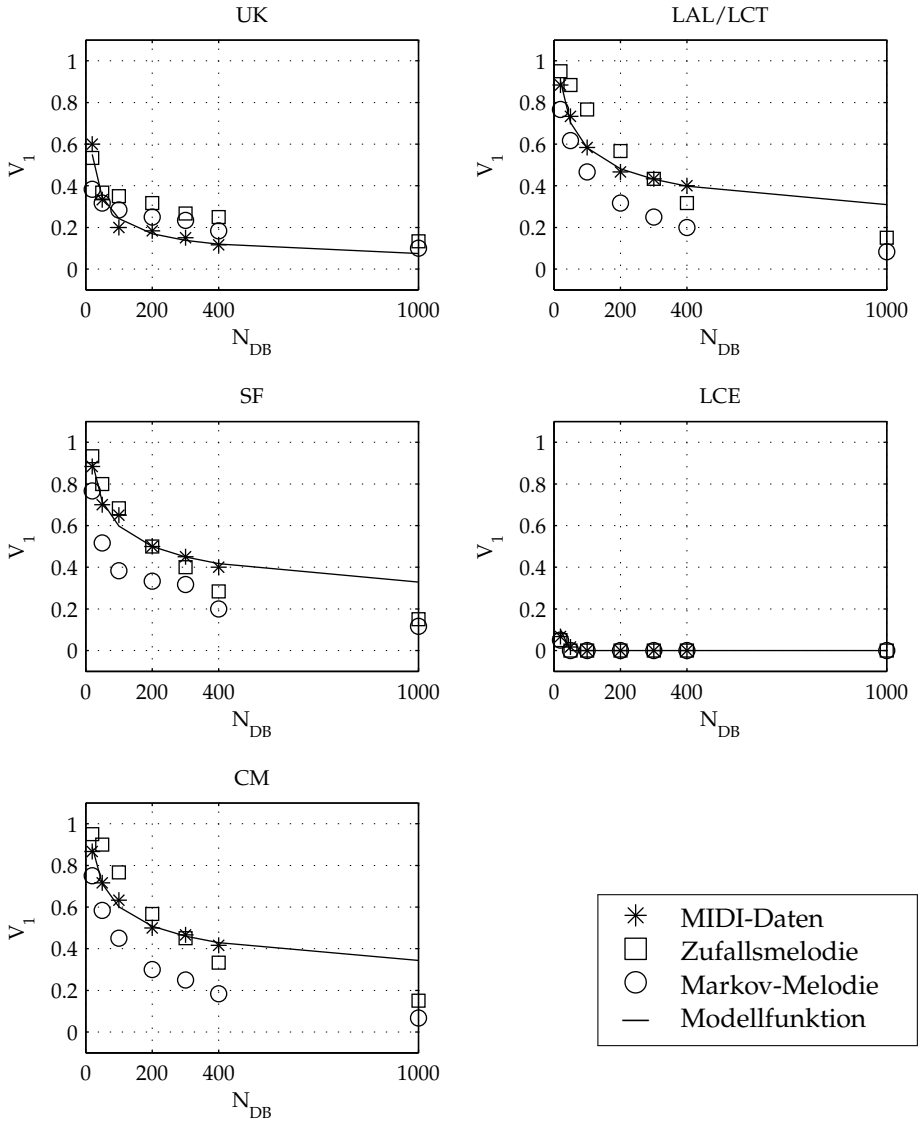


Abbildung 8.7: Ergebnisse für V_1 , sortiert nach Distanzmaßen für die drei untersuchten Melodiedatenbanken.

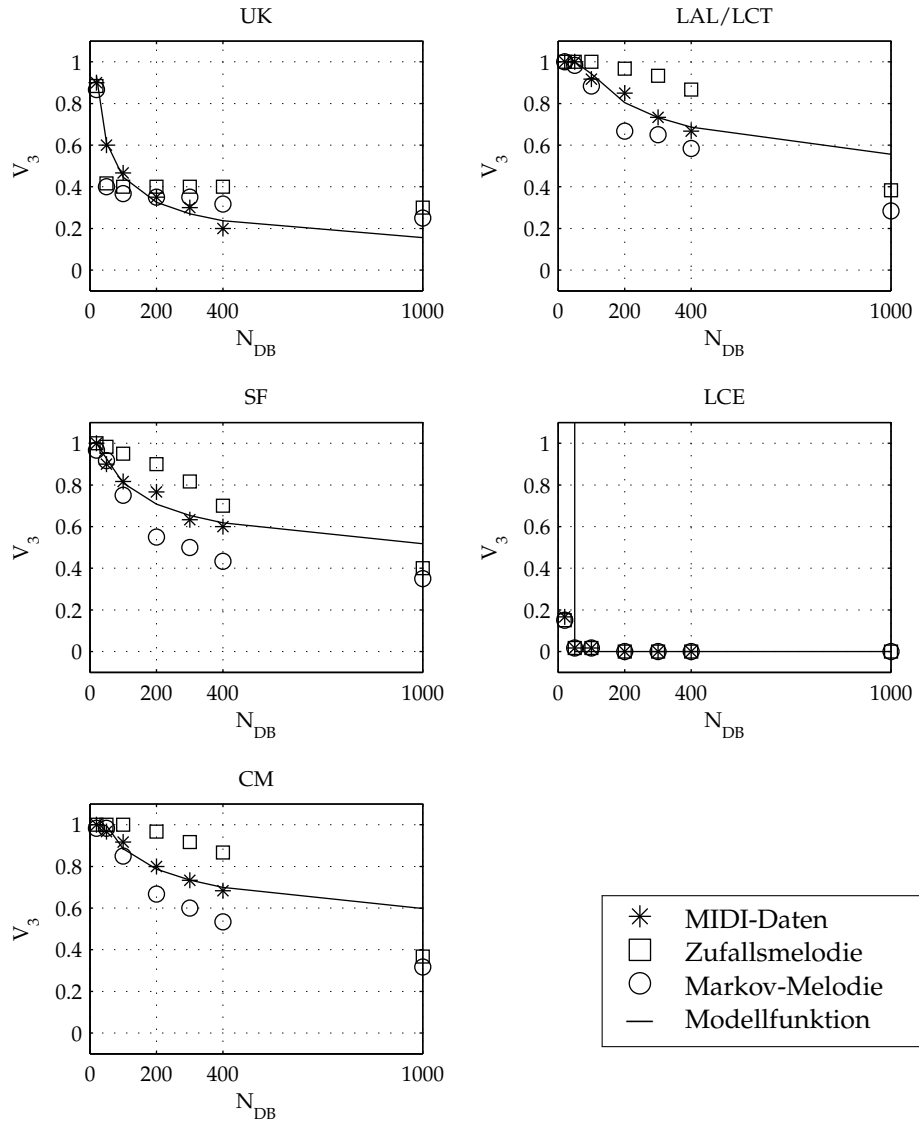


Abbildung 8.8: Ergebnisse für V_3 , sortiert nach Distanzmaßen für die drei untersuchten Melodiedatenbanken.

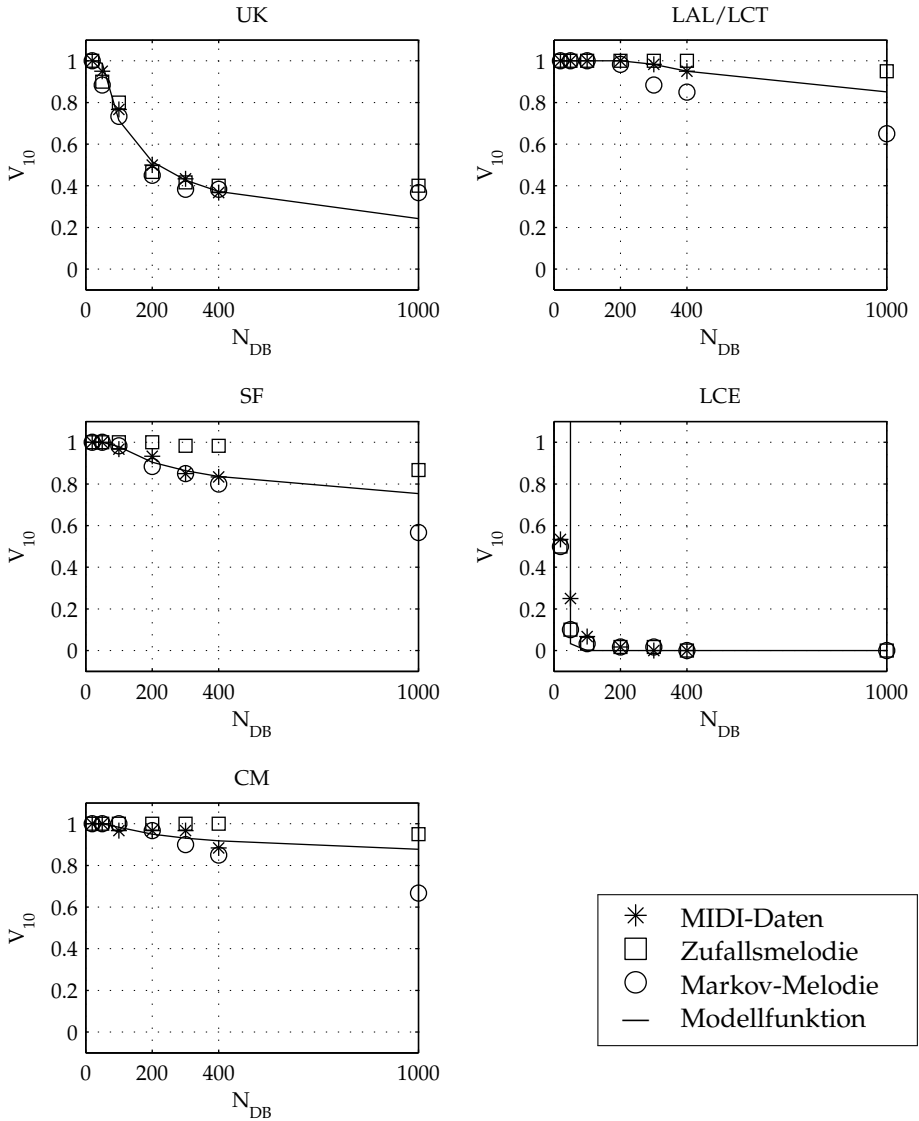


Abbildung 8.9: Ergebnisse für V_{10} , sortiert nach Distanzmaßen für die drei untersuchten Melodiedatenbanken.

mit den Konstanten c_1 und c_2 . Diese Konstanten werden als den Vollständigkeitswerten kleiner eins unter Minimierung des mittleren quadratischen Fehlers berechnet.

In den vorangegangenen Untersuchungen ist Gleichung (8.8) zur Extrapolation der Ergebnisse der MIDI-Daten verwendet worden. Die Güte der Modellierung der Ergebniskurve hängt von der Anzahl der verwendeten Stützstellen ab, die für den Fall $V_{\text{MIDI}} < 1$ zur Verfügung stehen.

8.3 Melodievergleich

Die Distanzmaße für den Melodievergleich lassen die Modifikation einiger Parameter zu; bei der Indizierung mit N-Grammen kann die Länge N variiert werden, bei den Verfahren zur Zeichenkettensuche können die Gewichte der Edieroperationen verändert werden. Für beide Vergleichstechniken ist die Normierung des ermittelten Ähnlichkeitswertes auf die Länge der untersuchten Symbolfolge von erheblichem Einfluss.

8.3.1 Indizierung

In diesem Versuch wird die N-Gramm-Länge variiert und im Zusammenspiel mit den verschiedenen Normierungen (siehe Abschnitt 7.5.2) untersucht. Für den Wertebereich findet man in der Literatur sehr verschiedene N-Gramm-Längen; von DANNENBERG und HU werden die Längen 1–4 untersucht [55], die Untersuchungen von UITDENBOGERD berücksichtigen Werte von 3–8 [186, 187]. DOWNIE untersucht N-Gramme mit 4, 5 und 6 Werten [60].

Bei dem Vergleich von Methoden der DP und N-Gramm-Techniken in [55] wird als beste N-Gramm-Länge $N = 3$ ermittelt, sehr schlechte Ergebnisse ergeben sich für $N = 1, 2$. Insgesamt wird die N-Gramm-Methode als sehr schwach gegenüber anderen Ähnlichkeitsmaßen dargestellt. In [186] liefern N-Gramme die besten Ergebnisse für $N = 5$ und $N = 7$.

Zur Ermittlung der optimalen N-Gramm-Länge werden wie in Abschnitt 8.2.2 60 Titel der MIDI-Datenbank herangezogen, darunter alle Titel der Top-10 aus Tabelle 8.1. Für die Anfrage werden jeweils 10, 15 und 20 Noten verwendet und die N-Gramm-Länge variiert. Es sollen die günstigste N-Gramm-Länge und die günstigste Normierung ermittelt werden. Es werden perfekte, d. h. fehlerfreie Anfragen und Anfragen mit zufällig eingefügten Fehlern verwendet. Für die folgenden Untersuchungen werden je ein oder zwei Einfügings-

Auslassungs- und Edierungsfehler generiert. Diese Fehlerkategorien werden in Abschnitt 8.4.2 weiter untersucht. Durch die Wahl dieser Parameter soll eine durchschnittliche Anfrage an ein QBH-System nachempfunden werden.

In Abbildung 8.10 sind die Ergebnisse, die sich mit diesen Anfragen erzielen lassen, in Vollständigkeitswerten V_{10} für die N-Gramm-Längen 3–8 dargestellt, die Ergebnisse sind das Ergebnis der Mittelung über alle Anfragetitel. Es werden alle in Abschnitt 7.5.2 dargestellten Normierungen untersucht: die Länge (len) der Melodie, der Logarithmus der Länge (log), 2-te und 9-te Wurzel (2rt, 9rt) und keine Normierung (non).

Je nach N-Grammlänge und Distanzmaß ist die optimale Normierung verschieden. Für die UK-Messung ist sehr leicht zu erkennen, dass lediglich mit der Normierung auf die Länge gute Ergebnisse (len) zu erzielen sind. Für SF ergeben sich über alle N-Gramm-Längen hinweg die besten Ergebnisse mit der Normierung auf die Länge, aber auch mit Wurzel 2 (2rt). Für CM ist die Normierung auf den Logarithmus, die zweite Wurzel (2rt) und die 9-te Wurzel (9rt) günstig. Um je Distanzmaß für alle N-Grammlängen eine geeignete Normierung zu finden, werden daher nun die Ergebnisse aller N-Gramm-Längen gemittelt.

Abbildung 8.11 zeigt die durchschnittliche Vollständigkeit über alle betrachteten N-Gramm-Längen. Optimal sind für UK und SF die Normierung auf die Länge, für CM ist der Logarithmus günstiger, dicht gefolgt von der Normierung auf die 9-te Wurzel.

Mit den ermittelten optimalen Normierungen soll nun untersucht werden, wie die Distanzmaße von der Länge der Melodieeingabe bzw. von Fehlern in der Melodie abhängen. Abbildung 8.12 zeigt die Vollständigkeitswerte V_{10} für Anfragen von 10, 15, und 20 Noten Länge, jeweils für fehlerfreie und fehlerhafte Eingaben. Die fehlerhaften Eingaben wurden mit jeweils ein bzw. zwei Auslassungs-, Einfügungs- und Edierungsfehlern gewählt. Für alle Ergebnisse gilt, dass mit steigender Anfragelänge das Ergebnis besser ausfällt. Bei fehlerfreien Anfragen sind mit Ausnahme von UK für die Melodielänge 10 große N-Gramm-Werte besser als kleine, wobei die Länge der Melodie eine natürliche Obergrenze darstellt. Aus Recheneffizienzgründen sind allerdings kleinere N-Gramm-Längen zu bevorzugen. Für wachsende Fehler in der Melodieanfrage zeigt sich, dass die optimale N-Gramm-Länge im Bereich kleiner 8 zu suchen ist. Um für die untersuchten Melodielängen 10, 15 und 20 die optimale N-Gramm-Länge festzustellen, werden daher nun die Ergebnisse der drei Fehlerszenarien gemittelt.

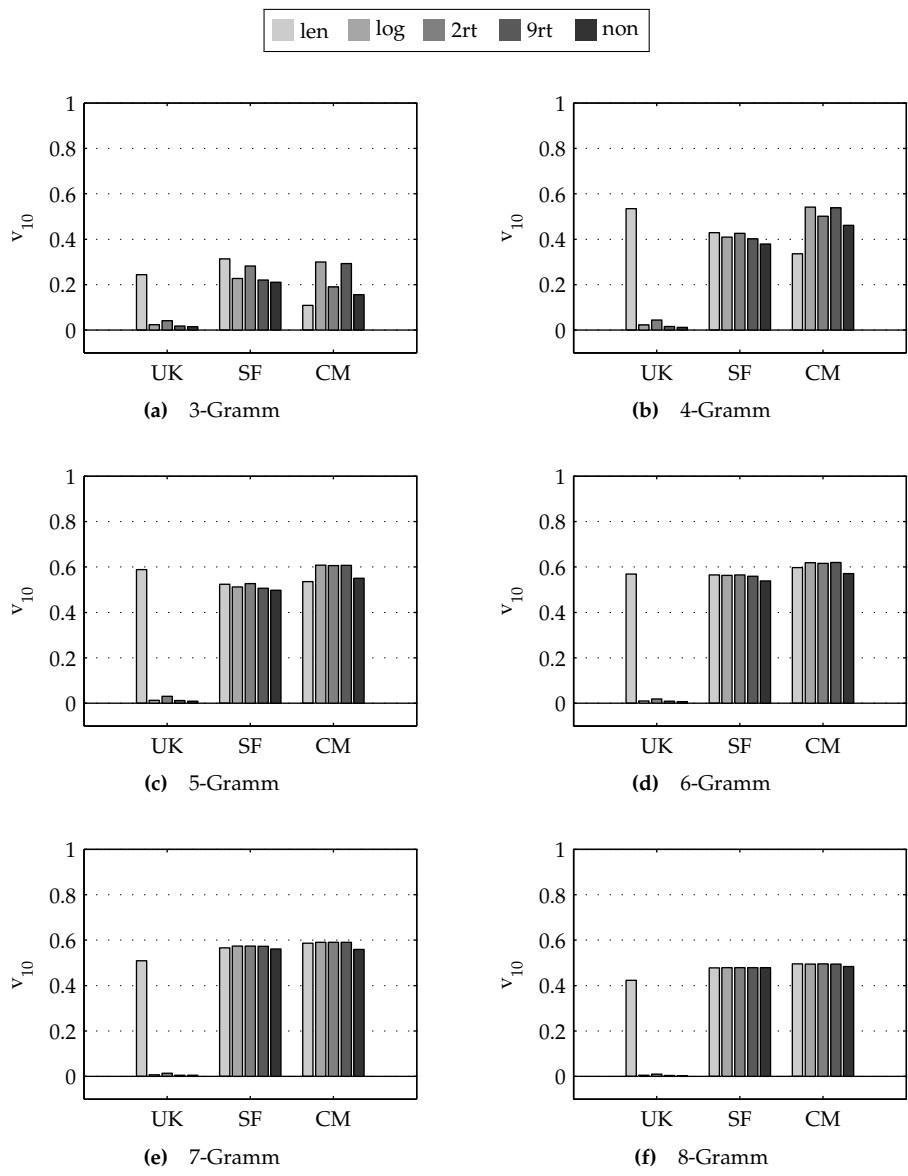


Abbildung 8.10: Vollständigkeit V_{10} der Top-10-Suche für verschiedene N-Gramm-Längen.

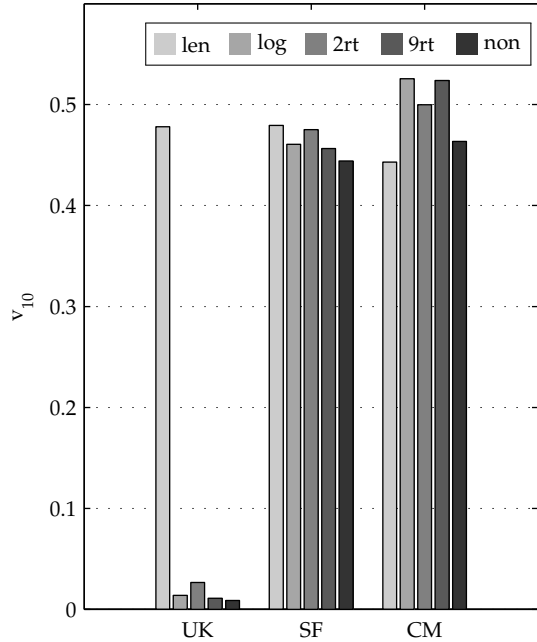


Abbildung 8.11: Die durchschnittliche Vollständigkeit V_{10} über alle betrachteten N-Gramm-Längen. Optimal sind für UK und SF die Normierung auf die Länge, für CM ist der Logarithmus günstiger.

Abbildung 8.13 zeigt den Mittelwert der Vollständigkeits über die in Abbildung 8.12 dargestellten Fehlerszenarien. Die optimale N-Gramm-Länge ist bei allen drei Distanzmaßen $N = 6$. Die Vollständigkeit V_{10} beträgt 0,5–0,6, genauso wie für die Ergebnisse mit $N = 5$ und $N = 7$. Damit werden die Ergebnisse von UITDENBOGERD [186] bestätigt. Die negative Bewertung der N-Gramm-Methode von DANNENBERG und HU in [55] wird ebenfalls offensichtlich – dort wurden nicht hinreichend große N-Gramme gewählt.

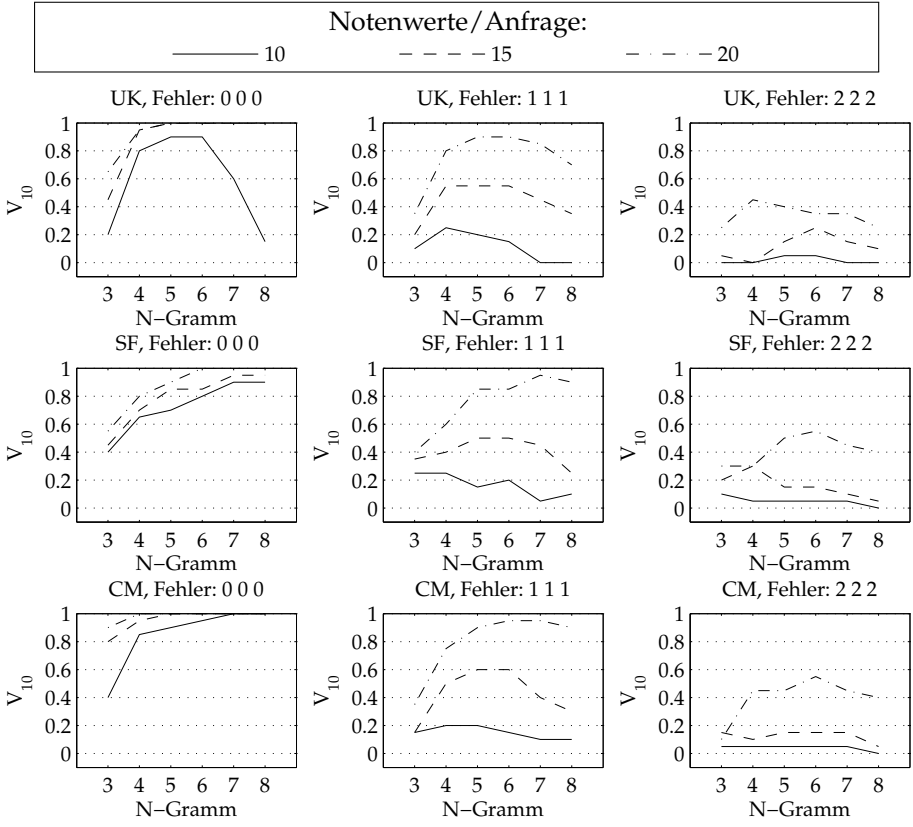


Abbildung 8.12: Die Vollständigkeit V_{10} in Abhängigkeit von der Länge der Anfrage unter dem Einfluss von Melodiefehlern (Einfügung/Auslassung/Edierung).

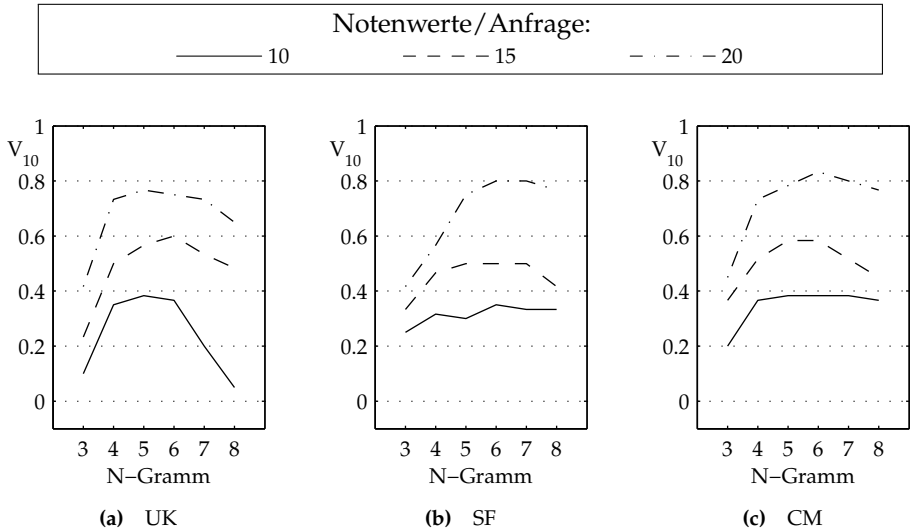


Abbildung 8.13: Mittelwert der Vollständigkeits über die in Abbildung 8.12 dargestellten Fehlerszenarien. Die optimale N-Gramm-Länge ist bei allen drei Distanzmaßen 6.

8.3.2 Zeichenkettensuche

Nun sollen die Verfahren der Zeichenkettensuche „lokaler Abgleich“ (LAL), „längste gemeinsame Teilsequenz“ (LCE) und „längste gemeinsame Zeichenkette“ (LCT) untersucht werden. Als wichtiger Parameter ist auch hier besonders die Normierung der errechneten Ähnlichkeiten zu betrachten, die starken Einfluss auf die Güte des Suchergebnisses hat.

Untersucht werden wieder 60 Titel als Anfragen mit den Top-10 aus Tabelle 8.1, diesmal begrenzt auf 10 bzw. 20 Noten. Zur Normierung der Ähnlichkeitswerte werden wie im vorigen Abschnitt die Varianten „len“, „log“, „2rt“, „9rt“ und „non“ verwendet. Als Fehler werden wie im vorigen Abschnitt Einfügings-, Auslassungs- und Edierungsfehler betrachtet.

Als Parameter der Distanzmaße der dynamischen Programmierung sind die Edierkosten möglich: d für Einfügung oder Löschung, e für Treffer und m , falls keine Übereinstimmung besteht. Die verwendeten Parametersätze WS1–WS3 sind in Tabelle 8.2 angegeben. Abbildung 8.14 zeigt die zugehörigen Ergebnisse. Die Variation der Ergebnisse durch die Edierkosten ist nicht besonders groß, in den weiteren Untersuchungen wird das in der Literatur übliche Set WS1 verwendet. Weiterhin kann man den Ergebnissen in Abbildung 8.14 entnehmen, dass als Normierung die 9-te Wurzel für alle Distanzmaße günstig ist.

Tabelle 8.2:
für die Distanzma-
ße der dynamischen
Programmierung

Parametersatz	d	e	m
WS1	-2	1	-1
WS2	-2	1	-2
WS3	-1	1	-1

Abbildung 8.15 zeigt den Einfluss von Fehlern in der Melodieanfrage. Während LAL und LCT bei fehlerfreien Anfragen gleich gut arbeiten (siehe Abbildung 8.15a), werden bei fehlerhaften Anfragen (Abbildung 8.15b) mit LAL bessere Vollständigkeitswerte bei langen Anfragen erzielt, bei den kurzen Anfragen ist LCT überlegen. LCE liefert keine brauchbaren Ergebnisse.

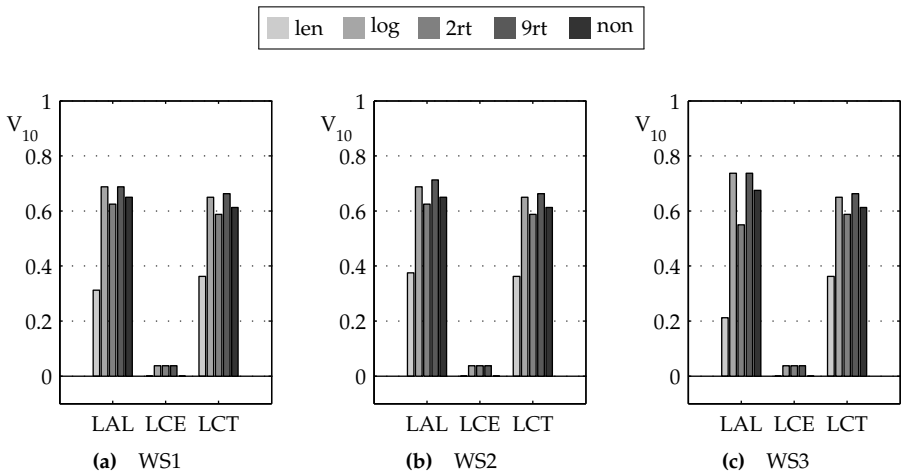


Abbildung 8.14: Vollständigkeit für Distanzmaße der Zeichenkettensuche vs. Normierung mit verschiedenen Parametersätzen für die Edierkosten.

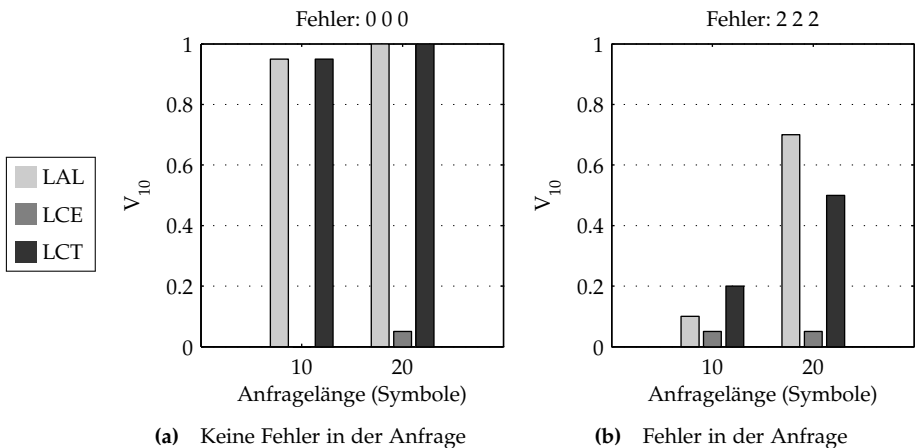


Abbildung 8.15: Vollständigkeit V_{10} vs. Anfragelänge bei Melodiefehlern. Fehler in der Anfrage führen bei kurzen Anfragen zu unbrauchbaren Ergebnissen. Für die Anfragelänge 20 erweist sich LAL als robuster gegenüber LCT.

8.4 Anfragefehler

Nachdem mit den Untersuchungen zu der Melodiedatenbank sowie den Distanzmaßen der „äußere“ Rahmen für die Melodiesuche abgesteckt ist, wird nun die Suchanfrage selbst in den Mittelpunkt der Beobachtung gestellt. Interessant ist neben der Frage, wie ausführlich, also wie lang eine Melodiesuchanfrage sein muss, der Einfluss der in der Suchanfrage enthaltenen Fehler. Um Fehler durch den Nutzer auszuschließen, werden zunächst Fehler in definierter Anzahl in die MIDI-Melodiedarstellung eingefügt und aus dieser eine fehlerhafte Suchanfrage generiert.

Fehler, die bei der Transkription von Melodien für die Datenbank entstehen können, wurden bislang in der Literatur nicht untersucht. Bisher werden nur Verfahren verwendet, die im Prinzip fehlerlos funktionieren wie die Extraktion von Melodien aus MIDI-Daten oder die als fehlerlos angenommen werden wie die Extraktion per Hand. Die Extraktion von Melodien durch Mehrfachgrundfrequenzanalyseverfahren, wie in Kapitel 6 beschrieben, wird bislang nicht verwendet. Prinzipiell sind Fehler zu erwarten, wie sie bei der Transkription der gesumten Anfrage entstehen.

8.4.1 Anfragelänge

In dem nun beschriebenen Experiment wird die Länge der Anfrage variiert. Wie bereits in Abschnitt 2.4 dargestellt, ist es die Tendenz der Nutzer eines QBH-Systems, nur kurze Anfragen zu stellen.

DOWNIE untersucht sehr kurze Anfragen mit 6, 8 und 10 Noten [60], die aus vorliegenden Melodien extrahiert werden. UITDENBOGERD untersucht Anfragelängen von 10–80 Noten [187]. Für QBH-System sind besonders kurze Anfragelängen von Interesse, da ein langer Gesangsvortrag nur von geübten Sängern zu erwarten sind. In Untersuchungen von UITDENBOGERD wird eine durchschnittliche Anfragelänge von 7 Konturwerten festgestellt [188]. Es soll nun die Güte der Suchergebnisse für verschiedene Anfragelängen untersucht werden. Als untere Grenze werden 6 Noten gewählt, damit für Suchmethoden der Indizierungstechnik ein 6-Gramm gebildet werden kann. Als Obergrenze wird 30 gewählt, was bei einer Melodie mit ausschließlich 8-tel-Noten bei Tempo MM = 60 gut 14s entspricht. Diese Dauer wurde als Durchschnittswert für Anfragen in den Versuchen von LESAFFRE et al. ermittelt [120] (siehe auch Abschnitt 2.4).

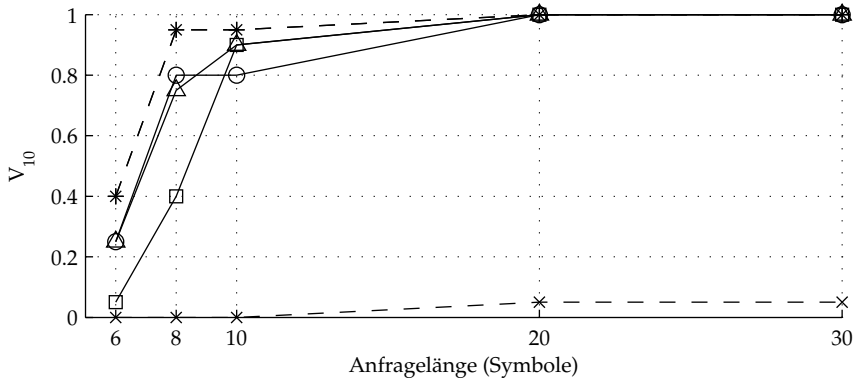


Abbildung 8.16: Vollständigkeit der untersuchten Suchanfragen über verschiedene Anfragelängen. Befriedigende Ergebnisse werden erzielt, wenn wenigstens 10 Noten in der Anfrage enthalten sind.

In Abbildung 8.16 ist die Vollständigkeit über verschiedene Anfragelängen dargestellt. Es werden die (fehlerfreien) Melodiekonturen der Top-10 aus Tabelle 8.1 verwendet und nach N Symbolen abgebrochen. Die Vollständigkeit V_{10} wird über die 10 Titel gemittelt. Das Ergebnis verschlechtert sich drastisch, sobald die Anfrage kürzer als 10 Notenwerte ist. Die gegen kurze Anfragen robustesten Distanzmaße sind SF und LAL. Mit LCE lassen sich für die untersuchten Anfragelängen überhaupt keine Treffer erzielen.

8.4.2 Melodiefehler

Fehler in der Anfrage können durch den Nutzer selbst, aber auch durch die monophone Transkription des Signals der Nutzeranfrage entstehen. Die Fehler eines Nutzers beim Singen einer Melodie werden in Abschnitt 2.4 dargestellt. Fehler durch die Signalverarbeitung der Transkriptionsstufe werden in Abschnitt 5.4.3 diskutiert. Das Ähnlichkeitsmaß sollte gegen diese Fehlerquellen möglichst robust sein, denn beide dieser Fehlerquellen haben Auswirkungen auf die Melodiekontur.

Auslassung Auslassungsfehler können durch den Nutzer verursacht werden, indem zur Melodie gehörige Töne vergessen werden, oder durch die mo-

nophone Transkriptionsstufe, wenn gesummte Töne nicht erkannt werden.

Ändert sich das Intervall zur nachfolgenden Note durch die Auslassung nicht, dann entsteht nur ein Fehler, die Kontur wird um einen Wert kürzer; Beispiel: $a_1a_1a_1$ wird a_1a_1 , damit wird aus der zugehörigen Kontur 0 0 eine 0.

Ändert sich das Intervall zur nachfolgenden Note, dann sind zwei Fehler möglich, falls sich ein neuer Konturwert ergibt; Beispiel: $a_1b_1c_1$ wird a_1c_1 , damit wird aus der Kontur 1 1 eine 2.

Einfügung Auch Einfügungen können durch den Nutzer verursacht werden, indem ein zusätzlicher Füllton gesummt wird. Durch die monophone Transkription können Töne eingefügt werden, wenn die Tonhöhe derartig schwankend gesummt worden ist, dass sie vom Transkriptionsverfahren nicht mehr korrigiert werden kann.

Durch eine Einfügung wird die Kontur länger und es entsteht dadurch ein Fehler, falls der Vorgängerton gleich ist; Beispiel: a_1b_1 wird $a_1a_1b_1$, damit wird aus der Kontur 1 eine 0 1.

Zwei Fehler werden in der Kontur verursacht, falls die Vorgängernote verschieden von der eingefügten Note ist; aus a_1b_1 wird $a_1c_1b_1$, damit wird aus 1 die Kontur 2 -1.

Zusammenfassung Werden zwei gleiche Töne zusammengefasst, so wirkt sich dies wie eine Auslassung aus, da eine Tonwiederholung wegfällt und nun ein Konturwert fehlt.

Ersetzung Eine Notenersetzung kann nur durch den Nutzer verursacht werden, indem eine Note mangels besseren Wissens oder Vermögens falsch gesummt wird.

Ersetzungen verursachen in der Kontur keine Fehler, wenn die Intervalle nicht die vorgegebenen Grenzen überschreiten; Beispiel: aus der Tonfolge $a_1c_1a_2$ wird $a_1e_1a_2$, die Kontur bleibt aber bei 2 2. Je nach Intervalländerung und Folge der ursprünglichen Intervalle können aber auch ein bis zwei Fehler verursacht werden. Beispiele: ein Fehler entsteht, wenn $a_1c_1c_1$ zu $a_1d_1c_1$, geändert wird, damit wird aus 2 0 die Kontur 2 -1. Zwei Fehler entstehen, wenn $a_1a_1a_1$ zu $a_1b_1a_1$ wird, dann wird aus 0 0 die Kontur 1 -1.

Um die beschriebenen Fehler gezielt untersuchen zu können, werden die fehlerfreien Melodien der Top-10 aus Tabelle 8.1 nun durch die beschriebenen Fehlertypen gestört. Die Fehler werden zufällig in der MIDI-Sequenz eingefügt, die sich daraus ergebenden Melodiekonturen werden für die zu untersuchende Anfrage verwendet. Für alle Fehlerkategorien werden 0, 1, 2, ... 10 eingefügte Fehler untersucht. Als Ergebnis ist die Vollständigkeit V_{10} angegeben und es wird betrachtet, bei wieviel Fehlern wenigstens die Hälfte aller Anfragen gefunden wurde.

Abbildung 8.17 zeigt die Vollständigkeit der Anfrageergebnisse für eine wachsende Anzahl von Auslassungsfehlern. Die Vollständigkeit des Ergebnisses sinkt mit wachsender Fehlerrate. Ab 4 Fehlern werden 50 % aller Titel nicht mehr gefunden, am besten funktioniert LAL von den DP-Techniken und UK von den N-Grammen. LCE liefert gar keine Treffer.

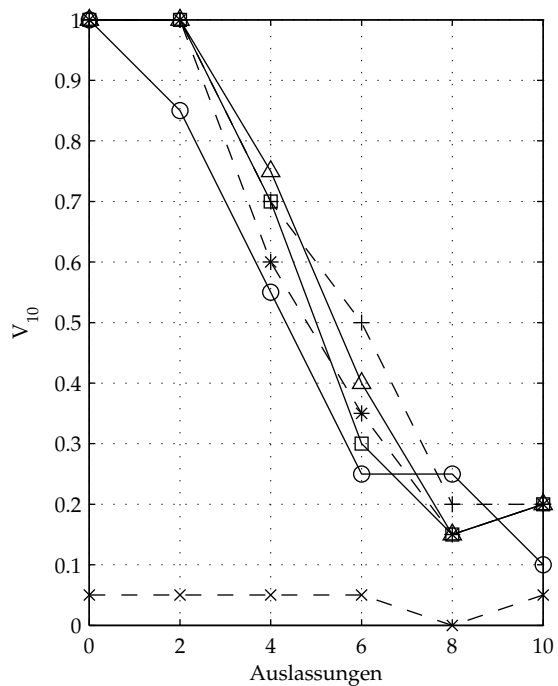
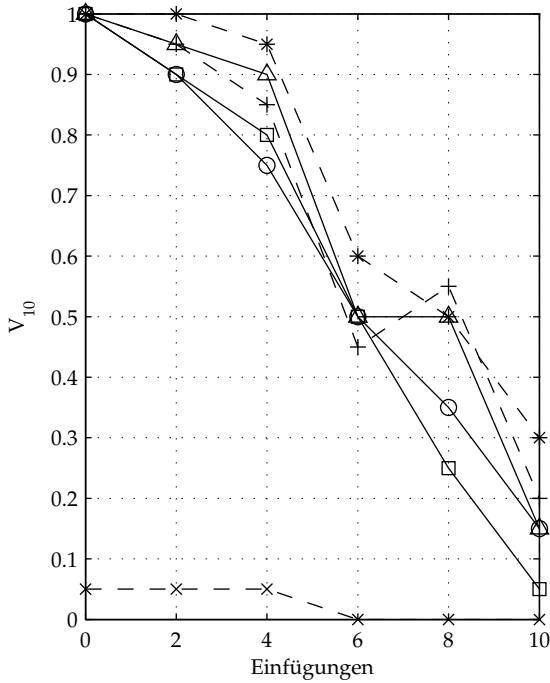


Abbildung 8.17: Auslassungen von Noten in der Melodie wirken sich vergleichsweise stark auf die Vollständigkeit des Suchergebnisses aus. Ab 4 Fehlern sind die Trefferquote insgesamt unter 50 %.

Abbildung 8.18 zeigt die Vollständigkeit bei Einfügungsfehlern. Die Fehlerquote steigt etwas weniger stark mit zunehmender Fehlerzahl, ab 6 Einfügungen wird für alle Distanzmaße keine Trefferquote über 50 % erzielt. Beste DP-Technik ist wieder LAL, bei den N-Gramm-Techniken liegen UK und CM etwa gleich. LCE versagt wieder völlig.

Abbildung 8.18:
Einfügungsfehler
sind etwas weniger
verheerend als Aus-
lassungsfehler. Ab
5 Fehlern wird bei
den meisten Distanz-
maßen nur noch die
Hälfte aller unter-
suchten Anfragen
gefunden.



Die Vollständigkeit bei Edierfehlern ist in Abbildung 8.19 dargestellt. Insgesamt liegt auch hier die Grenze zur 50 %-Fehlerquote bei 5 Fehlern, allerdings werden für das wieder am besten abschneidende LAL selbst für 10 Fehler noch gute Ergebnisse erzielt. Auch hier verhalten sich CM und UK ähnlich, wie es bei der Untersuchung der Einfügungsfehler schon der Fall war. LCE liefert erneut keine Treffer.

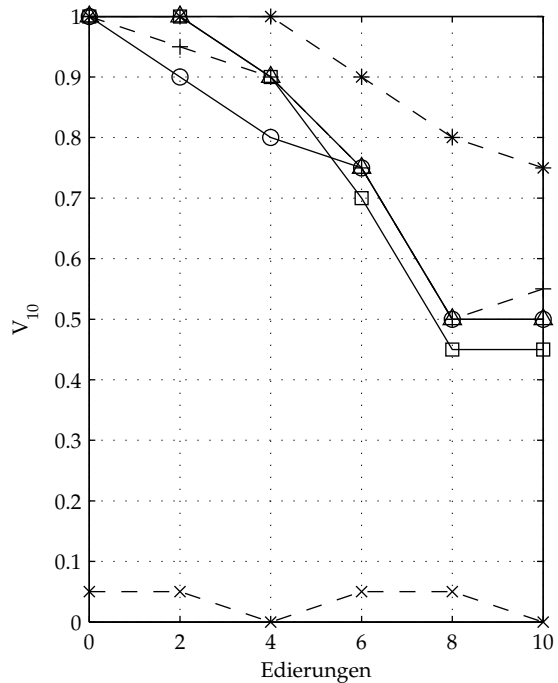
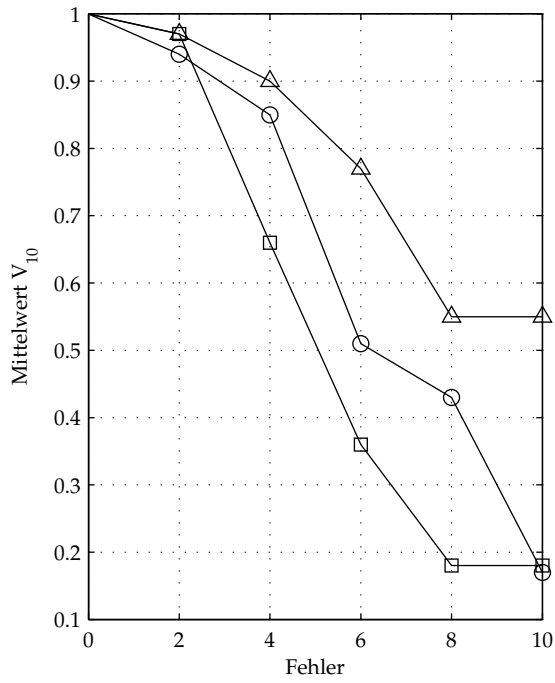


Abbildung 8.19: Edierfehler wirken sich am wenigsten fatal für den Sucherfolg aus. LAL findet auch bei 10 falschen Tönen noch 60 % aller gesuchten Titel.

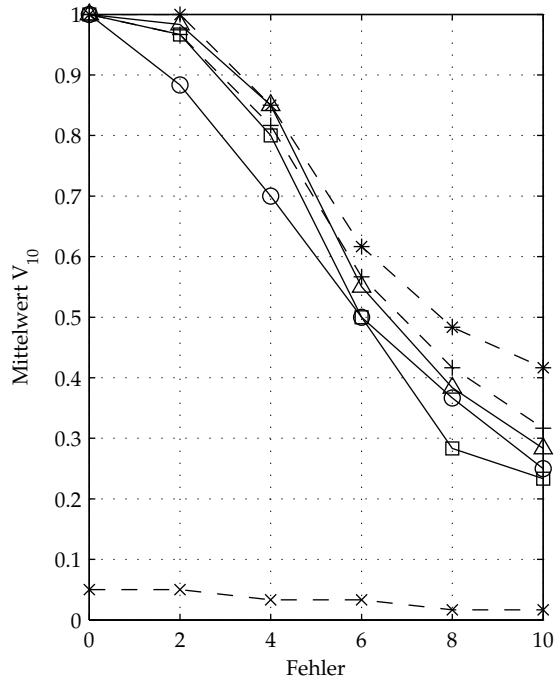
Abbildung 8.20 zeigt den Überblick über die diskutierten Fehlerkategorien Auslassung, Einfügung und Edierung. Die Vollständigkeitswerte sind in dieser Übersicht über alle Distanzmaße mit Ausnahme von LCE gemittelt. Edierungen mindern die Vollständigkeit am wenigsten, Einfügungen und Auslassungen wirken sich stärker aus.

Abbildung 8.20: Die Mittelwerte der Vollständigkeitswerte über alle Distanzmaße zeigen, dass Edierfehler am wenigsten kritisch sind, Auslassungen hingegen sich sehr schnell auf das Suchergebnis auswirken.



Um das fehlerrobusteste Distanzmaß auswählen zu können, werden nun alle Fehlerkategorien gemittelt und für alle untersuchten Distanzmaße dargestellt. Abbildung 8.21 zeigt, dass insgesamt LAL für fehlerbehaftete Anfragen am geeignetsten ist. Es folgen die Methoden LCT und CM.

Abbildung 8.21: Die Mittelwerte der Vollständigkeitswerte über die mittlere Anzahl der Fehler Auslassung, Einfügung und Edierung zeigen, dass LAL gegenüber den anderen Distanzmaßen überlegen ist.



Die Untersuchungen der Fehler in den Anfragen zeigen, dass die DP-Technik LAL für alle Fehlerkategorien am besten funktioniert, LCE bringt keine verwendbaren Ergebnisse. Die N-Gramm-Methoden UK und CM erweisen sich als ähnlich in ihrem Verhalten gegenüber Fehlern, sind aber insgesamt den Techniken LAL und LCT unterlegen.

8.4.3 Feldversuch

In den vorangegangenen Untersuchungen wurden Anfragen verwendet, die dem Bestand der Melodiedatenbank entnommen worden und den Bedürfnissen der Untersuchung entsprechend durch eingefügte Fehler modifiziert worden sind. Damit war die Kontrolle über Inhalt und Güte der Anfrage gegeben und es konnte das optimale Ähnlichkeitsmaß bestimmt werden. In diesem Abschnitt werden nun gesummte Anfragen verwendet. Um kontrollieren zu können, ob die gesummte Melodie gefunden worden ist, wurde der zu summende Titel vorgegeben. Damit ergibt sich folgendes Verfahren:

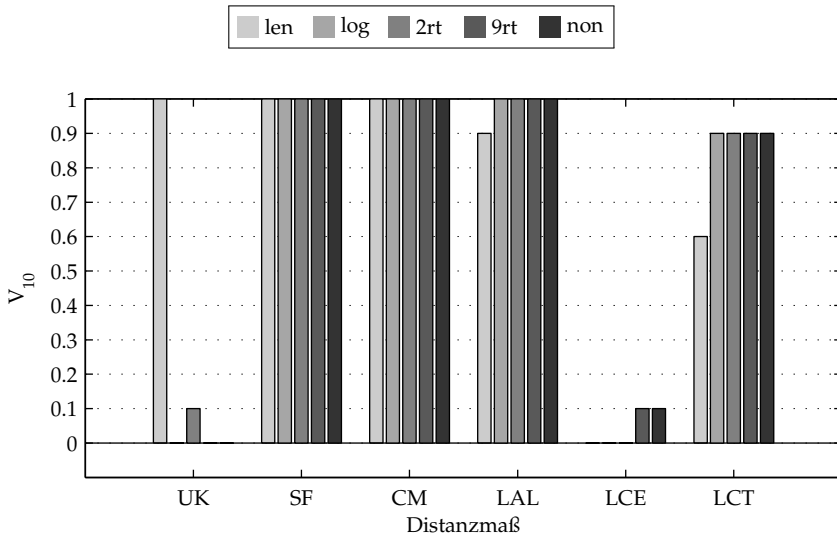
Training Der Teilnehmer der Untersuchung hört sich eine Darbietung der MIDI-Datei der Titel der Top-10 aus Tabelle 8.1 an; zusätzlich zum Vergleich steht die jeweilige Mp3-Version der Titel zur Verfügung.

Anfrage Nun singt der Teilnehmer den Titel nach und zeichnet ihn auf. Dabei kann er auswählen, ob er auf „na“ oder „da“ singt, wie schnell und wie lange er singt. Wiederholungen der Aufnahme sind erlaubt.

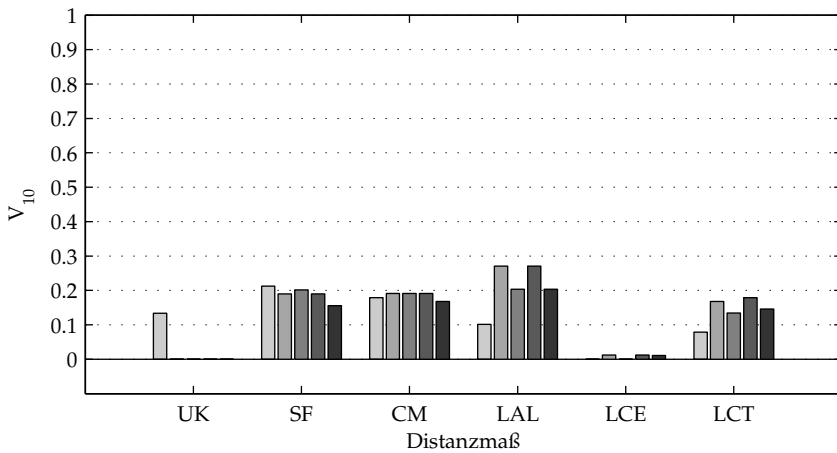
Beteiligt waren insgesamt 9 Probanden, davon drei weiblich und insgesamt fünf mit musikalischer Vorbildung. Die Fehler in der Anfrage können nun nicht mehr objektiv überprüft werden. Das Ergebnis einer genauen Prüfung auf Fehler beschrieb im Wesentlichen die Fähigkeit der Probanden zu singen, an dieser Stelle soll nur der Erfolg der Suchanfrage von Interesse sein.

Es wird nun untersucht, ob die vorher unter kontrollierten Bedingungen ermittelten optimalen Ähnlichkeitsmaße im Feldversuch tatsächlich wieder zu den besten Ergebnissen führen. Zur Kontrolle können die Ergebnisse herangezogen werden, die sich mit den MIDI-Versionen der zu singenden Top-10 ergeben.

Abbildung 8.22a zeigt die Vollständigkeitsmaße V_{10} für alle untersuchten Distanzmaße in Kombination mit allen Normierungen. Als N-Gramm-Länge wurde $N = 6$ gewählt, für die Verfahren der Zeichenkettesuche wurde WS1 aus Tabelle 8.2 verwendet. Alle Distanzmaße bis auf LCE und LCT erzielten ein vollständiges Ergebnis. Demgegenüber zeigt Abbildung 8.22b den Mittelwert aller Vollständigkeitsmaße für die Ergebnisse der Probanden. Die Vollständigkeit fällt auf ein Drittel ab, das höchste Vollständigkeitsmaß erzielt erwartungsgemäß LAL mit der Normierung 9rt bzw. log. Danach folgt die N-Gramm-Methode SF mit Normierung auf die Länge. Erst die nächstbesten Ergebnisse werden mit den Methoden Cm und LCT erreicht. Wie bei den Voruntersuchungen erzielt LCE keine verwertbaren Ergebnisse.



(a) MIDI



(b) Probanden

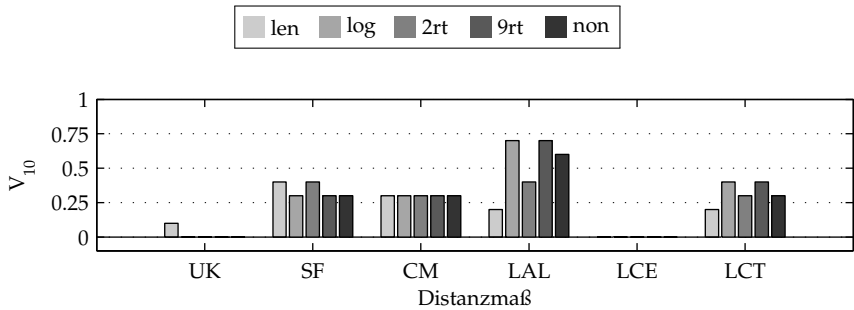
Abbildung 8.22: Die Ergebnisse für die MIDI-Eingabe und die durchschnittlichen Ergebnisse der Probanden für alle untersuchten Distanzmaße.

In Abbildung 8.23 sind die Vollständigkeitswerte für die vier erfolgreichsten der teilnehmenden Probanden dargestellt. Die höchsten Vollständigkeitswerte erzielten Proband 1 und 3, während Proband 2 besonders schlecht abschnitt. Bemerkenswert ist, dass bei Proband 1 der Unterschied zwischen den beiden erfolgreichsten Distanzmaßen LAL/9rt und SF/len relativ ausgeprägt ist, während bei Proband 3 der Unterschied geringer ausfällt. Bei dem erfolglosen Probanden 2 ist ebenso wie bei allen weiteren, nicht dargestellten Probanden abweichend vom Durchschnitt das Distanzmaß UK/len besonders gut.

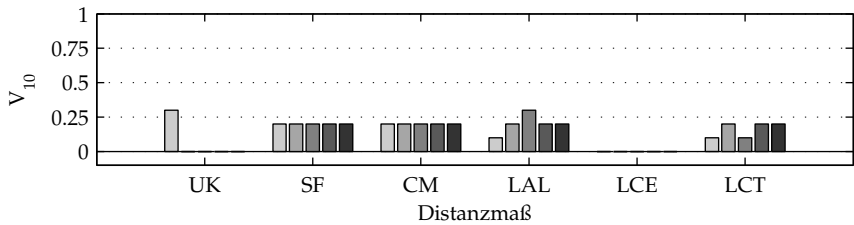
Aus diesen Betrachtungen lässt sich folgern, dass das in den Voruntersuchungen ermittelte beste Distanzmaß LAL/9rt tatsächlich am besten funktioniert. Es muss allerdings einschränkend festgestellt werden, dass abhängig von den Melodieanfragen bzw. dem Nutzer eines QBH-Systems Unterschiede im Sucherfolg möglich sind. Prinzipiell muss bei durch Summen eingegebenen Suchanfragen unter Verwendung der im Rahmen dieser Arbeit vorgestellten Transkriptionsmethoden mit einer hohen Fehlerquote gerechnet werden.

8.5 Zusammenfassung

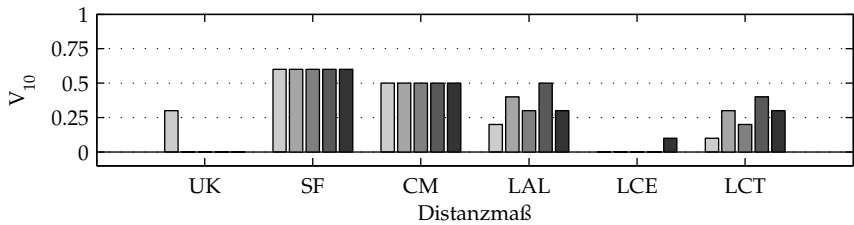
QBH-Systeme können durch die Angabe von Vollständigkeit und Präzision des Ergebnisses beurteilt werden, die dazu notwendigen Definitionen wurden in Abschnitt 8.1 angegeben. Bei den Untersuchungen des QBH-Systems *Queryhammer* wurde eine Melodiedatenbank herangezogen, welche die Top-10-Single-Charts vom März 2003 enthält; diese und alle weiteren Melodiekonturen wurden aus MIDI-Dateien extrahiert. Um eine von einem begrenzten Vorrat an MIDI-Dateien unabhängige Untersuchung von Melodiedatenbanken zu ermöglichen, wurden die statistischen Parameter Verteilungsdichte der Melodielängen, der Melodiekontursymbole und die Zustandsübergangsmatrix der Melodiesymbole der bestehenden MIDI-Datenbank gemessen. Mithilfe dieser Daten wurde eine neue Melodiedatenbank mit entsprechenden Zufallsprozessen generiert und ausführlich untersucht. Es konnte festgestellt werden, dass durch Modellierung mit den beschriebenen Parametern die Suche in Melodiesuchsystemen erfolgreich simuliert werden kann. Durch die Betrachtung des Einflusses der einzelnen statistischen Parameter wurde ein leistungsfähiges Werkzeug zur Untersuchung von Melodiedatenbanken entwickelt, das auch die Generierung sehr großer Datenbankbestände ermöglicht. Das verwendete Markov-Modell bietet die Möglichkeit zur Verfeinerung durch die Wahl höherer Ordnungen.



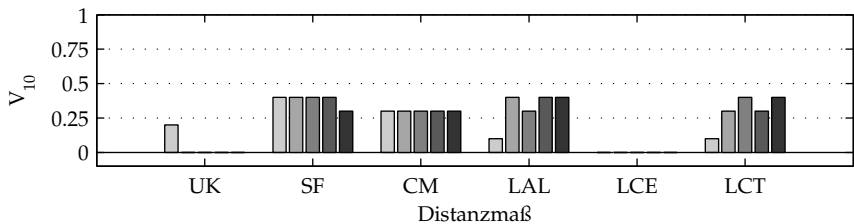
(a) Proband 1



(b) Proband 2



(c) Proband 3



(d) Proband 4

Abbildung 8.23: Die Ergebnisse der einzelnen Probanden.

Danach wurden die optimalen Parameter für die Ähnlichkeitsmaße mit Methoden der Indizierung und dynamischen Programmierung ermittelt. Für N-Gramm-Methoden erwies sich die Länge $N = 6$ als optimal. Weiterhin wurde für alle Ähnlichkeitsmaße die optimale Normierung bzgl. der Melodielänge experimentell festgestellt. Für die „Ukkonen-Messung“ (UK) und die „Summe der Häufigkeiten“ (SF) ist die Normierung auf die Länge (len) am günstigsten, für den „Koordinaten-Vergleich“ (CM) die 2-te Wurzel (2rt). Bei den Verfahren der Zeichenkettesuche „lokaler Abgleich“ (LAL), „längste gemeinsame Teilsequenz“ (LCE) und „längste gemeinsame Zeichenkette“ (LCT) ist in allen Fällen die 9-te Wurzel günstig.

Untersuchungen des Umfangs der Suchanfrage ergaben, dass wenigstens eine Melodielänge von 10 Noten vorgetragen werden sollte. Es wurde danach die Auswirkung von Fehlern in der Anfrage betrachtet und in Verbindung mit den einzelnen Komponenten des QBH-Systems gebracht.

Das größte Problem stellen Auslassungsfehler dar – sie werden durch die Signalverarbeitung der monophonen Transkriptionstufe verursacht, sofern diese mit der Eingabe des Nutzers überfordert ist. Dies kommt durch zu kurze oder undeutlich gesungene Noten zustande. Auslassungen führen zu einer starken Beeinträchtigung des Sucherfolgs. Einfügungen von Noten sind ebenfalls ein Problem für die Güte des Suchergebnisses. Sie entstehen durch Fehler der monophonen Transkription, die unsicher intonierte Töne in zwei oder mehr Noten teilt. Edierungen der Melodie erfolgen nur durch den Nutzer, sind aber auch am wenigsten kritisch für das Ergebnis.

*Was man sucht, findet man
immer zuletzt.*

anonym

Query-by-Humming-Systeme (QBH-Systeme) sind Musiksuchsysteme, die eine gesummte Melodie in eine symbolische Darstellung umwandeln und darüber in einer Melodiedatenbank nach einer ähnlichen Melodie suchen. In dieser Arbeit wurde die grundlegende Analyse und Bewertung der Funktionalität von QBH-Systemen untersucht. Praktische Untersuchungen wurden an dem im Rahmen dieser Arbeit erstellten Beispielsystem *Queryhammer* durchgeführt. Um eine möglichst allgemeingültige Aussage zu erhalten, sollten Analyse und Bewertung unabhängig von der Eingabe des Nutzers oder dem Inhalt der Melodiedatenbank erfolgen. Für die interne Darstellung der Melodie wurde die standardisierte Repräsentation MPEG-7-MelodyContour gewählt.

Die Beschreibung und das Verständnis der Funktionsweise eines QBH-Systems berührt verschiedene Disziplinen wie Musikwissenschaften, Elektrotechnik (speziell Signalverarbeitung) und Informatik. Anhand der Darlegung der wichtigsten Grundlagen der genannten Disziplinen wurden Fehlerquellen und Analysemöglichkeiten von QBH-Systemen in den einzelnen Kapiteln dargelegt. Diese werden nun abschließend zusammengefasst, im Ausblick werden künftige Entwicklungsmöglichkeiten aufgezeigt.

9.1 Zusammenfassung

In Kapitel 2 wurden einige Grundlagen der Musiktheorie und speziell der Melodiebegriff dargestellt, um das Verständnis und die Beschreibung für das Gesuchte zu ermöglichen, nämlich Information zu einer Melodie. Es wurden verschiedene symbolische Repräsentationen für die Melodie und die in QBH-Systemen üblicherweise verwendete Melodiekontur dargestellt. Zu nennen sind hier vor allem der PARSONS-Code und die MPEG-7-MelodyContour, die dem im Rahmen dieser Arbeit implementierten Beispielsystem *Queryhammer*

zugrunde liegt. Ebenfalls zur Beschreibung von Musik geeignete Begriffe wurden erläutert, z. B. das Tempo, Akkordfolgen oder die Tonart. Damit wurde verdeutlicht, dass Melodien eine signifikante, aber nicht die einzige Möglichkeit sind, um Musik zu beschreiben.

Kapitel 3 beginnt mit einer Einführung in das Thema Suchmaschinen im Allgemeinen und der Darstellung von Anforderungen an Musik- und Melodiesuchmaschinen im Speziellen. Mittels ausgewählter Beispiele bestehender Melodie- und Musiksuchmaschinen wurde der Stand der Technik dargestellt und es wurden kennzeichnende Merkmale solcher Systeme erarbeitet. Zu nennen sind die verschiedenen möglichen Anfrageformen (z. B. akustisch wie bei QBH-Systemen, aber auch in Form einer Stichprobe oder eines Beispiels, als Text oder in Notenform) und Datenbankaspekte (Art des Inhalts, Größe des Bestands, Format des Inhalts). Aus diesen Merkmalen wurden die Anforderungskriterien, verschiedene Nutzergruppen, Betriebsbedingungen und Funktionen von QBH-Systemen hergeleitet. Insbesondere wurden die Anforderungen an das Beispielsystem *Queryhammer* spezifiziert. Das System wurde in Funktionsblöcke mit den Aufgaben *monophone Transkription*, *polyphone Transkription* und *Melodievergleich* unterteilt. Diese Blöcke wurden in den entsprechenden Kapiteln ausführlich behandelt.

Die Diskussion verschiedener Multimediasstandards war Gegenstand des Kapitels 4. Besonders interessant für Anwendungen wie QBH-Systeme ist der Multimedia-Beschreibungsstandard MPEG-7, der in diesem Bezug ausführlich dargestellt wurde. Er stellt eine Vielzahl von Deskriptoren zur Inhaltsbeschreibung von Audio-, Bild- und Videodaten zur Verfügung. Zur symbolischen Melodiekonturbeschreibung ist der Deskriptor MPEG-7-MelodyContour vorgesehen, Definition und Verwendung wurden ausführlich erläutert. Weitere (nicht symbolische) Deskriptoren aus MPEG-7 beziehen sich auf die Beschreibung von Signalen durch Energie-, Grundfrequenzverlauf u. ä. und eignen sich daher als weitere Schnittstellenbeschreibung für QBH-Systeme. Durch die standardisierte Darstellung ermöglichen sie den Zugriff auf QBH-Systeme in Multimedia-Netzwerken. Innerhalb solcher Netzwerke sind in Bezug auf QBH-Systeme weiterhin Standards wie MPEG-21 und SMIL interessant; sie wurden daher am Ende des Kapitels 4 kurz vorgestellt.

Die Transkription der gesamten Anfrage wurde in Kapitel 5 behandelt. Diese Aufgabe wurde aufgeteilt in die Tonhöhenerkennung und Rhythmuserkennung. Es wurden relevante Verfahren für Tonhöhen- und Rhythmuserkennung vorgestellt und diskutiert. Die Auswahl des Verfahrens für die Tonhöhenerkennung im Beispielsystem *Queryhammer* fiel auf die Autokorrelationsme-

thode. Sie ist robust gegen Störungen und eine hinreichend hohe Frequenzauflösung zur Transkription von Musik kann einfach erreicht werden. Die in *Queryhammer* implementierte Rhythmuserkennung arbeitet auf der Basis des Grundfrequenzverlaufs, der Ergebnis der Tonhöhenerkennung ist. Damit handelt es sich um ein typisches System zur Gesangstranskription, das mit bestehenden QBH-Systemen vergleichbar ist.

Bei den Untersuchungen von gesummten Nutzereingaben wurden drei Fehlerarten in der Transkription unterschieden: Einfügungen, Zusammenfügungen und Auslassungen von Noten. Die Ursachen dieser Fehler konnten durch ein neu vorgestelltes Prüfverfahren durch vergleichende Untersuchungen mittels synthetisch erzeugter Signale genau zugeordnet werden. Einfügingsfehler sind den Intonationsschwankungen der Nutzer beim Summen zuzuschreiben, während das Zusammenfügen und Auslassen von Noten auch durch das Transkriptionsverfahren bedingt sein kann. Die einzelnen Töne sind in diesem Fall zu kurz oder schwach artikuliert gesummt.

Die Analyse des Transkriptionsverfahrens durch vergleichende Untersuchung von Beispielnutzern und synthetisch generierten Signalen lässt eine systematische Bewertung der Transkriptionsfehler zu. Da als Referenz für den tatsächlichen Inhalt einer gesummten Melodie üblicherweise die Transkription eines Musikers herangezogen wird, ist nicht nachvollziehbar, wie weit in dieser Referenz durch den Musiker für ihn leicht nachvollziehbare Fehler bewusst oder ggf. unwillkürlich korrigiert worden sind. Durch die Untersuchung einer Nutzeranfrage, die synthetisch generiert worden ist, wird nun eine klare Fehlerbetrachtung möglich.

Die in Kapitel 6 untersuchte polyphone Transkription wird dann benötigt, wenn eine Melodie aus einem polyphonen Audiosignal extrahiert werden soll. Im Rahmen dieser Arbeit wurde untersucht, ob es unter Verwendung der polyphonen Transkription möglich ist, eine Melodiedatenbank aus den Audio-dateien einer Musikdatenbank zu extrahieren. Die Übersicht bestehender Verfahren zur Transkription von Noten aus polyphonen Audiosignalen stellt die wichtigsten aktuellen Verfahren vor. Das von Goto in [79] vorgestellte Verfahren *PreFEst* zur Mehrfachgrundfrequenzanalyse wurde als besonders geeignet identifiziert und für das Beispielsystem *Queryhammer* implementiert.

Der für die Grundfrequenzschätzung wichtige Schritt der Momentanfrequenzberechnung wurde durch aktuelle Verfahren ersetzt, weiterhin wurde die Detektionsmethode der Grundfrequenzen für die vorgegebene Aufgabe optimiert. Trotz dieser Verbesserungen zeigen eigene Untersuchungen, dass abhängig vom Audiosignal einige Melodien zwar gut extrahiert werden kön-

nen, das Verfahren bei anderen Musiksignalen aber völlig versagt. Eine automatische Erstellung von Melodiedatenbanken aus Audiodateien ist mit aktuellen Verfahren der Signalverarbeitung daher nicht möglich.

Kapitel 7 behandelt den Melodievergleich, der die Kernaufgabe eines QBH-Systems ist. Er erfordert die Bestimmung der Ähnlichkeit zweier Melodien bzw. Melodiekonturen, wozu verschiedene Ähnlichkeitsmaße herangezogen werden können. Es eignen sich verschiedene Verfahren, die im Rahmen dieser Arbeit erstmals systematisch für die fünfstufige Melodiekontur gemäß des Standards MPEG-7 untersucht wurden. Es wurden Verfahren der Zeichenkettenuche dargestellt, die mit Hilfe der dynamischen Programmierung berechnet werden können. Dies sind der „lokaler Abgleich“ (LAL), die Suche nach der „längsten gemeinsamen Teilsequenz“ (LCE) und der „längsten gemeinsamen Zeichenkette“ (LCT). Weiterhin können Indizierungstechniken unter Verwendung von N-Grammen benutzt werden, die ebenfalls eine Aussage über die Ähnlichkeit machen. Es wurden die Ähnlichkeitsmaße „Ukkonen-Messung“ (UK), „Summe der Häufigkeiten“ (SF) und „Koordinaten-Vergleich“ (CM) dargestellt. Das speziell für MPEG-7 entwickelte Ähnlichkeitsmaß TPBM-I konnte nicht verwendet werden, da die Extraktion rhythmischer Information aus gesungenen oder gesummen Melodien mit bestehenden Verfahren der Signalverarbeitung nur schlecht oder gar nicht möglich ist und damit die für TPBM-I notwendige Taktinformation fehlt.

Kapitel 8 beschäftigt sich mit der Untersuchung der Melodiedatenbank. Weiterhin wurden praktische Untersuchungen am Beispielsystem *Queryhammer* durchgeführt. Allgemein lässt sich die Güte des Ergebnisses einer Suchanfrage nicht ohne weiteres angeben. Suchergebnisse von QBH-Systemen können durch die Angabe von Vollständigkeit und Präzision beurteilt werden, die dazu notwendigen Definitionen wurden zu Beginn des Kapitels angegeben und erörtert.

Ein wichtiger Parameter bei der Untersuchung einer Suchmaschine ist der Umfang der Datenbank. Üblicherweise werden bei den Untersuchungen von QBH-Systemen in der Literatur Melodiedatenbanken herangezogen, die aus einem gewissen Vorrat von MIDI-Dateien extrahiert oder sogar durch manuelle Transkription erstellt werden.

Um unabhängig von existierenden MIDI-Datensammlungen den Einfluss der Größe einer Melodiedatenbank untersuchen zu können, wurde im Rahmen dieser Arbeit eine synthetische Melodiedatenbank generiert, die mit einer bereits vorhandenen Melodiedatenbank aus MIDI-Daten verglichen wurde. Durch Messung der Verteilungsdichten der Melodielängen und der Kontur-

symbole konnte bereits eine Melodiedatenbank mit Zufallszahlen modelliert werden, deren Verhalten für Suchanfragen der bestehenden MIDI-Melodiedatenbank ungefähr entspricht. Eine Verbesserung war durch die Modellierung der Zufallszahlen mittels eines Markov-Prozesses erster Ordnung möglich. Durch vergleichende Untersuchungen mit MIDI-, Zufalls- und Markov-Datenbank konnte gezeigt werden, dass das Suchergebnis bei der Betrachtung der zehn besten Treffer gute Übereinstimmung von echten Daten und Zufallsdaten ergibt. Die Modellierung von Melodiedatenbanken mit Zufallszahlen ermöglicht den Test von QBH-Systemen für sehr große Datenbestände, wie sie bei realen Systemen vorkommen.

Die Parameter der verwendeten Ähnlichkeitsmaße haben großen Einfluss auf den Sucherfolg. In den weiteren Untersuchungen wurden die optimalen Parameter für Ähnlichkeitsmaße unter Verwendung von Methoden der Indizierung (N-Gramm-Techniken) und dynamischen Programmierung (Zeichenkettenvergleich) ermittelt, da diese in der Literatur bislang nicht speziell für die Konturdarstellung nach MPEG-7 behandelt wurden. Für N-Gramme erwies sich die N-Gramm-Länge $N = 6$ als optimal. Für die „Ukkonen-Messung“ (UK) und die „Summe der Häufigkeiten“ (SF) ist die Normierung auf die Länge (len) am günstigsten, für den „Koordinaten-Vergleich“ (CM) die 2-te Wurzel ($2rt$). Bei den Verfahren der Zeichenkettesuche LCE, LCT und LAL ist in allen Fällen die 9-te Wurzel günstig. Diese Ergebnisse stimmen für MIDI- wie Zufallsdatenbank überein.

Schließlich erfolgte die Untersuchung des Einflusses der Suchanfrage selbst auf das Suchergebnis. Untersuchungen der Länge der Suchanfrage ergaben, dass eine Melodie mit wenigstens 10 Noten vorgetragen werden sollte. Es wurde danach die Auswirkung von Fehlern in der Anfrage betrachtet und in Verbindung mit den einzelnen Komponenten des QBH-Systems gebracht. Das größte Problem für eine erfolgreiche Suche stellen Auslassungsfehler bedingt durch den Nutzer oder die monophone Transkription dar, diese führen zu einer starken Beeinträchtigung des Sucherfolgs. Einfügungen von Noten sind ebenfalls ein Problem für die Güte des Suchergebnisses. Einfügungen entstehen entweder durch den Nutzer oder durch Fehler der monophonen Transkription, die unsicher intonierte Töne in zwei oder mehr Noten teilt. Edierungen der Melodie erfolgen nur durch den Nutzer, sind aber auch am wenigsten kritisch für das Ergebnis.

Damit wurde eine Differenzierung von Fehlern vorgenommen und in Bezug auf Suchanfragen an QBH-Systeme systematisch untersucht. Durch die Konturdarstellung wirken sich Edierungsfehler am wenigsten aus, Auslassungs-

und Einfügingsfehler dagegen sehr stark. Diese Tatsache legt nahe, andere Melodiedarstellungen als die Konturdarstellung zu verwenden, da diese nicht ausreichend robust gegen alle Fehlerkategorien ist. Mögliche Lösungen werden abschließend im Ausblick diskutiert.

9.2 Ausblick

Für zukünftige Entwicklungen von QBH-Systemen können resultierend aus den Ergebnissen dieser Arbeit einige Konsequenzen angegeben werden.

- Der Begriff der Melodie ist sehr vielschichtig, eine Musikbeschreibung ist, wie erläutert wurde, auch auf anderen Wegen möglich. Wie und wie gut eine Melodie beschrieben werden kann, ist sehr stark vom Nutzer abhängig. Die Suche nach einem Musikstück über die Melodie allein stellt in vielen Fällen eine zu starke Einschränkung dar. Daher sollten künftige Musiksuchsysteme möglichst viele und sinnvoll kombinierbare Beschreibungsformen (also auch Text, Noten, Bilder) für die Suchanfrage bieten. Eine sog. multimodale Anfrage bietet durch sinnvolle Kombination mehrerer Merkmale eine bessere Anpassung an die Möglichkeiten des Nutzers und ermöglicht damit eine Steigerung des Sucherfolgs.
- Bisher werden Musiksuchmaschinen meist über eine Internetseite angeboten, die vom Nutzer interaktiv zu bedienen ist. Künftig denkbar sind jedoch auch Dienste wie *Google-Alert* für die Textsuche [6]. Der Dienst *Google-Alert* erlaubt das Speichern bestimmter Textsuchanfragen und durchsucht über die Textsuchmaschine Google unter Verwendung dieser Anfragen regelmäßig das Internet. Werden neue Suchergebnisse erzielt, wird der Nutzer darüber per E-Mail informiert.

Um einen Melodiesuchdienst im Sinne von *Google-Alert* anbieten zu können, wäre es notwendig, die Melodiesuchanfrage in einer bestimmten Form festzulegen und sie dem Dienst zu übergeben. Damit der Dienst verschiedene Musikdatenbanken abfragen kann, ist eine möglichst normierte Schnittstelle zur Beschreibung der Suchanfrage notwendig, mit der viele Musikdatenbanken abgefragt werden können. Der Multimedia-Standard MPEG-7 zeigt für dieses Anwendungsszenario wegweisende Aspekte, da er für verschiedene Medieninhalte geeignete Deskriptoren bereithält. Über die Melodiebeschreibung hinaus wären so auch multimodale Anfragen möglich. Beispielsweise könnten neben einer extrahier-

ten Melodiekontur weitere Beschreibungen wie das Bild eines Künstlers verwendet und durch MPEG-7-Deskriptoren standardisiert beschrieben werden.

- Die Transkription des Eingangssignals bei QBH-Systemen ist eine potentielle Fehlerquelle für die Melodieanfrage. Künftige Techniken sollten Modelle für typische Nutzerfehler bzgl. Intonationschwankungen, unklare Silben- und Notenanfänge usw. enthalten, um möglichst nutzerunabhängig eine robuste Transkription zu ermöglichen.
- Die Transkription von Melodien aus polyphonen Signalen ist bislang nur für einen sehr kleinen Teil von Musiksignalen möglich und muss weiter entwickelt werden. Diese Technik ist vor allem für bereits existierende Audioarchive interessant. Bei neuen Musikproduktionen ist es möglich, Metainformationen wie die Melodie und weitere Informationen mit abzuspeichern, da die Daten in der Regel explizit vorhanden sind. Neue Datenträgerformate bieten genügend Möglichkeiten, um solche Information unterzubringen.
- Für zukünftige QBH-Systeme sollte eine möglichst fehlerrobuste Melodierepräsentation verwendet werden. Die im Rahmen dieser Arbeit verwendete Konturdarstellung ist lediglich gegen Edierungsfehler des Nutzers robust, nicht aber gegen Auslassungen und Einfügungen. Hier ist die zusätzliche Auswertung von rhythmischen Informationen angezeigt. Die von der MPEG-7-MelodyContour ebenfalls enthaltene *Beat*-Kontur stellt in diesem Sinne bereits eine Verbesserung dar, ist aber wie gezeigt wurde für QBH-Systeme nicht zu handhaben. Eine mögliche Lösung des Problems besteht in einer Melodiedarstellung, die alle messbaren Parameter einer gesumten Suchanfrage darstellt. So kann zum Beispiel für alle erkannten Notenereignisse der Einsatzzeitpunkt und die absolute Dauer der Noten ermittelt werden, während auf die Taktinformation verzichtet wird. Zur Bewertung einer derartigen Melodiedarstellung werden neue Distanzmaße notwendig. Diese sollten nach Möglichkeit parametrisierbar sein, um eine Anpassung auf bestimmte Eigenschaften, ggf. bedingt durch Musik oder Nutzerverhalten, vornehmen zu können.
- Die Untersuchung an der synthetischen Melodiedatenbank mit Zufallszahlen hatten gezeigt, dass bisherige Verfahren unabhängig vom Inhalt der Datenbank sind. Für künftige Suchverfahren sind daher ggf. Erwei-

terungen des Modells notwendig, an denen sich Parameter zur Modellierung des Inhalts ablesen lassen. Durch einen Abgleich von tatsächlichen Melodien mit einem Modell aus Zufallszahlen ist eine Analyse und Optimierung künftiger Verfahren möglich.

Mit den im Rahmen dieser Arbeit vorgenommenen Analysen wurde eine Beurteilung von QBH-Systemen unabhängig von Nutzer und Musik angestrebt. Besonders mit der Untersuchung von synthetischen Datenbanken wurde eine Grundlage zur neutralen Beurteilung von Melodiesuchsystemen geschaffen. Ebenfalls wird durch diese Vorgehensweise die Analyse wichtiger statistischer Parameter einer Musikdatenbank möglich. Die Melodiesuche durch Summen ist in musikalischer Hinsicht sehr interessant – praktisch wichtig sind aber auch multimodale Systeme, die mehr Möglichkeiten zur Melodiebeschreibung bieten, etwa durch Text-, Genre- oder Bildbeschreibungen.

Titel der Melodiedatenbank

A

1000reasons.mid
500miles.mid
9to5.mid
achybreakyheart.mid
africa.mid
africa.mid
aintnomountainhigh.mid
allbymyself.mid
allihavetodoisdream.mid
alloutoflove.mid
alohao.mid
aloverfriend.mid
alwaysfaithful.mid
alwaysonmymind.mid
alwsonmm.mid
america.mid
americanpie.mid
angelinthemorning.mid
areyoulonesometonight.mid
astearsgoby.mid
atimeforus.mid
autumncomesandgoes.mid
autumnmapleleaf.mid
ave maria.mid
awayinamanger.mid
badboys.mid
badmoonrising.mid
bananaboatsong.mid
bandofgold.mid
barbiegirl.mid
beatit.mid
beautifulsunday.mid
beautybeast.mid
betchabygollywow.mid
beverlyhillscop.mid
bicyclebuiltfortwo.mid
biggirlsdontcry.mid
birdsandthebees.mid

bizare.mid
blackandwhite.mid
blackorwhite.mid
blessing.mid
blowinginthewind.mid
bluedanube.mid
blueonblue.mid
bornfree.mid
bothsidesnow.mid
brahms requiem 4th mvt.mid
brahms requiem 5th mvt.mid
britney_spears-
 baby_one_more_time.mid
butterflydream.mid
buttonsandbow.mid
calendargirl.mid
californiagirls.mid
candida.mid
candleinthewind.mid
candyman.mid
cantbuymelove.mid
canthelpfallinginlove.mid
cantsmilewithoutyou.mid
cantstandmissingyou.mid
canttakemyeyesoffofyou.mid
cathysclown.mid
cecelia.mid
celebration.mid
changingpartners.mid
chapeloflove.mid
chariotsoffire.mid
cherish.mid
cherrypinkapple-
 blossomwhite.mid
chimefrommonastery.mid
choices.mid
circleoflife.mid
climbeverymountain.mid

colorfulcarnival.mid
colorsofwind.mid
comealittlebitcloser.mid
comeandgetit.mid
confidenceinme.mid
countryroads.mid
cracklinrose.mid
crystalbluepersuasion.mid
cupid.mid
daddyshome.mid
dayafterday.mid
daydreambeliever.mid
dayswithoutlover.mid
delilah.mid
desperado.mid
devotedtoyou.mid
didntiblowyourmind.mid
dilemma.mid
discovery.mid
dixie.mid
dontbecruel.mid
dontcryformeargentina.mid
dontgo.mid
dontknowhowlovehim.mid
dontstandsoclosetome.mid
dontworrybehappy.mid
doremi.mid
downonthecorner.mid
downtown.mid
downunder.mid
doyouhearwhatihear.mid
duck.mid
dustinthewind.mid
easiersaidthandone.mid
ebonyandivory.mid
eidelweiss.mid
eightdaysaweek.mid
ein madchen oder weib-

chen.mid	howmuchdoggie.mid	lightmyfire.mid
elcondorpasa.mid	ialwaysloveyou.mid	limborock.mid
eldiaque.mid	iamwoman.mid	listeningtothewind.mid
elpaso.mid	ibethere.mid	littleredridinghood.mid
endlesslove.mid	icantdan.mid	locomotion.mid
endofworld.mid	idliketoteachtheworldtosing.mid	lola.mid
entertainer.mid	igetaround.mid	lonelybull.mid
erestu.mid	igotyoubabe.mid	lonelygoatherd.mid
everybm.mid	iknewilovedyou.mid	lookinbackdoor.mid
everythingisbeautiful.mid	ineverpromisedrosegarden.mid	loveblossoms.mid
everytimeyogoaway.mid	inthesummertime.mid	loveiseternity.mid
feelings.mid	inyear2525.mid	loveletters.mid
feliznavidad.mid	isawherstanding.mid	lovedmedo.mid
fincount.mid	isawthreeships.mid	lover.mid
firstlover.mid	iswear.mid	loversshirt.mid
flowerdrum.mid	ithinkiloveyou.mid	loverstear.mid
flowerswouldblossom.mid	itsallcomingback.mid	lovesong1990.mid
followyouforlife.mid	itsmyparty.mid	loveyoumore.mid
foronceinmylife.mid	itsnotunusual.mid	loveyoutenthousandyears.mid
forthelongesttime.mid	itsnowornever.mid	lovumore.mid
frommetoyou.mid	itssoeasytofallinlove.mid	mamasaid.mid
fruitfullife.mid	ivegottagetamessagetoyou.mid	mandarindream.mid
funkytwn.mid	iwalktheline.mid	maniac.mid
fur elise.mid	iwannaholdyourhand.mid	maria.mid
georgiegirl.mid	iwenttoyourwedding.mid	mariabnt.mid
girlonotherside3.mid	iwillfollowhim.mid	materialgirl.mid
gloryofbloodstain.mid	jambalaya.mid	mayflower.mid
godofwealtharrives1.mid	jasmine.mid	michelle.mid
goldentimes.mid	jesu joy of man's desiring.mid	midi_allbymyself.mid
greatestloveofall.mid	jinglebells.mid	midi_allout.mid
greengrassofhome.mid	jollyoldsaintnicholas.mid	midi_easyhurt.mid
greenislandserenade.mid	joytoworldjeremiah.mid	midi_iswear.mid
greensleeves.mid	justanoldfashionlovesong.mid	midi_lovelikewater.mid
guantanamera.mid	justonelook.mid	midi_moonriver.mid
halfmooncrescent.mid	killngmesoftly.mid	midi_nowandforever.mid
hangonsloopy.mid	kissgbye.mid	midi_savingallmyloveforu.mid
happytogether.mid	labamba.mid	midi_sayyoullbethere.mid
hardtoforget.mid	labamba.mid	midi_yesterday.mid
haveueverseenrain.mid	lacrymosa.mid	midi_yesterdayoncemore.mid
heaintheavy.mid	landdownunder.mid	midi_youhappy.mid
heartache.mid	leanonme.mid	minuet in g.mid
heyjude.mid	leavinonajetplane.mid	mondaymonday.mid
heytherelonelygirl.mid	legendofrejuvenate.mid	moneyfor.mid
highgreenmountains.mid	lemontree.mid	moongonghung.mid
holiday.mid	letitbe.mid	moonrepresentsmyheart.mid
hollyjollychristmas.mid	letmebethere.mid	morethanicansay.mid
hotelcal.mid	letsgoforabreeze.mid	moreword.mid

mostbeautifulgirl.mid
 mrsandman.mid
 myangelbaby.mid
 myfavoritethings.mid
 mygirl.mid
 myheartasnoreturnpath.mid
 mylife.mid
 neverfallinloveagain.mid
 neveronsundays.mid
 nicetobewithyou.mid
 nightfog.mid
 nightfragrance.mid
 nomatterwhat.mid
 nothingsgonnachange
 myloveforyou.mid
 nothingsgonnastopus.mid
 ocomeallyefaitful.mid
 ocomecomeemmanuel.mid
 odetojoy.mid
 ohbladi1.mid
 ohcarol1.mid
 oldflameslikedreams.mid
 olittletownofbethlehem.mid
 olivetree.mid
 onelaughatopensea.mid
 onetinsoldier.mid
 onlycareaboutyou.mid
 onlyifyouhaveme.mid
 onlyyou1.mid
 overtherainbow.mid
 particلمان1.mid
 physical.mid
 prettywoman.mid
 puffthefmagicedragon.mid
 puppetonastring.mid
 puppylove.mid
 putyourhand.mid
 putyourheadonmyshoulder1.mid
 quia fecit mihi magna.mid
 raindropskeepfalling.mid
 relax.mid
 rhythmofsea.mid
 rhythmoftherain.mid
 riverofdreams1.mid
 rockaroundclock.mid
 rosegarden.mid
 rosesarered.mid
 runaway.mid
 runawaytrain.mid
 santaclausiscoming.mid
 saturdayinthepark.mid
 savethelastdanceforme.mid
 sealedwithkiss.mid
 seasonsinthesun.mid
 separateinrainydays.mid
 shanghaibeach.mid
 shesatdistance.mid
 silentbonding.mid
 silverbells.mid
 sixteengoingseventeen.mid
 sixteentons.mid
 sohappytogether.mid
 solongfarewell.mid
 somekindoffeeling.mid
 somethinggood.mid
 somewheremylove.mid
 songsungblue.mid
 soundofmusic1.mid
 soundofmusic2.mid
 soundofsilence.mid
 soyoudontwantanything.mid
 spiritinthesky.mid
 standbyme.mid
 stayingalive.mid
 stillfriendsaftergoodbye.mid
 stoneinlove.mid
 storyofsmalltown.mid
 straycatstrut.mid
 strollingonpathoflife.mid
 sugershack.mid
 sukiyaki.mid
 surfinusa.mid
 swallowsseparate.mid
 sweetdreams.mid
 sweetie.mid
 takeabow.mid
 takeachanceonme.mid
 takecare.mid
 takemeballgame.mid
 tammy.mid
 teafortwo.mid
 teardropsofrain.mid
 tearsofacrown.mid
 telllaurailoveher.mid
 thatllbethe day.mid
 thecra1.mid
 thegreatpretender.mid
 thosewerethedays.mid
 thousandsofwords.mid
 tideishigh.mid
 timepassesbylikewater.mid
 tolovesomebody.mid
 top10_01_tatu-all
 _the_things_she_said.mid
 top10_02_scooter-
 weekend.mid
 top10_03_kate_ryan-
 desenchantee.mid
 top10_04_blue_feat.
 _elton_john-
 sorry_seems_be
 _the_hardest_word.mid
 top10_05_gareth_gates-
 anyone_of_us.mid
 top10_06_wolfsheim-
 kein_zurueck.mid
 top10_07_deutschland_sucht
 _den_superstar-
 weve_a_dream.mid
 top10_08_eminem-
 lose_yourself.mid
 top10_09_nena_and_friends-
 wunder_geschehen.mid
 top10_10_snap-
 rhythm_is_a_dancer_2003.mid
 topeofworld.mid
 tosirwithlove.mid
 trytoremember.mid
 turnturn.mid
 twistandshout.mid
 twoprinces.mid
 unabletoforgetyou.mid
 unchainedmelody.mid
 underthesea.mid
 unforgettablelove.mid
 unspokenseparation.mid
 uponthehousetop.mid
 upontheroof.mid
 uptowngirl.mid
 usingallmyheart.mid
 venus.mid

vulnerablewoman.mid	whydofools.mid	youandi.mid
walklikeaman.mid	windflowersnow.mid	youdontbringmeflowers.mid
waywewere.mid	windrain.mid	youlightupmylife.mid
wethreekings.mid	wipeout.mid	youmakemefeelbrandnew.mid
wewishyouamerry.mid	withlovefrommetoyou.mid	yourehappy.mid
whatafriendjesus.mid	withoutyou.mid	yourname.mid
whatawonderfulworld.mid	woodenheart.mid	youthminuet.mid
whatchildisthis.mid	words.mid	
wherehaveflowersgone.mid	yellowsubmarine.mid	
wholenewworld.mid	ymca.mid	

Literaturverzeichnis

- [1] *Encoders.* <http://www.sscnet.ucla.edu/geog/gessler/topics/encoders.htm>. Aufruf vom 11.3.2006. – Übersicht zu mechanischen Codierern
- [2] *mp3.com.* www.mp3.com/feature/aboutus. Aufruf vom 11.3.2006. – Informationsseite
- [3] *Andrang an der Musiktankstelle.* www.n24.de. Aufruf vom 11.6.2005. – Online-Artikel (N24.de, AP)
- [4] *Apple – iTunes – Musicstore.* www.apple.com/itunes/store. Aufruf vom 11.6.2005. – kommerzielles Musikportal
- [5] *mp3.de – musik im internet.* mp3.de. Aufruf vom 11.6.2005. – kommerzielles Musikportal
- [6] *Indigo Stream Technologies: GoogleAlert.* Aufruf vom 17.1.2006. – www.googlealert.com
- [7] *GUIDO Music Notation Format Page.* <http://www.informatik.tu-darmstadt.de/AFS/GUIDO/>. Aufruf vom 23.5.2005. – Notensatz-Software
- [8] *Harmony Central Homepage, MIDI Documentation.* www.harmony-central.com/MIDI/Doc/doc.html. Aufruf vom 23.5.2005
- [9] *LilyPond.* <http://lilypond.org/web/>. Aufruf vom 23.5.2005. – Notensatz-Software
- [10] *MIDI Manufacturers Association.* Aufruf vom 23.5.2005. – www.midi.org
- [11] *Musicline: Die ganze Musik im Internet.* Aufruf vom 23.5.2005. – QBH-System der phononet GmbH, www.musicline.de
- [12] *Musipedia, the Open Music Encyclopedia.* www.musipedia.org. Aufruf vom 23.5.2005

- [13] *Wikipedia, die freie Enzyklopädie.* de.wikipedia.org. Aufruf vom 23.5.2005
- [14] *notify!WhistleOnline.* www-mmdb.iai.uni-bonn.de/projects/nwo/index.html. Aufruf vom 25.5.2005. – Institut für Informatik III, Universität Bonn, Arbeitsgruppe Prof. Dr. Michael Clausen, Projektseite
- [15] *Suchmaschinen-Magazin @-web.* www.at-web.de. Aufruf vom 25.5.2005
- [16] *Vodafone-MusicFinder.* www.vodafone.de/dienste_kommunikation/infos_unterhaltung/56749.html. Aufruf vom 25.5.2005
- [17] *Apple iTunes Overview.* www.apple.com/itunes/overview. Aufruf vom 3.11.2006. – Informationsseite
- [18] *Wikipedia, the free encyclopedia.* en.wikipedia.org. last visit 05/23/2005
- [19] ABE, Toshihiko; KOBAYASHI, Takao ; IMAI, Satoshi: Harmonics Tracking and Pitch Extraction based on Instantaneous Frequency. In: *Proc. ICASSP*, 1995, S. 756–759
- [20] ABE, Toshihiko; KOBAYASHI, Takao ; IMAI, Satoshi: The IF Spectrogram: A New Spectral Representation. In: *Proc. ASVA*, 1997, S. 423–430
- [21] ALGHONIEMY, Masoud; TEWFIK, Ahmed H.: Rhythm and Periodicity Detection in Polyphonic Music. In: *Proc. IEEE third Workshop on Multimedia Signal Processing*, 1999, S. 185–190
- [22] ALONSO, M.; DAVID, B. ; RICHARD, G.: A Study of Tempo Tracking Algorithms from Polyphonic Music Signals. In: *4th COST 276 Workshop*, 2003
- [23] AUGER, Francois; FLANDRIN, Patrick: Improving the Readability of Time-Frequency and Time-Scale Representations by the Reassignment Method. In: *Trans. on Signal Processing* 43 (1995), Mai, Nr. 5, S. 1068–1089
- [24] AVARO, Olivier; SALEMBIER, Philippe: MPEG-7 Systems: Overview. In: *IEEE Transactions on Circuits and Systems for Video Technology* 11 (2001), Juni, Nr. 6, S. 760–764
- [25] BACH, Johann S.: *Partita A-Moll.* Musikverlag Zimmermann, 2000. – BWV 1013, für Flöte solo, Urtext (Werner Richter)

-
- [26] BAEZA-YATES, R.; PERLEBERG, C.: Fast and Practical Approximate String Matching. In: *Combinatorial Pattern Matching, Third Annual Symposium* (1992), S. 185–192
- [27] BATKE, Jan-Mark; EISENBERG, Gunnar ; WEISHAUPT, Philipp ; SIKORA, Thomas: Evaluation of Distance Measures for MPEG-7 Melody Contours. In: *International Workshop on Multimedia Signal Processing*. Siena, Italy, Oktober 2004
- [28] BATKE, Jan-Mark; EISENBERG, Gunnar ; WEISHAUPT, Philipp ; SIKORA, Thomas: A Query by Humming system using MPEG-7 Descriptors. In: *Proc. of the 116th AES Convention*. Berlin : AES, Mai 2004
- [29] BELLO, Juan P.; MONTI, Giuliano ; SANDLER, Mark: Techniques for Automatic Music Transcription. In: *International Symposium on Music Information Retrieval*, 2000
- [30] BELLO, Juan P.; SANDLER, Mark: Blackboard System and Top-Down Processing for the Transcription of simple Polyphonic Music. In: *Proceedings of the COST G-6 on DAFX-00*. Verona, Italy, Dezember 2000
- [31] BLACKBURN, Steven; DE ROURE, David: A Tool for Content Based navigation of Music. In: *ACM Multimedia 98*, 1998
- [32] BLACKBURN, Steven G.: *Content Based Retrieval and Navigation of Music Using Melodic Pitch Contours*, Univerity of Southampton, Diss., 2000
- [33] BOASHASH, Boualem: Estimating and Interpreting The Instantaneous Frequency of a Signal — Part 1: Fundamentals. In: *Proceedings of the IEEE* 80 (1992), April, Nr. 4, S. 520–538
- [34] BOERSMA, Paul: Accurate Short-Term Analysis of the Fundamental Frequency and the Harmonics-To-Noise Ratio of a Sampled Sound. In: *IFA Proceedings 17*, 1993
- [35] BREGMAN, Albert S.: *Auditory scene analysis: the perceptual organization of sound*. second. Cambridge : MIT Press, 1999
- [36] BROCHU, Eric; FREITAS, Nando de: Name That Song!: A Probabilistic Approach to Querying on Music and Text. In: *Neural Information Processing Systems*, 2002, S. 1505–1512

- [37] BRONSTED, Tom; AUGUSTENSEN, Soren ; FISKE, Brian ; HANSEN, Christian ; KLITGAARD, Jimmy ; NIELSEN, Lau W. ; RASMUSSEN, Thomas: A System for Recognition of Hummed Tunes. In: *Proc. of the COST G-6 Conf. on DAFX*, 2001
- [38] BROWN, Judith C.; PUCKETTE, Miller S.: Calculation of a Narrowed Autocorrelation Function. In: *J. Acoust. Soc. Am.* 85 (1989), April, Nr. 4, S. 1595–1601
- [39] BURBAT, Wolf: *Die Harmonik des Jazz*. 2. Auflage. Bärenreiter-Verlag, 1989
- [40] BURNETT, I.; WALLE, R. Van d. ; HILL, K. ; BORMANS, J. ; PEREIRA, F.: MPEG-21: goals and achievements. In: *IEEE Multimedia* 10 (2003), Oktober, Nr. 4, S. 60–70
- [41] CARRÉ, Matthieu; PHILIPPE, Pierrick ; APÉIAN, Christophe: New Query-By-Humming Music Retrieval System Conception and Evaluation based on a Query Nature Study. In: *Proceedings of the COST*, 2001, S. 1–5
- [42] CASEY, Michael A.; WESTNER, Alex: Separation of Mixed Audio Sources By Independent Subspace Analysis. In: *Proceedings of ICMC2000*, 2000
- [43] CASTAN, Gerd. *Music Notation Formats*. <http://www.music-notation.info/>. Aufruf vom 23.5.2005
- [44] CEMGIL, Ali T.; DESAIN, Peter ; KAPPEN, Bert. *Rhythm Quantization for Transcription*. w3. August 1999
- [45] CHAI, Wei; VERCOE, Barry: Melody Retrieval On The Web. In: *Proceedings of ACM/SPIE Conference on Multimedia Computing and Networking*, 2002
- [46] CHANG, Shih-Fu; SIKORA, Thomas ; PURI, Atul: Overview of the MPEG-7 Standard. In: *IEEE Transactions on Circuits and Systems for Video Technology* 11 (2001), Juni, Nr. 6, S. 688–695
- [47] CHARPENTIER, F. J.: Pitch Detection Using the Short-Term Phase Spectrum. In: *Proc. ICASSP*, 1986
- [48] CHEVEIGNÉ, Alain de: YIN, a fundamental frequency estimator for speech and music. In: *J. Acoust. Soc. Am.* 111 (2002), April, Nr. 4
- [49] CHIEN, Yu-Ren; JENG, Shyh-Kang: An Automatic Transcription System with Octave Detection. In: *Proc. of ICASSP*, 2002, S. 1865–1869

-
- [50] CLARISSE, L. P.; MARTENS, J. P. ; LESAFFRE, M. ; BAETS, B. D.; MEYER, H. D. ; LEMAN, M.: An Auditory Model Based Transcriber of Singing Sequences. In: *Proceedings of the ISMIR, 2002*, S. 116–123
- [51] CLAUSEN, M.; ENGELBRECHT, R. ; MEYER, D. ; SCHMITZ, J.: PROMS: A Web-based Tool for Searching in Polyphonic Music. In: *Proceedings Intl. Symp. on Music Information Retrieval*. Plymouth, M.A., USA, Oktober 2000
- [52] CLAUSEN, Michael; KURT, Frank ; ENGELBRECHT, Roland: Content-based Retrieval in MIDI and Audio. In: *ECDL WS Generalized Documents*, 2001
- [53] DAHLHAUS, Carl (Hrsg.); EGGEBRECHT, Hans H. (Hrsg.) ; OEHL, Kurt (Hrsg.): *Brockhaus Riemann Musiklexikon*. 2. Mainz : Atlantis-Schott Musikbuch-Verlag, 1995
- [54] DANNENBERG, Roger B.; BIRMINGHAM, William P. ; TZANETAKIS, George; MEEK, Colin ; HU, Ning ; PARDO, Bryan: The MUSART Testbed for Query-by-Humming Evaluation. In: *Computer Music Journal* (2003)
- [55] DANNENBERG, Roger B.; HU, Ning: Understanding Search Performance in Query-By-Humming Systems. In: *Proc. of the ISMIR, 2004*
- [56] DITTMAR, Christian; UHLE, Christian: Further Steps towards Drum Transcription of Polyphonic Music. In: *Proc. 116th AES Convention*, 2004
- [57] DIXON, Simon; GOEBL, Werner: Pinpointing the Beat: Tapping to expressive Performances. In: *Proceedings of the 7th International Conference on Music Perception and Cognition*. Sydney, 2002
- [58] DOWNIE, J. S.: *Evaluating a simple Approach to Music Information Retrieval: Conceiving Melodic N-Grams as Text*. London, Ontario, University of Western Ontario, Diss., Juli 1999
- [59] Kap. Music information retrieval (Chapter 7) In: DOWNIE, J. S.: *Annual Review of Information Science and Technology* 37. Medford, NJ : Information Today, 2003, S. 295–340. – Available from music-ir.org/downie_mir_arist37.pdf
- [60] DOWNIE, Stephen: Evaluation of a Simple and Effective Music Information Retrieval Method. In: *SIGIR, ACM*, 2000

- [61] DUXBURY, C.; BELLO, J. P. ; DAVIS, M. ; SANDLER, M.: A combined Phase and Amplitude Based Approach to Onset Detection for Audio Segmentation. In: *Proc. of WIAMIS*, 2003
- [62] *Der Brockhaus multimedial 2002*. DVD-ROM. 2001
- [63] EGGINK, Jana; BROWN, Guy J.: Application of Missing Feature Theory to the Recognition of Musical Instruments in Polyphonic Audio. In: *Proc. of ISMIR*, 2003
- [64] EGGINK, Jana; BROWN, Guy J.: Extracting Melody Lines from Complex Audio. In: *ISMIR*, 2004, S. 84–91
- [65] EIDENBERGER, Horst: Query by Humming. In: *iX* (2003), Nr. 6, S. 110–113
- [66] EISENBERG, Gunnar: *Rhythmuserkennung in Musiksignalen für inhaltsbezogene Datensuche nach MPEG-7*, Technische Universität Berlin, Diplomarbeit, 2003
- [67] EISENBERG, Gunnar; BATKE, Jan-Mark ; SIKORA, Thomas: Efficiently Computable Similarity Measures for Query by Tapping Systems. In: *Proc. of the 7th Int. Conference on Digital Audio Effects*. Naples, Italy : DAFx, Oktober 2004, S. 189–193
- [68] EISENBERG, Gunnar; BATKE, Jan-Mark ; SIKORA, Thomas: BeatBank – An MPEG-7 compliant Query by Tapping System. In: *Proc. of the 116th AES Convention*. Berlin : AES, Mai 2004
- [69] FENG, Yazhong; ZHUANG, Yueting ; PAN, Yunhe: A Hierarchical Approach: Query Large Music Database by Acoustic Input. In: *SIGIR'02*, 2002, S. 441–442
- [70] FERBER, Reginald: *Information Retrieval*. Heidelberg : dpunkt.verlag, März 2003
- [71] FERNÁNDEZ-CID, P.; CSAJÚS-QUIRÓS, F. J.: Multi-pitch Estimation for Polyphonic Musical Signals. In: *Proc. of the ICASSP*, 1998, S. 3565–3568
- [72] *Query by Humming Melodieerkennungssystem*. www.idmt.fraunhofer.de. – pdf-Datei, Produktblatt

-
- [73] FITZGERALD, D.; LAWLOR, R. ; COYLE, E.: Drum Transcription using Automatic Grouping of Events and prior Subspace Analysis. In: *Proc. of the WIAMIS*, 2003
- [74] FLANAGAN, J. L.; GOLDEN, R. M.: Phase Vocoder. In: *The Bell System Technical Journal* (1966), November, S. 1493–1509
- [75] FRANCU, C; NEVILL-MANNING, C. G.: Distance Metrics and Indexing Strategies for a Digital Library of popular Music. In: *Proceedings of the ICASSP*, 2000, S. 889–892
- [76] GHAS, Asif; LOGAN, Jonathan ; CHAMBERLIN, David ; SMITH, Brian C.: Query by Humming: Musical Information Retrieval in an Audio Database. In: *ACM Multimedia*, 1995, S. 231–236
- [77] GOLD, B.; RABINER, L.: Parallel Processing Techniques for Estimation Pitch Periods of Speech in the Time Domain. In: *Journal of the Acoustical Society of America* 46 (1969), Nr. 2, S. 442–448
- [78] GOMEZ, Emila; STREICH, Sebastian ; ONG, Beesuan ; PAIVA, Rui P.; TAPPERT, Sven ; BATKE, Jan-Mark ; BELLO, Graham Poliner Juan P.: A Quantitative Comparison of Different Approaches for Melody Extraction from Polyphonic Audio Recordings. In: *Computer Music Journal* (2005). – Artikel in Vorbereitung
- [79] GOTO, Masataka: A Robust Predominant-F0 Estimation Method for Real-Time Detection of Melody and Bass Lines in CD Recordings. In: *Proc. ICASSP*, 2000, S. 757–760
- [80] GOTO, Masataka: A Predominant-F0 Estimation Method for CD Recordings: MAP Estimation using EM Algorithm for Adaptive Tone Models. In: *Proc. ICASSP*, 2001, S. V–3365–3368
- [81] GOTO, Masataka: A Predominant-F0 Estimation Method for Real-world Musical Audio Signals: MAP Estimation for Incorporating Prior Knowledge about F0s and Tone Models. In: *Proceedings of CRAC*, CRAC, September 2001
- [82] GOUYON, F.; HERRERA, P.: A Beat Induction Method for Musical Audio Signals. In: *Proc. of the WIAMIS*, 2003

- [83] GOUYON, Fabien; PACHET, Francois ; DELERUE, Olivier: On the use of zero-crossing rate for an application of classification of percussive sounds. In: *Proceedings of the COST G-6 on DAFX-00*, 2000, S. 1–6
- [84] HAINSWORTH, Stephen; MACLEOD, Molcom: On Sinusoidal Parameter Estimation. In: *Proc. of the 6th Int. Conference on Digital Audio Effects, DAFx*, 2003
- [85] HAINSWORTH, Stephen W.; MACLEOD, Malcolm D.: *The Automated Music Transcription Problem*. 2004. – Submitted Journal Paper
- [86] HAINSWORTH, Stephen W.; MACLEOD, Malcom D. ; WOLFE, Patrick J.: Analysis of Reassigned Spectrograms for Musical Transcription. In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. Mohonk, N.Y. : IEEE, Oktober 2001
- [87] HAINSWORTH, Stephen W.: *Techniques for the Automated Analysis of Musical Audio*, University of Cambridge, Diss., Dezember 2003
- [88] HASHIGUCHI, Hiroki; NISHIMURA, Takuichi ; TAKITA, Junko; ZHANG, Xin ; OKA, Ryuichi: *Music Signal Spotting Retrieval by a Humming Query*. Proceedings of the 2nd Annual International Symposium on Music Information Retrieval (ISMIR 2001). Oktober 2001. – pp.211-218
- [89] HAUS, Goffredo; POLLASTRI, Emanuele: A Multimodal Framework for Music Inputs. In: *Proceedings of the 8th ACM International Conference on Multimedia*. Los Angeles : ACM, 2000
- [90] HAUS, Goffredo; POLLASTRI, Emanuele: An Audio Front End for Query-by-Humming Systems. In: *2nd Annual International Symposium on Music Information Retrieval*. Bloomington, Indiana, USA : ISMIR, 2001
- [91] HERNÁNDEZ COLL, Yorck: *Verfahren der Melodieerkennung im Standard MPEG-7*, Technische Universität Berlin, Diplomarbeit, 2002
- [92] HESS, Wolfgang; SCHROEDER, Manfred R. (Hrsg.): *Springer Series in Information Sciences*. Bd. 3: *Pitch Determination of Speech Signals*. Berlin : Springer-Verlag, 1983
- [93] HESS, Wolfgang: *Sprachsignalverarbeitung 1, 2*. Rheinische Friedrich-Wilhelms-Universität Bonn, 2002. – Vorlesungsskript

-
- [94] HÄNSLER, Eberhard: *Statistische Signale*. 2., neubearb. und erw. Auflage. Berlin : Springer, 1997
- [95] HODES, Karlheinrich: *Der Gregorianische Choral – Eine Einführung*. 2. Auflage. Bernardus Verlag, 1992
- [96] HOOS, Holger H.; RENZ, Kai ; GÖRG, Marko: GUIDO/MIR — an Experimental Musical Information Retrieval System based on GUIDO Music Notation. In: *Proceedings of the Second Annual International Symposium on Music Information Retrieval*, 2001
- [97] HU, Ning; DANNENBERG, Roger B.: A Comparison of Melodic Database Retrieval Techniques Using Sung Queries. In: *JCDL '02*, 2002, S. 301–307
- [98] HUNTER, Jane: An Overview of the MPEG-7 Description Definition Language (DDL). In: *IEEE Transactions on Circuits and Systems for Video Technology* 11 (2001), Juni, Nr. 6, S. 765–772
- [99] HYVÄRINEN, Aapo; KARHUNEN, Juha ; OJA, Erkki: *Independent Component Analysis*. John Wiley & Sons, 2001
- [100] ISO/IEC JTC 1/SC 29: *Information Technology – Multimedia Content Description Interface – Part 4: Audio*. ISO/IEC FDIS 15938-4:2001(E). Juni 2001
- [101] JANG, Jyh-Shing R.; LEE, Hong-Ru ; CHEN, Jiang-Chun: Super MBox: an efficient/effective content-based music retrieval system. In: *MULTIMEDIA '01: Proceedings of the ninth ACM international conference on Multimedia*. New York, NY, USA : ACM Press, 2001. – ISBN 1–58113–394–4, S. 636–637
- [102] JAYANT, N. S.; NOLL, Peter: *Digital Coding of Waveforms — Principles and Applications to Speech and Video*. Englewood Cliffs, New Jersey : Prentice-Hall, Inc., 1984
- [103] KAGEYAMA, T.; MOCHIZUKI, K. ; TAKASHIMA, Y.: Melody Retrieval with Humming. In: *Proc. International Computer Music Conference*, 1993
- [104] KASHINO, K.; MURASE, H.: A Sound Identification System for Ensemble Music based on Template Adaption and Music Stream Extraction. In: *Speech Communication* 27 (1999), S. 337–349

- [105] KAWAHARA, Hideki; KATAYOSE, Haruhiro ; CHEVEIGNÉ, Alain de ; PATTERSON, Roy D.: Fixed Point Analysis of Frequency to Instantaneous Frequency Mapping for Accurate Estimation of F_0 and Periodicity. In: *Sixth European Conference on Speech Communication and Technology*, 1999
- [106] KEDEM, Benjamin: Spectral Analysis and discrimination by zero-crossings. In: *Proceedings of the IEEE* 11 (1986), November, Nr. 74, S. 1477–1493
- [107] KIM, Hyoung-Gook; MOREAU, Nicolas ; SIKORA, Thomas: *MPEG-7 Audio and beyond*. West Sussex, England : Wiley & Sons, 2005
- [108] KIM, Youngmoo E.; CHAI, Wei ; GARCIA, Ricardo ; VERCOE, Barry: Analysis of a Contour-Based Representation for Melody. In: *Proc. International Symposium on Music Information Retrieval*, 2000
- [109] KLAPURI: Musical meter estimation and music transcription. In: *Cambridge Music Processing Colloquium*. Cambridge University, 2003
- [110] KLAPURI, Anssi: Means of Integrating Audio Content Analysis Algorithms. In: *110th Audio Engineering Society Convention*. Amsterdam, Netherlands, 2001
- [111] KLAPURI, Anssi: *Signal Processing Methods for the Automatic Transcription of Music*, Tampere University of Technology, Diss., 2004
- [112] KLAPURI, Anssi; ERONEN, Antti ; SEPPÄNEN, Jarno ; VIRTANEN, Tuomas: Automatic Transcription of Music. In: *Symposium on Stochastic Modeling of Music*, 2001
- [113] KLAPURI, Anssi P.: Automatic Transcription of Music. In: *Proceedings of the Stockholm Music Acoustics Conference*. Stockholm, Sweden, August 2003
- [114] KLAPURI, Anssi P.; ASTOLA, Jaakko T.: Efficient Calculation of a Physiologically-motivated Representation for Sound. In: *IEEE International Conference on Digital Signal Processing*. Santorini, Greece, 2002
- [115] KODERA, Kunihiko; GENDRIN, Roger ; VILLEDARY, Claude de: Analysis of Time-Varying Signals with Small BT Values. In: *IEEE Trans. on Acoustics, Speech and Signal Processing* ASSP-26 (1978), Februar, Nr. 1, S. 64–76

-
- [116] KOSUGI, Naoko; NISHIHARA, Yuichi ; SAKATA, Tetsuo ; YAMAMURO, Masashi ; KUSHIMA, Kazuhiko: A Practical Query-By-Humming System for a Large Music Database. In: *Proceedings of the Eighth ACM International Conference on Multimedia*, 2000, S. 333–342
- [117] KRAMER, André: Erkennen Sie die Melodie? – Audio-Retrival: Was Rechner von Musik verstehen. In: *c't* (2004), Nr. 7, S. 178–181
- [118] KURTH, Frank; RIBBROCK, Andreas ; CLAUSEN, Michael: Efficient Fault Tolerant Search Techniques for Full-Text Audio Retrieval. In: *Proc. of the 112th AES Convention*, AES, Mai 2002
- [119] LEMSTRÖM, Kjell: *String Matching Techniques for Music Retrieval*, University of Helsinki, Diss., 2000
- [120] LESAFFRE, Micheline [u. a.]: User Behavior in the Spontaneous Reproduction of Musical Pieces by Vocal Query. In: *Proc. of the 5th Triennial ESCOM Conference*, 2003, S. 208–211
- [121] LI, Weiping (Hrsg.): *Special issue on MPEG-7*. IEEE Circuits and Systems Society, 2001
- [122] LINDSAY, Adam T.: *Using Contour as a Mid-Level Representation of Melody*, Massachusetts Institute of Technology, Diplomarbeit, September 1996
- [123] LOUGHLIN, Patrick J.; TACER, Berkant: Comments on the Interpretation of Instantaneous Frequency. In: *IEEE Signal Processing Letters* 4 (1997), Mai, Nr. 5, S. 123–125
- [124] LU, Lie; YOU, Hong ; ZHANG, Hong-Jiang: A new approach To Query By humming In Music Retrieval. In: *IEEE International Conference on Multimedia and Expo*, 2001
- [125] MANJUNATH, B. S. (Hrsg.); SALEMBIER, Philippe (Hrsg.) ; SIKORA, Thomas (Hrsg.): *Introduction to MPEG-7*. West Sussex, England : Wiley & Sons, 2002
- [126] MAROLT, Matija: SONIC : transcription of polyphonic piano music with neural networks. In: *Proceedings of Workshop on Current Research Directions in Computer Music*. Barcelona, nov 2001

- [127] MAROLT, Matija: A Connectionist Approach to Transcription of Polyphonic Piano Music. In: *IEEE Transactions on Multimedia* 6 (2004), jun, Nr. 3, S. 439–449
- [128] MARTÍNEZ, José M.: MPEG-7 Overview / International Organisation For Standardisation. Pattaya, März 2003 (ISO/IEC JTC1/SC29/WG11N5525). – Forschungsbericht. Version 9; <http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm>
- [129] MARTINS, Luís Gustavo P. M.; FERREIRA, Aníbal J. S.: PCM to MIDI Transposition. In: *112th AES Convention*, 2002
- [130] McNAB, Rodger J.; SMITH, Lloyd A.: Evaluation of a melody transcription system. In: *Proceedings of the ICASSP*, 2000, S. 819–822
- [131] McNAB, Rodger J.; SMITH, Lloyd A. ; BAINBRIDGE, David ; WITTEN, Ian H.: The New Zealand Digital Library MELody inDEX. In: *D-Lib Magazine* (1997), Mai
- [132] McNAB, Rodger J.; SMITH, Lloyd A. ; WITTEN, Ian H.: Signal Processing for Melody Transcription. In: *Proceedings of the 19th Australasian Computer Science Conference*, 1996
- [133] McNAB, Rodger J.; SMITH, Lloyd A. ; WITTEN, Ian H. ; HENDERSON, Clare L. ; CUNNINGHAM, Sally J.: Towards the Digital Music Library: Tune Retrieval from Acoustic Input. In: *Proceedings of the first ACM international conference on Digital libraries*. Bethesda, Maryland, United States : ACM, 1996, S. 11–18
- [134] MEDDIS, R.; HEWITT, M. J.: Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification. In: *J. Acoust. Soc. Am.* 89 (1991), Nr. 6, S. 2866–2882
- [135] MEEK, Colin; BIRMINGHAM, William P.: The danger of parsimony in query-by-humming applications. In: *International Conference on Music Information Retrieval*. Baltimore, Oktober 2003
- [136] MICHELS, Ulrich: *dtv-Atlas Musik*. Bd. 1: *Systematischer Teil, Musikgeschichte von den Anfängen bis zur Renaissance*. 18. Auflage. München : Deutscher Taschenbuch Verlag, 1998

-
- [137] MICHELS, Ulrich: *dto-Atlas Musik*. Bd. 2: *Musikgeschichte vom Barock bis zur Gegenwart*. 11. Auflage. München : Deutscher Taschenbuch Verlag, 1999
- [138] MÜLLENSIEFEN, Daniel; FRIELER, Klaus: Optimizing Measures of Melodic Similarity for the Exploration of a Large Folk Song Database. In: *Proc. ISMIR*, 2004
- [139] Moeck Musikinstrumente + Verlag: *MOECK Note Nr. 1*. 2005
- [140] MONGEAU, Marcel; SANKOFF, David: Comparison of Musical Sequences. In: *Computers and the Humanities* 24 (1990), S. 161–175
- [141] MOTTE, Diether de l.: *Melodie – Ein Lese- und Arbeitsbuch*. Deutscher Taschenbuch Verlag, 1993
- [142] MULDER, Tom D.; MARTENS, Jean-Peiree ; LESAFFRE, Micheline; LEMAN, Marc ; BAETS, Bernard D. ; MEYER, Hans D.: An Auditory Model Based Transcriber of Vocal Queries. In: *Proc. of the ISMIR*, 2003
- [143] MUSMANN, Hans G. *Statistische Methoden der Nachrichtentechnik*. Vorlesung an der Universität Hannover. 1993
- [144] NISHIMURA, Takuichi; HASHIGUCHI, Hiroki ; SEKIMOTO, Nobuhiro; ZHANG, J.Xin ; GOTO, Masataka ; OKA, Ryuichi. *Music Signal Spotting Retrieval by a Humming Query Using Start Frame Feature Dependent Continuous Dynamic Programming*. IPSJ SIGNotes Music and Computer. 2001
- [145] NOLL, A. M.: Cepstrum Pitch Determination. In: *The Journal of the Acoustical Society of America* 41 (1967), Nr. 2, S. 293–309
- [146] NOLL, Peter: *Signale und Systeme*. TU Berlin, FG Fernmeldetechnik, 1999. – Vorlesungsskript
- [147] OPPENHEIM, Alan V.; SCHAFFER, Ronald W.: *Zeitdiskrete Signalverarbeitung*. 2. München, Wien : R. Oldenburg Verlag, 1995
- [148] PANITZ, Lina: Musikkauf übers Internet kommt in Mode. In: *Die Welt* (2005), Zeitungsartikel, 17. Mai
- [149] PAUWS, Steffen: CubyHum: A Fully Operational Query by Humming System. In: *Proc. of the 3rd ISMIR*, 2002

- [150] PAUWS, Steffen: Effects of song familiarity, singing training and recent song exposure on the singing of melodies. In: *ISMIR*, 2003
- [151] PEETERS, Geoffroy; RODET, Xavier: SINOLA: A New Analysis/Synthesis using Spectrum Peak Shape Distortion, Phase and Reassigned Spectrum. In: *Proc. ICMC99*, 1999
- [152] PLANTE, F.; MEYER, G. ; AINSWORTH, W.A.: Improvement of speech spectrogram accuracy by the method of reassignment. In: *IEEE Transactions on Speech and Audio Processing* 6 (1998), Mai, Nr. 3, S. 282–287
- [153] POLLASTRI, Emanuele: A Pitch Tracking System Dedicated to Process Singing Voice for Music Retrieval. In: *Proc. of the ICASPP*, 2002, S. 341–344
- [154] PRECHELT, Lutz; TYPKE, Rainer: An interface for melody input. In: *ACM Transactions on Computer-Human Interaction* 8 (2001), Nr. 2, S. 133–149
- [155] PRESS, William H. [u. a.]: *Numerical Recipes in C*. 2nd edition. Cambridge University Press, 1992
- [156] QUACKENBUSH, Schuyler; LINDSAY, Adam: Overview of MPEG-7 Audio. In: *IEEE Transactions on Circuits and Systems for Video Technology* 11 (2001), Juni, Nr. 6, S. 727–729
- [157] QUANTZ, Johann J.: *Versuch einer Anweisung die Flöte traversiere zu spielen*. Berlin, 1752
- [158] RABINER, Lawrence; SAMBUR, Marvin ; SCHMIDT, Carol: Applications of nonlinear smoothing algorithm to speech processing. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 23 (1975), S. 552–557
- [159] RABINER, Lawrence R.; CHENG, Michael J. ; ROSENBERG, Aaron E. ; MCGONEGAL, Carol A.: A comparative Performance Study of Several Pitch Detection Algorithms. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* Assp-24 (1976), Oktober, Nr. 5, S. 399–419
- [160] RAJU, M. A.; SUNDARAM, Bharat ; RAO, Preeti: Tansen: a Query-by-Humming based Musical Retrieval System. In: *National Conference of Communications (NCC)*, 2003

-
- [161] RAO, Preeti; RAJU, M. A.: Building a Melody Retrieval System. In: *Proc. of the National Conference on Communications*. Bombay : I.I.T., 2002
- [162] RAPHAEL, Christopher: Automatic Transcription of Piano Music. In: *Proc. ISMIR*, 2002
- [163] REIN, Stephan; REISLEIN, Martin ; SIKORA, Thomas: Audio Content Description with Wavelets and Neural Nets. In: *Proc. of the ICASSP*. Montreal, Canada : IEEE, Mai 2004
- [164] ROEDERER, Juan G.: *Physikalische und psychoakustische Grundlagen der Musik*. Springer-Verlag, 1975
- [165] ROLLAND, Pierre-Yves; RASKINIS, Gailius ; GANASCIA, Jean-Gabriel. *Musical Content-Based Retrieval*. 0000
- [166] ROSSI, L.; GIROLAMI, G.: Instantaneous Frequency and Short Term Fourier Transforms: Application to Piano Sounds. In: *J. Acoust. Soc. Am.* 110 (2001), November, Nr. 5, S. 2412–2420
- [167] RYNNÄNEN, Matti P.; KLAUPURI, Anssi P.: Modelling of Note Events of Singing Transcription. In: *Proc. of ISCA Tutorial and Workshop on Statistical and Perceptual Audio Processing*, 2004
- [168] SAGAYAMA, Shigeki; TAKAHASHI, Keigo ; KAMEOKA, Hirokazu ; NISHIMOTO, Takuya: Specmurt Anasylis: A Piano-Roll-Visualization of Polyphonic Music Signals by Deconvolution of Log-Frequency Spectrum. In: *Proc. SAPA*, 2004
- [169] SAKURABA, Yohei; KITAHARA, Tetsuro ; OKUNO, Hiroshi G.: Comparing Features for Forming Music Streams in Automatic Music Transcription. In: *Proc. ICASSP Bd. IV*, 2004, S. 273–277
- [170] SCHEIRER, Eric D.: Tempo and beat analysis of acoustic musical signals. In: *J. Acoust. Soc. Am.* 103 (1998), Januar, Nr. 1, S. 588–601
- [171] SCHROTH, Alexander: *Implementierung eines effizienten Verfahrens zur Änderung der Grundfrequenz von Sprachsignalen auf einem digitalen Signalprozessor*, Technische Universität Berlin, Diplomarbeit, 2004
- [172] SEDGEWICK, Robert: *Algorithmen*. 2. korrigierter Nachdruck. Addison-Wesley, 1995

- [173] SHALEV-SHWARTZ, Shai; DUBNOV, Shlomo ; FRIEDMAN, Nir ; SINGER, Yoram: Robust Temporal and Spectral Modeling for Query By Melody. In: *SIGIR'02*, 2002, S. 331–338
- [174] SHANDILYA, Saurabh; RAO, Preeti: Pitch Detection of the Singing Voice in Musical Audio. In: *Proc. of the 114th AES Convention*. Amsterdam : AES, 2003
- [175] SHIH, Hsuan-Huei; NARAYANAN, Shrikanth ; KUO, Jay: Multidimensional humming transcription using a statis humming. In: *Proc. of the ICASSP*, 2003
- [176] SIKORA, Thomas: The MPEG-7 Visual Standard for Content Description — An Overview. In: *IEEE Transactions on Circuits and Systems for Video Technology* 11 (2001), June, Nr. 6, S. 696–702
- [177] SMARAGDIS, Paris: *Redundancy Reduction for Computational Audition, a Unifying Approach*, Massachusetts Institue of Technology, Diss., 2001
- [178] SMARAGDIS, Paris; BROWN, Judith C.: Non-Negative Matrix Factorization for Polyphonic Transcription. In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, IEEE, Oktober 2003, S. 177–180
- [179] STUCKENSCHMIDT, H. H.: *Zwischen den beiden Kriegen*. Bd. 2: *Neue Musik*. Berlin : Suhrkamp Verlag, 1951
- [180] TAPPERT, Sven: *Transkription von Melodiekonturen gemäß MPEG-7 aus Musiksignalen*, Technische Universität Berlin, Diplomarbeit, 2004
- [181] TAUPIN, Daniel; MITCHELL, Ross ; EGLER, Andreas: *MusiX_{TEX}*. – Notensatz-Macropaket für $\text{T}_{\text{E}}\text{X}$, siehe <http://icking-music-archive.org/software/indexmt6.html>
- [182] TEKALP, A. M. (Hrsg.): *Special Issue on MPEG-7 technology*. Elsevier, 2000
- [183] TYPKE, Rainer. *MIR Systems: A Survey of Music Information Retrival Systems*. mirsystems.info. Aufruf vom 23.5.2005
- [184] TYPKE, Rainer; GIANNOPOULOS, Panos ; VELTKAMP, Remco C.; WIERING, Frans ; OOSTRUM, René van: Using Transportation Distances for Measuring Melodic Similarity. In: *International Conference on Music Information Retrieval*. Baltimore, Oktober 2003

-
- [185] UHLE, Christian; DITTMAR, Christian ; SPORER, Thomas: Extraction of Drum Tracks from Polyphonic Music Using Independent Subspace Analysis. In: *ICA 2003*. Nara, JP, April 2003
- [186] UITDENBOGERD, A.; ZOBEL, J.: Music ranking techniques evaluated. In: OUDSHOORN, M. (Hrsg.): *Proceedings of the Australasian Computer Science Conference*. Melbourne, Australia, Januar 2002, S. 275–283
- [187] UITDENBOGERD, Alexandra L.: *Music Information Retrieval Technology*, Royal Melbourne Institute of Technology, Diss., Mai 2002
- [188] UITDENBOGERD, Alexandra L.; YAP, Yaw W.: Was Parson right? An experiment in usability of music representations for melody-based music retrieval. In: *Proc. ISMIR*, 2003
- [189] Kap. Sweep the Music! In: UKKONEN, Esko; LEMSTRÖM, Kjell ; MÄKINEN, Veli: *Computer science in perspective*. New York, NY, USA : Springer-Verlag, 2003, S. 330–342
- [190] VAIDYANATHAN, P. P.: *Multirate Systems and Filter Banks*. Prentice Hall, 1993
- [191] VEIT, Ivar: *Technische Akustik*. 4. Auflage. Vogel, 1988
- [192] VETRO, Anthony; DEVILLERS, Sylvain: Delivery Context in MPEG-21. In: *W3C Workshop on Delivery Context*, 2002
- [193] VETTERLI, Martin: A Theory of Multirate Filter Banks. In: *IEEE Trans. on Acoustics, Speech and Signal Processing* ASSP-35 (1987), März, Nr. 3, S. 356–372
- [194] VIITANIEMI; KLAURI ; ERONEN: A probabilistic model for the transcription of single-voice melodies. In: *Finnish Signal Processing Symposium, FINSIG*. Tampere University of Technology, Mai 2003, S. 59–63
- [195] WEISHAUP, Philipp: *Distanzmaße für die melodiebezogene Datensuche nach MPEG-7*, Technische Universität Berlin, Diplomarbeit, 2004
- [196] WILGEN, Arne van: *Tonhöhenbestimmung für Verfahren der Melodieerkennung im Standard MPEG-7*, Technische Universität Berlin, Diplomarbeit, 2003

- [197] ZHU, Yongwei; KANKANHALLI, Mohan: Music Scale Modelling for Melody Matching. In: *Proc. MM*, 2003, S. 359–362
- [198] ZHU, Yunyue; SHASHA, Dennis ; ZHAO, Xiaojian: Query by Humming — in Action with its Technology Revealed. In: *SIGMOD 2003*, 2003, S. 675
- [199] ZIEGENRÜCKER, Wieland: *ABC Musik — Allgemeine Musiklehre*. 3. Auflage. Wiesbaden : Breitkopf & Härtel, 1997
- [200] ZWICKER, Eberhard: *Psychoakustik*. Berlin : Springer-Verlag, 1982

Lebenslauf Johann-Markus Batke

- 1970 geboren in Mülheim an der Ruhr, aufgewachsen allerdings in Lohne in Oldenburg
- 1989 Abitur am Gymnasium Lohne
- 1989–1990 Wehrdienst als Saxophonist im Heeresmusikkorps 11, Bremen
- 1990–1997 Studium der Elektrotechnik an der Universität Hannover
- 1998–1999 Software-Entwickler, Tiscon AG Infosystems, Ulm
- 1999–2005 Wissenschaftlicher Mitarbeiter, Fachgebiet Nachrichtenübertragung, Technische Universität Berlin
- seit 2006 Entwicklungsingenieur, Deutsche Thomson OHG, Hannover

Nachwort

*Es hat ja auch kein Mensch
behauptet, dass Promovieren
Spaß macht.*

Dr.-Ing. Andreas Willig

Die vorliegende Arbeit entstand während meiner Tätigkeit als wissenschaftlicher Mitarbeiter am Fachgebiet Nachrichtenübertragung der Technischen Universität Berlin.

Mein Dank geht zuerst an Herrn Prof. Dr.-Ing. THOMAS SIKORA, der das Thema Query-by-Humming-Systeme unter Verwendung des Multimedia-Standards MPEG-7 angeregt hat und mir große Freiheiten bei der Ausführung der Arbeit ließ. Herzlicher Dank geht an Prof. Dr.-Ing. PETER NOLL, bei dem ich meine Tätigkeit am Fachgebiet Fernmeldetechnik begonnen habe und der als 2. Gutachter die Arbeit konstruktiv begleitet hat. Dem 3. Gutachter Prof. Dr.-Ing. WOLFGANG HESS danke ich herzlich für die detaillierten Hinweise zur Endfassung der Arbeit. Prof. Dr.-Ing. REINHOLD ORGLMEISTER danke ich für die Übernahme des Vorsitzes des Promotionsausschusses.

Im angelsächsischen Raum wird die Bezeichnung PhD gelegentlich als Abkürzung für "permanent head damage" verstanden. Auch wenn an dieser Stelle nicht geklärt werden kann und soll, ob es sich um eine Folge oder notwendige Voraussetzung handelt, liegt nahe, dass eine Dissertation der Hilfe weiterer Personen bedarf. Daher möchte ich zuerst meiner Frau MONI danken, die mir in Momenten des Zweifels (nicht nur in Bezug auf den o. g. Sachverhalt) wichtige seelische und moralische Stütze war.

Wichtig war die Unterstützung meiner laufenden und nicht laufenden Kollegen am Fachgebiet Nachrichtenübertragung. Besonders bedanken möchte ich mich bei (in Reihenfolge der revidierten Kapitel) TILMAN LIEBCHEN, JUAN JOSÉ BURRED, GUNNAR EISENBERG und meinem Zimmerkollegen KAI CLÜVER. Für die umfassende organisatorische Unterstützung danke ich BIRGIT BOLDIN. Die musikwissenschaftliche Rückversicherung besorgten mein Freund, der Musiker und Komponist HANNES FAHLING und mein Klavierkollege, der Musik-

wissenschaftler MAXIMILIAN RAUSCHER. Für die orthographische Orientierung danke ich CHRISTIANE STEPHANI.

Auch den hier nicht namentlich genannten Kollegen und Freunden danke ich herzlich für ihre Hilfestellung, denn das Erstellen eines Druckwerks ist nicht ausschließlich Quelle innerer Freude. Das gilt sogar für den Druck von Melodien, schon 1613 stellte MICHAEL PRAETORIUS dazu fest:

Denn was ich wegen des Druckens und Correctoris für Wunder,
Mühe Arbeit und Unlust aussgestanden kan alhier dergestalt nicht
erzehlet werden ...

Das kann man gar nicht anders sagen.