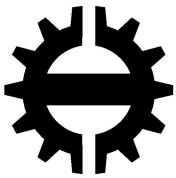


TAMPERE UNIVERSITY OF TECHNOLOGY
Department of Information Technology

Toni Heittola

Automatic Classification of Music Signals

Master of Science Thesis

The subject was approved by the Department of
Information Technology on the 14th February 2003.
Thesis supervisors: Professor Jaakko Astola (TUT)
M.Sc. Anssi Klapuri (TUT)
M.Sc. Antti Eronen (Nokia)

Preface

This work was carried out at the Institute of Signal Processing, Department of Information Technology, Tampere University of Technology, Finland.

I wish to express my gratitude to my thesis advisors and examiners Mr. Anssi Klapuri and Mr. Antti Eronen, for they constant guidance, advice, and patience throughout this work. Further, I thank the examiner Professor Jaakko Astola for his advice and comments.

I am grateful to Ms. Paula Keltto for her great effort to collect and thoroughly annotate all the pieces in the music database. I want also to thank Mr. Vesa Peltonen for his advice and for the database access tools, and Mr. Tuomas Virtanen for his sinusoids-plus-noise spectrum model used in this thesis. I am grateful to Mr. Juha Tuomi for all the valuable discussions during the development of the system.

I want to thank the members of the Audio Research Group of Tampere University of Technology for their valuable knowledge and support and for providing a stimulating working atmosphere.

I wish to thank my mother and grandparents for all understanding, support and endless love. Big thanks also to my friends for understanding my little time for them during this work.

Tampere, December 2003

Toni Heittola

Contents

Tiivistelmä

TAMPEREEN TEKNILLINEN YLIOPISTO
Tietotekniikan osasto
Signaalinkäsittelyn laitos

HEITTOLA, TONI: Musiikkisignaalien automaattinen luokittelu
Diplomityö, 100 s.
Tarkastajat: Prof. Jaakko Astola, DI Anssi Klapuri, DI Antti Eronen
Rahoittajat: Tampereen teknillinen yliopisto, Signaalinkäsittelyn laitos, Nokia Research Center.
Joulukuu 2003
Avainsanat: musiikin sisältöanalyysi, musiikin luokittelu, automaattinen musiikkityylin tunnistus, kuuntelukoe, soitinten havaitseminen.

Digitaaliset musiikkikokoelmat ovat yleistyneet viime vuosina. Samalla kun nämä kokoelmat laajenevat tulee digitaalisen sisällön hallinta yhä tärkeämmäksi. Tässä diplomityössä tutkitaan sisältöpohjaista akustisten musiikkisignaalien luokittelua musiikkityylin (esim. klassinen, rock) sekä käytettyjen soitinten pohjalta. Ihmisen kykyä tunnistaa musiikkityylejä tutkittiin kuuntelukokeella. Tämä diplomityö kattaa kirjallisuustutkimuksen ihmisen musiikkityylin tunnistuksesta, nykyaikaisista musiikkityylin tunnistusjärjestelmistä sekä muista aiheeseen liittyvistä aloista. Lisäksi esitellään monikäyttöinen musiikkitietokanta joka koostuu äänityksistä sekä niiden manuaalisesti tehdyistä annotaatioista. Myös työssä käytettyjen piirteiden sekä luokittimien teoria selostetaan ja tehtyjen simulaatioiden tulokset esitellään.

Kehitetty musiikkityylin tunnistusjärjestelmä käyttää mel-taajuus kepstrikertoimia esittämään musiikkisignaalin aikamuuttuvaa magnitudispektriä. Musiikkityylikohtaiset piirrejakaumat mallinnetaan piilotetuilla Markov malleilla. Soittimen havaitsemista musiikista tutkitaan vastaavalla järjestelmällä. Lisäksi tässä työssä esitetään menetelmä rumpusoitinten havaitsemiseen. Rumpusoitinten läsnäolo musiikissa tunnistetaan havainnoimalla jaksollisuutta signaalin alikaistojen amplitudiverhokäyrissä.

Suoritettu kuuntelukoe osoittaa, että musiikkityylin tunnistus ei ole yksiselitteistä edes ihmiselle. Ihmiset pystyvät tunnistamaan oikean musiikkityylin keskimäärin 75 % tarkuudella (viiden sekunnin näytteillä). Lisäksi tulokset osoittavat, että ihmiset pystyvät tunnistamaan musiikkityylin melko tarkasti ilman pitkän aikavälin temporaalisia piirteitä, kuten rytmi. Kuudelle musiikkityylille kehitetty automaattinen musiikkityylin tunnistusjärjestelmä saavutti noin 60 % tarkkuuden, joka on vertailukelpoinen muiden vastaavien tunnistusjärjestelmien kanssa. Rumpusoitinten havainnoimisessa saavutettiin 81 % tarkkuus käyttäen esitettyä menetelmää.

Abstract

TAMPERE UNIVERSITY OF TECHNOLOGY
Department of Information Technology
Institute of Signal Processing

HEITTOLA, TONI: Automatic Classification of Music Signals
Master of Science Thesis, 100 pages.
Examiners: Prof. Jaakko Astola, M.Sc. Anssi Klapuri, M.Sc. Antti Eronen
Funding: Tampere University of Technology, Institute of Signal Processing, Nokia Research Center.
December 2003
Keywords: music content analysis, music classification, automatic musical genre recognition, listening experiment, musical instrument detection.

Collections of digital music have become increasingly common over the recent years. As the amount of data increases, digital content management is becoming more important. In this thesis, we are studying content-based classification of acoustic musical signals according to their musical genre (e.g., classical, rock) and the instruments used. A listening experiment is conducted to study human abilities to recognise musical genres. This thesis covers a literature review on human musical genre recognition, state-of-the-art musical genre recognition systems, and related fields of research. In addition, a general-purpose music database consisting of recordings and their manual annotations is introduced. The theory behind the used features and classifiers is reviewed and the results from the simulations are presented.

The developed musical genre recognition system uses mel-frequency cepstral coefficients to represent the time-varying magnitude spectrum of a music signal. The class-conditional feature densities are modelled with hidden Markov models. Musical instrument detection for a few pitched instruments from music signals is also studied using the same structure. Furthermore, this thesis proposes a method for the detection of drum instruments. The presence of drums is determined based on the periodicity of the amplitude envelopes of the signal at subbands.

The conducted listening experiment shows that the recognition of musical genres is not a trivial task even for humans. On the average, humans are able to recognise the correct genre in 75 % of cases (given five-second samples). Results also indicate that humans can do rather accurate musical genre recognition without long-term temporal features, such as rhythm. For the developed automatic recognition system, the obtained recognition accuracy for six musical genres was around 60 %, which is comparable to the state-of-the-art systems. Detection accuracy of 81 % was obtained with the proposed drum instrument detection method.

List of acronyms and symbols

ACF	Autocorrelation Function
BER	Band Energy Ratio
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
EER	Equal Error Rate
ESACF	Enhanced Summary Autocorrelation Function
FA	False Acceptance
FR	False Rejection
GMM	Gaussian Mixture Model
HMM	Hidden Mixture Model
k-NN	k-Nearest Neighbour
MFCC	Mel-Frequency Cepstral Coefficients
MIR	Music Information Retrieval
ML	Maximum Likelihood
MMI	Maximum Mutual Information
PCA	Principal Component Analysis
ROC	Receiver Operation Characteristic
SACF	Summary Autocorrelation Function

$\mathbf{x}$	vector $\mathbf{x}$
A'	transpose of matrix A
A^{-1}	inverse of matrix A
λ	parameter set of an HMM
$\boldsymbol{\mu}_{i,m}$	mean vector for the m^{th} mixture component of the i^{th} state
$\boldsymbol{\Sigma}_{i,m}$	covariance matrix for the m^{th} mixture component of the i^{th} state
$w_{i,m}$	mixture weight for the m^{th} mixture component of the i^{th} state
a_{ij}	state transition probability from state i to j
$b_j(\mathbf{x})$	probability of observing feature vector $\mathbf{x}$ in state j
$p(X \mid \lambda)$	likelihood of data X given model λ
$\arg\max_b a$	the value of b which maximizes a

1 Introduction

Personal and public collections of digital music have become increasingly common over the recent years. The amount of digital music available on the Internet has increased rapidly at the same time. As the amount of data increases, efficient management of the digital content is becoming more and more important. However, most of the indexing and labelling of the music is currently performed manually, which is time-consuming and expensive.

Computers are being used to organise information in various domains, but the use of computers to organise music has still been fairly limited. For example, the currently available music search engines on the Internet rely on file names and embedded metadata, and do not make any use of the acoustic signal itself. In this thesis, we are studying content-based classification of acoustic musical signals according to their musical genre (e.g., classical, jazz, rock) and the instruments used.

Music Information Retrieval (MIR) in general has gained increasing research attention over the recent years. Large-scale robust MIR systems will have, apart from their academic merits, important social and commercial implications. They will create significant value added to the existing music libraries by making the entire collection of music easily accessible. The automatic analysis of the music information will enable automatic classification, organisation, indexing, searching, and retrieval of the music in these libraries.

Music information retrieval can be divided into symbolic MIR, where structured signals (MIDI) are processed, and audio MIR, where audio signals are processed. Some research issues are common to both symbolic MIR and audio MIR, since some audio MIR methods transform an audio signal into some sort of structured form before analysis. The audio MIR domain includes various acoustic signal analysis problems, e.g., musical instrument identification, beat tracking, music transcription (to notes), similarity estimation of music, song identification, and segmentation and recognition of music signal based on different musical properties.

In the development of MIR systems, an appropriate music database is essential for testing and evaluation. This thesis introduces a general-purpose music database to be used in the evaluations in different areas of MIR. All the music pieces collected for the database were manually annotated.

1.1 Musical genre recognition

Music can be classified according to its genre, which is probably one of the most often used descriptors for music. Musical genres are widely used for categorising music in the record stores, radio stations, in all sorts of music libraries, and nowadays increasingly on the Internet. Classifying music into a particular musical genre is a useful way of describing qualities that it shares with other music from the same genre, and separating it from other music. Generally, music within the same musical genre has certain similar characteristics, for example, similar types of instruments used, similar rhythmic patterns, or similar harmonic/melodic features.

We conducted a listening experiment to study human abilities to recognise musical genres. The experiment shows that people without any formal musical training can quite accurately and easily classify music into musical genres, even from short excerpts. In general, people can easily make a clear distinction between classical music and, say, rock music. However, musical genres do not have clear boundaries or definitions since music is an art form that evolves constantly. A musical genre can be considered always to be a subjective category with regard to both the listener and the cultural background.

Automatic musical genre recognition has been a growing area of research in the last few years [Jiang02, Tzanetakis02a, Burred03, Li03]. The research in the field is mostly in its initial phase, and there are various subproblems that need to be explored and studied. Practical applications of automatic musical genre recognition include agents to search and select music (e.g. from a database or a radio station), and generating playlists based on musical genre. Information about the musical genre can also be utilised to control more specific MIR tasks, such as transcription, segmentation, tempo estimation, and beat tracking. If we have some high-level information about the music, such as the musical genre, we might be able to make some assumptions and predictions about the music.

This thesis studies the automatic recognition of musical genres according to the spectral characteristics of the music signal. Furthermore, the performance of the developed system is compared with the human abilities in order to have a realistic baseline for the performance.

1.2 Instrument detection

Instrumentation is an important high-level descriptor of music, thus providing useful information for many MIR related tasks. In many cases, exactly expressible descriptors are more efficient for information retrieval than more ambiguous concepts, such as the musical genre. For example, someone might search for classical music by requesting a piece with string instruments and without drums. Furthermore, instrument detection is useful for the automatic musical genre recognition, since some of the instruments are more characteristic for some genres. For example, electric guitar is quite a dominant instrument in rock music, but is hardly ever used in classical music.

In this thesis, we study the detection of instruments and vocals in musical signals according to the spectral characteristics of the signal. Furthermore, we study segmentation of musical signals according to the presence or absence of drum instruments. We propose a new approach to detect the presence of drum instruments by detecting periodicities in the amplitude envelopes of the signal at subbands. The presence and absence of drum instruments in music can be used in audio editing, or in further analysis, e.g. as a front-end for music transcription, metrical analysis, or rhythm recognition.

1.3 Organisation of the thesis

This thesis is organised as follows. Chapter 2 describes a literature review on automatic musical genre recognition and related fields of interest. Chapter 3 presents the music database used in this thesis. Chapter 4 describes the listening experiment studying human abilities in recognising musical genres. The experiment is later used as a baseline for the performance evaluation of the developed automatic musical genre recognition system. Chapter 5 reviews the features and the classification techniques studied. In Chapter 6, the automatic musical genre recognition system is evaluated and the performance is compared to human abilities. We also evaluate the instrument detection and the system locating segments with drums from music signals. Chapter 7 summarises the observations made in this study and suggests some directions for future work. Appendix A describes the musical genre hierarchy used and Appendix B lists the pieces in the music database used in this research.

2 Literature review

This chapter reviews previous work in the fields that are relevant to this thesis.

2.1 Musical genre

The concept of musical genre is a dual one. On one hand, genre is an interpretation of a music piece through a cultural background. It is a linguistic category providing common term to talk about the music. On the other, it is a set of music pieces sharing similar properties of music, like tempo and instrumentation. Musical genres themselves are ill-defined: attempts to define them precisely will most probably end up in circular structures. [Aucouturier03]

In order to keep things simple (within this thesis), musical genre can be considered as a specific category of music, which shares common characteristics and properties that in the perception of the average listener distinguish a music performance from performances belonging to other genres. There are many other factors, besides the instrumentation and rhythmic structure, that have effect on musical genre recognition done by humans.

2.1.1 Genre taxonomy

Different genre taxonomies can be applied when categorising music pieces into logical groups within a hierarchical structure. Pachet *et al.* analysed in [Pachet00] existing taxonomies of musical genres used by the music industry and Internet-based vendors. Most of the taxonomies used by the music industry are album-oriented, while albums often contain titles of various genres. This has an impact on the structure of the taxonomies. They compared closely taxonomies used by a three Internet-based vendor: allmusic.com (with 531 genres), amazon.com (with 719 genres) and mp3.com (with 430 genres). The taxonomies have no shared structure, and there is no consensus in the genre labels used. Large and commonly used genres like "rock" or "pop" do not have a common definition and do not contain same pieces. Semantics of forming the logical groups may vary even within the given taxonomies. Within same taxonomy, there might be genres formed based on origin of music (e.g., latin music), period of music (e.g., 60's pop) or topic of music (e.g., love song). This is a rather poor description scheme for the automatic musical genre recognition system.

Because of these obvious problems in existing genre taxonomies, Pachet *et al.* proposed in [Pachet00] the definition of new genre taxonomy purely for the purposes of music information retrieval. The proposed taxonomy was designed to be as objective as possible and consistent throughout the taxonomy. The genre descriptors were kept independent from other descriptors used in their metadatabase. Specific descriptors were introduced to describe other musical features, e.g. instrumentation, audience location, and danceability. In addition, the taxonomy was designed to support similarity links between genres. Genres could have relations based on musical features (e.g. rhythm type, instrumentation, or orchestration), inheritance (e.g. rock/alternatif/folk and rock/alternatif/ska), or string-matching (e.g. rock/latino and world/latino).

2.2 Human performance for genre classification

There have not been many studies concerning the human ability to classify music into genres. One of the studies was conducted by Perrot *et al.* [Perrot99]. They used a ten-way forced-choice paradigm (with genres: blues, classical, country, dance, jazz, latin, pop, r&b, rap, and rock) in their study. From each genre they had eight pieces, four pieces with vocals and four without vocals. The pieces were labelled according to the leading web-based record vendors. From each piece, five excerpts with different duration were used in the study (250 ms, 325 ms, 400 ms, 475 ms, and 3000 ms). The subjects for the study were 52 first-year psychology students at the college level. The subjects used around 24 hours per week to listening to music on an average. Excerpts were played in random order and subjects were asked to decide on one of the ten genres for each excerpt.

For three-second samples subjects performed with 70 % accuracy compared to the record companies' original classification. For 250 ms samples the recognition accuracy was still around 40 %. The recognition accuracy for instrumental samples was slightly better for each time interval. These results are quite interesting because they show that humans can in fact recognise musical genres without using any higher-level abstractions, like rhythm. The shortest 250 ms excerpts are by far too short to enable perception of the rhythm, melody or to comprehend the structure of music. This indicates that genre recognition can be done, at least to some extent, only with spectral and timbral features (tone colour).

Soltau presented in his thesis a study of human abilities to recognise four musical genres (classical, pop, rock, and techno) [Soltau97]. A total of 37 subjects with a wide age-range were gathered for the experiment. For three-second samples, the subjects achieved 85 % accuracy on average, which is comparable to the results presented in [Perrot99], although the number of genres used was different. Most of the confusions were made between pop and rock. Musical training was concluded to have positive influence on the recognition accuracy. Listening habits of the subjects were found to have an effect on the genre recognition abilities. The subjects who had been exposed to a wide range of different music had better recognition abilities.

Furthermore, the importance of short-term and long-term characteristics of the music to the genre recognition was studied by properly modifying the samples. The long-term characteristics of the music were destroyed by dividing the music signals in 50 ms frames and playing them in a random order. The short-term characteristics were attenuated by adding white noise to the signal. Based on the short-term characteristics recognition accuracy was 70 % and based on the long-term characteristics it was 80 %. Techno and rock music were recognised better based on the short-term characteristics.

2.3 Automatic musical genre recognition

Recently, there has been more and more research focusing on the automatic musical genre recognition. Some of the most recent contributions are reviewed and the approaches used are compared to give view of a state-of-the-art automatic musical genre recognition to a given taxonomy. In [Aucouturier03], Aucouturier *et al.* provides a quite extensive review on some of the earlier contributions.

All the reviewed methods have three basic stages of a pattern recognition system: frame-based feature extraction, training of the classifier based on examples, and classification. First, the music signal is split into frames, and certain characteristics of the music signal within the frames are represented with a feature vector. The classifier is then trained with examples from each of the musical genres. Finally the trained classifier is used to assign the feature vectors of the test data on the most probable genres.

2.3.1 State-of-the-art

Table 2.1 summarises the reviewed automatic musical genre recognition systems. The meanings of the acronyms used in the table are given in Table 2.2. A direct comparison between systems is impossible due to the different genre taxonomies and the different evaluation database. However, at least some conclusions can be drawn by comparing the amounts of genre labels in taxonomy used and the amount of individual pieces used in evaluations. Since there is so much variation in properties of music even within a genre, a large evaluation database is required for proper evaluation. In general, classical music seems to be easily distinguishable for all the reviewed systems. Rock, pop, and jazz are more likely to be confused with each other because of the similar instrumentation.

The selection of musical genre labels in the reviewed systems was narrow and the genre labels were inconsistent. The systems are rather like a proof of concept than complete and ready musical genre recognition systems. The number of genre labels used varied between 4 and 13, and the selection of genres seems to be rather arbitrary. However, there are also some similarities, most of the system used labels to describe general genres like classical, rock, pop, jazz, and blues. Some systems also used categories to identify speech and background noise in addition to music [Tzanetakis02a, Burred03].

Table 2.1: Summary of genre recognition systems. Number of genres used denoted with G and amount of individual pieces in the evaluation database with DB.

Article	G	DB	Features	Classifiers	Accuracy
Soltau *et al.*, 1998, [Soltau98]	4	360	Cepstrum	HMM ETM-NN	79% 86%
Pye, 2000, [Pye00]	6	350	MFCC	GMM TreeQ	92% 90%
Casey, 2002 [Casey02]	8	many hours	MPEG-7 Low-Level Descriptor (LLD)	HMM	95%
Jiang *et al.*, 2002 [Jiang02]	5	1500	Spectral Contrast	GMM	82%
Tzanetakis *et al.*, 2002, [Tzanetakis02a]	10	1000	Timbral texture Beat histogram Pitch content	GMM	61%
Burred *et al.*, 2003, [Burred03]	13	850	Timbral Beat histogram MPEG-7 LLD Other	GMM	52%
Li *et al.*, 2003, [Li03]	10	1000	Daubechies Wavelet Coefficient Histograms	GMM k-NN LDA SVM	64% 62% 71% 79%
Xu *et al.*, 2003, [Xu03]	4	100	MFCC LPC-derived cepstrum Spectrum power ZCR Beat spectrum	GMM HMM k-NN SVM	88% 88% 79% 93%

Table 2.2: The acronyms used in Table 2.1.

ETM-NN	Explicit Time Modelling with Neural Network	LPC	Linear Prediction Coding
GMM	Gaussian Mixture Model	MFCC	Mel Frequency Cepstral Coefficients
HMM	Hidden Markov Model		
k-NN	k-Nearest Neighbour clasifier	SVM	Support Vector Machine
		TreeQ	Tree-based Vector Quantization
LDA	Linear Discriminant Analysis	ZCR	Zero Crossing Rate

Tzanetakis *et al.* proposed feature sets for representing the timbral texture, rhythmic content, and pitch content of music [Tzanetakis02a]. By combining these features they achieved acceptable recognition accuracy (61 %) for ten musical genres. The recognition accuracy based on just the rhythmic content or the pitch content was quite poor (28% and 23%). However, they were still better than a random recognition and therefore provided at least some information about the musical genre. These results can think to be a well-evaluated baseline for the automatic musical genre recognition. In [Li03], a thorough and extensive comparative study was performed between proposed Daubechies Wavelet Coefficient Histogram (DWCH) feature and the features used in [Tzanetakis02a]. They used the same evaluation database than was used in [Tzanetakis02a] to enable a direct comparison between these systems. Based on their experiment for the features proposed in [Tzanetakis02a], they concluded that the timbral texture is more suitable than rhythmic or pitch content for musical genre recognition. The DWCH improved the recognition accuracy significantly to 79 % (compared to 61 % obtained in [Tzanetakis02a]). Different classification methods were also compared, but the choice of features seemed to be more important than the choice of classifiers. The selected features have a much larger effect to the recognition accuracy than the selected classifiers have. However, the Support Vector Machine (SVM) was observed to be the best classifier for musical genre recognition.

Jiang *et al.* proposed a new feature to represent the relative spectral characteristics of music, spectral contrast feature [Jiang02]. Based on their evaluations, it performed better in the genre recognition task than Mel-Frequency Cepstral Coefficients (MFCC). In general, the selection of genre labels used was quite "easy", and this partly explains the rather high performance (82 %). Burred *et al.* obtained interesting results with a hierarchical classification approach and genre dependent feature sets [Burred03]. In spite of high number of genres used, they achieved an acceptable accuracy (52%) at the lowest level of hierarchy. The rather good recognition performances (above 85 %) reported in [Soltau98, Pye00, Casey02, Xu03] could be partly explained with the narrow evaluation databases used, and partly with the rather "easy" set of genres used.

2.3.2 Features

Three main types of features have been explored in the automatic genre recognition: timbre-related, rhythm-related, and pitch-related. The pitch is a perceptual attribute of sound, defined as the frequency that is obtained by adjusting the frequency of a sine wave of an arbitrary amplitude to match to the target sound [Hartmann96]. The timbre can be defined as a feature of sound that enables us to distinguish it from other sounds with the same pitch, loudness, and duration [Houtsma97].

Timbre-related features

The instrumentation of a music performance is an important factor in genre recognition. The timbre of constituent sound sources is reflected in the spectral distribution of a music signal. Thus most of the features used in reviewed systems describe the spectral distribution of a signal.

Cepstrum coefficients, used in [Soltau98, Xu03], and MFCC, used in [Pye00, Jiang02, Tzanetakis02a, Burred03, Xu03], are a compact way to represent the coarse shape of the spectrum. The Cepstrum is defined as the inverse Fourier transform of the logarithm of the magnitude spectrum. Other definition for the cepstrum is based on a Linear Prediction Coding (LPC), where the signal is approximated as a linear combination of past signal samples with an all-pole filter. This source-model is partially valid for the periodic signals produced by instruments or vocals. The poles of this all-pole filter correspond the peaks in the power spectrum of the signal. In MFCC, the frequencies are scaled non-linearly to the Mel-frequency scale before taking the inverse Fourier transform. The Mel-frequency scale is used to approximate the non-linear frequency resolution of the human ear by using the linear scale at the low frequencies and the logarithmic scale at the higher frequencies [Houtsma95]. MFCC will be defined in detail in Chapter 5. In order to model spectral variation of the data, the differential of the feature values between consecutive frames can be added to the feature set (delta coefficients) (as used in [Pye00]). MFCCs are widely used in many audio signal classification applications. For instance, they have proven to be useful in speech recognition [Rabiner93], speaker recognition [Reynolds95], and in musical instrument recognition [Eronen03a].

Much of the music available is in a compressed format nowadays. The computation of MFCC and the compression of audio (MPEG-1 layer III) share a few common steps: dividing the signal into Mel-frequency scale subbands, and decorrelating the coefficients with the Discrete Cosine Transform (DCT). In order to avoid fully decompressing the audio before extracting MFCCs, a computationally efficient way of deriving MFCC-like features, MP3CEP, from a partially decompressed MPEG audio has been proposed [Pye00]. Approximately six times faster feature extraction was achieved for MP3CEP with an acceptable decrease in recognition accuracy compared to the MFCC.

The MPEG-7 standard [MPE01] includes standardised spectrum-based features for sound recognition. The use of these for automatic musical genre recognition has been studied in [Casey02, Burred03]. The extraction method of the MPEG-7 audio

features closely resembles the extraction method of MFCC with few major differences. The MPEG-7 audio features use the logarithmic frequency scale as opposed to the Mel-frequency scale used in MFCC. The dimensionality of the feature vectors is reduced in the MFCC feature extraction with the DCT using same DCT bases for all the audio classes. In MPEG-7 audio feature extraction, the dimensionality is reduced with the Principal Component Analysis (PCA) performed on distinct PCA space derived from the training examples of each audio class. The MPEG-7 audio features have shown to have performance comparable to MFCC in more general sound classification tasks [Xiong03].

The MFCC averages the spectral energy distribution in each subband and thus may produce similar average spectral characteristics for two different spectra. A Spectral Contrast feature has been proposed to represent the relative distribution of the spectral energy instead of the average spectral envelope [Jiang02]. The strength of spectral peak and spectral valley, and their difference are considered separately in each octave-scale subband. This feature roughly reflects the relative distribution of the harmonic and non-harmonic components in the spectrum, since often the harmonic components correspond to the strong spectral peaks and non-harmonic components to the spectral valleys. In the experiments made by Jiang *et al.*, the Spectral Contrast feature achieved better performance than the MFCC in the musical genre recognition [Jiang02].

The wavelet transform can be used to provide simultaneously a good frequency and time resolutions. With the transform, the signal is represented as a linear combination of the scaled and shifted versions of the wavelet function. A Daubechies Wavelet Coefficient Histogram (DWCH) has been proposed to represent the local and global information of the music signals simultaneously [Li03]. The set of subband signals are produced by wavelet decomposition and a histogram of the wavelet coefficients at each subband is constructed. The wavelet decomposition of the signal highly resembles the octave-band decomposition of audio signal [Li00]. The histogram provides an approximation of the waveform variations at each subband. The average, the variance, and the skewness of the histograms are used as features along with the energy of each subband. The experiments showed that the DWCH improved the recognition accuracy significantly compared to the feature set used to represent the timbral texture, rhythmic content, and pitch content in [Tzanetakis02a].

In addition to the previously presented features, some characteristics of the spectrum can be represented with Spectral Centroid (used in [Tzanetakis02a, Burred03]), Spectral Flux (used in [Tzanetakis02a, Burred03]), Spectral Roll-Off (used in [Tzanetakis02a, Burred03]), Low Energy (used in [Tzanetakis02a, Burred03]), and Zero Crossing Rate (used in [Tzanetakis02a, Xu03, Burred03]). *Spectral Centroid* measures the brightness of the signal and is defined as the balancing point of the magnitude spectrum. *Spectral Flux* is used to represent the degree of change in the spectral shape, and it is defined as the frame-to-frame magnitude spectral difference. *Spectral Roll-Off* is defined as the frequency below that a certain amount (e.g. 85 %) of the power spectrum resides, and especially percussive sounds and other transients can be detected with this feature. *Low Energy* measures the amplitude distribution of the signal, and it is defined as percentage of frames to have energy less than

the average energy over the whole signal. *Zero Crossing Rate* (ZCR) measures the number of time-domain zero crossings within a frame. In [Tzanetakis02a], low order statistics (mean, variance, skewness, and kurtosis) are computed for these features and for MFCC over a larger analysis segment (30 seconds).

Rhythm-related features

The rhythmic structure of music gives valuable information about the musical genre. The complexity of the beat can be used to distinguish, for example, between straight rock music and rhythmically more complex jazz music. Thus, besides just taking into account the global timbre, the rhythm-related features have also been used in musical genre recognition systems [Soltau98, Tzanetakis02a, Burred03, Xu03].

Tzanetakis *et al.* [Tzanetakis02a] proposed the concept of beat histogram, a curve describing beat strength as a function of tempo values, to be used to gain information about the complexity of the beat in the music. The regularity of the rhythm, the relation of the main beat to the subbeats and the relative strength of subbeats to the main beat, is used as one of the features in their musical genre recognition system. The Discrete Wavelet Transform (DWT) is used to divide the signal into octave bands and, for each band, full-wave rectification, low pass filtering, downsampling and mean removal are performed in order to extract an envelope. The envelopes of each band are summed up and the autocorrelation is calculated to capture the periodicities in the signal's envelope. The dominant peaks in autocorrelation function are accumulated over the whole audio signal into a beat histogram.

In [Burred03], the beat histogram is extracted with a method proposed in [Scheirer98]. The music signal between 200 Hz and 3200 Hz is divided into six subbands with a filterbank. In each subband the envelope is extracted, a first-order difference function is calculated, and the signal is half-wave rectified. The periodic modulation in each subband is examined with a filterbank of comb filter resonators in order to produce an overall tempo estimate. The beat histogram is collected over time with this analysis. The beat strength is captured by calculating mean, standard deviation of the derivative, skewness, kurtosis, and entropy of the obtained beat histogram. The high rhythmic regularity shows in the histogram with periodically spaced strong peaks, which can be detected as clear peaks in the normalised autocorrelation of the histogram.

In [Xu03], the beat spectrum is formed directly from the extracted spectrum-based features. A similarity matrix is formed by calculating similarity with a distance measure between all pairwise combinations of features. The final beat spectrum is obtained using autocorrelation for this similarity matrix.

Soltau *et al.* proposed an architecture called Explicit Time Modelling with Neural Networks (ETM-NN) to be used to extract information about the temporal structure of a musical performance [Soltau98]. An autoassociative neural network is trained for cepstral coefficients extracted from the music signal. The activation strength of the hidden units in this neural network is used to determine the most significant abstract musical events within each frame. The temporal structure of these events

is represented with unigram-counts, bigram-counts, trigram-counts, event durations, and statistics of event activation. A second neural network is trained to recognise the musical genre based on the temporal pattern of the abstract events.

Pitch-related features

The pitch content of music can be used to characterise particular musical genres. For example, jazz music tends to have more chord changes, classical music has a large variability of harmonic properties and in rock music high-pitched notes are mainly absent and they seldom exhibit a high degree of harmonic variation.

The pitch histogram was proposed by Tzanetakis *et al.* as a way of representing the pitch content of music signals both in symbolic and audio form [Tzanetakis02a]. For polyphonic audio they used a multipitch detection algorithm presented in [Tolonen00] to extract the pitch content. The multipitch detection algorithm is defined as follows. The signal is first decomposed into two frequency bands (below and above 1000 Hz) and the amplitude envelope is extracted for both bands. A summary autocorrelation (SACF) is computed for these envelopes. Autocorrelation is enhanced (ESACF) by subtracting integer multiples of the peak frequencies.

The three most dominant peaks (which correspond here to the pitches in the music signal) of the ESACF are accumulated over the whole audio signal into a pitch histogram. This so-called unfolded histogram is useful for determining the pitch range. In addition, a folded pitch histogram is created, by mapping all notes to a single octave. This yields a representation similar to Wakefield's chromagram [Wakefield99]. The folded histogram is used to describe the overall harmonic content of the music. The following features are extracted from the unfolded and folded pitch histograms:

- Period of maximum peak of the unfolded histogram, which corresponds to the dominant octave range (e.g. flute pieces have higher octave range than bass pieces).
- Period of maximum peak of the folded histogram, which corresponds to the most common pitch class (all pitches existing in an octave relationships).
- Amplitude of maximum peak of the folded histogram, which corresponds to the frequency of main pitch class occurrence.
- Interval between the two highest peaks of the folded histogram, which corresponds to the main tonal interval relation, whether piece have simple (fifth or fourth interval) or complex harmonic structure.
- Overall sum of the histogram, which measures the strength of the pitch detection.

Interestingly, there is only a small decrease in the recognition accuracy between the use of real pitch information taken from the symbolic data (MIDI) and the use of

multipitch detection for the audio synthesised directly from the same symbolic data [Tzanetakis02b]. However, the genre recognition accuracy for audio signals based on just the pitch content was still quite poor (23 %) [Tzanetakis02a].

2.3.3 Classification

Two different types of classification approaches have been used in the automatic musical genre recognition systems. The most common approach is to classify each frame separately, and to combine these classification results over an analysis segment in order to get the global classification result. The second approach is to also take into account the temporal relationships between frames in the classifier. For humans, order of frames is an important property. The recognition accuracy has been observed to decrease from 85 % to 70 % when the order of frames was mixed [Soltau97].

One of the simplest classifiers is the k-Nearest Neighbour Classifier (k-NN), used in [Li03, Tzanetakis02a, Xu03]. The distance between the tested feature vector and all the training vectors from different classes is measured. The classification is done according to the k nearest training vectors. The Gaussian Mixture Model (GMM) is used in [Pye00, Jiang02, Tzanetakis02a, Burred03, Li03, Xu03]. Based on the available training data, the distributions of feature values in different musical genres are modelled as a weighted sum of Gaussian density functions. This mixture is then used to determine the probability of a test feature vector to belong to a particular musical genre. The mathematical definitions and descriptions of the k-NN and the GMM will be given in Chapter 5.

Tree-based Vector Quantization (TreeQ) is used in [Pye00]. Instead of modelling the class densities, the Vector Quantizer models the discrimination function between classes defined by a set of labelled codebook vectors. A quantization tree is formed to partition the feature space into regions with maximally different class populations. The tree is used to form a histogram template for a musical genre and the classification is done by matching template histograms of training data to the histograms of the test data. The classification can also be done with a Feed-forward neural network, as used in [Soltau98]. A neural network is trained with examples from different classes so as to map the high-dimensional space of feature vectors onto the different classes. The Linear Discriminant Analysis (LDA) finds a linear transformation for the feature vectors that best discriminates them among classes. The classification is done in this transformed feature space with some distance metric, e.g., Euclidean distance as in [Li03].

In addition to the previous multi-class learning methods, a binary classification approach using the Support Vector Machine (SVM) is studied in [Li03, Xu03]. Feature vectors are first non-linearly mapped into a new feature space and a hyperplane is then searched in the new feature space to separate the data points of the classes with a maximum margin. In [Li03], the SVM is extended into multi-class classification with one-versus-the-rest, pairwise comparison, and multi-class objective functions. In the one-versus-the-rest method, binary classifiers are trained to separate one class

from rest of the classes. The multi-class classification is then carried out according to the maximal output of these binary classifiers. In pairwise comparison, a classifier is trained for each possible pair of classes and the unknown observation is assigned to the class getting the highest number of classification "votes" among all the classifiers. In the multi-class objective-function method, the objective function of a binary SVM is directly modified to allow the simultaneous computation of a multi-class classifier. In [Xu03], the SVM is extended into multi-class classification with a hierarchical classification approach.

The only approach to also take into account the temporal order of frames is the Hidden Markov Model (HMM), used in [Soltau98, Casey02, Xu03]. The HMM can be considered to consist of several GMMs and the probabilities describing the transitions between them. The definition of the HMM will be given in Chapter 5.

Hierarchical classification

Most of the reviewed systems classify music from all the genres according to the same genre independent feature set [Soltau98, Pye00, Casey02, Jiang02, Tzanetakis02a, Li03]. However, it is clear that different genres have different classification criteria, thus some features are more suitable than others in separating some genres from others. For example, the beat strength is likely to perform better in discriminating between classical and pop than between rock and pop music. A hierarchical classification scheme enables us to use different features for different genres. It is also more likely to produce more acceptable misclassifications within higher-level genre.

In [Xu03], music is first classified into two metaclasses (pop/classical or rock/jazz) according to the beat spectrum and LPC-derived cepstrum. After this the pop /classical music is further classified (into pop or classical) according to LPC-derived cepstrum, spectrum power, and MFCCs. The rock/jazz music is classified (into rock or jazz) according to ZCR and MFCCs.

Burred *et al.* compared the hierarchical classification scheme with direct classification [Burred03]. In hierarchical classification, signals were first classified into speech, music and background noise. Music was then classified with a three-level structure, first classifying it into classical and non-classical music, and after that classifying it into chamber or orchestral music, or into rock, electronic/pop, and jazz/blues. At the third level, music was classified further, e.g. rock into hard rock or soft rock. They extracted a wide selection of different features and used a feature selection algorithm to select the best performing set of features for each particular subgenre classification. As a result, they achieved very similar recognition accuracies for hierarchical and direct classification schemes.

2.4 Related work

2.4.1 Musical signal processing

Many ideas and solutions presented in music signal processing in general are also relevant for the automatic musical genre recognition. Some of the most relevant fields are presented in the following.

Music similarity

As more and more music is available in digital libraries, new query mechanisms to find music from these libraries are becoming necessary. For instance, a user may want to search the library for similar or almost similar music to the given piece (Query-By-Example). Many dimensions of music are perceptually important for characterising and for making judgements about the similarity, including tempo, rhythm, melody, instrumentation, voice qualities, and musical genre.

Foote proposed in [Foote97] a music indexing system based on histograms of MFCC features derived from discriminatively trained vector quantizer. In [Logan01], a signature was formed for each piece based on the clustered spectral features. This signature was then compared to find similar signatures.

Welsh *et al.* proposed a system capable of performing similarity queries in a large digital library [Welsh99]. They used frequency histograms, tonal transitions, relative noise level, volume, and tempo as features. Rhythmic information was extracted with the algorithm proposed in [Scheirer98] (described earlier). Similarity was determined with the Euclidean distance between the feature vectors. The effectiveness of the system was evaluated with musical genre queries. One hundred manually labelled albums from seven musical genres (classical, electronic, folk, indie, pop, rock, and soul) were used in the evaluations. On average, they achieved quite poor query accuracies (average performance being around 39 %).

Music segmentation

In music segmentation, parts of a music signal are discriminated based on pitch changes [Raphael99], timbres (transients and steady parts) [Rossignol98], instruments [Herrera00], vocals [Berenzweig01], or musical structures (verse and chorus) [Foote00]. In general, two different approaches can be observed in segmentation systems. In the first one, features are extracted frame-by-frame and the segment boundaries are detected by looking for abrupt changes in the feature trajectories. In the other one, the feature trajectories are matched with a model (HMM) of each possible type of segment to allow more intelligent segmentation [Aucouturier01].

Vocals and the singer' voice are an important aspect of music for humans when identifying and classifying it. Berenzweig *et al.* have proposed a detector to locate the segments of music signal with vocals [Berenzweig01]. They used a traditional HMM-based speech recognition system as a detector for speech-like sounds (vocals), and achieved approximately 80 % accuracy at the frame level.

Rhythm

Rhythmic content of music is an important factor when determining the musical genre. Jazz music has usually a more complex rhythmic structure than e.g. rock music. Therefore beat tracking and rhythmic pattern detection are very interesting fields with respect to musical genre recognition.

Most humans do not have any difficulties to tap their foot in time with music. However, this has proven to be a challenging problem for automatic systems, since it requires a real understanding of the rhythm by finding the fine hierarchical structure of the timing relationships in the music. This hierarchical structure has several levels: tatum is the lowest metrical level, bar, measure, and beat are at the higher level. The different levels of rhythmic measurements have been studied in [Dixon01, Seppänen01, Klapuri03].

Paulus *et al.* proposed a system to measure the similarity of two arbitrary rhythmic patterns [Paulus02]. Based on a probabilistic musical meter estimation music signal is first segmented into patterns. The fluctuation of loudness and brightness is then modelled within the pattern and the dynamic time warping is used to align the patterns. In the evaluations they achieved quite promising results.

2.4.2 Speech/music discrimination and general audio classification

Obviously, a music information retrieval system cannot give usable results when the input signal does not represent music. It does not make sense to recognise musical genre for speech signal or to recognise speech in music. Therefore the ability to distinguish between music and the speech is a valuable front-end for either of the systems. It is also essential for generic sound-analysis systems. The front-end discriminators pass the signals on to the specific back-end audio classification system, which is designed to handle especially music or speech signals. In [Saunders96], a good discrimination rate (98 %) between music and speech is achieved by simply thresholding the average ZCR and energy features. In [Scheirer97], multiple features and statistical pattern recognition approaches were carefully evaluated for the task.

Zhang *et al.* proposed a heuristic rule-based system for the real-time segmentation of audio signals from TV programs and movies [Zhang01]. The segmentation is performed based on the time-varying properties of simple features, including the energy function, the average ZCR, the fundamental frequency and the spectral peak tracks. In their experiments they achieved a good accuracy (above 90 %) for basic audio categories, e.g., like pure speech, pure music, speech with music on the background, and sound effects.

2.4.3 Speaker recognition and verification

Various algorithms used in speaker recognition are applicable to MIR-related tasks, too. Speaker recognition and verification have a various security applications, where

it is used to recognise the person or to determine whether the person belongs to a list of approved people. The problem of speaker verification shares some issues with instrument detection, discussed later on this thesis.

As an example, Reynolds *et al.* presented a robust text-independent speaker identification system [Reynolds95]. MFCCs were used to represent the spectral content of speech and the speaker identity was modelled by representing general speaker dependent spectral shapes with individual Gaussian components of a GMM. The proposed framework is quite well operative in music classification as such.

3 Music database

An appropriate database is needed to evaluate a classification system. Since there are diverse types of music signals, the database has to be carefully designed. The objective was to form a general-purpose music database to be used in different areas of MIR. Altogether 505 musical pieces were collected to get a representative set of pieces from different musical genres. Every piece in the database was thoroughly annotated, and part of the annotated information remains unused in this thesis. The database consists of the actual acoustic signals and their manually annotated content in textual format.

3.1 Collecting the database

The aim was to collect a database that would adequately cover different musical genres. To make sure that we take into account all major musical genres, a hierarchical categorisation system had to be established for musical genres. Seven primary genres: classical, electronic/dance, hip hop, jazz/blues, rock/pop, soul/RnB/funk, and world/folk (explained in detail in the section 3.4), were considered when collecting pieces for the database. The relative amount of the pieces for each genre were approximated based on what people nowadays listen to. Majority of the pieces were collected during the year 2001 and during the summer of 2002 the database was updated with some pieces.

Representative pieces have to be carefully selected due to the somewhat large variation within each genre. Some compromises had to be made in order to keep the database as segmented as possible. Pieces clearly exploiting characteristic elements from various musical genres at the same time, for instance heavy rock music played with cellos, were mainly excluded. Appendix B lists all the pieces in the music database.

3.2 Format of acoustic data

All the pieces were transferred directly from CDs in digital form and the Pulse-Code Modulated (PCM) signals were stored with 44.1 kHz sampling rate and a 16 -bit resolution. The pieces were converted into monophonic form by averaging the left and the right channel into one channel. The audio files were stored along their textual annotation files.

InstrumentRep	=	vocal;guitar	@	61	-	71	#	text
label		value		start (s)		end (s)		comment
compulsory			optional					

Figure 3.1: Annotation file format.

3.3 Annotations

Pieces were annotated manually by a third party person, not working in the development of the system, but having a good basic knowledge of music theory and music in general.

3.3.1 File format

The annotation file format is identical to one used by Peltonen in [Peltonen01]. This enabled the use of database access tools developed during that research. These tools allowed easy searching from the database using different search criteria, e.g. musical genre.

The annotation file consists of annotation fields. Each field represents a single feature and occupies one line in the annotation file. The format of the field is presented in Figure 3.1. *Label* represents the type of the annotated data and *value* is a numeric or verbal description of the data. The time interval is an optional parameter used to determine a time segment where the value is valid. The start and end points for the interval are annotated in seconds with respect to the piece beginning. Absence of this interval data denotes that value is valid for the entire audio file. A hash mark is used to denote a comment at the beginning of each comment, and the text after the hash mark will be ignored when making queries to the database.

3.3.2 Annotation fields

An example of an annotation file is presented in Figure 3.2. The following labels were used in the annotations:

- **Artist**, the performer.
- **Title**, the title of the piece.
- **Album**, the name of the album from which the piece was taken.
- **Genre**, the musical genre. If the piece clearly belongs to two different higher-level genres, label **Genre2** can be used in addition. This was the case only for one piece in the database. Genre was annotated hierarchically with three levels. Every level adds the accuracy of the description of genre. Only the first

```
Artist= Kool and the gang
Title= Funky stuff
Album= The best of
Genre= soul/rnb/funk; funk
Instruments= electric guitar, bass guitar, brass section, saxophones,
             drums, male vocals
Drums= set; comp
Acoustic= both acoustic and electronic
Vocals= yes
Number of players= band
Key= major
Time signature= 4/4
Tempo= 100
Recording environment= studio
Recording year= 1973
Composer= Kool & the gang
Media= CD
Representative= sample @ 57-105
InstrumentRep= male vocals;drums,bass guitar,electric guitar @ 70-80
Length= 187
DrumsOn=  sample @ 0 - 4.1375
DrumsOff= sample @ 4.1375 - 7.765
DrumsOn=  sample @ 7.765 - 182.7181
```

Figure 3.2: Example of an annotation file.

level was required. (notation: first level genre; second level genre, third level genre)

- **Instruments**, a list of the instruments used, in the order of decreasing dominance.
- **Drums**, the type of drums used in the piece. (notation: type of drums; type of playing. Possible values for the former: set, percussion only, orchestral and for the latter: partial, comp)
- **Acoustic**, the degree of acoustic instruments in the piece. (possible values: strictly, mostly, both acoustic and electronic, electronic)
- **Vocals**, the presence of vocals in the piece. (possible values: yes, none, speech)
- **Number of players**, the estimate of the number of players described verbally. (possible values: solo, duet, band, big band, choir, orchestra, orchestra and choir, ensemble, symphonic orchestra)
- **Key**, the key signature defines the diatonic scale used in piece. (possible values: minor, major, modern).
- **Time signature**, a predominant time signature. Describes how many and what kind of notes there are per measure. (notation: the number of notes per measure / what kind of note, possible values: e.g. 3/4, 4/4).

- **Tempo**, an average tempo of the piece. If tempo clearly changes within the piece, different tempo values were annotated for the corresponding intervals.
- **Recording environment**, the environment where the recording was done. (possible values: studio, live)
- **Recording year**
- **Composer**, the composer of the piece.
- **Media**, media from which the piece was transferred to the database. Currently all are from CDs.
- **Representative**, about an one-minute representative excerpt from the piece, starting and ending points annotated in seconds.
- **Instrument Representative**, a ten-second excerpt within the representative excerpt, which is as homogeneous as possible with regard to the instruments involved. Instruments used within the excerpt were annotated. Instruments that played the melody and the accompanying instruments are annotated separately in the order of decreasing dominance. (notation: melody instruments; accompanying instruments @ interval)
- **Length**, the length of the whole piece in seconds rounded down to the nearest integer number.
- **DrumsOn** and **DrumsOff**, intervals with and without drum instruments.

These fields have been annotated for all pieces with few exceptions. The instrument representative excerpt was annotated for 98 % of the pieces and the time segments with and without drums were annotated for 79 % of the pieces. There are also some minor defects with other fields, because the structure of the annotation file was changed at an early stage and some of the old annotations remained. About 6 % of the pieces are lacking the proper "instruments"-field and "acoustic"-field annotations.

Representative Excerpt

For the genre classification purposes, approximately one-minute interval within each piece was manually annotated to represent it. Interval was annotated for all the pieces in the database and was used to represent the piece in simulations.

Instrument Representative Excerpt

In addition to the representative excerpt, a section of exactly ten-seconds within the representative part was annotated, which is homogeneous from the point of view of instruments used. No big structural changes were allowed in music during the chosen interval. Furthermore, the instruments used within the excerpt were

annotated to give an exact view of the used instrumentation. Instruments used to play the melody and the accompanying instruments were annotated separately in the order of decreasing dominance. The interval was used in the instrument detection simulations.

Segments with drum presence

"The drum presence" was defined to include the requirement that the drum is played in a rhythmic role. However, some clear individual drum strokes were also transcribed. These were excluded later on in order to get explicit evaluation data. Presence of drums was decided based on following guidelines:

- Drum instrument is used to produce a rhythm pattern that repeats itself. For example, cymbal or kettledrum can be used as a sort of an effect instrument, in these cases it was not consider as a drum instrument.
- Drum pattern is repeated a few times within a five seconds.

A special care had to take with classical music. Kettledrum is used in many classical pieces, but not always in a rhythmic role. Kettledrum has to play a clear repeating pattern to be accepted as a drum instrument, not just to be used to emphasise a certain part of the piece. In modern electronic dance music, the amplitude of a drum track may increase gradually. In these cases, the boundary was chosen based on when a human listener first perceived the presence of the drums.

3.4 Musical genre hierarchy

In order to make a proper genre hierarchy, we explored different hierarchical music catalogues used in some of the Internet based record vendors (allmusic.com, amazon.com, audiogalaxy.com, and mp3.com). First we mapped all the genres and looked the relationships between them. As it was previously discussed in section 2.1, musical genre is not a clear concept and genres have rather fuzzy boundaries. For this reason, the objective was to have as few higher-level genres as possible in the genre hierarchy and at the same time to make manual classification according to this hierarchy as straightforward as possible. To achieve this, some of the closest and clearly overlapping genres had to be combined. Minor genres were merged with the closest broader genres. Eventually, we ended up with seven primary genres; classical, electronic/dance, hip hop, jazz/blues, rock/pop, soul/RnB/funk and world/folk. We had to make some compromises while forming the hierarchy. However, we considered that the classification, at least for humans, with these genre categories should be rather unambiguous.

Genre hierarchy consists of three genre-levels. Under the seven primary genres are two levels of subgenres. Each of these subgenre levels adds the accuracy of the description. The two highest levels of the genre hierarchy are presented in Figure 3.3. Appendix A presents the complete genre hierarchy.

- **classical** (107)
  - chamber music (24)
  - classical general (20)
  - crossover
  - film music
  - general instrumental
  - solo instruments (12)
  - symphonic (10)
  - vocal (6)
- **electronic / dance** (71)
  - ambient (7)
  - breakbeat/breaks/drum'n'bass (11)
  - dance (9)
  - electronica
  - house (11)
  - industrial (5)
  - techno / trance (22)
- **hip hop** (37)
- **jazz / blues** (96)
  - blues (32)
  - jazz (62)
- **rock / pop** (121)
  - alternative (1)
  - country (12)
  - easy listening (6)
  - metal (12)
  - pop (30)
  - rock (55)
  - rock-n-roll oldies
- **soul / RnB / funk** (56)
  - funk (13)
  - gospel (6)
  - RnB (15)
  - rhythm and blues (4)
  - soul (18)
- **world / folk** (17)
  - African
  - Asia
  - Caribbean
  - Celtic
  - ceremonial/chants
  - European
  - folk (8)
  - Latin American (7)
  - Mediterranean
  - Middle East
  - North American
  - Northern Europe
  - Oceania
  - old dance music
  - Scandinavian
  - South Pacific
  - World Pacific
  - World Beat
  - World Fusion
  - World Traditions

Figure 3.3: Genre hierarchy with first two of the levels used. The numbers in parentheses indicate the amount of pieces from particular genre in the music database.

3.4.1 Descriptions of musical genres

In order to give some idea what is characteristic for the seven primary genre labels used, a short description will be given about each of them. These descriptions are superficial and do not even try to be exact. Descriptions are based on the discussions presented in [Clarke89, Crowther95, Sadie01].

Classical music. In the strictly historical sense, classical music refers to music written during the so-called Classical period around the years 1770-1830. However, within this thesis, we will use the term in a much broader sense to include all the music derived from the same educated European musical tradition. Classical music is very often composed rather precisely for every instrument used. Any suitable group of musicians can perform it; still only little nuances of interpretation are left for performers. Classical music is usually performed by an orchestra, a large group of musicians playing a variety of different instruments, or by a chamber group, a smaller group of musicians.

Electronic / dance music. In the 1970's, avant-garde electronic musical experimentations evolved by bands like Kraftwerk and introduced the heavy use of electronic instruments into popular music. The development of synthesizers made it possible for virtually everybody to create this new type of music. The heavy use of all kinds of electronic instruments (e.g. synthesizers, drum machines, samplers, computers, effects, and record players) is characteristic for this music. Most of the electronic music is based on a strong rhythmic structure and a strong beat. Although this genre is relatively new, it has been fragmented into an infinite number of subgenres.

Hip hop music. The hip hop culture evolved from the inner-city Afro-American communities in the late 1970's and early 1980's. Part of this culture is to rhyme and rap over the beats and breaks played from records (disco, funk, jazz, rock, or soul) by a disc jockey or specially made music with electronic instruments (drum machines, samplers, and synthesizers). The way of performing rhythmic speech phrases over the beat is characteristic for the hip hop music. The hip hop music uses influences from many different musical genres, but still remaining the rhyming as a vital part of the act.

Jazz / blues music. The blues is a traditional folk music that originated among Afro-Americans at the beginning of the 20th century, mixing African musical traditions with European folk traditions. One of the defining characteristics of the blues music is the use of *blue notes*. It is a scale resembling the minor scale with flatted (lowered) notes on the third, the seventh and the fifth scale-degree, thus producing a melancholy sound [Sadie01]. The structure of the pieces is often uniform. For example, a form (twelve-bar) where three four-measure long phrases are repeated is widely used in blues [Sadie01]. Jazz grew out of the blues during the beginning of

the 20th century, and it emphasises improvisation. A propulsive rhythm, melodic freedom, and improvisation solos are characteristic for jazz music.

Rock / pop music. This category is a bit problematic, since both rock and pop are vague terms describing the majority of modern popular music and thus creating a too wide and meaningless category. Nevertheless, to have a compact and unambiguous set of higher-level genres these genre labels have to be used. Originally, rock music evolved from country and blues music. A characteristic for rock music is the heavy use of electric guitars and a strong beat. Nowadays rock music has fragmented into subgenres, and many of these subgenres have further grown a genre of their own. Characteristic for pop music are short and simple melodic pieces having some catching tune to make them easily memorable.

Soul / RnB / funk music. Rhythm and blues (R&B) is, in its wide sense, an "umbrella category" to hold majority of the "black music", which grew out of the urbanisation of the blues. Almost all R&B music made in the 1960's and 1970's can be labelled as soul. Vocal intensity is characteristic for this music. Funk grew out of the soul music by adding strong groove and influences from rock music on it. Contemporary R&B, denoted here as RnB, is nowadays more and more close to hip hop and pop music. However, there are still some characteristic features like smooth vocals with a bouncy beat.

World / folk music. This category was introduced in order to have a complete set of higher-level genres. It will not be used in the automatic genre classification tasks, because of its miscellaneous nature. Virtually any music, which does not originate from Western popular music traditions, can be labelled as world music. World music use native instrumentation and traditional styles, at least to some extent. Traditional Western music is called folk music.

3.5 Statistics of the database

Table 3.1 shows the number of pieces from different musical genres in the database. There is a quite representative set of pieces for classical, rock/pop and jazz/blues. Amount of hip hop pieces is rather small, but it is also a quite narrow genre in the real life, too. A more detailed view of the amount of different genres in the database was presented in Figure 3.3.

Table 3.2 shows the occurrence frequencies of certain instruments in the ten-second "instrument representative" field of the annotations. These are shown separately for each higher-level genre. In order to be accounted for the statistics, the instrument has to be the most predominant or the second most predominant accompanying instrument or one of the melody instruments. The instrument class "bowed" holds all the instruments that are played with a bow, e.g. the violin and the cello. Due to the different instrumentation in the musical genres these instruments do not occur

Table 3.1: Amount of pieces from the different musical genres in the music database.

Musical genre	%	#
classical	21 %	107
electronic / dance	14 %	71
hip hop	7 %	37
jazz / blues	19 %	96
rock / pop	24 %	121
soul / rnb / funk	11 %	56
world / folk	3 %	17
Total (pieces)	505 pieces	
Total (duration)	1d:13h:22m:44s	

Table 3.2: Amounts (number of pieces) of certain instruments annotated into "instrument representative" field.

Musical genre	bowed	electric guitar	piano	saxophone	vocal
classical	70	-	25	-	6
electronic / dance	-	1	-	1	28
hip hop	-	9	1	-	37
jazz / blues	2	23	35	26	29
rock / pop	6	46	14	1	105
soul / rnb / funk	2	11	7	5	49
world / folk	6	-	4	2	7
Total (pieces)	**86**	**90**	**86**	**35**	**261**

Table 3.3: Statistics of the drum presence evaluation database.

Musical genre	%	#	Drums absent	Drums present
classical	27%	107	89%	11%
electronic / dance	7%	27	18%	82%
hip hop	3%	12	5%	95%
jazz / blues	16%	64	10%	90%
rock / pop	29%	115	11%	89%
soul / rnb / funk	11%	45	8%	92%
world / folk	7%	27	56%	44%
Total (pieces)		**397**	**32%**	**68%**

evenly among the musical genres. This had to be taken into consideration when using the database for the evaluations.

Table 3.3 presents the annotated drum presence in the music database. The drums were present only seldom in classical music. For the rest of the genres drums were present most of the time, only exception being world/folk with equal amounts of segments with and without drums. However, this kind of imbalance was expected, since drums are a basic element in Western music.

4 Listening experiment

Musical genre is probably the most often used music descriptor. Distinguishing between musical genres is typically a rather trivial task for humans. Humans can do a coarse classification already within few hundred milliseconds (see Section 2.2). A good real-world setting to observe this capability in action is to scan the radio dial. We can quite easily detect the musical genre and make a decision what to listen to. This chapter presents a listening experiment to determine the level of human accuracy in recognising musical genres. The experiment results can be used as a baseline for the performance evaluation of the developed automatic musical genre recognition system.

4.1 Experiment design

The listening experiment has to be carefully designed in order to minimise the errors and to be able to draw reasonable conclusions from the outcome of the experiment [Levitin99].

4.1.1 Objectives and assumptions

The main goal of this experiment was to investigate the accuracy of human musical genre recognition. A secondary goal was to study the relationship between the length of a heard excerpt and the recognition accuracy. In Chapter 6, human abilities are compared with the developed automatic musical genre recognition system. Overall performance and classification confusions are compared in order to better understand the differences between humans and the computational recognition system.

Since musical genre is an ambiguous concept, the number of genre categories had to be limited. Only six higher-level genres were chosen for this experiment. These were classical, electronic/dance, hip hop, jazz/blues, rock/pop, soul/RnB/funk. By using as few genre categories as possible we tried to minimise the possibility of confusion and ambiguity. Although all musical genres overlap each other to some extent, the mentioned primary genres are much better segregated. A more detailed description of genre hierarchy was presented in Section 3.4.

Some assumptions were made concerning the test subjects and the human perception of musical genres. Human capability to recognise musical genres was considered relatively independent of sex, age and educational background of the test subjects, thus

we did not try to balance these out in our test population. Contrary to [Perrot99], no special attempt was made to find differences between recognition accuracies for segments with and without vocals.

4.1.2 Method

The experiment was a six-way forced-choice classification task. One sample was presented at a time and classified alone. Although a few example samples from each musical genre were presented in the familiarisation stage, the method primarily relied on the subjects' internal reference about the musical genres. The internal reference is not constant over the time; it is constantly updated, even during the experiment itself. However, the method is very suitable for a large number of stimuli required to perform a reliable experiment with such a variable material as music.

Test subjects were evenly divided in two separate groups each including the same number of stimuli and identical musical genres, but having separate sets of test pieces. This allowed the use of a larger amount of samples in experiment while keeping the total duration of the experiment tolerable.

4.1.3 Stimuli

Selecting test pieces and stimuli

Test pieces were chosen from the music database described in Chapter 3. For each of the six genres, ten pieces were randomly selected leading to a total of 60 pieces for both test groups. Special attention had to be paid to genres having a strong lower-level hierarchy (e.g. jazz/blues) to ensure balanced groups and to have as representative samples as possible. In these cases, a representative amount of pieces was selected randomly from each subgenre. The amounts of test samples from different subgenres within the primary categories are shown in Table 4.1.

Four excerpts with different durations (250 ms, 500 ms, 1500 ms, and 5000 ms) were extracted from every piece selected. For every piece in the database, there is a representative one-minute long interval, selected manually to represent the whole piece as well as possible. These manually annotated segments were used while selecting stimuli. Starting point of the stimulus was randomly selected within this segment separately for each subject and duration. This procedure may also produce rather silent or indistinct excerpts. However, these effects belong to the randomisation, which ensures that the selection of the excerpts does not produce systematic errors.

Preprocessing of stimuli

The level of all samples was normalised, so that the level would not have an influence to the classification decision. Amplitude-scaling was done by scaling the variance to unity for the whole piece. Since we were also using rather short sample durations

Table 4.1: The amounts of test samples from different subgenres within the higher-level.

classical	**10**
chamber music	2
classical general	2
solo instruments	2
symphonic	2
vocal	2
electronic / dance	**10**
ambient	2
breakbeat/breaks/drum'n'bass	2
dance	2
house	2
techno/trance	2
hip hop	**10**
jazz / blues	**10**
blues	5
jazz (latin jazz, jazz fusion)	5
rock / pop	**10**
rock (alternative)	3
country	2
metal	2
pop (easy listening)	3
soul / RnB / funk	**10**
funk	3
RnB	4
soul (rhythm & blues, gospel)	3

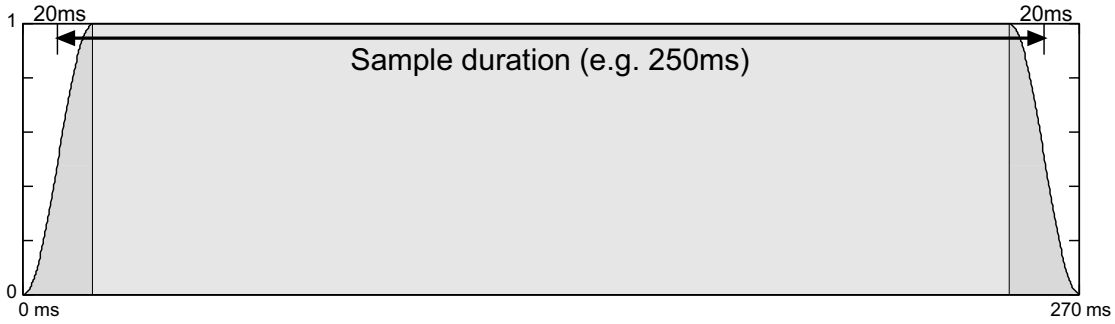


Figure 4.1: Windowing of the stimuli.

(250 ms and 500 ms), we used windowing to avoid unwanted clicks and masking on the sample start and end. The beginning and the end of a sample were smoothed with a half-Hanning window as illustrated in Figure 4.1. An equal-length (40 ms) Hanning window was used for each sample durations.

4.1.4 Equipment and facilities

The experiment was conducted in a listening room in order to minimise the interference of background noise. The subjects were seated in front of a projector screen presenting the experiment interface. The subjects had also a full access to the genre hierarchy (see Appendix A).

Subjects listened to the samples with headphones (Beyerdynamic DT931) ensuring

that reverberation did not affect the experiment. Every subject could set the amplification at a comfortable listening level. Besides this, there was no other calibration of the sound reproduction system.

4.1.5 Subjects

An ideal situation would have been to use some sort of music labelling professionals as test subjects. However, it would have been practically impossible to gather a representative set of such professionals, thus we recruited amateur listeners. Ten test subjects were recruited for both test groups, totalling 20 test subjects. Since we were using amateur listeners, we had to ensure that they knew musical genres sufficiently to be able to reliably recognise them. In subject selection, subjects who had some musical background were preferred. Furthermore, subjects who had been exposed to the used music database were excluded from the experiment.

Every subject had to take a short questionnaire, which was intended to map the knowledge about different musical genres. Subjects had to name a characteristic feature, an adjective, or an artist to be representative for a particular genre. The questionnaire form used is shown in Figure 4.2. If two or more of the genres were clearly unknown to the subject, he or she was excluded from the experiment. Experiment controller conducted the questionnaire and decided whether the subject met the requirements.

4.2 Test procedure

Before the test itself, subject candidates had to pass the previously described selection procedure. The subjects selected were put through a two-stage procedure for the experiment: a familiarisation stage and the main experiment.

4.2.1 Preparation

Before the experiment, the experiment controller explained the whole test procedure to the subject. Every stage was explained in order to motivate subjects. A short written paper was also provided, with the main points of the experiment explained in both English and Finnish. The musical genre hierarchy used was explained and shown to the subject (see Appendix A). Genre hierarchy was also available during the experiment. This minimised errors in cases where the subject clearly knew the right subgenre, but could not assign it on any of the six primary genres.

In addition, some background information was collected about the subjects. This information was gathered to possibly be used to explain exceptionally good or bad recognition abilities in the result analysis stage. A graphical user interface used to collect the data is shown in Figure 4.3.

All the user interfaces were in English. This should not be a problem for non-native speakers, since terms were pretty clear and well understandable. Nevertheless, if

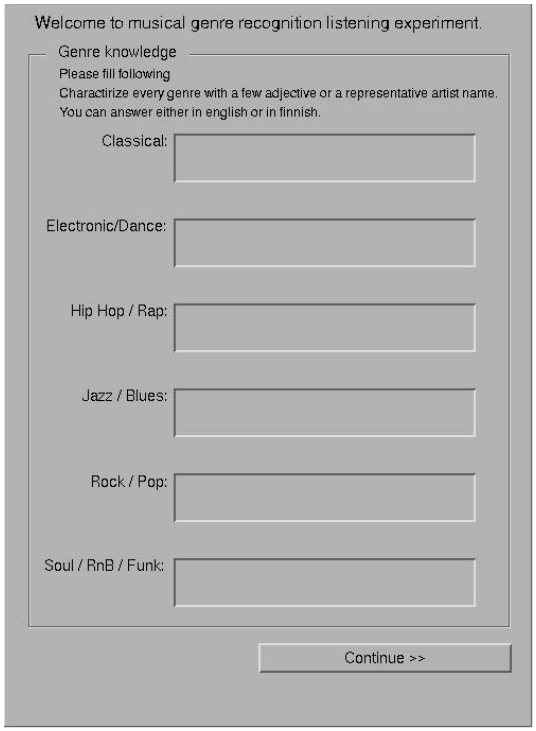


Figure 4.2: A questionnaire to map subject's knowledge about different musical genres.

this was found to be a problem, the experiment controller explained the terms to the test subject.

4.2.2 Familiarisation stage

The familiarisation stage had two functions. Firstly, to train the subjects for the task and to familiarise them with the test setup, procedure and user interface. Secondly, to make sure that the subjects are at least somehow familiar with the musical genre labels used. Familiarisation was also used to provide the subject some kind of internal reference for the musical genres.

Familiarisation was done with the experiment controller and the correct musical genre was shown after subjects' own classification. From each of the musical genres, two pieces were randomly selected for the familiarisation. These familiarisation pieces were excluded from the actual experiment. The first six samples (one from each genre) were 5000 ms long, and the next six samples were randomly either 250 ms or 500 ms long. This way the subject also got an idea of how short the samples could be.

Amplification of the headphones was set to a comfortable listening level during the familiarisation stage and was held fixed during the rest of the experiment. The listening level was allowed to vary between subjects.

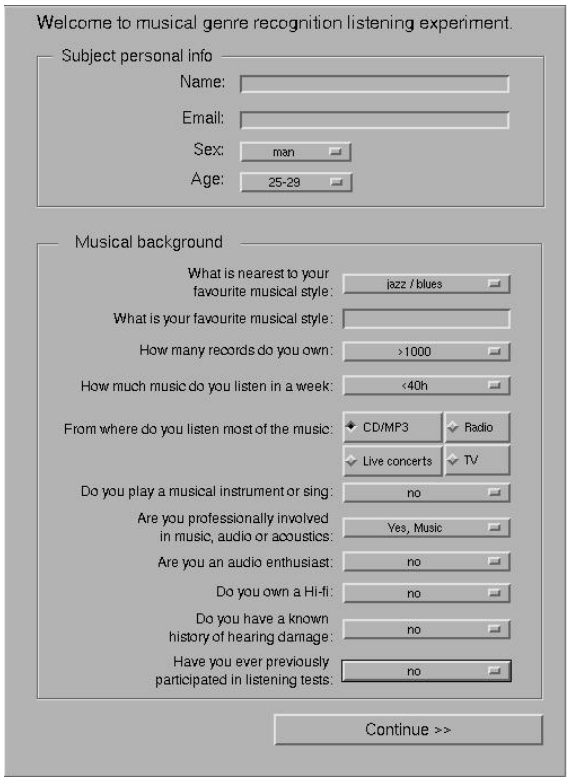


Figure 4.3: Background information query.

4.2.3 Main experiment

Before the experiment, subjects were divided into two test groups. Subjects in each of the test groups classified 240 samples (10 pieces $\times$ 6 genres $\times$ 4 durations). In order to minimise the learning effect, subjects were allowed to listen to each test sample only once, and contrary to the familiarisation stage, the correct musical genre was not shown after the classification. Because the samples used were rather short, the test subjects were allowed to trigger them by themselves. This way the subjects could control the pace of the experiment and carefully listen to all the samples. Since the subject had to press a button to play the sample, his attention was distracted for a while. After the sample was triggered, there was a 500 ms delay before it was actually played, so that the subject could focus all his or her attention only on the sample. The subjects were encouraged to proceed at their own pace, taking breaks whenever necessary. A typical experiment took about 50 minutes to complete.

In order to reduce order-effects, the stimuli were presented in a random order, different for each subject. However, it would have been confusing to mix up all four very different length of stimuli used in the experiment. Therefore, stimuli of an equal length were randomly collected into 20-stimulus blocks, and these blocks were then presented in a random order.

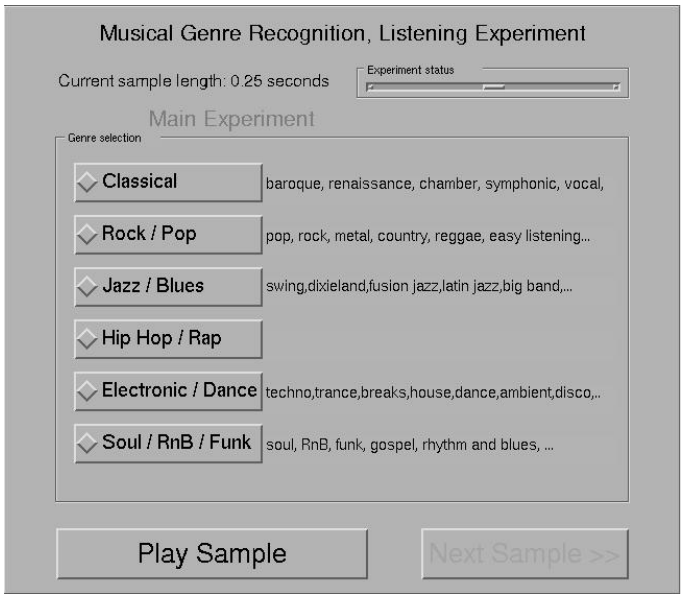


Figure 4.4: User interface of the main experiment.

4.2.4 User interface for data collection

Figure 4.4 shows the main user dialogue for the experiment. In order to minimise the effect caused by the ordering of genre-selection buttons, the buttons were ordered randomly for every test subject.

When the subject clicked the *Play sample* button, the current stimulus was played after a 500 ms delay. After the music was completed, the subject clicked a genre (radio button) as he or she felt suitable for the current stimulus. Clicking the *Next Sample* button approved the genre selection. Duration of the next sample was displayed, so that the subject could prepare him or herself for the shorter durations. A graphical bar was used to indicate the progress of the experiment to the subject.

The classification task was divided into three stages:

- sample has not been played yet
- sample has been played, but the user has not selected a genre yet
- user has selected a genre and is ready to move to the next sample

In every stage, the actions of the users were guided by shadowing all the unnecessary buttons.

4.3 Results

Results of the experiment for both the test groups and for different sample durations are shown in Table 4.2. The random guess rate for this experiment would be 16.6 %.

Table 4.2: The recognition accuracies for both the test groups. Overall recognition accuracies obtained by combining performances of both the test groups. Entries are expressed as mean recognition rate (percentages) with 95 % confidence interval. The mean performances are calculated over all sample durations.

Sample duration	Group 1	Group 2	Overall
250 ms	57±5	56±5	**57±6**
500 ms	62±5	63±5	**63±6**
1500 ms	62±5	68±5	**69±6**
5000 ms	72±4	77±4	**75±7**
Mean performance	65±4	66±4	**66±5**

Table 4.3: Confusion matrix for the genre recognition with 250 ms excerpts. Also the total number of responses for each of the genres is presented. All entries are expressed as percentages.

Presented \ Responded	Class	Electr	HipH	JazzB	RockP	SoulR
Classical	**88**	3	-	5	3	2
Electronic/Dance	3	**49**	5	10	15	19
Hip Hop	1	6	**66**	5	13	10
Jazz/Blues	7	5	1	**49**	28	11
Rock/Pop	4	14	3	12	**57**	11
Soul/RnB/Funk	2	10	14	19	25	**31**
Total responded	17	14	15	17	23	14

Pooling across all subjects, the right musical genre was selected on 57 % of trials with 250 ms long samples and the performance gradually increased with longer sample durations.

Table 4.3 shows a confusion matrix for the 250 ms excerpt stimuli, pooled across all subjects. In the confusion matrix, the rows correspond to the actual genre and the columns to the genre responded by subject. For example, the cell of row 1, column 5 with value 3 means that 3 % of the classical music was wrongly classified as rock/pop music. The percentages of correct classification for every genre lie in the diagonal of the confusion matrix.

Especially with shorter excerpts, one of the problems was that within the piece there are short intervals that could be interpreted out of the context into a different genre than as a part of the piece. In some cases this may have caused clearly acceptable confusions, but overall, the effect to the results is minimal. Table 4.4 shows the confusion matrix for 5000 ms long stimuli. Much less confusions occur, as can be seen. Some of the confusions are quite understandable, since in some cases musical genres overlap each other, or, at least they are rather close to another. The most distinct genre was classical, with very few false responses. Rock/pop got the most

Table 4.4: Confusion matrix for the genre recognition with 5000 ms excerpts. Also the total number of responses for each of the genres is presented. All entries are expressed as percentages.

Presented \ Responded	Class	Electr	HipH	JazzB	RockP	SoulR
Classical	**99**	1	-	-	-	1
Electronic/Dance	3	**70**	2	4	11	12
Hip Hop	-	3	**82**	1	8	7
Jazz/Blues	4	-	-	**66**	19	12
Rock/Pop	2	3	2	12	**77**	5
Soul/RnB/Funk	-	4	7	12	26	**52**
Total responded	18	13	15	16	23	15

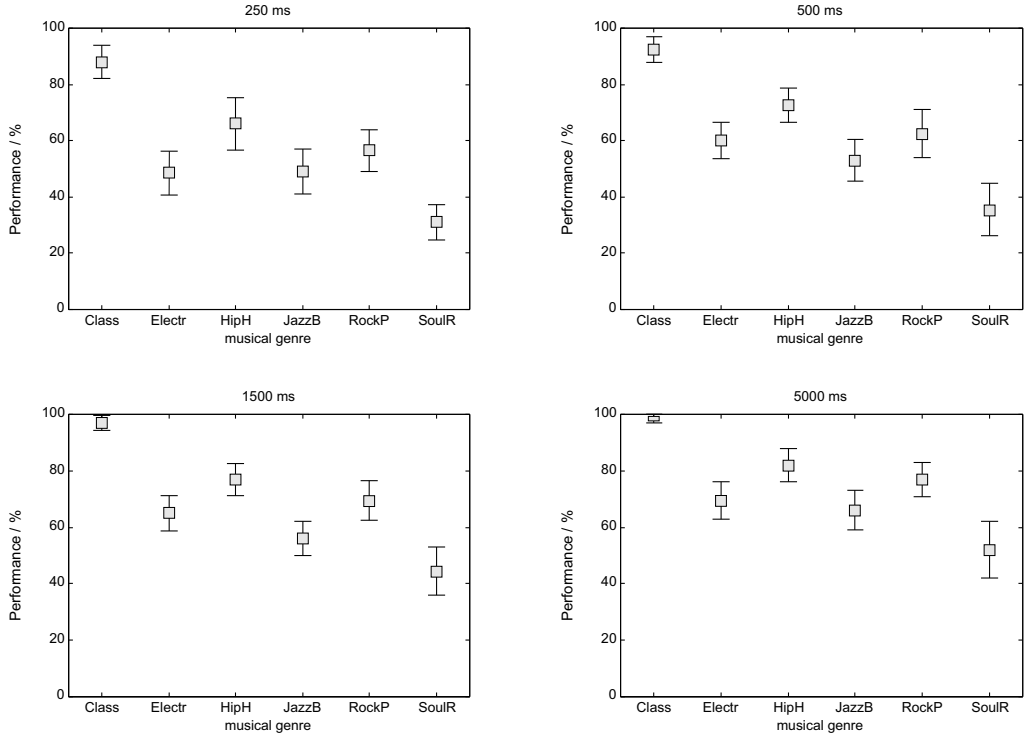


Figure 4.5: Pooled recognition results for musical genres. Confidence limits (95 %) are marked with whiskers.

false responses. The most ambiguous genres were soul/RnB/funk, rock/pop, and jazz/blues.

Figure 4.5 shows the recognition accuracy with confidence limits for the individual genres. For classical music, the recognition accuracy was excellent with all sample lengths. Soul/RnB/funk was the hardest genre to recognise throughout the experiment.

Table 4.5: F statistics for testing whether the individual test groups had equal recognition accuracies.

Duration	F	p value
250 ms	0.01	0.91
500 ms	0.11	0.75
1500 ms	0.21	0.65
5000 ms	4.23	0.05

4.3.1 Analysis

A compact statistical analysis was performed for the experiment results. The main question that arises after the presented experiment results is: are the results reliable and do they give a genuine picture of human abilities to recognise musical genres.

A chi-squared test is used to determine whether the experiment results are due to genuine difference, or whether it is just due to chance [Fisher70, pp.78-113]. For the 250 ms test excerpts confusion matrix presented in Table 4.3, for example, we got $\chi^2(6,6) = 1733$, $p \ll 0.001$. Based on this we can say with a high degree of certainty that the differences between the values in the confusion matrix are a true reflection of a variation and not due to chance. The overall results and results for the individual subjects were all strongly significant using this test.

One-way Analysis of Variance (ANOVA) is used to test the null-hypothesis that the means of the independent variable among the tested data groups were equal, under the assumption that sampled populations are normally distributed [Fisher70, pp.213-249]. If the means of the tested data groups differ significantly, a null-hypothesis is rejected, and we conclude that at least one of the groups was from a population with a different mean. In order to perform the analysis, population variances have to be assumed equal.

At first, we tested whether the test groups had equal recognition accuracies. If they were in fact equal, it would imply that the recognition accuracies of the groups were independent on the selection of the pieces used and the particular division of test subjects into the groups. The null-hypothesis is defined as follows:

H_o: Both of the two test groups had the same average performance.

Hypothesis testing is done individually for each excerpt-length. Test results are presented in Table 4.5. For the lengths 250 ms and 500 ms one can accept the null-hypothesis with some confidence. For the length 5000 ms, one has to reject the null-hypothesis, since an observed F statistic would occur by chance only once in 20 times if the means were truly equal. The different test pieces and subjects in the groups showed in the results only with the longest sample length. Thus one can conclude that the selection of the pieces or the different abilities of the test subjects had some significance in this experiment.

Table 4.6: Recognisability scores for the musical genres.

Musical genre	250 ms	500 ms	1500 ms	5000 ms	ranking
classical	43	46	48	47	1.
electronic / dance	26	32	35	37	3.
hip hop	35	38	40	42	2.
jazz / blues	25	28	30	34	4.
rock / pop	24	26	27	32	5.
soul / RnB / funk	17	19	24	27	6.

Next, we will study the differences in recognition accuracies caused by the length of the sample. The null-hypothesis is defined as follows:

> H_o: Length of the presented test samples did not have significant influence on the recognition accuracy.

Test results clearly show that the length of the sample had a significant effect on the recognition accuracy ($F = 27.75$, $p = 3.1921e - 12$).

Finally, it is tested whether the musical genre had an effect on the recognition rate in the experiment. The null-hypothesis is defined as follows:

> H_o: Accuracy of musical genre recognition is equal regardless of the presented musical genre.

Test results for each sample length show clearly that the musical genre had a significant effect on the recognition accuracy ($F_{250ms} = 28.29$, $p = 0$, $F_{500ms} = 30.78$, $p = 0$, $F_{1500ms} = 36.05$, $p = 0$, $F_{5000ms} = 28.89$, $p = 0$).

In order to further investigate the differences in the recognition accuracies with musical genres, a "recognisability" score was calculated for each of the genres. The score is calculated as the number of trials in which the genre was correctly identified, divided by the total number of trials in which the genre appeared as either a stimuli or a response. Table 4.6 presents scores for each of the sample lengths. Based on the scores, the ranking of the musical genres was the same regardless of the sample length.

4.4 Discussion and conclusions

The genre recognition rate of humans was 57 % for 250 ms sample duration and 75 % for 5000 ms sample duration in the experiment. As the analysis showed, the recognition accuracy depends on the length of the presented sample. Although the shortest sample lengths used (250 ms and 500 ms) are very short, fairly good recognition accuracy were achieved for them. This shows that humans can do rather accurate musical genre recognition without long-term temporal features. Since 250

ms and 500 ms are too short to represent any rhythmic aspects of the music, a subject may have to recognise, for example, some of the instruments and make a classification based on them. With the longer sample lengths, human can also use rhythm information and other long-term aspects of music to recognise the musical genre more reliably.

Different musical genres are recognised with different accuracies. Some genres are more distinguishable than others, e.g. classical music proved to be highly distinguishable. In some cases, genres like rock/pop, hip hop, and soul/RnB/funk might be rather confusing, as they often share common elements, like instrumentation.

Comparison between the experiment and the previous studies are hard to make, since the genre selection used and the test stimuli differ. Nevertheless, one can conclude that results presented here are essentially similar to those presented in [Perrot99, Soltau97].

5 System description

This chapter presents the basic structure of the used pattern recognition system. Signal classification is done based on two leading principles. Musical genre and instrumentation is recognised based on the overall tone colour of signals. Features and classifiers studied in our experiments will be presented in detail. In addition, the presence of drum instruments in music is detected by measuring the periodicity of stochastic energy at subbands.

5.1 General structure of pattern recognition system

An overall view of our system is presented in Figure 5.1. At the general block-diagram level, all modern pattern recognition systems share the same basic structure. The system consists of four parts: preprocessing, feature extraction, pattern learning, and classification. At the preprocessing stage, the input signal is normalised before extracting acoustic features, which are used to characterise the signal. In order to train the recognition system, we need to have a sufficient amount of training examples from each class to be recognised. In practice, this is usually achieved by assigning 70% of the manually labelled evaluation data into training set and the rest to the test set. In the pattern learning stage, a representative pattern will be determined for the features of the actual class. This can be done with a model that uses statistical information concerning the features in the training data. In the classification stage, the test data is compared to previously calculated models and classification is done by measuring the similarity between the test data and each model, and assigning the unknown observation to the class whose model is most similar to the observation.

5.2 Extraction of spectral features

It is difficult to determine what particular features allow us to distinguish between musical genres. It is even more challenging to find a compact numerical representation for a segment of audio that would retain those distinguishing properties, and at the same time lose the irrelevant information. The use of right features is essential for the classification process. There are a wide variety of different features that can be used to characterise audio signals. Features can be divided generally into time-domain and frequency-domain (spectral) features.

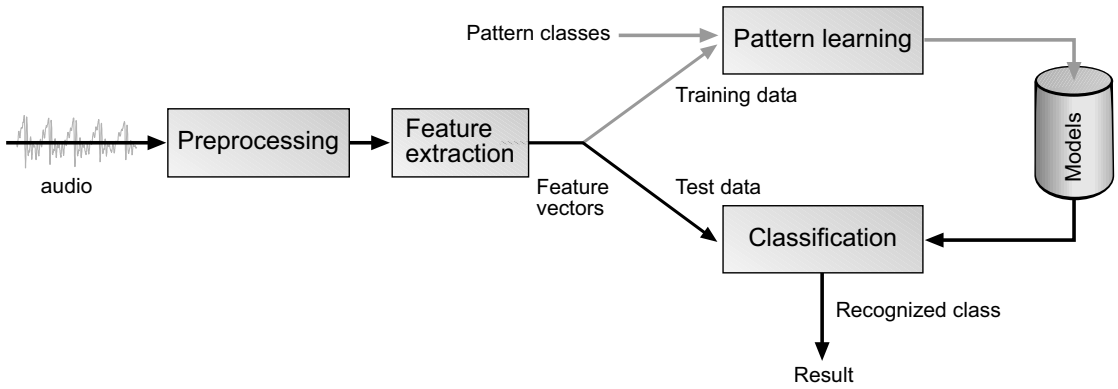


Figure 5.1: A block diagram of the system.

Many different features were preliminarily tested, but only a few of the most promising ones were used in the final experiments. Mel-Frequency Cepstral Coefficients have been successfully used in many audio classification problems [Reynolds95, Eronen03a], and they proved to be a good choice for our application, as well. Another feature, Band Energy Ratio was used in the extraction of the rhythm features.

Before the feature extraction, the time-domain signals are normalised to have zero mean and unity variance. After that the audio signals are divided into 20 ms frames. The frame blocking causes edge-effects (spectral leakage), which are minimised by using a windowing function for the frame. This gives more weight to the samples that are located at the centre of the window. Successive windows overlap each other 5 ms.

5.2.1 Mel-frequency cepstral coefficients

Mel-Frequency Cepstral Coefficients (MFCC) is the most widely-used feature in speech recognition [Rabiner93]. They give a good discriminative performance with reasonable noise robustness. MFCC is a short-term spectrum-based feature, which represents the amplitude spectrum in a compact form. Figure 5.2 shows the steps of extracting the MFCC features. These steps are motivated by perceptual and computational considerations.

The preprocessing step involves pre-emphasising the audio signal, dividing the signal into frames and windowing it. Pre-emphasis is done using a first-order finite impulse response (FIR) filter $1 - 0.97z^{-1}$ to increase the relative energy of high-frequency spectrum. The aim of frame blocking is to segment the signal into statistically stationary blocks. Hamming window is used to weight the pre-emphasised frames. Next, the Discrete Fourier transform (DFT) is calculated for the frames. Since the human auditory system does not perceive pitch linearly, a perceptually meaningful frequency resolution is obtained by averaging the magnitude spectral components over Mel-spaced bins. This is done by using a filterbank consisting of 40 triangular

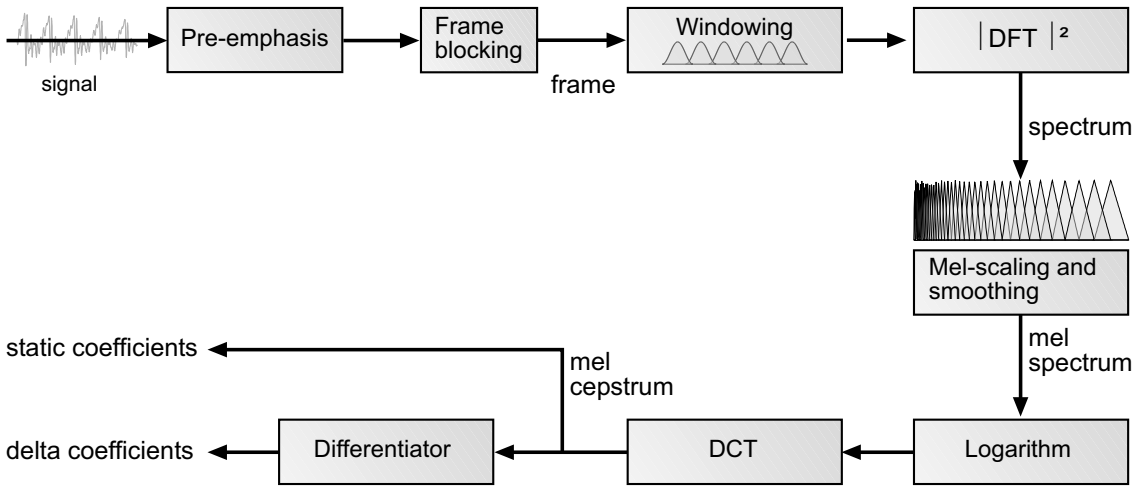


Figure 5.2: Overview of the MFCC feature extraction system.

filters occupying the band from 80 Hz to half the sampling rate, spaced uniformly on the Mel-scale. An approximation between a frequency value f in Hertz and in Mel is defined by [Houtsma95]:

$$Mel\,(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{5.1}$$

After the Mel-scale filterbank, logarithm is applied to the amplitude spectrum, since the perceived loudness of a signal has been found to be approximately logarithmic. The Mel-spectral components are highly correlated. This is an unwanted property especially for features to be used with Gaussian mixture models, since it increases the number of parameters required to model the features. The decorrelation of the Mel-spectral components allows the use of diagonal covariance matrices in the subsequent statistical modelling. The Mel-spectral components are decorrelated with the DCT, which has been empirically found to approximate the Karhunen-Loeve transform, or, equivalently, the Principal Component Analysis in case of music signals [Logan00]. The DCT is calculated by

$$c_{mel}\,(i) = \sum_{j=1}^{M} (\log S_j) \cdot \cos\left(\frac{\pi i}{M}\left(j - \frac{1}{2}\right)\right) \quad , i = 1, 2, \ldots, N\,, \tag{5.2}$$

where $c_{mel}\,(i)$ is the i^{th} MFCC, M is the total number (40) of channels in filterbank, S_j is the magnitude response of the j^{th} filterbank channel, and N is the total number of the coefficients. For each frame, 17 cepstral coefficients are obtained using this transform and the first coefficient is discarded, as it is a function of the channel gain. The final number of cepstral coefficients is 16, which was found in preliminary evaluations to give sufficient representation of the amplitude spectrum.

Transitions in music carry relevant information and consecutive feature vectors correlate, thus it is important to consider time domain dynamics in feature representa-

tion. In addition to the static coefficients, their differentials are also estimated by using a linear regression over consecutive cepstral coefficients. The first-order time derivatives are approximated with a three-point first-order polynomial fit as follows

$$\Delta c_{mel}(i,u) = \frac{\sum_{k=-1}^{1} k \cdot c_{mel}(i,u+k)}{\sum_{k=-1}^{1} k^2}, \tag{5.3}$$

where $c_{mel}(i,u)$ denotes the i^{th} cepstral coefficient in time frame u [Rabiner93, pp. 116-117].

Many authors have used MFCCs to model music simply because they have been so successful for speech recognition. Logan examined how some of the assumptions made merely based on speech hold with music [Logan00]. Like speech, music is non-stationary and phase-independent, so frame-based analysis and amplitude spectrum are also justified for music. Furthermore, the perception of loudness is still logarithmic with music. Suitability of Mel-scaling was tested by comparing the performance of linear and Mel-scaled cepstral features in speech/music discrimination. They found the Mel-scaling to at least not be harmful for the discrimination. Usefulness of DCT to approximate the Karhunen-Loeve transform for music was proven by observing that the eigenvectors of Karhunen-Loeve were also "cosine-like" for music.

5.2.2 Band Energy Ratio

Band energy ratio (BER) is defined as the ratio of the energy at a certain frequency band to the total energy. Thus the BER for the i^{th} subband in time frame u is:

$$F_{BER}(i,u) = \frac{\sum_{n \epsilon S_i} |X_u(n)|^2}{\sum_{n=0}^{M} |X_u(n)|^2}, \tag{5.4}$$

where X_u is the DFT of the time domain signal within the frame u, M is the index for the highest frequency sample (half of the DFT order), and S_i is the set of Fourier transform coefficients belonging to the i^{th} subband [Li01]. The Mel-scale filterbank is also applied here to obtain a perceptually meaningful frequency resolution.

5.3 Rhythm feature extraction

The features described above are used to represent a coarse spectral shape of the music signal. Besides the spectral shape, the rhythm is also an important property of the music. We present an approach to detect the presence of drum instruments in music by measuring the signal's long-term periodicity. This approach has been previously presented in [Heittola02].

The aim was to develop a drum detection system, which would be as generic as possible. The problem of drum detection in music is more difficult than what it seems at a first glance. For a major part of techno or rock/pop music, for example, detection

is more or less trivial. However, detection systems designed for these musical genres do not generalise to other genres. Music contains a lot of cases that are much more ambiguous. Drums go easily undetected in jazz/big band music, where only hihat or cymbals are softly played at the background. On the other hand, erroneous detections may pop up for pieces with acoustic steel-stringed guitar, pizzicato strings, cembalo, or staccato piano accompaniment, to mention some examples.

Earlier work in the area of the automatic analysis of musical rhythms has mostly concentrated on metrical analysis [Scheirer98]. Most of the work in this field is done with MIDI data, however there are a few exceptions. Alghoniemy *et al.* used a narrowband filter at low frequencies to detect periodicities in polyphonic music [Alghoniemy99]. Tzanetakis *et al.* used the Discrete Wavelet Transform to decompose the signal into a number of bands and the autocorrelation function to detect the various periodicities of the signal's envelope [Tzanetakis01]. They used this structure to extract rhythm-related features for musical genre classification. Soltau *et al.* used Neural Networks to represent temporal structures and variations in musical signals [Soltau98].

5.3.1 Preprocessing with sinusoidal modelling

In Western music, drum instruments typically have a clear stochastic noise component. The spectral energy distribution of the noise component varies, being wide for the snare drum, and concentrated to high frequencies for cymbal sounds, for example. In addition to the stochastic component, some drums have strong harmonic vibration modes, and they have to be tuned. In the case of tom toms, for example, approximately half of the spectral energy is harmonic. Nevertheless, these sounds are still recognisable based on the stochastic component only. While most other musical instruments produce chiefly harmonic energy, an attempt was made to separate the stochastic and harmonic signal components from each other.

A sinusoids-plus-noise spectrum model was used to extract the stochastic parts of acoustic musical signals. The model, described in [Serra97], estimates the harmonic parts of the signal and subtracts them in time domain to obtain a noise residual. Although some harmonic components are not detected and beginning transients of other instruments leak through, in general the residual signal has significantly better "drums-to-other" ratio than the input signal.

5.3.2 Periodicity detection

Periodicity is characteristic for musical rhythms. Drum events typically form a pattern which is repeated and varied over time. As a consequence, the time-varying power spectrum of the signal shows clear correlation with a time shift equal to the pattern length in the drum track. We propose that the presence of drums can be detected by measuring this correlation in musical signals. This relies on the assumption that periodicity of stochastic signal components is a universal characteristic of

musical signals with drums. In order to alleviate the interference of other musical instruments, periodicity measurement is performed in the residual signal after preprocessing with a sinusoidal model.

Feature stream

A signal model was employed which discards the fine structure of signals, but preserves their rough spectral energy distribution. BER was used as the feature. Feature vectors are extracted from the preprocessed signal. Features were extracted in 10 ms analysis windows (Hanning windowing) and with 50% overlap. Short window length was preferred to achieve a better time resolution in the autocorrelation calculations later on. Using 16 frequency bands in BER ensured a sufficient frequency resolution. The obtained feature vectors form a feature stream $F_{BER}(i,u)$, which is subjected to autocorrelation function calculations.

Summary autocorrelation function

At each frequency band, an autocorrelation function (ACF) is calculated over the BER values within a sliding analysis window. A three-second analysis window was chosen to capture a few patterns of even the slowest rhythms. Autocorrelation function of a U-length excerpt of $F_{BER}(i,u)$ at band i is given by:

$$r_i(\tau) = \frac{1}{U} \sum_{u=0}^{U-\tau-1} F_{BER}(i,u) \cdot F_{BER}(i,u+\tau) \ , \tag{5.5}$$

where τ is the lag, and $F_{BER}(i,u)$ is the calculated BER value at band i in frame u. Peaks in the autocorrelation function correspond to the delays in which the time-domain signal shows high correlation with itself.

Despite the preprocessing, other instruments also cause peaks to the bandwise autocorrelation functions. Fortunately, however, the spectrum of the other instruments tends to concentrate to the mid-bands, whereas drums are more prominent at the low or high frequencies (there are exceptions from this rule, e.g. the violin or the snare drum). Based on this observation, we will weight bands differently before forming the summary autocorrelation function (SACF). Autocorrelation functions are weighted and then summed up to form the SACF

$$s(\tau) = \sum_{i=1}^{Bands} W_i \cdot r_i(\tau) \ . \tag{5.6}$$

This overall structure bears a close resemblance to the mechanisms of human pitch perception, as modelled in [Meddis91]. A major difference here is that processing is done for subband amplitude envelopes instead of the signal fine structure. The SACF is then mean-normalised to get real peaks step out better from the SACF. Mean normalisation was done with the following equation [Cheveigné02]:

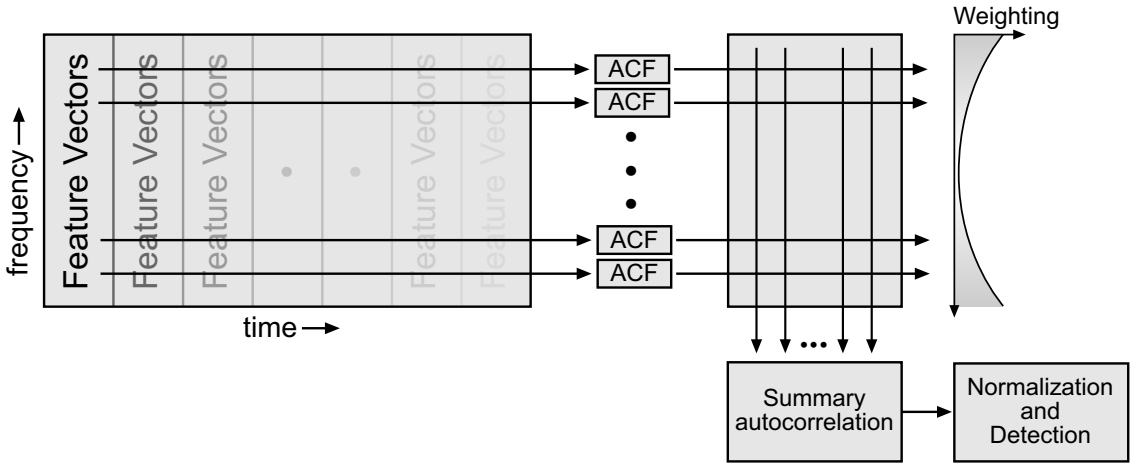


Figure 5.3: System overview.

$$\begin{cases} \widehat{S}(0) &= 1 \\ \widehat{S}(\tau) &= \frac{s(\tau)}{\frac{1}{\tau}\sum_{j=1}^{\tau} s(j)} \end{cases} . \tag{5.7}$$

Overview of the whole system is shown in Figure 5.3.

Detection

Since a quite short analysis frame (10 ms) was used in extracting the feature stream, the lowest frequency components cause slight framing artefacts. These appear as a low-amplitude and high-frequency ripple in the SACF, which is easily removed using moving averaging. Further, a long-term trend caused by differences in signal levels within the ACF analysis window will be detrended from SACF using high pass filtering. Thus obtained SACFs for different types of music are shown in Figure 5.4.

As can be seen in Figure 5.4, periodic drum events also produce a periodic SACF. In order to robustly detect this, SACF has to be enhanced in a manner illustrated in Figure 5.5. The original SACF curve is time-scaled by a factor of two and three and these two stretched curves are added to the original, resulting in the enhanced summary autocorrelation function (ESACF). Thus peaks at integer multiples of fundamental tempo are used to enhance the peaks of a slower tempo. If the original SACF is periodic in nature, this technique produces clearer peaks. Idea for this technique has been adopted from [Tolonen00], where subharmonic cancellation was done by subtracting stretched curves from the original one.

The region of interest in the ESACF is determined by reasonable tempo limits. Lower limit was fixed to lag of 1.7 seconds, which corresponds to the tempo of 35 beats per minute. The higher limit was fixed to 120 beats per minute. Whereas the upper limit may seem too tight, it should be noted that due to the enhancement procedure, these limits actually correspond to 35 and 360 beats per minute in the original SACF. This wide tempo range is essential, since certain drum instruments (e.g. hihat) are

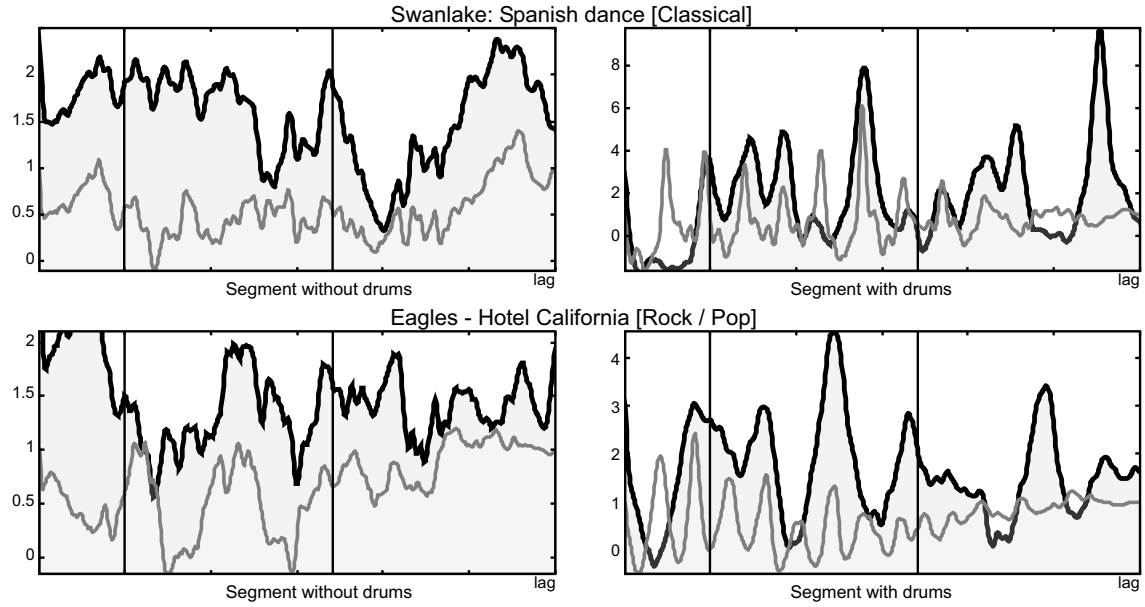


Figure 5.4: Representative SACFs (grey line) and ESACFs (black line) from different type of music. Tempo limits marked in the plots.

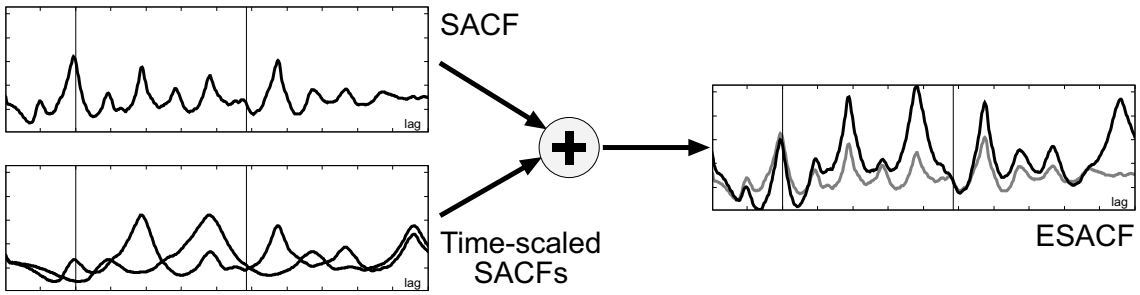


Figure 5.5: Enhancing the summary autocorrelation function.

typically played at an integer multiple of the tempo. Final detection is carried out by measuring the absolute maximum value within the given tempo limits. Maximum value distributions for segments with and without drums are presented in Figure 5.6. Based on these distributions a threshold value for maximum value within periodicity limits was defined and detection was done according to this threshold value. The threshold value was chosen to produce equal error-probability for segments with and without drums. Distributions overlap to some extent, but nevertheless enable robust classification.

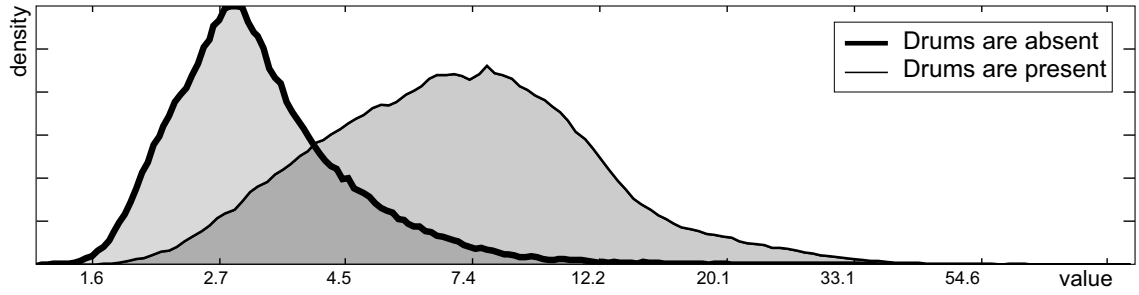


Figure 5.6: Unit area normalised feature value distributions for music with and without drums, as indicated in the legend.

5.4 Classification

The next stage after feature extraction is to assign each object to a specific category based on features extracted. There are many different classification techniques to choose from. Two types of classifiers are described in the following: the (distance-based) k-Nearest Neighbour classifier and the (probabilistic) hidden Markov models.

5.4.1 K-Nearest Neighbour classifier

In this classification approach, a data vector to be classified is compared to training data vectors from different classes and classification is performed according to distance to the k nearest neighbouring data points. In this thesis, Mahalanobis distance is used in determining the nearest neighbours. The distance between the vectors $\mathbf{x}$ to be classified and the training vectors $\mathbf{y}$ is measured by:

$$D = (\mathbf{x} - \mathbf{y})^T \mathbf{C}^{-1} (\mathbf{x} - \mathbf{y}) \tag{5.8}$$

where $\mathbf{C}$ is the covariance matrix of the training data. The classification is done by picking the k points nearest to the current test point, and the class most often picked is chosen as classification result [Cover67]. It is challenging to find the best value for k. The neighbours should be close to the test data point $\mathbf{x}$ to get accurate estimate, but still the number of neighbours should be large enough to get a reliable estimate for *a posteriori* probability of the data point $\mathbf{x}$ belonging to the class ω_j, i.e. $P(\omega_j \mid \mathbf{x})$.

The implementation of k-NN classifier is straightforward. The computational load is high with a large set of training data, since all the training data is stored and a distance between the every test point and all the training data is calculated.

5.4.2 Hidden Markov Models

The distance-based k-NN classifier takes into account only the average of feature distributions. The hidden Markov model (HMM) is a widely used method of statis-

tical modelling, which also takes into account the shapes of the feature distributions. The parameters for the class distributions are estimated based on the training data. For each individual class, a model is trained independently and by maximising the posterior probability the recognition accuracy is also assumed to be maximised. For the tested observation sequence $\mathbf{O} = (\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T)$, the HMM parameters are used to calculate the *a posteriori* probabilities $P(\omega_j \mid \mathbf{O})$ for each class, and the class corresponding to the model with highest probability is chosen. This principle is also known as Bayes's rule.

An HMM is a finite state machine, where a series of states represents the target to be modelled. Each state is associated with a probability distribution and transitions between the states are determined by a set of transition probabilities. Each state also has an output probability distribution that determines the likelihood of observing certain feature values in certain states. Only the outcome is visible to the observer, not the state that generated the observation, i.e. states are "hidden". HMMs provide a good parametric model for time-varying audio signals. The output distribution can be matched with the distributions of target feature vectors by varying the parameters of an HMM.

In contemporary speech recognition, HMMs are the dominant tool to model speech as acoustic units (a phone, a syllable, a word, a sentence, or an entire paragraph) [Rabiner93, pp. 435-481]. In recent years, HMMs have also been increasingly used in other audio content analysis applications [Aucouturier01, Casey02, Eronen03a].

First, we review the theory of Markov Models, and then extend them to HMM. Further, we will take a closer look at two fundamental problems for HMM design: the model parameter estimation, and the evaluation of the probability (classification). Two training methods are presented: conventional maximum-likelihood estimation using the Baum-Welch algorithm, and discriminative training.

Markov Models

A finite Markov model or a Markov process is defined as a random process where the probability of transitions to a next state depends only on the current state. Since we are dealing with finite models, the states of the process can be enumerated as $\{1, 2, \ldots N\}$ and each state q has a probability distribution for each time t, denoted by $P(q_t)$. Further, if we know the state of the process at time t, we know exactly the probabilities of the states for time $t+1$, i.e. $P(q_t \mid q_{t-1}, q_{t-2}, q_{t-3}, \ldots) = P(q_t \mid q_{t-1})$. If we assume that these conditional probabilities are time-invariant, we have $P(q_t = j \mid q_{t-1} = i) = a_{ij}$, where $1 < i, j < N$. These probabilities a_{ij} can be presented as an $N \times N$ state transition probability matrix,

$$A = \begin{bmatrix} P(1 \mid 1) & P(2 \mid 1) & \cdots & P(N \mid 1) \\ P(1 \mid 2) & P(2 \mid 2) & \cdots & P(N \mid 2) \\ \vdots & & \ddots & \vdots \\ P(1 \mid N) & P(2 \mid N) & \cdots & P(N \mid N) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2N} \\ \vdots & & \ddots & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NN} \end{bmatrix}. \quad (5.9)$$

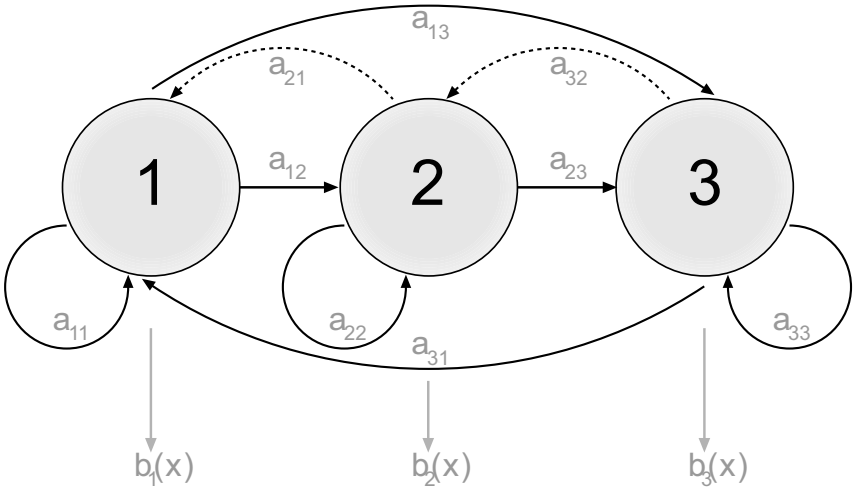


Figure 5.7: Representing state transition matrix as graph. Nodes represent the states and weighted edges indicate the transition probabilities. Two model topologies are presented: left-to-right and fully-connected (i.e. ergodic HMM), the dotted transitions have zero-probability in the left-right model.

The matrix A determines the allowable transitions within a model. A zero probability disables the transition. Finite-state time-invariant Markov processes can also be conveniently represented with graph as shown in Figure 5.7. Two different model topologies, fully-connected and left-right, are depicted in the figure.

Hidden Markov Models

A finite-state hidden Markov model is similar to Markov model, but the states which produce the outputs are not observable, i.e. the state information is hidden. The HMM can be completely defined with the initial state probability distribution Π, state transition probabilities A, and output distributions of the states B. The initial state probability distribution defines the probability of being in state i at the beginning of the process, $\pi_i = P(q_1 = i)$, $\Pi = [\pi_1, \ldots, \pi_i, \ldots \pi_N]$. Outputs of the process, the observations, are the outputs of the states. The probability of the observation x in the state j is denoted by $P(\mathbf{o}_t = x \mid q_t = j) = b_j(x)$. These output distributions can be modelled with multinormal distributions defined by the mean vector and covariance matrix. The parameters of the state distributions for all the states are denoted by B. [Rabiner93, pp. 329-330]

In this thesis, a mixture of multivariate Gaussian density functions is used in modelling the state-conditional densities. By means of multiple Gaussians, one can improve the modelling accuracy provided that there is enough training data available to estimate the mixture parameters. A multivariate Gaussian density function

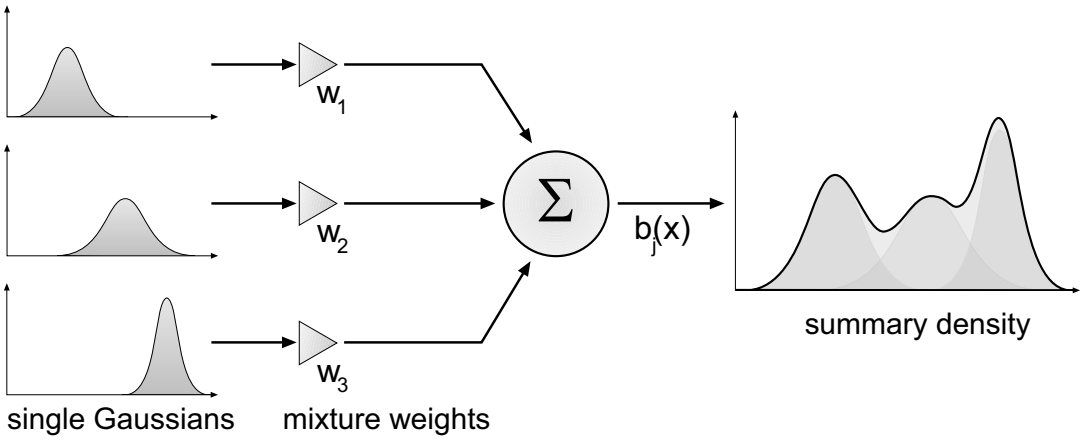


Figure 5.8: Representing a Gaussian mixture density.

is defined as

$$\mathcal{N}(\mathbf{o}_t,\, \boldsymbol{\mu}_m,\, \boldsymbol{\Sigma}_m) = \frac{1}{\sqrt{(2\pi)^n\,|\boldsymbol{\Sigma}_m|}} e^{-\frac{1}{2}(\boldsymbol{o}_t-\boldsymbol{\mu}_m)^T\boldsymbol{\Sigma}_m^{-1}(\boldsymbol{o}_t-\boldsymbol{\mu}_m)}, \tag{5.10}$$

with mean vector $\boldsymbol{\mu}_m$ and diagonal covariance matrix $\boldsymbol{\Sigma}_m$. The notation $|\,|$ denotes matrix determinant. Arbitrarily shaped densities can be approximated with a linear combination of Gaussian basis functions (see Figure 5.8).

With the mixture distributions, the probability that the observation vector $\mathbf{x}$ has come from state j is denoted by

$$b_j(\mathbf{x}) = \sum_{m=1}^{M} w_{j,m}\mathcal{N}\left(\mathbf{o_t},\, \boldsymbol{\mu}_{j,m},\, \boldsymbol{\Sigma}_{j,m}\right), \tag{5.11}$$

where M is the total number of Gaussian densities in state j, w_m's are positive mixture weights which sum to unity, x is an n-dimensional observation vector, and $\mathcal{N}\left(\mathbf{x},\, \boldsymbol{\mu}_{j,m},\, \boldsymbol{\Sigma}_{j,m}\right)$ is a multinormal distribution [Rabiner93, p. 350].

A single-state HMM where output distributions are modelled with Gaussian densities is also called a Gaussian Mixture Model (GMM). GMMs have been successfully used for a variety of audio classification tasks such as speaker recognition [Reynolds95] and musical genre recognition [Tzanetakis02a].

HMM Parameter Estimation

The estimation of the HMM parameters for the classes to be modelled is a challenging task. Based on training data, the HMM parameters are estimated according to some criterion. However, there is no universally optimum solution for this optimisation problem. An iterative training procedure to find a local maximum of the Maximum Likelihood (ML) objective function, known as the Baum-Welch re-estimation algorithm, is widely used, and it will generally find a good set of parameters. In ML

estimation, a set of model parameters which maximises the likelihood of the HMM given the training data is found. [Rabiner93, p. 342]

First, we have to make a rough guess about parameters of an HMM, and based on these initial parameters more accurate parameters can be found by applying the Baum-Welch re-estimation algorithm. The re-estimation procedure is sensitive to the selection of initial parameters. The model topology is specified by an initial transition matrix, where the disabled transitions are assigned to zero. In this thesis, only fully-connected and left-to-right topologies are evaluated (see Figure 5.7). The state means and variances can be initialised by clustering the training data into as many clusters as there are states in the model with the K-means clustering algorithm, and estimating the initial parameters from these clusters. [Rabiner93, pp. 370, 382-384]

The basic idea behind the Baum-Welch algorithm (also known as Forward-Backward algorithm) is to iteratively re-estimate the parameters of a model, and to obtain a new model with a better set of parameters $\overline{\lambda}$ which satisfies the following criterion for the observation sequence $\mathbf{O} = (\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T)$:

$$P\left(\mathbf{O} \mid \overline{\lambda}\right) \geq P\left(\mathbf{O} \mid \lambda\right), \tag{5.12}$$

where the given parameters are $\lambda = (\Pi,\, A,\, \mu_i,\, \Sigma_i)$. By setting $\lambda = \overline{\lambda}$ at the end of every iteration and re-estimating a better parameter set, the probability of $P\left(\mathbf{O} \mid \lambda\right)$ can be improved until some threshold is reached. The re-estimation procedure is guaranteed to find in a local optimum.

To be able to evaluate the probability $P\left(\mathbf{O} \mid \lambda\right)$, we need to define the forward and the backward probability, $\alpha_t\left(i\right)$ and $\beta_t\left(i\right)$, respectively. In order to calculate the probability $P\left(\mathbf{O} \mid \lambda\right)$, the probability of each possible state sequence which could produce the desired output has to be summed as follows:

$$P\left(\mathbf{O} \mid \lambda\right) = \sum_{q} b_{q_1}\left(\mathbf{o}_1\right) P\left(q_2 \mid q_1\right) b_{q_2}\left(\mathbf{o}_2\right) \times \cdots \times P\left(q_T \mid q_{T-1}\right) b_{q_t}\left(\mathbf{o}_T\right), \tag{5.13}$$

where $\mathbf{q} = (q_1, q_2, \ldots, q_T)$ denotes the state sequence, and $b_{q_1}\left(\mathbf{o}_1\right)$ is the probability that output $\mathbf{o}_1$ is observed in state q_1. This sum would be practically impossible to calculate, since there are N^T different sequences to be summed. A fast, but equally accurate, recursive algorithm can be used for the probability evaluation. The probabilities of partial observation sequences to end in specific state i, the forward probability $\alpha_t\left(i\right)$ for parameters λ, is denoted by

$$\alpha_t\left(i\right) = P\left(\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_t, q_t = i \mid \lambda\right). \tag{5.14}$$

The HMM produces the output sequence $(\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_{t+1})$ and ends in state i exactly when it first produces the sequence $(\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_t)$ ending in any state, and then moves into state i and outputs $\mathbf{o}_{t+1}$. This joint probability of having the partial observation sequence ending in the state i at time t can be efficiently calculated by the following forward procedure:

1. Initialisation
$$\alpha_1(j) = \pi_j b_j(\mathbf{o}_1), \quad 1 \le j \le N . \tag{5.15}$$

2. Induction
$$\alpha_t(i) = b_i(\mathbf{o}_t)\left[\sum_{j=1}^{N} \alpha_{t-1}(j)\, a_{ji}\right], \quad \begin{matrix} 2 \le t \le T \\ 1 \le i \le N \end{matrix} . \tag{5.16}$$

3. Termination
$$P(\mathbf{O} \mid \lambda) = \sum_{j=1}^{N} \alpha_T(j). \tag{5.17}$$

The main idea of this recursion is that the probability of being in state i at time t while observing $\mathbf{o}_t$ can be obtained by summing the forward probabilities of all possible preceding states j weighted by the transition probability a_{ji} and multiplying with $b_i(\mathbf{o}_t)$. [Rabiner93, pp. 335-337]

We also have to define the backward probability. The backward probability $\beta_t(i)$ is the probability of generating the partial observation sequence from time t to time T. This can be derived in a fashion similar to the forward probability [Rabiner93, p. 337]:

1. Initialisation
$$\beta_T(i) = 1, \quad 1 \le i \le N. \tag{5.18}$$

2. Induction
$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(\mathbf{o}_{t+1})\, \beta_{t+1}(j), \tag{5.19}$$
where $i = 1, 2, \ldots, N$ and $t = T-1, T-2, \ldots, 1$.

3. Termination
$$P(\mathbf{O} \mid \lambda) = \sum_{i=1}^{N} \beta_1(i)\, \pi_i b_i(\mathbf{o}_1). \tag{5.20}$$

The joint probability of generating the observation sequence and ending in state i at time t is the product of the forward and backward probabilities, $\alpha_t(i)\,\beta_t(i)$. The probability $P(\mathbf{O} \mid \lambda)$ is obtained simply by summing all the forward and backward products:

$$P(\mathbf{O} \mid \lambda) = \sum_{i=1}^{N} \alpha_t(i)\, \beta_t(i). \tag{5.21}$$

Before forming the re-estimation formulas of the Baum-Welch algorithm, we define $\xi_t(i,j)$ to be the probability of the process being in state i at time t and in state j at time $t+1$:
$$\xi_t(i,j) = P(q_t = i, q_{t+1} = j \mid \mathbf{O}, \lambda). \tag{5.22}$$

With the definition of the forward-backward procedure this can be expressed in the form [Rabiner93, p. 342]:

$$\begin{aligned} \xi_t(i,j) &= \frac{P(q_t = i, q_{t+1} = j, \mathbf{O} \mid \lambda)}{P(\mathbf{O} \mid \lambda)} \\ &= \frac{\alpha_t(i)\, a_{ij} b_j(\mathbf{o}_{t+1})\, \beta_{t+1}(j)}{P(\mathbf{O} \mid \lambda)} \\ &= \frac{\alpha_t(i)\, a_{ij} b_j(\mathbf{o}_{t+1})\, \beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_t(i)\, a_{ij} b_j(\mathbf{o}_{t+1})\, \beta_{t+1}(j)}. \end{aligned} \tag{5.23}$$

The probability of being in the m^{th} mixture component of the i^{th} state at time t can be expressed as [Rabiner93, p. 352]:

$$\gamma_t(i,m) = \left[\frac{\alpha_t(i)\, \beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\, \beta_t(i)} \right] \left[\frac{w_{i,m} \mathcal{N}(\mathbf{o_t},\, \boldsymbol{\mu}_{i,m},\, \boldsymbol{\Sigma}_{i,m})}{\sum_{m=1}^{M} w_{i,m} \mathcal{N}(\mathbf{o}_t,\, \boldsymbol{\mu}_{i,m},\, \boldsymbol{\Sigma}_{i,m})} \right]. \tag{5.24}$$

For the m^{th} mixture component of the i^{th} state in an HMM, the re-estimation equations become as follows [Rabiner93, pp. 343, 351-352]:

$$\overline{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \sum_{m=1}^{M} \gamma_t(i,m)} \tag{5.25}$$

$$\overline{c}_{i,m} = \frac{\sum_{t=1}^{T} \gamma_t(i,m)}{\sum_{t=1}^{T} \sum_{m=1}^{M} \gamma_t(i,m)} \tag{5.26}$$

$$\overline{\boldsymbol{\mu}}_{i,m} = \frac{\sum_{t=1}^{T} \gamma_t(i,m)\, \mathbf{o}_t}{\sum_{t=1}^{T} \gamma_t(i,m)} \tag{5.27}$$

$$\overline{\boldsymbol{\Sigma}}_{i,m} = \frac{\sum_{t=1}^{T} \gamma_t(i,m) \cdot (\mathbf{o}_t - \boldsymbol{\mu}_{i,m})(\mathbf{o}_t - \boldsymbol{\mu}_{i,m})^T}{\sum_{t=1}^{T} \gamma_t(i,m)}. \tag{5.28}$$

Discriminative training algorithm

Musical signals are highly varying material, thus most likely the acoustic models used are not able to sufficiently model the observation statistics. It is very unlikely that a single HMM could capture all the acoustic variation in classical or jazz music, for example. Moreover, our training database is still rather small and does not suffice to reliably estimate the parameters for complex models with high amounts of component densities.

However, in classification tasks the class-conditional densities do not even have to be accurately modelled, efficient modelling of class boundaries is sufficient [Bilmes02]. Models can be trained to focus on the differences between classes using discriminative training of model parameters instead of conventional ML training. The aim of discriminative training methods, such as the maximum mutual information (MMI), is maximising the ability to distinguish between the observation sequences generated by the model of the correct class and those generated by models of other classes [Rabiner93, pp. 363-364]. This differs from the ML criterion where the aim is to maximise the likelihood of the data given the model for each class separately, see Eq. 5.12.

Different discriminative algorithms have been proposed in the literature. In this thesis, we are using the algorithm recently proposed by Ben-Yishai *et al.* [Ben-Yishai]. It is based on an approximation of the MMI criterion, and one of its benefits is a straightforward implementation. The MMI objective function is given as

$$M(\Theta) = \sum_{r=1}^{R} \left\{ \log \left[p(l^r)\, p(\mathbf{O}^r \mid l^r) \right] - \log \sum_{k=1}^{K} p(k)\, p(\mathbf{O}^r \mid k) \right\},$$

where Θ denotes model parameters for all classes, $\mathbf{O}^r$ denotes the sequence of feature vectors extracted from recording r and l^r denotes an associated class labels for it, $p(l^r)$ denotes the prior probability for the associated class labels, and $p(k)$ for class k. There exists no simple optimisation method for this problem, although an approximation can be used. The approximated maximum mutual information (AMMI) criterion was defined in [Ben-Yishai] as

$$J(\Theta) = \sum_{k=1}^{K} \left\{ \sum_{r \in A_k} \log \left[p(k)\, p(\mathbf{O}^r \mid k) \right] - \eta \sum_{r \in B_k} \log \left[p(k)\, p(\mathbf{O}^r \mid k) \right] \right\}, \tag{5.29}$$

where A_k is the set of indices of training recordings that are from class k, and B_k is the set of indices of training recordings assigned to class k by a maximum posteriori classification. The "discrimination rate" is controlled with the parameter $0 \leq \eta \leq 1$. Optimisation of Eq. 5.29 can be done for each class separately. Thus, we maximise the objective functions:

$$J_k(\Theta) = \sum_{r \in A_k} \log p(\mathbf{O}^r \mid k) - \eta \sum_{r \in B_k} \log p(\mathbf{O}^r \mid k) \tag{5.30}$$

for all the classes $1 \leq k \leq K$. This means that the parameter set of each class can be estimated separately, thus leading to a straightforward implementation.

For the general case of an HMM parameter ν, the re-estimation procedure with the ML estimation (Eqs. 5.25 to 5.28) takes the form

$$\overline{\nu}_{ML} = \frac{N(\nu)}{G(\nu)}, \tag{5.31}$$

where $N(\nu)$ and $G(\nu)$ are accumulators that are computed according to the set A_k. Correspondingly, the re-estimation procedure of the algorithm presented in [Ben-Yishai] is

$$\overline{\nu}_{AMMI} = \frac{N(\nu) - \eta N_D(\nu)}{G(\nu) - \eta G_D(V)}, \tag{5.32}$$

where $N_D(\nu)$ and $G_D(\nu)$ are the discriminative accumulators computed according to the set B_k, obtained by recognition on the training set. The recognition is done only at the first iteration, after which the set B_k stays fixed.

Based on the AMMI criterion, the following re-estimation equations for HMM parameters can be obtained [Ben-Yishai]:

$$\overline{a}_{ij} = \frac{\sum_{r\in A_k}\sum_{t=1}^{T_r-1}\xi_t(i,j) - \eta\sum_{r\in B_k}\sum_{t=1}^{T_r-1}\xi_t(i,j)}{\sum_{r\in A_k}\sum_{t=1}^{T_r-1}\sum_{m=1}^{M}\gamma_t(i,m) - \eta\sum_{r\in B_k}\sum_{t=1}^{T_r-1}\sum_{m=1}^{M}\gamma_t(i,m)} \tag{5.33}$$

$$\overline{c}_{i,m} = \frac{\sum_{r\in A_k}\sum_{t=1}^{T_r}\gamma_t(i,m) - \eta\sum_{r\in B_k}\sum_{t=1}^{T_r}\gamma_t(i,m)}{\sum_{r\in A_k}\sum_{t=1}^{T_r}\sum_{m=1}^{M}\gamma_t(i,m) - \eta\sum_{r\in B_k}\sum_{t=1}^{T_r}\sum_{m=1}^{M}\gamma_t(i,m)} \tag{5.34}$$

$$\overline{\boldsymbol{\mu}}_{i,m} = \frac{\sum_{r\in A_k}\sum_{t=1}^{T_r}\gamma_t(i,m)\,\mathbf{o}_t - \eta\sum_{r\in B_k}\sum_{t=1}^{T_r}\gamma_t(i,m)\,\mathbf{o}_t}{\sum_{r\in A_k}\sum_{t=1}^{T_r}\gamma_t(i,m) - \eta\sum_{r\in B_k}\sum_{t=1}^{T_r}\gamma_t(i,m)} \tag{5.35}$$

$$\overline{\boldsymbol{\Sigma}}_{i,m} = \frac{\sum_{r\in A_k}\sum_{t=1}^{T_r}\gamma_t(i,m)\left(\mathbf{o}_t - \overline{\boldsymbol{\mu}}_{i,m}\right)^2 - \eta\sum_{r\in B_k}\sum_{t=1}^{T_r}\gamma_t(i,m)\left(\mathbf{o}_t - \overline{\boldsymbol{\mu}}_{i,m}\right)^2}{\sum_{r\in A_k}\sum_{t=1}^{T_r}\gamma_t(i,m) - \eta\sum_{r\in B_k}\sum_{t=1}^{T_r}\gamma_t(i,m)}, \tag{5.36}$$

where $\xi_t(i,j)$ is the probability of being in state i at time t and in state j at time $t+1$ (defined in Eq. 5.23), and $\gamma_t(i,m)$ is the probability of being in m^{th} mixture component of the state i at time t (defined in Eq. 5.24).

This discriminative re-estimation can be iterated. Typically five iterations are enough, since the improvement in recognition accuracy is small beyond that. In many cases, first iteration is enough and typically gives the greatest improvement. The following iterations still increase the AMMI objective function and increase the accuracy at least in the training set. However, continuing iterations too long may cause the algorithm to overfit the parameters to the training data, leading to poor generalisation over the training data.

Classification with HMMs

Each class $k = \{1, 2, \ldots, K\}$ is represented by an HMM with model parameters λ_k. The classification is done with a maximum-likelihood classifier by finding the model which has the maximum *a posteriori* probability for the given observation sequence $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_U)$, i.e.,

$$\overline{k} \;=\; \arg \max_{1 \le k \le K} p\left(\lambda_k \mid \mathbf{Y}\right). \tag{5.37}$$

According to the Bayes's Rule this can be written as

$$\overline{k} \;=\; \arg \max_{1 \le k \le K} \frac{p\left(\mathbf{Y} \mid \lambda_k\right) p\left(\lambda_k\right)}{p\left(\mathbf{Y}\right)}. \tag{5.38}$$

By assuming all the classes equally likely, $p(\lambda_k) = \frac{1}{K}$, and since $p\left(\mathbf{Y}\right)$ is independent of k, the classification rule can be defined as

$$\overline{k} \;=\; \arg \max_{1 \le k \le K} p\left(\mathbf{Y} \mid \lambda_k\right). \tag{5.39}$$

The forward and backward procedures can be used to estimate the probability $p\left(\mathbf{Y} \mid \lambda_k\right)$. In practice, however, this probability is approximated by the probability of the most likely state sequence. This can be efficiently obtained using the Viterbi algorithm [Viterbi67]. The Viterbi algorithm is based on a similar recursive structure as the forward and backward procedures, but instead of summing the probabilities of all the possible state transitions only the transition with maximum probability is used. Multiplications of small probabilities within the algorithm require very large dynamic range, thus log-likelihood scores are computed to avoid numerical problems.

6 System evaluation

In this chapter, we describe the examined classification approaches and present the obtained results. This chapter is divided into three parts. In the first part, recognition accuracies obtained for different classification approaches are evaluated for the automatic musical genre recognition. In addition, the recognition accuracies are compared with the human ability studied earlier in Chapter 4. In the second part, detecting the used instruments in music signals is evaluated. In the third part, the detecting drum segments in music signals is evaluated. All the algorithms and the evaluations were implemented in Matlab, which is an efficient tool for testing and evaluating signal processing algorithms.

6.1 Musical genre recognition

The automatic recognition of real-world audio signals according to musical genres is studied in this section. Mel-frequency cepstral coefficients are used to represent the time-varying magnitude spectrum of the music signal, and the recognition is done based on these features. The music database described in Chapter 3 is used to train statistical pattern recognition classifiers. We use the optimal Bayes classifier, and assume equal prior probabilities for the classes. Different parametric models for modelling the class-conditional densities are experimented. Two classification schemes were studied, one using only higher-level genres as classes, and one using also subgenres and allowing misclassifications at the level of higher-level genres.

6.1.1 Method of evaluation

Database

The utilised music database was described in detail in Chapter 3. Pieces from six higher-level genres (classical, electronic/dance, hip hop, jazz/blues, rock/pop, and soul/RnB/funk) were used in the evaluations, totalling 488 pieces for evaluating the developed system. Pieces annotated as a world/folk were excluded from the evaluations due to the miscellaneous nature of the genre. Manually annotated approximately one-minute long interval within each piece was used to represent the piece in the simulations.

Procedure

The feature set used in the simulations included 16-dimensional MFCC and ΔMFCC feature vectors. Features were extracted in 20 ms windows. In order to avoid numerical problems caused by small variances of the feature values, all the features were normalised to have zero mean and unity variance. The normalisation was done based on statistics calculated from the whole training data.

Typically, the features used (MFCC and ΔMFCC) would have been catenated to form one feature vector, but it is possible that one model cannot efficiently model feature distributions of the both features. Thus two separate models were trained, one for the MFCC and one for the ΔMFCC. The likelihoods were calculated at first individually for both of them. For the HMM, the likelihoods were calculated for the whole analysis segment with the Viterbi algorithm. For the GMM, the likelihoods were obtained by pooling all likelihoods of the frames within the analysis segment. The feature streams were assumed independent, thus the likelihoods obtained for the MFCC and for the ΔMFCC were joined together by multiplying them. Further, the feature vector dimension and the analysis segment length were equal for the feature streams, so no normalisation was required for the likelihoods. The classification was performed by selecting the class with the highest likelihood. In preliminary evaluations this classification scheme was found to give better results than using a shared model for both features. A 25-second long analysis segment, from the beginning of the annotated interval, was used in the classification unless otherwise reported.

The pieces in the database were randomly divided into two separate sets, the training set and the test set. Sets were formed by assigning 70 % of the pieces into the training set and 30% into the test set. In order to ensure that the recognition accuracy would not be biased because of a particular partitioning of training and testing, this random division was iterated five times. The overall recognition rate was obtained as the arithmetic mean of recognition rates of the individual iterations. Due to the varying number of pieces in different classes (genres), the same weight was given to all the classes by computing the recognition rate as follows:

1. For each class, the recognition rate was calculated as a percentage of correctly classified pieces among all the classified pieces.
2. The overall recognition rate was calculated as the arithmetic mean of recognition rates of the individual classes.

6.1.2 Results

GMM

The use of the GMM for modelling the class-conditional densities was evaluated by varying the number of Gaussian densities in the mixture model. The recognition accuracies as a function of the number of Gaussians for the both classifiers, one

Table 6.1: GMM classification accuracy mean and standard deviation with different model orders.

Model order	MFCC	ΔMFCC	combined
2	49±3	48±5	**54**±4
4	51±2	50±4	**57**±4
8	53±3	55±3	**59**±4
16	56±4	56±2	**61**±4
32	55±3	58±3	**61**±4
64	56±3	59±3	**62**±4

Table 6.2: Genre confusion matrix for GMM (model order 64) trained with MFCC. Entries are expressed as percentages and are rounded to the nearest integer.

Tested \ Recognised	class	electr	hipH	jazzB	rockP	soulR
classical	**89**	1	1	8	0	1
electronic/dance	3	**30**	30	11	12	14
hip hop	-	2	**68**	3	18	8
jazz/blues	8	2	5	**50**	27	7
rock/pop	4	6	4	19	**52**	15
soul/RnB/funk	6	4	11	11	24	**44**

trained for the MFCC and one for the ΔMFCC, are presented in Table 6.1. The recognition results obtained by joining output likelihoods of these classifiers are also presented in the table. The ± part shows the standard deviation of the recognition accuracy for the train/test set iterations. The GMMs were trained using 40 iterations of the Baum-Welch algorithm.

Feasible recognition accuracies were obtained already with classifiers trained for MFCCs and ΔMFCCs alone. The recognition accuracies seemed to saturate when the model order was 16 and only slightly increased after that. Interestingly, the accuracy improved consistently while the model order was increased, and no overfitting of the models to the training data was observed. In general, the best performance was obtained using 64 Gaussian distributions in the mixture model. The classification accuracy of a random guess is 16.6 % with six classes. Quite surprisingly, the classifier trained for ΔMFCCs performed a bit better than one trained for MFCCs. One possible reason for the rather high performance obtained using only ΔMFCCs is transient-like sounds, generally present in music, which are likely to show in delta coefficients. Tables 6.2 and 6.3 show the confusion matrix for the best-performing MFCC and ΔMFCC configuration. The confusions made by the classifiers were somewhat different, thus indicating that the classifiers used separate information about the signal.

Table 6.3: Genre confusion matrix for GMM (model order 64) trained with ΔMFCC. Entries are expressed as percentages and are rounded to the nearest integer.

Tested \ Recognised	class	electr	hipH	jazzB	rockP	soulR
classical	**83**	0	-	5	12	-
electronic/dance	6	**48**	13	3	20	9
hip hop	-	7	**72**	-	7	15
jazz/blues	4	6	1	**53**	27	8
rock/pop	7	7	-	14	**56**	16
soul/RnB/funk	1	6	17	17	19	**40**

Table 6.4: Genre confusion matrix for system joining classifiers for MFCC and ΔMFCC (both GMM with model order 64). Entries are expressed as percentages and are rounded to the nearest integer.

Tested \ Recognised	class	electr	hipH	jazzB	rockP	soulR
classical	**88**	-	-	8	2	3
electronic/dance	3	**41**	25	5	14	12
hip hop	-	2	**83**	3	7	5
jazz/blues	4	2	3	**54**	27	10
rock/pop	2	6	2	21	**50**	20
soul/RnB/funk	3	6	15	8	17	**51**

The recognition accuracy was improved consistently by joining the likelihoods of these two classifiers. The confusion matrix for the best-performing configuration is presented in Table 6.4. Classical and hip hop are both recognised accurately. Rock/pop is often misclassified as jazz/blues or soul/RnB/funk. The confusion matrix also shows that some of the misclassifications are very similar to those of humans. For example, soul/RnB/funk is in some cases very close to popular and hip hop music.

HMM

Next, we studied the use of the HMM for modelling the class-conditional densities. Unlike the GMM, also the temporal structure of the data is taken into account with the HMM, since the HMM exploits transition probabilities between states. In order to get some kind of idea what was modelled by different HMM states, the Viterbi segmentation was visually studied after the training. In Figures 6.1 and 6.2, a three-state HMM has been trained using one Gaussian component per each state. The top panel shows the audio signal and the bottom panel shows the resulting Viterbi

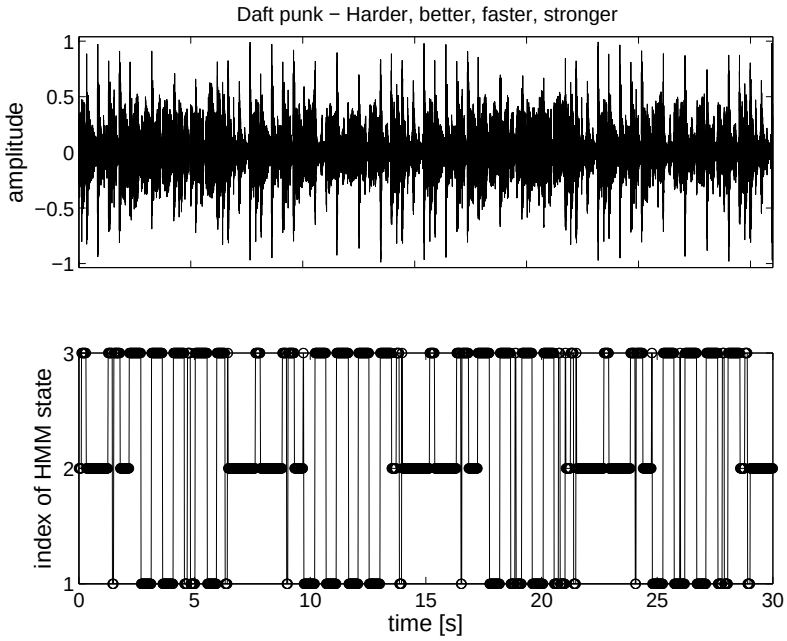


Figure 6.1: The top panel shows the time-domain signal of a piece and the bottom panel shows the Viterbi segmentation through a three-state HMM trained with the piece.

segmentation into the three states.

Figure 6.1 depicts a quite representative piece of modern dance music; its repeating structure can even be seen from the audio signal. The HMM states seem to be modelling parts of the repeating lower-level structure and the structure can be seen nicely from the state transitions. A more broader instrumentation is used in the piece shown in Figure 6.2, and the HMM states seem to model much larger structural parts in the piece. However, states can be interpreted to model also properties of individual sound events.

The states can be modelling some predominant instruments, some other properties of sound, or the structure of the music piece. In general, forcing the states to model certain properties of the music is difficult. Music should be segmented before training and we do not have a clear view what segmentation scheme would be advantageous, for example, in the musical genre recognition.

The genre recognition with the HMM was evaluated with a varying number of states and Gaussians used to model the state output distributions. Table 6.5 shows recognition accuracies as a function of the model order and number of states obtained by joining the output likelihoods of separate classifiers trained for MFCCs and ΔMFCCs. The training was done in an unsupervised manner, because we do not know what are the underlying acoustic classes that are modelled with different HMM states. The number of iterations in the Baum-Welch training algorithm was 10. Two model topologies were studied: fully-connected and left-right. On the whole, left-right topology performed better. There is no clear consistency on the results, partly

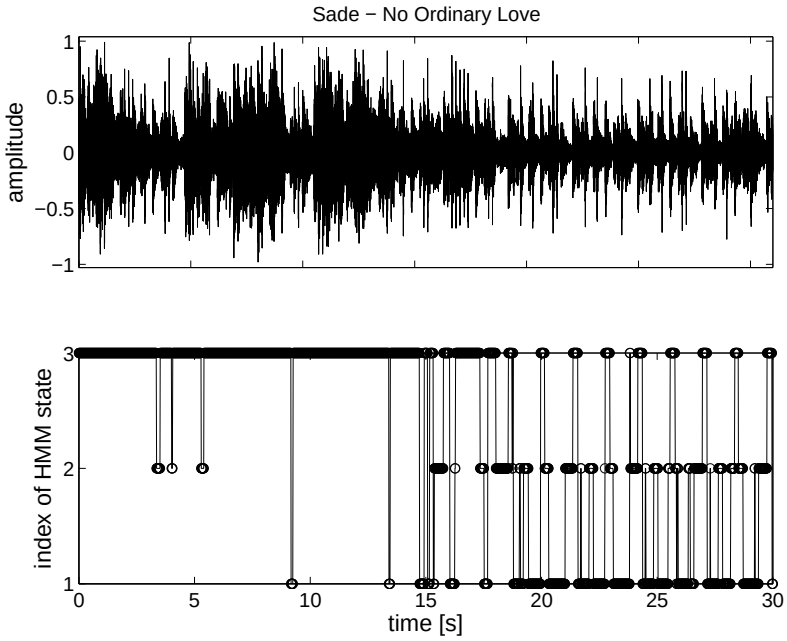


Figure 6.2: The top panel shows the time-domain signal of a piece. The bottom panel shows the Viterbi segmentation through a three-state HMM trained with the piece.

Table 6.5: Mean and standard deviation of classification accuracies with varying HMM model topology and complexity. NS denotes the number of states, and NC denotes the number of component densities used to model output distribution of a state.

	Fully-connected					Left-right				
NC \ NS	2	3	4	5		NC \ NS	2	3	4	5
1	52±4	56±3	55±3	58±4		1	52±4	57±4	58±3	58±3
2	57±3	58±3	58±3	58±2		2	57±3	59±5	59±3	59±3
3	58±2	59±3	59±4	59±3		3	58±2	61±4	59±5	61±3
4	59±3	59±3	61±3	60±2		4	58±2	61±4	60±3	61±3
5	59±2	60±3	59±3	61±2		5	59±4	60±3	61±4	61±3
6	60±3	59±4	59±3	59±2		6	61±3	60±4	61±5	61±3
7	59±3	60±4	60±2	60±2		7	60±3	60±4	61±4	61±4
8	59±3	62±3	60±2	60±3		8	60±3	60±4	60±4	61±4

due to the sensitivity of the Baum-Welch algorithm for the initial parameters. On average, incrementing the model order only slightly increased performance. The number of states had not direct influence on the recognition accuracy.

A confusion matrix for one of the best-performing configurations is shown in Table 6.6. The results were obtained with a five state left-right model topology, and

Table 6.6: Genre confusion matrix for HMM. Entries are expressed as percentages and are rounded to the nearest integer.

Tested \ Recognised	class	electr	hipH	jazzB	rockP	soulR
classical	**88**	-	-	8	2	3
electronic/dance	3	**41**	25	5	14	12
hip hop	-	2	**83**	3	7	5
jazz/blues	4	2	3	**54**	27	10
rock/pop	2	6	2	21	**50**	20
soul/RnB/funk	3	6	15	8	17	**51**

using four Gaussians to model the output distribution of the states. The initial parameters for the Baum-Welch algorithm were obtained using K-means clustering initialised randomly. In order to get a realistic view of the recognition accuracy of the HMM, the training process was also iterated five times. The overall recognition accuracy was obtained as an arithmetic mean of recognition accuracies for the individual iterations. The recognition accuracy was 61±3%, and it is very close to the results obtained with the GMM. In addition, the misclassifications are very similar to the ones made by the GMM based recognition system.

HMM with discriminative training

The purpose of this set of simulations was to evaluate the usefulness of discriminative training for HMMs presented in [Ben-Yishai]. These results have been earlier presented in [Eronen03b].

The discriminative training algorithm presented in Section 5.4.2 is rather time-consuming because the classification has to be done also for the training set. In order to reduce the computational load, evaluations were done only for one train/test set division and only for the MFCCs. The Baum-Welch algorithm was used to train the baseline HMMs, which were further trained with the discriminative training algorithm. A maximum of five iterations of the discriminative training algorithm with a fixed discrimination rate ($\eta = 0.3$) was found to give an improvement in most cases without much danger of overfitting.

Table 6.7 shows the recognition accuracies as a function of the model order and number of states with varying model topologies and training methods. As one can see, discriminative training gives an improvement of only a few percentage points. However, improvement is observed almost consistently across the tested model orders, number of states, and topologies. Results presented here are not comparable with previously presented results, since we are using only static MFCCs as features. Furthermore, it as previous results showed, the joint use of the MFCCs and the ΔMFCCs should give the best recognition accuracies.

Table 6.7: Genre recognition accuracies for both training methods with varying model topology and complexity. NS denotes the number of states, and NC denotes the number of component densities used to model the output distribution of a state.

Fully-connected

NS \ NC	1	2	3	4
	Baum-Welch			
3	52.9	53.8	53.8	55.7
4	54.3	54.6	55.0	56.8
	Discriminative			
3	53.4	56.9	56.1	58.5
4	56.7	54.4	57.6	59.6

Left-right

NS \ NC	1	2	3	4
	Baum-Welch			
3	52.4	53.7	54.7	55.2
4	54.0	55.0	56.2	56.1
	Discriminative			
3	54.7	55.5	58.1	58.2
4	55.7	55.9	58.0	57.9

For HMMs with a small number of states and component densities, the discriminative training clearly improved the recognition accuracy. For more complex state densities no improvement was observed. This is due to the overfitting of the models to the training data, which leads to a poor generalisation to unseen test data.

Genre hierarchy in the classification

In the previous evaluations, each higher-level genre was treated as a single class. Since the higher-level genres used are formed hierarchically from subgenres, we also studied the use of these subgenres in the recognition. When we are using class models only for higher-lever genres, for example, jazz piece might be misclassified as a soul/RnB/funk due to the poor fit of the model for the jazz/blues genre. However, by using also the subgenres in the classification this might be avoided. The class model just for the jazz or even for blues might fit better than one for jazz/blues. The classification results are still examined only at the level of the higher-level genres by allowing misclassification among subgenres.

The main idea was to use the higher-level genres and the subgenres under those at the same time in the classification by relying on the assumption that most of the misclassification will occur within a higher-level genre. The subgenres under the higher-level genre are more likely to have common characteristics with each other rather than with subgenres under some other higher-level genre. For this reason, subgenres are more likely to be misclassified with the subgenres under the same higher-level genre. So the model was trained not only for higher-level genre e.g. jazz/blues, but also for subgenres blues and jazz, and in the classification misclassifications are allowed among all these classes (jazz/blues, blues and jazz). The classes used in the classification are shown in Table 6.8 along with the amount of the available training and test data.

Only the use of the GMM for modelling the class-conditional densities was evaluated in this set of simulations. The train/test set division was iterated five times as earlier.

Table 6.8: The classes used in the classification. The higher-level genres are denoted as bold text and the boxes indicate higher-level genres. The number of training and test pieces for every genre are also presented.

Musical genre	#train	#test
classical	70	36
chamber music	16	8
classical general	26	14
solo instruments	8	4
symphonic	7	3
vocal	4	2
electronic / dance	45	26
ambient	4	3
breakbeat/drum'n'bass	7	4
dance	9	6
house	7	4
techno	18	9
hip hop	25	12
jazz / blues	63	33
blues	22	10
jazz	28	16
fusion jazz	7	4
latin jazz	6	3
rock / pop	74	47
country	7	5
metal	10	6
pop	22	14
rock	35	22
soul / RnB / funk	37	19
funk	9	4
RnB	10	5
soul	18	10

Although we had very different amounts of training material for each class, the evaluations were done with a fixed number of Gaussians for all classes. The obtained results are presented in Table 6.9. For the classification scheme also using subgenres, the results were calculated as a mean of recognition accuracies for individual higher-level genres. The recognition accuracy for a higher-level genre was calculated as a percentage of pieces correctly classified either directly into higher-level genre or into one of its subgenres. The recognition accuracy was improved only by a few percentage points in comparison to the classification scheme using only higher-level genres. However, the improvement can be observed consistently across the tested model orders. The improvement was biggest with lower model orders. The small improvement can be partly explained with the insufficient amount of training data for some of the subgenres. The best performance was obtained using 64 Gaussian distributions in the mixture model.

In Table 6.10, a confusion matrix at the level of the higher-level genres is presented for the best-performing configuration (model order 64). More detailed information about the performance for all the classes used in the evaluations is shown in Table 6.11. The recognition rate at the subgenre level was 25 % while the random-guess rate is 4.3 %. This classification scheme seems to work nicely for genres like classical and rock/pop. For the electronic/dance the scheme did not work as intended, but fortunately, pieces from other genres did not confuse with its subgenres. Some of the misclassifications were expected, for example, RnB is quite easily confused with

Table 6.9: Genre recognition with different classification schemes. Mean recognition accuracies with standard deviation are presented for different model orders of the GMM.

NC	higher-level genres	including subgenres
2	54±4	56±4
4	57±4	60±4
8	59±4	60±3
16	61±4	62±3
32	61±3	62±4
64	62±3	63±4

Table 6.10: Genre confusion matrix at the level of the higher-level genres for the best-performing hierarchical classification scheme. Entries are expressed as percentages and are rounded to the nearest integer.

Tested \ Recognised	class	electr	hipH	jazzB	rockP	soulR
classical	**91**	1	-	6	1	2
electronic/dance	3	**41**	22	6	19	8
hip hop	-	2	**78**	2	12	7
jazz/blues	4	1	1	**58**	30	7
rock/pop	2	5	1	15	**63**	13
soul/RnB/funk	4	4	15	15	18	**44**

hip hop and electronic/dance.

Recognition rate as a function of analysis excerpt duration

We also studied the recognition rates as a function of the analysis excerpt length. There is a remarkable uncertainty of the performance measurement with the shorter analysis excerpt durations since the performance greatly depends on the starting point of the excerpt. The beginning of the one-minute representative part was used as a starting point in these simulations. When the duration of analysis excerpt increases the uncertainty of the performance measurement decreases.

In Figure 6.3, the recognition rates of different classifiers are shown as a function of the analysis excerpt duration. The graph was created by increasing the excerpt duration gradually from 0.15 seconds to 28 seconds. The behaviour of three different classifiers was studied: two using GMM to model class-conditional densities with 64 Gaussians, and one using HMM modelling with five-state left-right topology and four Gaussian components. For the GMM, two classification schemes were studied: one using only higher-level genres, and one using also subgenres. Increasing the

Table 6.11: Genre confusion matrix including subgenres. Entries are expressed as percentages and are rounded to the nearest integer. Higher-level genres are denoted as bold text. The boxes indicate higher-level genres.

tested \ rec.	**classical**	chamber	general	solo	symphonic	vocal	**electronic/dance**	ambient	breaks	dance	house	techno	**hip hop**	**jazz/blues**	blues	jazz	jazzFusion	jazzLatin	country	metal	pop	rock	**rock/pop**	country	metal	pop	rock	**soul/RnB/Funk**	funk	RnB	soul
classical	**64**	8	-	-	-	-	4	-	-	-	-	-	-	8	-	-	-	-	-	-	-	-	4	-	-	-	-	-	8	-	4
chamber	32	**58**	8	-	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
general	70	-	**14**	3	1	-	-	-	-	-	-	-	-	9	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-
solo	15	-	-	**85**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
symphonic	47	-	13	-	**40**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
vocal	90	-	-	-	-	-	-	-	-	-	-	-	-	10	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ambient	-	-	-	7	-	-	27	-	-	-	-	-	20	13	-	-	-	-	-	-	-	-	-	7	-	-	27	-	-	-	-
breaks	-	-	-	-	-	-	5	-	-	15	-	-	40	10	-	-	-	-	-	-	10	-	-	-	10	-	-	-	-	-	-
dance	-	-	-	-	-	-	30	-	-	-	-	-	30	-	-	-	-	-	-	-	-	20	-	-	-	-	-	-	-	-	-
house	-	-	-	-	-	-	20	-	-	-	-	-	35	5	15	-	-	-	-	-	-	-	5	-	-	-	15	-	-	-	-
techno	4	-	2	-	-	-	47	-	-	2	4	**18**	4	-	-	-	-	-	-	-	4	2	-	7	-	4	-	-	-	4	-
hip hop	-	-	-	-	-	-	2	-	-	-	-	-	**78**	2	-	-	-	-	-	-	-	-	8	7	-	-	-	-	-	-	-
blues	4	-	-	2	-	-	2	-	-	-	-	-	2	38	**4**	-	-	-	-	-	-	-	10	10	2	-	-	-	-	-	-
jazz	4	-	1	-	-	-	-	-	-	-	-	-	-	36	-	**25**	1	-	-	-	-	3	8	6	-	-	1	-	-	-	-
jazzFusion	-	-	-	-	-	-	-	-	-	-	-	-	-	40	-	10	**25**	-	-	-	-	-	-	-	-	-	-	-	-	-	-
jazzLatin	-	-	-	-	-	-	-	-	-	-	-	-	-	53	-	7	-	-	-	-	-	-	7	-	-	-	-	-	-	-	-
country	-	-	-	-	-	-	-	-	-	-	-	-	-	28	-	8	-	-	-	-	-	4	4	32	**24**	-	4	4	-	-	-
metal	-	-	-	-	-	-	10	-	-	-	-	-	-	-	3	-	-	-	-	-	**50**	7	-	30	-	**50**	7	-	-	-	-
pop	4	1	-	-	-	-	7	-	-	-	-	-	3	3	-	1	1	-	-	-	4	**19**	3	26	-	4	**19**	3	14	-	11	1
rock	1	-	-	-	-	-	2	-	-	-	-	1	1	18	-	2	-	-	-	-	-	-	**26**	33	6	-	-	**26**	8	-	-	3
funk	-	-	-	-	-	-	10	-	-	-	-	-	25	20	-	-	-	-	-	-	-	-	-	40	-	-	-	-	-	**5**	-	-
RnB	-	-	-	-	-	-	4	-	-	-	-	-	28	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	28	-	**40**	-
soul	6	-	2	-	-	-	2	-	-	-	-	-	4	20	-	-	-	-	-	-	-	-	2	16	-	-	-	-	36	4	-	**8**

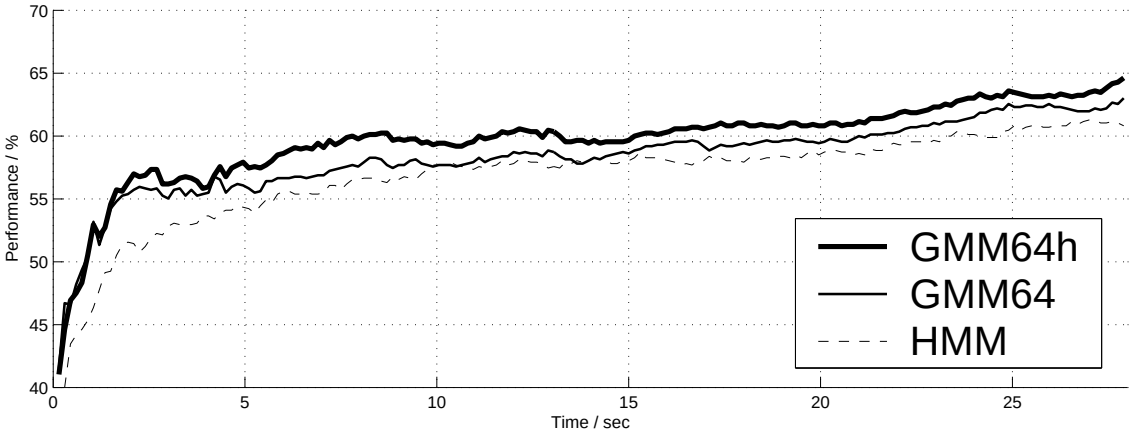


Figure 6.3: Recognition rates as a function of analysis excerpt duration for following classification approaches: GMM and HMM with six higher-level genres, and GMM with the use of genre hierarchy (denoted with GMM64h).

excerpt duration improves the overall recognition accuracy as expected. It should be noted that the recognition accuracy grows rapidly until the analysis duration reaches two seconds and further lengthening of the duration does not give such a significant improvement anymore. For both of the classifiers using GMM, the recognition accuracy reaches reasonable level (over 50%) already within 0.5 seconds. For the classifier using the hierarchical classification scheme, the recognition curve seems to converge around 7 seconds to the level of 60%. A small increase can be noticed after 20 seconds, ending up to the level of 64% at 27 seconds. The learning curves of studied classifiers are rather close to each other, and the trend of them is ascending on average. The recognition accuracy for the classifier using HMM did not increase as rapidly as with classifiers using the GMM.

6.1.3 Comparison with human abilities

An experiment was conducted to enable a direct comparison with human abilities (see Chapter 4). Since musical genre recognition is not a trivial task, this is essential in order to get a realistic comparison of the recognition rates. The full stimulus set used in the listening experiment was employed in the simulations. In the listening experiment, the test subjects were divided into two groups each having a separate set of stimuli. In training of the computational model, separate training sets were formed for both of these groups from the music database, excluding only the pieces used in that group in the listening experiment. This makes good use of the available training data and guarantees that system has not heard the piece being tested before.

Table 6.12 shows the main results of the experiment. The recognition was done with the developed recognition system and the results were pooled as in Chapter 4. Two classification schemes were examined, basic classification with only higher-level

Table 6.12: Direct comparison to human abilities. Mean and standard deviation of the recognition rates of classifiers using GMM and HMM.

Sample duration	Human	GMM				HMM
		NC16	NC32	NC64	NC64 hierarchy	NC4 NS5
250 ms	**57**	48±7	48±7	47±7	46±8	47±7
500 ms	**63**	53±8	52±5	52±6	52±5	50±8
1500 ms	**69**	56±6	54±6	55±6	55±6	54±6
5000 ms	**75**	57±7	57±7	59±6	59±6	59±6

Table 6.13: Confusion matrix for the genre recognition with 250 ms excerpts. Entries are expressed as percentages.

Tested \ Recognised	class	electr	hipH	jazzB	rockP	soulR
classical	**72**	6	2	14	4	2
electronic/dance	8	**33**	14	14	17	14
hip hop	-	11	**55**	7	11	17
jazz/blues	8	8	4	**40**	20	21
rock/pop	1	7	5	29	**41**	17
soul/RnB/funk	3	18	9	16	16	**38**
Totals	15	14	15	20	18	18

genres and classification including also the subgenres. The use of the GMM was evaluated with a varying model order. The use of the HMM was evaluated only with the best-performing left-right model configuration. The evaluated classifier configurations performed in the same way throughout the simulation. Only a small improvement was obtained by increasing the sample duration. Human listeners outperformed the developed system by 10-16% depending on the sample duration. The performance gap between humans and the developed system is larger with longer samples. Increasing the excerpt duration does not bring any essentially new information to the recognition, since the developed system only uses frequency domain features excluding all the long-term (rhythmic) information of the music.

Tables 6.13 and 6.14 show the confusion matrices for the best-performing configuration. If we compare them with earlier presented human confusions (see Tables 4.3 and 4.4) we can see that confusions are rather similar. The recognition of classical music is easy for both, and soul/RnB/funk is confused with many other genres.

"Agreement scores" were calculated in order to more precisely evaluate the similarity of the confusions made by the developed recognition system and human listeners. The agreement is defined here as a similar classification result among the developed recognition system and human listeners. Two different "agreement scores" are calculated:

Table 6.14: Confusion matrix for the genre recognition with 5000 ms excerpts. Entries are expressed as percentages.

Tested \ Recognised	class	electr	hipH	jazzB	rockP	soulR
classical	**91**	3	-	3	3	2
electronic/dance	9	**35**	16	9	19	13
hip hop	-	2	**72**	3	10	14
jazz/blues	6	3	1	**50**	30	11
rock/pop	-	5	3	20	**61**	12
soul/RnB/funk	5	12	7	5	22	**51**
Totals	18	10	16	15	24	17

Table 6.15: "Agreement scores" between the developed recognition system and humans. Agreement among right classifications is denoted with A and among wrong classifications is denoted with B. Entries are expressed as percentages and are rounded to the nearest integer.

Duration		Overall	class	electr	hipH	jazzB	rockP	soulR
250 ms	A	70	94	59	81	55	67	33
	B	21	11	28	23	19	19	21
500 ms	A	76	93	73	89	63	74	34
	B	23	9	22	27	27	20	23
1500 ms	A	79	99	72	92	61	79	44
	B	20	8	17	27	27	17	17
5000ms	A	84	99	81	95	74	86	49
	B	18	0	15	23	19	25	14

- Percentage of agreement among samples that were classified correctly by the developed recognition system (denoted here with A).
- Percentage of agreement among samples that were classified wrong by the developed recognition system (denoted here with B).

The calculated scores for the best-performing configuration are shown in Table 6.15. Humans agreed with the developed system in over 70 % of the cases where the developed system recognised the right genre, and in around 20 % of the cases where the developed system failed. Based on these scores one can conclude that the developed system and the human listeners agreed most of the time with the correct classifications. However, with wrong classifications the developed system mostly picked a different genre than the human listeners. When the sample duration increased, the agreement among correct classifications gradually increased, but the agreement among false classifications remained almost the same. Agreement is high among the correct classifications especially for classical and hip hop. For classical,

this can be explained with high recognition accuracy for humans and the developed system. There is low agreement among the correct classification of soul/RnB/funk, highlighting the fact that the genre is rather close to the other genres.

6.1.4 Discussion

Despite the rather fuzzy nature of the musical genres, automatic musical genre recognition could be performed with an accuracy much beyond the random-guess rate. Moreover, the obtained results were comparable to the human abilities to recognise musical genres. Reasonable recognition rate (60 %) was already obtained with seven-second long analysis excerpts.

The performance differences between the evaluated classifiers were only marginal. However, the parameter estimation is more straightforward for the GMM classifier. The discriminative training for HMM provided minor improvements with small model orders. The classification scheme using also subgenres proved to be the best approach. However, this approach needs a much larger and wider database in order to accurately model smaller subgenres.

6.2 Instrument detection

Instrument recognition has proved to be a demanding task even for monophonic music played by a single instrument [Martin99, Eronen01]. In this section, the detection of instruments used in polyphonic music signals is studied. The basic idea of "detection" differs from classification. Unlike classification, we have a prior hypothesis about the instrument present in the music and we are verifying that hypothesis.

Detection of the instruments used provides useful information for many MIR related tasks, e.g. for automatic musical genre recognition. Some of the instruments are characteristic for some genres. For example electric guitar is a pretty dominant instrument in rock/pop, but is hardly ever used in classical music. Detection of the following five instruments was attempted:

- bowed (including all bowed instruments, e.g. violin, viola, cello, and string section)
- electric guitar
- piano
- saxophone (including all saxophones: tenor, alto, soprano, and baritone)
- vocals (including choir)

The magnitude spectrum of the music signal was represented with MFCCs and ΔMFCCs. The class-conditional densities were modelled either with a GMM or an HMM. Two models were trained, one for music including the instrument and one for music including all kind of instruments. The detection was done based on the likelihood-ratio of these models for the given music. The likelihoods were obtained by multiplying the likelihoods for the separate models trained for the MFCCs and the ΔMFCCs.

6.2.1 Detection

The task of instrument detection is to determine if the hypothesised instrument I is played in a given segment of music, Y. The instrument detection can be also referred to as instrument verification. The segment of music will most probably contain other instruments too, and these are just considered as noise. The detection methods used here are adopted from speaker detection [Campbell97]. The detection task can be restated as a test between the hypotheses:

H_0: The hypothesised instrument I is played in Y
H_1: The hypothesised instrument I is not played in Y

H_0 is represented by a model denoted λ_{hyp} that characterises the hypothesised instrument I in the feature space of the music, x. The test to decide between these two hypothesis for feature vector $\mathbf{x}$ is a log-likelihood ratio test given by:

$$\log p\left(\mathbf{x} \mid \lambda_{hyp}\right) - \log p\left(\mathbf{x} \mid \lambda_{\overline{hyp}}\right) \quad \begin{cases} \geq \theta, & accept\, H_0 \\ < \theta, & reject\, H_0 \end{cases}, \tag{6.1}$$

where $p\left(\mathbf{x} \mid \lambda_{hyp}\right)$ is the likelihood of the hypothesis H_0 evaluated for the observed features $\mathbf{x}$, and $p\left(\mathbf{x} \mid \lambda_{\overline{hyp}}\right)$ is likelihood of the hypothesis H_1. The decision threshold for accepting or rejecting H_0 is θ. The evaluated classifiers were used to produce values for the two likelihoods, $p\left(\mathbf{x} \mid \lambda_{hyp}\right)$ and $p\left(\mathbf{x} \mid \lambda_{\overline{hyp}}\right)$.

The model for the hypothesis H_0 is clearly defined, and can be estimated using training material with instrument I present. The estimation of the model $\lambda_{\overline{hyp}}$ is not straightforward, since it potentially must represent the entire space of possible alternatives to the hypothesised instrument. The alternative hypothesis is modelled here by pooling music with several instruments and training a single background model. In preliminary studies, it was found to give better results to also include music with the hypothesised instrument in the training set of the background model. The advantage of this is that a single instrument-independent model is trained only once and then used as background model for all hypothesised instruments in the study.

6.2.2 Method of evaluation

Database

The utilised music database was described in Chapter 3. A homogeneous ten-second interval with annotated instruments was used in the evaluations. In order that the segment was accounted for the instrument had to be the most predominant or the second most predominant accompanying instrument or one of the melody instruments. Due to different instrumentation in the musical genres instruments did not occur evenly among the musical genres (see Table 3.2). This had to be taken into consideration when analysing the obtained results.

Procedure

Since there was only a rather small amount of evaluation material available, 60 % of it was assigned into a training set and the rest 40% to the test set. In order to ensure that the detection accuracy would not be biased because of a particular partitioning into train and test sets, evaluation was repeated for five random divisions.

There are two types of errors made by the detection system: false acceptance (FA) and false rejection (FR). Either one of the errors can be reduced at the expense of the other. A single performance measure is inadequate to represent the capabilities of the system, since the system has many levels of sensitivity. The ideal tradeoff between the two types of errors depends on the needs of the application, whether a low FA or a low FR is more critical. The levels of both types of errors are represented by a performance curve when measuring the overall system performance. The rates of the FA and the FR are adjusted by changing the value of the decision threshold θ. The equal error rate (EER) is an operating point having equal rates of FA and FR. The EER was used to produce the detection rate in the evaluations.

6.2.3 Results

The instrument detection was evaluated by varying parameters of the classifiers. In the preliminary evaluations, it was noticed that neither the number of states nor the model order had any significant effect on the detection accuracy while using HMMs. So HMMs were tested only with the fixed parameter set and only the model topology was varied. A three-state HMM using six Gaussians to model the output distribution of states was selected to be used in the evaluations. The decision threshold was instrument-specific and the common threshold was used throughout the train/test set iterations. The detection accuracies for the tested classifiers are presented in Table 6.16. The detection rates obtained with the HMMs were very similar to the ones obtained with the GMMs.

The highest detection accuracy was achieved for bowed instruments. The bowed instruments are used mainly in classical music, thus the detector was making the detection mainly based on the musical genre and not the instruments used. This

Table 6.16: Instrument detection results. Entries are expressed as percentages.

		bowed	electric guitar	piano	saxophone	vocals
GMM	NC16	84	65	74	64	78
	NC32	85	63	73	65	78
	NC64	86	63	71	67	77
HMM	fullyC	84	64	73	65	77
NS3 NC6	leftR	84	63	74	68	76

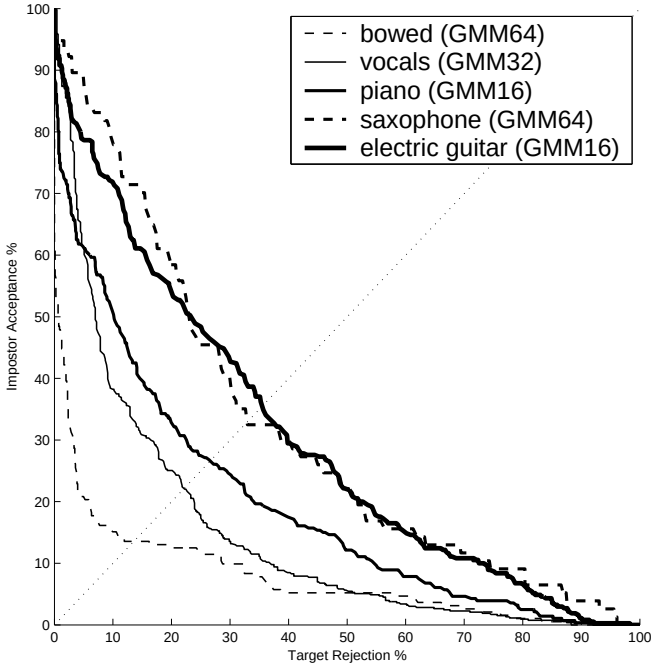
partly explains the good performance for bowed instruments. Acceptable detection accuracies were obtained for piano and vocals. Piano is widely used in many different genres, and thus the detection of piano was mainly based on the presence of the instrument. Although vocals are present in many genres, the problem here was that pure instrumental pieces are very rare in some genres, e.g. in hip hop or in soul/RnB/Funk. Nevertheless, the detection was mainly done based on the presence of vocals. Very low (around 65 %) detection accuracies were achieved for electronic guitar and saxophone. However, the accuracies were still better than pure chance (50 %).

The Receiver Operating Characteristic curve (ROC) is used to depict the tradeoff between FA and FR for the detection system [Egan75]. The probability of the FA versus probability of the FR is plotted by changing the threshold of acceptance for the log-likelihood ratio. The equal error probability is denoted by a dotted diagonal line in the ROC curve. Figure 6.4 shows the ROC curves for the best-performing detector configurations. The closer to the origin the curve is the better the performance of the detection system.

6.2.4 Discussion

Despite the demanding conditions (polyphonic multi-instrument real-world music) the detection for the evaluated instruments could be performed with rates significantly better than chance. Acceptable detection accuracy was obtained for bowed, vocals, and piano. The achieved detection accuracy for vocals is comparable to detection results (80 %) presented in [Berenzweig01]. One of the problems in the simulations was that some of the instruments are mainly used in particular genres only. Thus it was hard to form a well-balanced evaluation database and this caused the detector to detect eventually the musical genre rather than the instrument.

Figure 6.4: Plot of ROC curves for the best-performing configuration of the instrument detectors. EER is denoted by a dotted diagonal line.



6.3 Locating segments with drums

The objective here was to segment musical signals according to the presence or absence of drum instruments. Two different approaches were taken to solve the problem. The other one was based on periodicity detection in the amplitude envelopes of the signal at subbands, as discussed in Section 5.3. The other mechanism applied straightforward acoustic pattern recognition approach with MFCCs as features and a GMM and k-nearest neighbour classifiers. Approaches and the simulations results have been previously presented in [Heittola02].

6.3.1 Test setup

A subset of the music database described in Chapter 3 was used to evaluate the two drum detection schemes. Detailed statistics of the used database was shown in Table 3.3. The annotations of segments with and without drums were used with a precision of one second in the simulations, and only more than five second long stable segments were included in the simulations. The evaluation database was not nicely balanced from the point of view of the amount of material with and without drums in each individual genre. This was expected, since drums are a basic element in many Western musical genres.

In order to assure that we have our train and test sets as balanced as possible, the

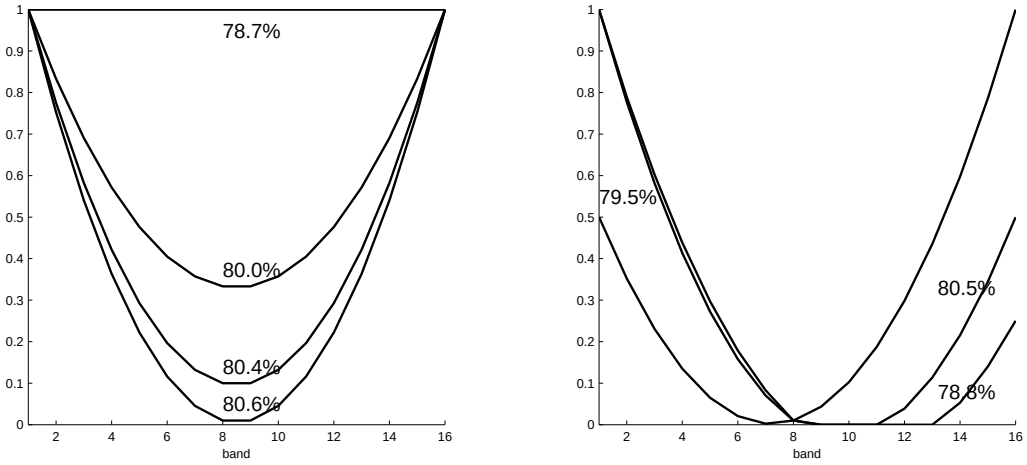


Figure 6.5: Effect of weighting before SACF.

following scheme was used:

1. Pieces were divided into the seven higher-level genres (see Table 3.3).
2. These genres were further divided into three sub-categories: pieces containing only segments where drums are present, pieces containing only segments where drums are absent, and pieces containing both kind of segments.
3. Fifty percent of the pieces in each sub-category were randomly selected to the training set, and the rest to the test set.
4. An individual piece may appear only in the test or in the train set, but not in both.

6.3.2 Results

Periodicity Detection Approach

Despite the preprocessing, also other instruments may cause peaks to the bandwise autocorrelation functions. However, drum instruments tend to be more prominent at the low or high frequencies, and based on this observation frequency bands are weighted differently. An optimal weight vector W_i, has to be determined to be used in the SACF formulation, defined in Eq. 5.6. For this sake, a smaller simulation was carried out. Test set was formed using scheme described above, but only 30 % of pieces were chosen. Results are presented in Figure 6.5. Performance difference between the flat line (78.7 %) and steep parabola (80.6 %) was quite small. However, the best performance is reached with equally weighted lower and higher band and attenuation for centre bands. So a fixed unit weight for both the highest and the lowest band, and 1/100 weight for centre band was used in final simulations.

Table 6.17: Results using periodicity detection.

Musical genre	Performance	Drums absent	Drums present
classical	83 %	84 %	78 %
electronic / dance	91 %	61 %	96 %
hip hop	87 %	70 %	88 %
jazz / blues	75 %	38 %	79 %
rock / pop	83 %	82 %	83 %
soul / RnB / funk	78 %	80 %	78 %
world / folk	69 %	52 %	92 %
Total	**81 %**	77 %	83 %

Fifty percent of the pieces were used to estimate feature value distributions for intervals with and without drums. Division between this distribution estimation set and final test set was done using scheme described above. Obtained feature value distributions were presented earlier in Figure 5.6. The feature value distributions of the two classes overlap each other somewhat, because the stochastic residual contains harmonic components and beginning transients from other instruments, too, and in some cases these show very much drum-like periodicity. Thus the starting hypothesis that periodic stochastic components reveal drum events was still mainly right. More attention should be paid for the preprocessing system in order to make concluding remarks. Based on these feature value distributions, a threshold was chosen to produce an equal error rate for segments with and without drums. Detection results obtained with this threshold value are shown in Table 6.17. Overall performance was 81 % and the performance is rather well-balanced between segments with and without drums.

Acoustic Pattern Recognition Approach

As was discussed earlier in Section 5.3, drums have characteristic spectral energy distributions. The spectral energy of a bass drum is concentrated to lower frequencies. Cymbals and hihats occupy a wide frequency band, mainly concentrated to the treble end. The highest frequencies of the cymbals and hihats are so high that there are only a few other instruments to have prominent frequency components in the same range (e.g. strings). Therefore drums make a significant contribution to the overall tone colour of musical signals. Based on this, an ability of acoustic features to indicate the presence of drums in musical signals was studied. MFCCs alone and catenated with ΔMFCC were evaluated as features. For catenated features a single classifier was trained, unlike in the classification approach used in the musical genre recognition where separate classifiers were trained for the features. In order to avoid numerical problems, the features were normalised to have zero mean and unit variance.

The results obtained with the GMM are shown in Table 6.18. Separate models were

Table 6.18: Results for the GMM with a varying model order.

Number of Gaussians	MFCC with preprocessing	MFCC+ΔMFCC with preprocessing	MFCC+ΔMFCC without preprocessing
4	82 %	86 %	86 %
8	83 %	86 %	87 %
12	84 %	86 %	86 %
16	83 %	86 %	86 %
24	84 %	87 %	87 %

Table 6.19: The best-performing GMM configuration with two global model.

Musical genre	Performance	Drums absent	Drums present
classical	90 %	97 %	39 %
electronic / dance	89 %	49 %	96 %
hip hop	94 %	26 %	98 %
jazz / blues	74 %	58 %	76 %
rock / pop	92 %	68 %	95 %
soul / RnB / funk	91 %	77 %	93 %
world / folk	68 %	48 %	95 %
Total	**87 %**	84 %	88 %

estimated for two classes, one for music with drums and another for music without drums. A three-second long analysis excerpt was used in the evaluations. As one can see, the overall performance was slightly better than with periodicity detection approach. The performance was improved by adding the ΔMFCC in the feature set. There was only marginal performance difference between original signal and preprocessed signal (stochastic residual signal from the sinusoidal modelling) with this approach.

If we take a closer look to the results of the best-performing configuration presented in Table 6.19, we will see that performance was not evenly distributed within different musical genres. Although a high performance was obtained for one class (e.g. drums present), the other failed within the individual musical genre. In other words, the system starts to recognise the musical genre rather than the drums. This is clearly seen for classical music, for example. Due to the small amount of training material for classical music with drums, GMM was unable to model it effectively with one generic model for all genres with drums present.

In order to prevent this, the number of GMM models was increased. Two models were estimated for each musical genre: one for intervals with drums and one for interval without. The musical genre was ignored in the classification, and the classification was done based on the set of models (models calculated for intervals with

Table 6.20: The best-performing GMM configuration with two models for each of the musical genre.

Musical genre	Performance	Drums absent	Drums present
classical	86 %	89 %	61 %
electronic / dance	89 %	63 %	95 %
hip hop	90 %	25 %	94 %
jazz / blues	71 %	67 %	71 %
rock / pop	89 %	77 %	90 %
soul / RnB / funk	92 %	85 %	93 %
world / folk	66 %	46 %	93 %
Total	**84 %**	80 %	86 %

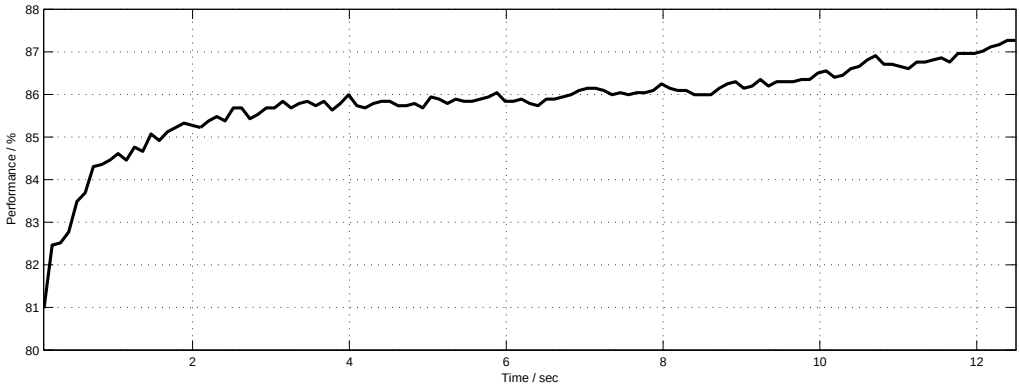


Figure 6.6: Detection accuracy as a function of analysis excerpt duration for the best-performing GMM configuration.

drums and for intervals without drums) that gave the highest likelihood. The results were somewhat better balanced than those obtained with just two joint models for all the genres, as shown in Table 6.20.

In Figure 6.6, the overall performance of the best-performing configuration is shown as a function of the analysis excerpt duration used in the classification. A reasonable performance (81 %) was achieved already with 100 ms analysis excerpts. The performance seems to converge to the level of 85.5% around 3 seconds and after 9 seconds the performance is gradually increasing ending up to the level of 87% at 12 seconds.

In addition, a k-NN classifier was also used in order to evaluate differences between the classifiers. The classification is done by calculating the distance between every test point and all the training data. The features were processed before using them with the classifier in order to reduce amount of calculations needed. The mean and the standard deviation of each feature were calculated within frames of 0.5 seconds, and these were used in place of the original features. This doubled the amount of features, but significantly reduced the amount of feature vectors over time. The

Table 6.21: Detection results for k-NN classifier with k being 5.

Musical genre	Performance	Drums absent	Drums present
classical	83 %	88 %	44 %
electronic / dance	86 %	25 %	96 %
hip hop	95 %	11 %	99 %
jazz / blues	77 %	47 %	80 %
rock / pop	89 %	46 %	93 %
soul / RnB / funk	89 %	46 %	93 %
world / folk	60 %	32 %	95 %
Total	**83 %**	71 %	89 %

performance was slightly improved by increasing the number of "voting" points, k. When only the closest neighbour was considered performance was 80 % and with the five closest neighbours it was 83%. The results obtained with the five closest neighbours are presented in Table 6.21. The performances obtained with k-NN were between the ones obtained with the GMM and the ones obtained with the periodicity detection approach. The performance was more imbalanced than with other approaches.

Combination of the two approaches

The presented two drum detection systems are based on different information, one on periodicity and one on spectral features. One would thus expect that a combination of the two systems would perform more reliably than either of them alone. Fusion of the two systems was realized by combining their output likelihoods. For periodicity detection, the likelihood is obtained from the feature value distributions presented in Figure 5.6. The results are presented in Table 6.22. Only a minor improvement (1-2 %) was achieved. This is due to the fact that both of the systems typically misclassified within the same intervals. For example, jazz pieces where drums were played quite softly with brush, or ride cymbal was continually tapped were likely to be misclassified with both systems. However, the misclassification might be acceptable in some cases, since the drums are difficult to detect even for a human listener.

6.3.3 Discussion

The obtained results of different approaches are rather close to each other and, somewhat surprisingly, the combination performs only slightly better. This highlights a fact which was also validated by listening, both system fail in borderline cases that are difficult, not just due to algorithmic artefacts. Achieved segmentation accuracy of the integrated system was 88 % over a database of varying musical genres. The misclassified intervals are more or less ambiguous by nature, and in many cases a user

Table 6.22: Comparison of the results obtained earlier and by combining the best-performing GMM configuration and periodicity detection.

Detection system	Overall performance	Drums absent	Drums present
Periodicity detection	81 %	77 %	83 %
GMM	87 %	84 %	88 %
Combined detection	**88 %**	**84 %**	**90 %**

might tolerate the misclassifications. However, drum instrument detection was the best-performing instrument detection system studied (see results in Table 6.16). In order to construct a substantially more accurate system, it seems that more complicated sound separation and recognition mechanism would be required. In non-causal applications, longer analysis excerpts and the global context can be used to improve the performance.

7 Conclusions

We have studied the automatic classification of music signals according to their genres and instrumentation. Furthermore, a listening test was conducted to determine the level of human accuracy in recognising musical genres. An important part of the work was to collect and annotate a general-purpose music database to be used in different areas of MIR. This was a time-consuming task but worth the effort, since the amount of the training data is an important factor when evaluating classification systems.

The listening test showed that recognition of musical genres is not a trivial task even for humans. On the average, humans were able to recognise the correct genre in 75 % of cases (given 5000 ms samples). The recognition accuracy was found to depend on the length of the presented sample. Fairly good recognition accuracy was achieved even for the shortest-sample lengths, 250 ms and 500 ms, indicating that humans can do rather accurate musical genre recognition without long-term temporal features such as rhythm.

Based on our studies, automatic musical genre recognition can be done with accuracy significantly above chance, despite the rather fuzzy nature of musical genres. MFCCs and ΔMFCCs were used to represent the time-varying magnitude spectrum of music signals and the genre-conditional densities were modelled with HMMs. The best accuracy (around 60 %) is comparable to the state-of-the-art systems. A slight improvement was obtained by using subgenres along with six primary musical genres in the classification. The results are comparable to the human genre recognition abilities, especially in the case of the shorter sample lengths.

Musical instrument detection was studied for a few pitched instruments using MFCCs and ΔMFCCs as features and modelling instrument-conditional densities with HMMs. Despite the demanding conditions with polyphonic multi-instrument music, fairly acceptable detection accuracies were achieved for some instruments. A novel approach for drum instrument detection was proposed. The presence of drum instruments in music was determined by periodicity detection in the amplitude envelopes of the signal at subbands. Rather good detection accuracy (81 %) was obtained with this approach.

There is still a lot of work to be done before we have a complete and reliable music classification system. Using a genre hierarchy and genre-dependent features in the classification seems a promising approach. Additionally, the use of rhythmic information in genre recognition may provide better performance. For instrument detection, instrument-specific features and possibly sound separation need to be studied to make the detection more accurate.

Bibliography

[Alghoniemy99] M. Alghoniemy and A. Tewfik. *Rhythm and Periodicity Detection in Polyphonic Music.* In IEEE third Workshop on Multimedia Signal Processing, pages 185–190, September 1999.

[Aucouturier01] J.-J. Aucouturier and M. Sandler. *Segmentation of Musical Signals using Hidden Markov Models.* Amsterdam, The Netherlands, May 2001. 110th Convention of the Audio Engineering Society (AES).

[Aucouturier03] J.-J. Aucouturier and F. Pachet. *Representing Musical Genre: A State of the Art.* Journal of New Music Research, vol. 32, no. 1, pages 83–93, March 2003.

[Ben-Yishai] A. Ben-Yishai and D. Burshtein. *A Discriminative Training Algorithm for Hidden Markov Models.* Submitted to IEEE Transactions on Speech and Audio Processing.

[Berenzweig01] A. Berenzweig and D. Ellis. *Locating singing voice segments within music signals.* pages 119–122. IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics, October 2001.

[Bilmes02] J. Bilmes. *What HMMs Can Do.* Technical Report UWEETR-2002-0003, University of Washington, Department of Electrical Engineering, January 2002.

[Burred03] J. J. Burred and A. Lerch. *A Hierarchical approach to automatic musical genre classification.* pages 308–311, London, UK, September 2003. International Conference on Digital Audio Effects (DAFx-03).

[Campbell97] J. P. Campbell. *Speaker recognition: A tutorial.* In Proceedings of the IEEE, vol. 85, pages 1437–1461. IEEE, September 1997.

[Casey02] M. Casey. *General sound classification and similarity in MPEG-7.* Organized Sound, vol. 6, no. 2, pages 153–164, 2002.

[Cheveigné02] A. de Cheveigné and H. Kawahara. *YIN, a fundamental frequency estimator for speech and music.* Journal of the Acoustical Society of America, vol. 111, pages 1917–1930, 2002.

[Clarke89] D. Clarke, editor. The Penguin Encyclopedia of Popular Music. Viking, London, 1989.

[Cover67] T. Cover and P. Hart. *Nearest Neighbor Pattern Classification.* In IEEE Transactions on Information Theory, vol. 13, pages 21–27, January 1967.

[Crowther95] J. Crowther, editor. Oxford Advanced Learner's Dictionary of Current English, 5th edition. Oxford University Press, 1995.

[Dixon01] S. Dixon. *Automatic Extraction of Tempo and Beat from Expressive Performances.* Journal of New Music Research, vol. 30, no. 1, pages 39–58, 2001.

[Egan75] J. Egan. Signal Detection Theory and ROC analysis. Cognitition and Perception. Academic Press, New York, NY., 1975.

[Eronen01] A. Eronen. *Comparison of features for musical instrument recognition.* pages 19–22. IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (WASPAA), August 2001.

[Eronen03a] A. Eronen. *Musical instrument recognition using ICA-based transform of features and discriminatively trained HMMs.* vol. 2, pages 133–136. Seventh International Symposium on Signal Processing and Its Applications, July 2003.

[Eronen03b] A. Eronen and T. Heittola. *Discriminative Training of Unsupervised Acoustic Models for Non-speech Audio.* pages 54–58. Tampere University of Technology, The 2003 Finnish Signal Processing Symposium, Finsig'03, May 2003.

[Fisher70] R. Fisher. Statistical Methods for Research Workers, 14th ed. Oliver and Boyd, Edinburgh, 1970.

[Foote97] J. T. Foote. *Content-based retrieval of music and audio.* In Proceedings of SPIE, vol. 3229, pages 138–147. Multimedia Storage and Archiving Systems II, 1997.

[Foote00] J. Foote. *Automatic audio segmentation using a measure of audio novelty.* vol. 1, pages 452–455. IEEE International Conference on Multimedia and Expo (ICME), 2000.

[Hartmann96] W. Hartmann. *Pitch, Periodicity and Auditory Organization.* Journal of the Acoustical Society of America, vol. 100, no. 6, pages 3491–3502, 1996.

[Heittola02] T. Heittola and A. Klapuri. *Locating Segments with Drums in Music Signals.* pages 271–272, Paris, France, September 2002. International Conference on Music Information Retrieval (ISMIR).

[Herrera00] P. Herrera, X. Amatriain, E. Batlle and X. Serra. *Towards instrument segmentation for music content description: a critical review of instrument classification techniques.* International Symposium on Music Information Retrieval (ISMIR), 2000.

[Houtsma95] A. J. M. Houtsma. Hearing, handbook of perception and cognition, chapter 8: Pitch Perception, pages 267–296. Academic Press Inc., San Diego, CA, USA, 2nd edition edition, 1995.

[Houtsma97] A. J. M. Houtsma. *Pitch and Timbre: Definition, Meaning and Use.* Journal of New Music Research, vol. 26, no. 2, pages 104–115, 1997.

[Jiang02] D.-N. Jiang, L. Lu and H.-J. Zhang. *Music Type Classification by Spectral Contrast Features.* vol. 1, pages 113–116, Lausanne Switzerland, August 2002. IEEE International Conference on Multimedia and Expo (ICME).

[Klapuri03] A. Klapuri. *Musical meter estimation and music transcription.* Paper presented at the Cambridge Music Processing Colloquium, Cambridge University, UK, 2003.

[Levitin99] D. Levitin. Music, Cognition and Computerized Sound : An Introduction to Psychoacoustics., chapter 23: Experimental design in psychoacoustic research, pages 299–328. M.I.T. Press, Cambridge, MA, 1999.

[Li00] G. Li and A. A. Khokhar. *Content-based indexing and retrieval of audio data using wavelets.* vol. 2, pages 885–888. IEEE International Conference on Multimedia and Expo (ICME), July 2000.

[Li01] D. Li, I. Sethi, N. Dimitrova and T. McGee. *Classification of general audio data for content-based retrieval.* Pattern Recognition Letters, vol. 22, no. 5, pages 533–544, April 2001.

[Li03] T. Li, M. Ogihara and Q. Li. *A comparative study on content-based music genre classification.* In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, pages 282–289. ACM, ACM Press, July 2003.

[Logan00] B. Logan. *Mel Frequency Cepstral Coefficients for music modeling.* International Symposium on Music Information Retrieval, 2000.

[Logan01] B. Logan and A. Salomon. *A music similarity function based on signal analysis.* pages 745–748. IEEE International Conference on Multimedia and Expo (ICME), August 2001.

[Martin99] K. Martin. *Sound-source recognition: A theory and computational model.* PhD thesis, Massachusetts Institute of Technology, 1999.

[Meddis91] R. Meddis and M. J. Hewitt. *Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification.* Journal of the Acoustical Society of America, vol. 89, no. 6, pages 2866–2882, June 1991.

[MPE01] *Information Technology - Multimedia Content Description Interface - Part 4: Audio.* ISO/IEC FDIS 15938-4, 2001.

[Pachet00] F. Pachet and D. Cazaly. *A Taxonomy of Musical Genres.* Paris, France, April 2000. Content-Based Multimedia Information Access Conference (RIAO).

[Paulus02] J. Paulus and A. Klapuri. *Measuring the similarity of rhythmic patterns.* pages 150–156, Paris, France, September 2002. International Conference on Music Information Retrieval (ISMIR).

[Peltonen01] V. Peltonen. Computational Auditory Scene Recognition. Master's thesis, Department of Information Technoly, Tampere University Of Technology, 2001.

[Perrot99] D. Perrot and R. Gjerdigen. *Scanning the dial: An exploration of factors in the identification of musical style.* page 88(abstract). Society for Music Perception and Cognition, 1999.

[Pye00] D. Pye. *Content-based methods for the management of digital music.* vol. 4, pages 2437–2440. IEEE International Conference on, Acoustics, Speech, and Signal Processing, 2000.

[Rabiner93] L. Rabiner and B.-H. Juang. Fundamentals of Speech Recognition. PTR Prentice-Hall Inc., New Jersey, 1993.

[Raphael99] C. Raphael. *Automatic Segmentation of Acoustic Musical Signals Using Hidden Markov Models.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 21, no. 4, pages 360–370, 1999.

[Reynolds95] D. Reynolds and R. Rose. *Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models.* IEEE Transactions on Speech and Audio Processing, vol. 3, no. 1, pages 72–83, 1995.

[Rossignol98] S. Rossignol, X. Rodet, J. Soumagne, J.-L. Collette and P. Depalle. *Feature Extraction and Temporal Segmentation of Acoustic Signals.* pages 199–202, Ann Arbor, USA, 1998. International Computer Music Conference (ICMC).

[Sadie01] S. Sadie, editor. The New Grove Dictionary of Music and Musicians. MacMillan Publishing Company, 2001.

[Saunders96] J. Saunders. *Real-time discrimination of broadcast speech/music.* vol. 2, pages 993–996, Atlanta, GA, May 1996. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

[Scheirer97] E. Scheirer and M. Slaney. *Construction and evaluation of a robust multifeature speech/music discriminator*. vol. 2, pages 1331–1334. IEEE International Acoustics, Speech, and Signal Processing (ICASSP), April 1997.

[Scheirer98] E. Scheirer. *Tempo and beat analysis of acoustic musical signals*. Journal of the Acoustical Society of America, vol. 103, no. 1, pages 558–601, 1998.

[Seppänen01] J. Seppänen. *Tatum grid analysis of musical signals*. pages 131–134, New Paltz, NY, USA, October 2001. IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (WASPAA).

[Serra97] X. Serra. Musical Signal Processing, chapter Chapter 3: Musical Sound Modeling with Sinusoids plus Noise, pages 91–122. Swets & Zeitlinger Publishers, Lisse, the Netherlands, 1997.

[Soltau97] H. Soltau. Erkennung von Musikstilen. Master's thesis, Institute of Logic, Complexity and Deductive Systems, University of Karlsruhe, 1997.

[Soltau98] H. Soltau, T. Schultz, M. Westphal and A. Waibel. *Recognition of Music Types*. Seattle, WA, 1998. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[Tolonen00] T. Tolonen and M. Karjalainen. *A Computationally efficient multipitch analysis model*. IEEE Transactions on Speech and Audio Processing, vol. 8, no. 6, pages 708–716, November 2000.

[Tzanetakis01] G. Tzanetakis, G. Essl and P. Cook. *Automatic Musical Genre Classification Of Audio Signals*. Bloomington, Indiana, 2001. International symposium on musical information retrieval (ISMIR).

[Tzanetakis02a] G. Tzanetakis and P. Cook. *Musical genre classification of audio signals*. IEEE Transactions on Speech and Audio Processing, vol. 10, no. 5, pages 293–302, July 2002.

[Tzanetakis02b] G. Tzanetakis, A. Ermolinskyi and P. Cook. *Pitch Histograms in Audio and Symbolic Music Information Retrieval*. pages 31–39, Paris, France, September 2002. International Conference on Music Information Retrieval (ISMIR).

[Viterbi67] A. Viterbi. *Error bounds for convolutional codes and an asymptotically optimum decoding algorithm*. IEEE Transactions on Information Theory, vol. 13, pages 260–269, April 1967.

[Wakefield99] G. Wakefield. *Mathematical Representation of Joint Time-Chroma Distributions*. Denver, Colorado, 1999. SPIE International Symposium on Optical Science, Engineering and Instrumentation.

[Welsh99] M. Welsh, N. Borisov, J. Hill, R. von Behren and A. Woo. *Querying large collections of music for similarity.* Technical Report UCB/CSD00 -1096, U.C. Berkeley Computer Science Division, Berkeley, CA, November 1999.

[Xiong03] Z. Xiong, R. Radhakrishnan, A. Divakaran and T. Huang. *Comparing MFCC and MPEG-7 audio features for feature extraction, maximum likelihood HMM and entropic prior HMM for sports audio classification.* vol. 5, pages 628–631. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2003.

[Xu03] C. Xu, N. C. Maddage, X. Shao, F. Cao and Q. Tian. *Musical Genre Classification Using Support Vector Machines.* vol. 5, pages 429–432. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2003.

[Zhang01] T. Zhang and C.-C. Kuo. *Audio content analysis for online audiovisual data segmentation and classification.* IEEE Transactions on Speech and Audio Processing, vol. 9, no. 4, pages 441–457, November 2001.

A Musical genre hierarchy

The musical genre hierarchy (taxonomy) used in this thesis is listed here.

Classical

Chamber Music
- Chamber Brass
- Chamber Orcestra
- Chamber Percussion
- Chamber Strings
- Chamber Woodwinds

Classical General
- 20th Century & Contemporary
  - Classical Crossover
  - Electronic Classical
  - Experimental Classical
  - Expressionism
  - Impressionism
  - Minimalist
  - Nationalism
  - Serialism
  - Third Stream
- Baroque
- Classical Era
- Medieval
- Renaissance
- Romantic
  - 19th Century Romantic
  - 20th Century Romantic

Crossover

Film Music

General Instrumental
- Marches
  - Polka
  - Waltzes

Solo Instruments
- Solo Brass
  - Solo Guitar
  - Solo Percussion
  - Solo Piano & Keyboards
  - Solo Strings
  - Solo Woodwinds
  - Solo Instruments w/ Accompaniment
    - Brass w/ accompaniment
    - Guitar w/ accompaniment
    - Percussion w/ accompaniment
    - Piano & Keyboards w/ accompaniment
    - Strings w/ accompaniment
    - Woodwinds w/ accompaniment

Symphonic

Vocal
- Choral
  - Opera
  - Small Vocal Ensembles
  - Solo Vocal

Electronic / Dance

Ambient
- Abstract
- Ambient Breakbeat
- Ambient Dub
- Ambient House
- Ambient Ragga
- Ambient Techno
- Dark Ambient
- Darkside Ambient

Dance
- Alternative Dance
- Club
- Dancehall Dance
- Disco Dance
- Euro Dance

- Freestyle Dance
- Gabber
- High-NRG

Breakbeat/breaks/Drum n' Bass
- Ambient Drum N' Bass
- Big Beat
- Down Tempo
  - Illbient
  - Trip Hop
- Funky Breaks
- Future Funk
- Jump-Up
- Jungle Drum-N-Bass
- Tech Step

Electronica
- Progressive Electronica
- Symphonic Electronica

House
- Acid House
- Deep House
- Funk House
- Garage House
- Hard House
- House Acid
- Latin House
- Madchester
- Newbeat
- Progressive House
- Speed Garage
- Vocal House

Industrial
- Dark Techno/Darkwave
- EBM
- Electro
- Industrial Dance
- Industrial Rock

Techno/Trance
- Acid Techno
- Detroit Techno
- Gabber Techno
- Garage Techno
- Electro
- Experimental
  - Minimalist Experimental
  - Noise
- Happy Hardcore
- Hardcore Techno
- Minimal Techno
- Intelligent Techno
- Neo-Electro
- Old Skool Techno
- Rave
- Techno Dub
- Trance
  - Goa
  - Hard Trance/Acid
  - Melodic Trance
  - Progressive/Dream
  - Psytrance
- Tech Trance
- Traxx
- Tribal

Hip Hop / Rap

- Alternative
- Bass Assault
- Bass Music
- British Rap
- Christian Rap
- Comedy Rap
- Crossover Rap
- Dirty Rap
- East Coast
- Electric Funk
- Foreign Rap
- Freestyle Rap
- G-Funk
- Gangsta Rap
- Hardcore Rap
- Horror Core
- Jazz-Rap
- Latin Rap
- Miami Bass
- New Jack R&B Rap
- New School
- Old School
- Party Rap
- Pop-Rap
- Rap Core
- Rap Metal
- Rap-Rock
- South West
- Southern Rap
- Suburban Rap
- Turntablist-Solo
- Turntablist-Group
- Underground Rap
- West Coast

Jazz / Blues

Blues
- Acoustic Blues
- Blues Country
- Blues Rock
- Blues Vocalist
- Chicago Blues
- Classic Female Blues
- Delta Blues
- Early American Blues
- Electric Blues
- General Blues
- Jump Blues
- Modern Blues
- Spiritual Blues
  - Christian Blues

Improvised

Jazz
- Acid Jazz
- Afro-Cuban Jazz
- Ballad Jazz
- Bebop
- Big Band Jazz
  - Modern Big Band Jazz
- Boogie-Woogie
- Bossa Nova
- Contemporary Acoustic Jazz
- Cool Jazz
- Crossover Jazz
- Dixieland
- Electronic Jazz
- Experimental Jazz
- Free Jazz
- General Jazz
- Hard Bop
- Jazz Blues
- Jazz Folk
- Jazz Fusion
  - Contemporary Jazz Fusion
  - Experimental Jazz Fusion
  - World Jazz Fusion
- Jazz Organ
  - Jazz Organ Blues
- Jazz Piano
  - Jazz Piano Blues
  - Stride Piano Jazz
- Jazz Vocals
- Latin Jazz
- New Orleans Style Jazz
- R&B Jazz
- Ragtime
- Smooth Jazz
- Swing
- Traditional Jazz
- West Coast Jazz
- World Jazz
  - Brazilian Jazz
  - Indian Jazz

Rock / Pop

Alternative
- Alternative General
- Ambient Alternative
- Experimental Alternative

Country
- Alt Country
- Adult Country Themes
- Alternative Country
- Cowpunk
- Bluegrass
  - Contemporary Bluegrass
  - New Grass
  - Progressive Bluegrass
  - Traditional Bluegrass
- Country Blues
  - Skiffle
- Country General
- Country Rock
- Country Soul
- Honky-Tonk
- New Country
- Progressive Country
- Rockabilly Country
- Spiritual Country
  - Christian Country
- Square Dance
- Tejano Country
  - Tex/Mex
- Western Swing

Easy Listening
- Adult Contemporary
  - Chamber Pop
- Easy Listening General
- Lounge
- Crooners/Vocal Stylists
- Love Songs
  - Ballads
- Mood Music
- Musicals/Broadway
- New Age
  - Adult Alternative Country
  - Contemporary Instrumental
  - Meditation
  - New Age Electronic
  - New Classical

- Relaxation
- Self-Help
- Sounds of Nature
- Space
- Spiritual
- Singer-Songwriter
- Spiritual Easy Listening
  - Children's Christian
  - Christian Easy Listening
  - Contemporary Christian

Leftfield

Metal
- Alternative Metal
- Black Metal
- British Metal
- Dark Ambient/Noise
- Death Metal
- Doom/Stoner Metal
- Gothic Metal
- Grindcore
- Hair Metal
- Hard Core Metal
- Heavy Metal
- Industrial Metal
- Instrumental Metal
- Metalcore
- Metal Rap
- Power Metal
- Progressive Metal
- Speed Metal
- Spiritual Metal
  - Christian Metal
- Thrash

New Wave

Pop
- Acoustic Pop
- Adult Conremporary
- Alternative Brit Pop
  - Mod
  - Shoegazer
- Beach Pop
- British Pop
- Dream-Pop
- EuroPop
- French Pop
- Grunge-Pop
- Holiday Pop
- J-Pop
- Jangle Pop
- Latin Pop
- Noise Pop
- Polka Pop
- Pop Vocals
- Power Alternative Pop
- Power Pop
- Reggae
- Spiritual Pop
  - Christian Pop
- Synth Pop
- Teen Idol

Punk
- Emo
- Folk Punk
- Garage
- Hardcore Punk
  - Post Hardcore
  - Oi
- Lo-Fi/Garage
- Old School Punk
- Pop Punk
- Post Punk
- Proto-Punk
- Psycho-Billy
- Riot Grrr
- Ska Punk
- Skate Punk
- Surf Punk
- Straight Edge Punk
- Twee Cuddle Core

Rock
- AAA/Adult Alternative
- Acid Rock
- Acoustic
- Adult Alternative Rock
- Adventure Rock
- Americana
- Arena Rock
- Art/Progressive Rock
- Alternative Rock
- Alternative Space Rock
- Avant-Rock
- Beach Rock
- Boogie Rock
- British Invasion
- British Rock
- British Traditional Rock
- Classic Rock
- Classical Rock
- Comedy Rock
- Electronica Rock
- Experimental Rock
  - No Wave
- Folk Rock
- Funk Rock
- Garage Rock
- Glam Rock
- Goth Rock
  - Alternative Gothic Rock
  - Christian Gothic
  - Industrial Gothic
- Groove Rock
- Guitar Rock
  - Guitar Virtuoso
- Grunge
  - Neo Grunge
  - Post Grunge
- Hard Rock
- Improv Rock

- Indie
  - Indie Pop/Lo Fi
- Instrumental Rock
- Latin Rock
- Math Rock
- Neo-Psychedelia
- New Wave
  - New Romantic Rock
- Noise Rock
- Progressive Rock
- Psychedelic
- Rock & Roll
- Rock En Espanol
- Rockabilly
- Soft Rock
  - Acoustic Soft Rock
- Southern Rock
- Space Rock
- Spiritual Rock
  - Christian Rock
- Surf Rock
- The Adult Arena

Rock-n-Roll Oldies

Seasonal/Holiday
- Christmas
  - Contemporary
  - International
  - Traditional
- Hanukkah
- Other Holidays

Soul / RnB / Funk

Funk
- Acid Funk
- P-Funk

Gospel
- Alternative CCM
- CCM Contemporary
- Gospel Hymns
- Inspirational
- Instrumental Gospel
- Jewish
- Reggae Gospel
- Spirituals
- Traditional Gospel

R&B
- Contemporary R&B
- Doo-Wop
- Motown
- New Jack R&B

Rhythm and Blues

Soul
- Black-Eyed Soul
- Deep Soul
- Philly Soul
- Pop Soul
- Quiet Storm
- Retro Soul
- Soul Country
- Southern Soul
- Swamp Soul

World/Folk

- African
- Asia
- Caribbean
- Celtic
- ceremonial/chants
- European
- folk
- Latin American
- Mediterranean
- Middle East
- North American
- Northern Europe
- Oceania
- old dance music
- Scandinavian
- South Pacific
- World Pacific
- World Beat
- World Fusion
- World Traditions

B Pieces in the music database

The following lists the names of the artists and pieces in the music database in detail. Pieces are divided according to the annotated first-level genres. The pieces used in the listening experiment are also indicated.
List notation: <artist> - <*title*> [second-level genre] (G1 or G2 according to the test group the piece was assigned in the listening experiment)

Classical

- *Deck the halls*, (G2)
- *Etude, op.25 No.9 in G flat major*, [Solo Instruments], (G2)
Academy chamber ensemble - *Sonata in A-dur op.5/1: allegro*, [chamber music]
Academy chamber ensemble - *Sonata in A-dur op.5/1: andante-adagio*, [chamber music]
Academy chamber ensemble - *Sonata in A-dur op.5/1: gavotte (allegro)*, [chamber music], (G1)
Academy chamber ensemble - *Sonata in A-dur op.5/1: larghetto - allegro*, [chamber music]
Academy chamber ensemble - *Sonata in e minor: allegro*, [chamber music]
Academy chamber ensemble - *Sonata in e minor: allemande (andante allegro)*, [chamber music]
Academy chamber ensemble - *Sonata in e minor: andante larghetto - adagio*, [chamber music]
Academy chamber ensemble - *Sonata in e minor: gavotte (allegro)*, [chamber music]
Academy chamber ensemble - *Sonata in e minor: rondeau*, [chamber music]
Academy chamber ensemble - *Sonata in e minor: sarabande (largo assai)*, [chamber music], (G2)
Agustin Anievas - *Etude in C minor ("Revolutionary")*, [solo instruments]
Armenia Philharmonic orchestra (conducted by Loris Tseknavorian - *The Sabre dance*
Avanti-kvartetti - *Jousikvartetto*, [classical general]
Camarata Labacensis - *Eine kleine nachtmusik: menuetto*, [chamber music]
Choralschola der Wiener Hofburgkapelle - *Kyrie Eleison*, [vocal], (G2)
City of London Sinfonia - *Suite in F major: Menuet*, [classical general]
City of London Sinfonia, conductor Richard Hickox - *Suite in F major: Air*, [classical general]
Consortium classicum - *Introduktion und elegie für klarinette, zwei violinen, viola und violoncello: rondo: allegro scherzando*
Covent Garden royal opera choir and orchestra - *Heprealaisten orjien kuoro*, [vocal], (G2)
Dallas brass - *Carol of the bells*, (G2)
Daniel Barenboim - *Lieder ohne worte op.19, no.2 a-moll: andante espressivo*, [solo instruments]
Daniel Barenboim - *Lieder ohne worte op.19, no.5 fis-moll: piano agitato*, [solo instruments]
Daniel Barenboim - *Lieder ohne worte op.30, no.6 fis-moll: allegretto tranquillo*, [solo instruments]
Daniel Barenboim - *Lieder ohne worte op.67, no.2 fis-moll: allegro leggiero*, [solo instruments]
Das salonorchester Cölln - *Albumblatt*, [chamber music]
Das salonorchester Cölln - *Notturno no.3, Liebestraum*, [chamber music]
Das salonorchester Cölln - *Ungarischer tanz no.5*, [chamber music]
Dubravka Tomsic - *Italian concert F major: Andante*
Dubravka Tomsic - *Sonate no.14 C sharp minor op27 no 2: adagio sostenuto (Moonlight sonata)*, [solo instruments]
Erkki Rautio (cello), Izumi Tateno (piano) - *Berceuse*, [solo instruments], (G1)
Éva Maros - *Pavana con su glosa*, [classical general]
Gudrun Derler, Barbara Hölzl - *Der Fischer*, [vocal], (G1)
György Geiger (trumpet), Éva Maros (harp) - *Le Coucou*
Hamburg chamber orchestra - *The four seasons concerto op.8 no.1: Spring, allegro*
Hamburg chamber orchestra - *The four seasons concerto op.8 no.2: summer, presto*
Hamburg chamber orchestra - *The four seasons concerto op.8 no.3: autumn allegro*
Hamburg chamber orchestra - *The four seasons concerto op.8 no.4: winter, allegro non molto*
Hamburg radio symphony - *Ouverture Fidelio, op.72*
Hans Fagius - *Toccata in D minor*, [classical general]
Hungarian state opera chamber orchestra, solo trumpet
Ede Inhoff - *Sonata no.10 for Trumpet and strings: allegro/presto*, [classical general]
I Salonisti - *Kuolema op.44: Valse Triste*, [chamber music], (G1)
I Salonisti - *Préludes: La plus que lente*, [chamber music]
I Salonisti - *Serenata*, [chamber music]
Ida Czernecka - *Mazurka no.47 in a minor op.68 no.2*, [solo instruments]
Ida Czernecka - *Nocturne no.1 in Bb minor op.9 nr.1*, [solo instruments], (G2)
Ida Czernecka - *Prelude no.3 in Db major op.28, Raindrops*, [solo instruments]
Ida Czernecka - *Waltz no.12 in f minor op.70 nr.2*, [solo instruments], (G1)

John Ogdon, Brenda Lucas - *En bateau*, [classical general]
Kamariorkesteri Vox Artis, conductor Lev Markiz - *Serenade for strings in C major, op.48: II Walzer (moderato tempo di valse)*, [chamber music]
Lars Henning Lie, Barbara Hölzl - *Zur Johannisnacht*, [vocal], (G1)
London concert orchestra, conducted by Sir Anthony Arthur - *Swanlake: Hungarian dance - Czardas*
London concert orchestra, conducted by Sir Anthony Arthur - *Swanlake: Scene*
London concert orchestra, conducted by Sir Anthony Arthur - *Swanlake: Spanish dance*
London festival orchestra - *Bolero*
London philharmonic orchestra - *Die Zauberflöte: ouverture*, [symphonic], (G1)
London philharmonic orchestra - *Symphony no.41 in C major, Jupiter: Allegro*, [symphonic]
London philharmonic orchestra - *The marriage of Figaro: ouverture*, [symphonic]
London symphony orchestra - *Faust (ballet): adagio*
Marian Lapsansky, Peter Toperczer, Czechoslovak Radio Symphony Orchestra - *Carnival of the animals*, [symphonic]
Marian Lapsansky, Peter Toperczer, Czechoslovak Radio Symphony Orchestra - *Peter and the Wolf*, [symphonic], (G2)
Marián Lapsansky (solo), Slovak Philharmonic Orchestra - *Piano Concerto in A minor: Allegro vivace*
Mozart Festival Orchestra, conductor Alberto Lizzio, horn solo Josef Dokupil - *Horn concerto nr.1 D major: Allegro*, (G1)
Mozart Festival Orchestra, conductor Alberto Lizzio, horn solo Josef Dokupil - *Horn concerto nr.2 Es Major Andante*
Munich chamber ensemble - *Brandenburg concerto no.2 F major : Allegro*, [chamber music]
Munich chamber orchestra - *Brandenburg concerto no.5 D major: Affettuoso*, [chamber music]
New York philharmonic orchestra conducted by Bruno Walter - *Hungarian dance number 1 in G minor*
New York trumpet ensemble - *Rondeau fron Symphonies de fanfares*, [classical general]
New philharmonia orchestra London - *Symphony no.6, Pathetique, in Bb minor op.74: Finale. Adagio lamentoso*, [symphonic]
Philharmonia quartett Berlin, soloist Dieter Klöcker - *Quintett Es-dur: allegro moderato*
Philharmonic ensemble pro musica - *Peer Gynt suite no.1 op.46: Anitra's dance*
Philharmonic ensemble pro musica - *Peer Gynt suite no.1 op.46: Death of ase*
Philharmonic ensemble pro musica - *Peer Gynt suite no.2 op.55: Solveij's song*
Philharmonic ensemble pro musica - *Peer gynt suite no.1 op.46: In the hall of the mountain king*
Philharmonica Hungarica, conductor Richard P. Kapp - *La Damnation de Faust: Hungarian dance*
Piffaro - *Ave regina caelorum*, [classical general], (G1)
Piffaro - *Entre du fol*, [classical general]
Piffaro - *Gaillarde*, [classical general]
Piffaro - *J'ay pris amours*, [classical general]
Piffaro - *Passe et medio & reprise*, [classical general]
Piffaro - *Pavane&Gaillarde "la Dona"*, [classical general]
Pro musica antiqua - *Fireworks music, Concerto grosso no.26 D major: La paix*, [chamber music]
Radio symphony orchestra Ljubljana - *Symphony no.5in C major: allegro con brio*
Radio symphony orchestra Ljubljana - *Symphony no.8 Bb minor, The unfinished symphony: allegro moderato*, [symphonic], (G2)
Royal Danish symphony orchestra - *Hungarian march*
Royal Danish symphony orchestra - *Tales of Hoffman: barcarole*
Rudolf Heinemann - *Sonate 1 f-moll: allegro moderato e serioso*
Rudolf Heinemann - *Sonate 2 c-moll: adagio*
Rudolf Heinemann - *Sonate 4 B-dur: allegretto*
Rudolf Heinemann - *Sonate 5 D-dur: allegro maestoso*
Soile Viitakoski (vocals), Marita Viitasalo (piano) - *Solveig's song*, [vocal]
Symphonic orchestra Berlin, conductor Kurt Wöss - *Love to the 3 oranges: march*
Süddeutsche philharmonic - *A midsummer night's dream. Dance of the clowns,* [symphonic]
Süddeutsche philharmonic - *A midsummer night's dream. Notturno. Con moto tranquillo*, [symphonic], (G1)
Süddeutsche philharmonic - *A midsummer night's dream. Wedding march*, [symphonic]
Südwestdeutsches kammerorchester - *Sarabande op.93*, [chamber music], (G2)
Südwestdeutsches kammerorchester - *Serenade nr.2 F-dur für streichorchester:Allegro moderato*, [chamber music]
Südwestdeutsches kammerorchester - *Zwei elegische melodien nach gedichten von A.O.Vinje für Streichorchester:Letzter Frühling*, [chamber music]
The Candomino Choir - *Soi Kiitokseksi Luojan* , [vocal]
The New York trumpet ensemble - *Canzon no.1, 1615*, [classical general]
The New York trumpet ensemble - *Sonata à 7*, [classical general]
The New York trumpet ensemble - *Toccata*, [classical general]
The Philharmonia orchestra - *Concerto for trumpet and orchestra II-nocturne andantino* , [classical general]
The Philharmonia orchestra - *Concerto no.2 for trumpet: II-grave*, [classical general]
The River Brass Band - *Muistojen bulevardi*

Electronic/dance

666 - *Bomba*, [dance]
Aphex Twin - *Ageispolis*, [ambient]
Aphex Twin - *Heliosphan*, [ambient]
Aphex Twin - *Pulsewidth*, [ambient], (G2)
Armand van Helden - *Alienz*, [house]
Armand van Helden - *Mother earth*, [house]
Armand van Helden - *The boogie monster*, [house]
Art of Noise - *Something always happens*, [breakbeat/breaks/drum'n'bass], (G1)
Artful dodger feat.Craig David - *Re-rewind*, (G2)
Artful dodger feat.Lynn Eden - *Outrageous*
Artful dodger feat.MC Alistair - *R u ready*
Blümchen - *Schmetterlinge*, [techno/trance]
Blümchen - *Übermorgenland*, [techno/trance]
Chain reaction - *Dance Freak*, [dance], (G1)
Chemical Brothers - *Let forever be*, [breakbeat/breaks/drum'n'bass], (G1)
Daft punk - *Harder, better, faster, stronger*, [house], (G2)
Daft punk - *Voyager*, [house], (G1)
Deisix - *Scream bloody core*, [industrial]
Delerium - *Silence (DJ Tiesto mix)*, [techno/trance]
Dune - *Can't stop raving*, [techno/trance]

Energy 52 - *Café del Mar*, [techno/trance]
Fatboy Slim - *Star 69*, [breakbeat/breaks/drum'n'bass]
Gloria Gaynor - *I will survive*, [dance]
Goldie - *Angel*, [breakbeat/breaks/drum'n'bass]
Hardy Hard - *Everybody shake your body (electro mix)*, [techno/trance]
Hyper - *Noise alert*, [house]
Hypnotist - *Live in Berlin*, [techno/trance], (G1)
Jeff Mills - *The Bells*
KC and the sunshine band - *That's the way (I like it)*, [dance], (G1)
Les Rythmes Digitales - *Jacques your body (make you sweat)*, [techno/trance], (G1)
Les Rythmes Digitales - *Music makes you lose control*, [house], (G2)
Marusha - *Somewhere over the rainbow*, [techno/trance]
Members of Mayday - *The day X*, [techno/trance]
Moodymann - *Long hot sexy nights*, [house], (G1)
Mr. Velcro fastener - *Phlegmatic*, [techno/trance]
Mr. Velcro fastener - *Real robots don't die*, [techno/trance]
Nebulla II - *Peacemakers*, [techno/trance], (G2)
Neuroactive - *Inside your world*, [industrial]
Neuroactive - *Space divider*, [industrial], (G2)
New Order - *Everything's gone green (Advent remix)*, [industrial]
Orbital - *Forever*, [ambient], (G1)
Orbital - *Science friction*, [ambient]
Pansoul - *Ezio*, [house]
Paradise 3001 - *Long distance call to heaven*, [house]
Photek - *Minotaur*, [breakbeat/breaks/drum'n'bass]
Photek - *Smoke rings*, [breakbeat/breaks/drum'n'bass]
Plastikman - *Konception*, [techno/trance]
Plastikman - *Marbles*, [techno/trance]
Prodigy - *Charly*, [techno/trance]
Queen Yahna - *Ain't it time*, [dance], (G2)
RMB - *Spring*, [techno/trance]
Richard D. James - *Fingerbib*
Sash! feat.Rodriguez - *Ecuador*, [dance]
Scooter - *Hands up!*, [techno/trance]
Scooter - *How much is the fish*, [techno/trance]
Sirius B feat. Afrika Bambaataa and Hardy Hard - *If you Techolectro*, [techno/trance]
Skylab - *The trip (Roni Size mix)*, [breakbeat/breaks/drum'n'bass]
Stardust - *Music sounds better with you*
Sunbeam - *Outside world*, [techno/trance]
Sunship - *The Original Sun*, [breakbeat/breaks/drum'n'bass], (G2)
Sunship - *The Unseen*, [breakbeat/breaks/drum'n'bass], (G2)
TDR - *Squelch*, [house]
Terminal choice - *Totes Fleisch*, [industrial]
The Jacksons - *Can you feel it*, [dance]
The Weather Girls - *It's raining men*, [dance]
Tricky - *Contradictive*, [breakbeat/breaks/drum'n'bass]
Tricky - *Hot like a sauna*, [breakbeat/breaks/drum'n'bass]
Tufàan - *Probe* (the Green Nuns of Revolution Mix), [techno/trance]
Ultra Naté - *Free*, [dance]
William Orbit - *Cavalleria rusticana*, [ambient], (G1)
William Orbit - *L'Inverno*, [ambient], (G2)

Hip Hop

Beastie boys - *No sleep till Brooklyn*
Beastie boys - *Rhymin & stealin*
Busta Rhymes - *One*, (G1)
Busta Rhymes - *Turn it up (Remix) Fire it up*, (G2)
Busta Rhymes - *When disaster strikes*, (G2)
Cameo - *She's strange (12" rap version)*, [electric funk], (G2)
Ceebrolistics - *Jalat maassa*, (G1)
Ceebrolistics - *aie/i try*, (G1)
Coolio - *2 minutes & 21 seconds of funk*
Coolio - *Hit 'em*
Coolio - *The devil is dope*
Cypress Hill - *How I could just kill a man*, [gangsta rap], (G2)
Dead Prez - *Be healthy*, (G2)
Dead Prez - *Mind sex*, (G2)
Dr.Dre feat. Hittman, Kurupt, Nate Dogg, Six Two - *Xxplosive*
Dr.Dre feat. Snoop Dogg - *The next episode*
Eminem - *Stan*
Eminem - *The way I am*
Grunge is dead (Butch Vig) / M.O.P. - *How bout some hardcore*, [rap-rock], (G2)
Jay-Z - *Hard knock life*
Jay-Z feat.DMX - *Money, cash, hoes*
Jay-Z feat.Foxy Brown - *Paper chase*, (G1)
Jermaine Dupri and Mariah Carey - *Sweetheart*
Jermaine Dupri feat. Jay-Z - *Money ain't a thing*, (G2)
Missy "Misdemeanor" Elliott - *She's a bitch*
Petter - *En resa*, (G1)
Petter - *Minnen*, (G2)
Petter feat.Kaah - *Ut och in på mig själv*
Public enemy - *Crayola*, (G1)
Public enemy - *Last mass of the caballeros*, (G1)
Run DMC - *What's it all about*
Run DMC - *Word is born*, (G1)
Static-X / Dead Prez - *Hip hop*
System of a down/Wu-Tang clan - *Shame*, [gangsta rap], (G1)
The Roots - *The next movement*
The Roots feat. Erykah Badu - *You got me*, (G2)
Wyclef Jean - *Gone till November*, (G1)

Jazz/Blues

Abraham Laboriel - *Dear friends*, [jazz], (G1)
Abraham Laboriel - *Look at me*, [jazz]
Abraham Laboriel - *My joy is you,* [jazz]
Ahmad Jamal - *Autumn in New York,* [jazz]
Ahmad Jamal - *The girl next door*, [jazz]
Al DiMeola - *Dark eye tango*, [jazz], (G2)
Al DiMeola - *Mediterranean sundance,* [jazz]
Alex Welsh - *Maple leaf rag*, [jazz]
Antero Jakoila - *Pieni tulitikkutyttö*, [jazz]
B.B King - *How blue can you get*, [blues]
B.B King - *The thrill is gone*, [blues]
B.B.King - *Hummingbird*, [blues]
Billy Boy Arnold - *Don't stay out all night*, [blues]
Billy Boy Arnold - *How long can this go on?*, [blues]
Billy Boy Arnold - *Lowdown thing or two*, [blues]
Bob Wilber and Antti Sarpila - *Lester's bounce*, [jazz]
Bob Wilber and Antti Sarpila - *Moon song*, [jazz]
Bob Wilber and Antti Sarpila - *Rent party blues*, [jazz]
Brian Green's dixie kings - *Tiger rag*, [jazz]
Chick Corea elektric band - *Child's play,* [jazz]
Chick Corea elektric band - *Inside out*, [jazz]
Claudio Roditi - *Rua Dona Margarida*, [jazz]

Dan Stewart - *New Orleans blues*, [blues]
Dave Grusin - *Old bones*, [jazz]
Dave Grusin - *Punta del soul*, [jazz]
Dizzy Gillespie and Arturo Sandoval - *Rimsky*, [jazz]
Dizzy Gillespie and Arturo Sandoval - *Wheatleigh Hall*, [jazz]
Don Byron - *Charley's Prelude*, [jazz]
Don Byron - *Frasquita serenade*, [jazz]
Gary Moore - *I loved another woman*, [blues]
Gary Moore - *The supernatural*, [blues], (G2)
George Gee and the jump, jive and wailers - *720 in the books!*, [jazz], (G2)
George Gee and the jump, jive and wailers - *Buzzin' baby*, [jazz]
Glenn Miller - *In the mood*, [jazz]
Glenn Miller - *Over the rainbow*, [jazz]
Headhunters - *Frankie and Kevin*, [jazz], (G2)
Headhunters - *Skank it*, [jazz]
Herbie Hancock - *You've got it bad girl*, [jazz]
Hilton Ruiz - *Something grand*, [jazz]
Howlin' Wolf - *Back door man*, [blues]
Humphrey Lyttleton - *Black& blue*, [jazz]
James Cotton - *Straighten up baby*, [blues], (G1)
Jimmy Johnson - *Black Night*, [blues]
Jimmy Johnson - *My baby by my side*, [blues]
John Lee Hooker - *Ground hog blues*, [blues]
John Lee Hooker - *I love you baby*, [blues]
Johnny Adams - *Neither one of us (wants to be the first to say goodbye)*, [blues]
Johnny Adams - *Room with a view*, [blues], (G2)
Johnny Copeland - *Love song*, [blues]
Johnny Copeland - *San Antone*, [blues], (G2)
Lars Edegran Orchestra & New Orleans Jazz Ladies - *Panama*, [jazz]
Lars Edegran orchestra and the New Orleans jazz ladies - *Oh papa*, [jazz]
Lee Ritenour - *Starbright*, [jazz]
Lee Ritenour - *Tush*
Little Jimmy King & the Memphins soul survivors - *Wild woman*, [blues]
Memphis Slim - *Really got the blues*, [blues]
Memphis Slim - *Tiajuana*, [blues], (G1)
Miles Davis - *'Round midnight*, [jazz]
Miles Davis - *Human nature*, [jazz], (G2)
Miles Davis - *Seven steps to heaven*, [jazz]
Miles Davis - *So what*, [jazz]
Miles Davis - *Someday my prince will come*, [jazz]
Miles Davis - *Time after time*, [jazz]
Muddy Waters - *Baby please don't go*, [blues]
Muddy Waters - *Forty days and forty nights*, [blues], (G2)
Muska Babitzin - *Cry your heart out*, [blues]
Närpes skolmusikkår-Närpes youth band - *Malagueña*, [jazz]
Närpes skolmusikkår-Närpes youth band - *The pink panther*, [jazz]
Närpes skolmusikkår-Närpes youth band - *Watermelon man*, [jazz], (G2)
Paco de Lucia - *Chanela*, [jazz]
Paco de Lucia - *Solo quiero caminar*, [jazz]
Paco de Lucia - *Zyryab*, [jazz]
Pat Metheny group - *Follow me*, [jazz]
Pat Metheny group - *Too soon tomorrow*, [jazz]
Pelle Miljoona - *13 bar blues*
Pepe Ahlqvist and Jarkka Rissanen - *Bad, bad whiskey*, [blues], (G2)
Pepe Ahlqvist and Jarkka Rissanen - *Sip of tequila*, [blues]
Rory Block - *The spirit returns*, [blues], (G1)
Sanne - *Where blue begins*, [blues], (G1)
Spyro Gyra - *Heart of the night*, [jazz]
Spyro Gyra - *Surrender*, [jazz]
Spyro Gyra - *Westwood moon,* [jazz], (G1)
Terry Lighfoot - *Summertime*, [jazz]
The Brecker brothers - *Sponge*, [jazz]
The Brecker brothers - *Squish*, [jazz]
The Dave Weckl band - *Mud sauce*, [jazz]
The Dave Weckl band - *Song for Claire*, [jazz], (G1)
The Dave Weckl band - *The zone*, [jazz]
The Dutch swing college band - *Savoy blues*, [jazz]
The Erstrand-Lind quartet - *Avalon*, [jazz], (G1)
The Erstrand-Lind quartet - *I got rhythm*, [jazz], (G1)
The Jeff Healey band - *Roadhouse blues*, [blues], (G1)
Turner Parrish - *Ain't gonna be your dog no more*, [blues]
Weather report - *Birdland*, [jazz]
Weather report - *Harlequin*, [jazz]
Willie Harris - *West side blues*, [blues]

Rock/Pop

Abba - *Lay all your love on me*, [pop]
Abba - *S.O.S*, [pop]
Abba - *Waterloo*, [pop]
Beatles - *Love me do*, [rock]
Beatles - *Misery*, [rock]
BeeGees - *Alone*, [rock]
BeeGees - *Closer than close*, [rock]
BeeGees - *Still waters run deep*, [rock]
Benitez - *Mariposa (part 1)*, [rock]
Black sugar - *Viajecito*, [rock]
Bob Marley - *Sun is shining*, [pop]
Boo Radleys - *Lazarus*, [rock]
Boo Radleys - *Leaves and sand*, [rock]
Boo Radleys - *Upon 9th and Fairchild*, [rock]
Boris Gardiner - *I Wanna Wake Up With You*, [pop]
Britney Spears - *Lucky*, [pop]
Britney Spears - *Oops! I did it again*, [pop]
Béla Fleck and the Fleckstones - *Cheeseballs in Cowtown*, [country]
Béla Fleck and the Fleckstones - *Lochs of dread*, [country], (G2)
Béla Fleck and the Fleckstones - *Shubbee's doobie*, [country], (G1)
Béla Fleck and the Fleckstones - *Stomping grounds*, [country]
CMX - *Palvonnan eleitä*, [rock], (G1)
CMX - *Rautakantele*, [rock]
Celine Dion - *My heart will go on*, [pop], (G2)
Celine Dion - *River deep, mountain high*, [pop]
Chango - *Mira pa'ca*, [rock]
Chicago - *25 or 6 to 4*, [rock]
Chicago - *Colour my world*, [rock]
Chicago - *Saturday in the park*, [rock]
Children of Bodom - *Towards Dead End*, [metal], (G1)
Cradle of filth - *A gothic romance (red roses for the devil's whore)*, [metal], (G2)
Cradle of filth - *Beauty slept in Sodom*, [metal]
Cradle of filth - *Heaven torn asunder*, [metal], (G2)
Cream - *Sunshine of your love*, [rock]
Creedence clearwater revival - *(wish I could) Hideaway*, [rock]
Creedence clearwater revival - *Have you ever seen the rain*, [rock]
Creedence clearwater revival - *It's just a thought*, [rock]
Crosby, Stills, Nash &Young - *Dream for him*, [country]
Crosby, Stills, Nash &Young - *Heartland*, [country]

Crosby, Stills, Nash &Young - *Looking forward*, [country]
Depeche Mode - *It's no good*, [pop]
Depeche Mode - *Personal Jesus*, [pop]
Depeche mode - *Enjoy the silence*, [pop]
Desmond Dekker - *You Can Get It If You Really Want*, [pop]
Dire straits - *Money for nothing*, [rock]
Dire straits - *Ride across the river*, [rock]
Eagles - *Hotel California*, [rock]
Eagles - *Take it away*, [rock]
Faith No More - *Epic*, [rock]
Frank Sinatra - *Bad, bad Leroy Brown*, [easy listening]
Frank Sinatra - *Strangers in the night*, [easy listening]
Frank Sinatra and Nancy Sinatra - *Somethin' stupid*, [easy listening], (G1)
Gladys Knight - *You*, [pop]
Gladys Knight and the Pips - *It's gonna take all our love*, [pop]
Gladys Knight and the Pips - *Love overboard*, [pop]
HIM - *Bury Me Deep Inside Your Arms*, [metal]
Inner circle - *Mary Mary*, [pop]
Inner circle - *Standing firm*, [pop]
Inner circle - *We 'a' rockers*, [pop]
Jane's addiction - *Been caught stealing*, [rock]
Jane's addiction - *Jane says*, [rock]
Jane's addiction - *Kettle whistle*, [rock]
Jesus Jones - *The devil you know*, [rock]
Jesus Jones - *Your crusade*, [alternative], (G1)
Jimi Hendrix - *Machine gun*, [rock]
Jimi Hendrix - *Voodoo child*, [rock]
Joe Cocker - *That's all I need to know*, [rock]
Joe Cocker - *That's the way her love is*, [rock]
Joe Cocker - *Tonight*, [rock]
Kenny Rogers - *Ain't no sunshine*, [country], (G2)
Kenny Rogers - *Love don't live here anymore*, [country]
Kenny Rogers - *Three times a lady*, [country], (G1)
Kiss - *Journey of 1,000 years*, [rock]
Kiss - *Psycho circus*, [rock], (G1)
Kiss - *Within*, [rock]
Korn - *Got the Life*, [rock]
Life of agony - *Drained*, [metal]
Life of agony - *Other side of the river*, [metal]
Lynyrd Skynyrd - *Free bird*, [rock]
Lynyrd Skynyrd - *Swamp music*, [rock]
Madonna - *Into the groove*, [pop]
Madonna - *Like a virgin*, [pop]
Malo - *Street man*, [rock]
Mariah Carey - *My all*, [pop]
Marilyn Manson - *Sweet Dreams*, [metal]
Marisa Monte - *Danca da solidao*, [pop], (G2)
Marisa Monte - *Segue o seco*, [pop]
Michael Jackson - *Bad*, [pop], (G1)
Michael Jackson - *Black or white*, [pop], (G1)
Paula Abdul - *Opposites Attract*, [pop]
Pet Shop Boys - *Always on my mind*, [pop]
Pet Shop Boys - *Being boring*, [pop]
Red Hot Chili Peppers - *Parallel Universe*, [rock]
Robert Wells - *Bumble-bee boogie*
Robert Wells - *Rhapsody in rock IV*
Robert Wells - *Spanish rapsody*, (G2)
Santana - *Black magic woman*, [rock]
Santana - *She's not there*, [rock], (G1)
Sapo - *Been had*, [rock]
Saxon - *Dogs of war*, [metal]
Saxon - *The great white buffalo*, [metal]
Shania Twain - *Man! I feel like a woman*, [country]
Shania Twain - *You're still the one*, [country]
Skunk Anansie - *Brazen (Weep)*, [rock]
Steppenwolf - *Magic carpet ride*
Stone - *Empty corner*, [metal]
Stone - *Mad hatter's den*, [metal]
Suede - *Trash*, [pop]
The Golden Nightingale Orchestra - *Annie's song*, [easy listening]
The Golden Nightingale Orchestra - *Love story*, [easy listening], (G2)
The Golden Nightingale Orchestra - *The sound of silence*, [easy listening]
The Move - *Flowers in the rain*, [rock], (G2)
The Police - *It's alright for you*, [rock]
The Police - *Message in a bottle*, (G2)
Travis - *Turn*, [rock]
U2 - *Last night on earth*, [rock]
U2 - *Staring at the sun*, [rock]
Zucchero - *Eppure non t'amo*, [rock]
Zucchero - *Menta e Rosmarino*, [rock]
Zucchero - *Senza una donna*, [rock]

Soul/RnB/Funk

Al Green - *Let's stay together*, [soul], (G2)
Al Green - *Tired of being alone*, [soul], (G2)
All-4-One - *I turn to you*, [rnb]
Cameo - *I just want to be*, [funk], (G1)
Cassandra Wilson - *Last train to Clarksville*, [gospel]
Cassandra Wilson - *Memphis*, [gospel]
D'Angelo - *I found my smile again*, [rnb], (G1)
Defunkt - *Dogon A.D.*, [funk]
Defunkt - *Without justice*, [funk]
Dr.Funkenstein and the brides of Funkenstein - *Rat kissed the cat*, [funk], (G1)
Harry van Walls - *Tee nah nah*, [rhythm and blues]
Herbie Hancock - *Watermelon man*, [funk]
James Brown - *It's time to love (put a little love in your heart)*, [soul]
James Brown - *Show me*, [soul]
James Brown - *Standing on higher ground*, [soul], (G1)
Joe Morris - *The applejack*, [rhythm and blues], (G1)
Joe Turner - *Sweet sixteen*, [rhythm and blues]
Johnnie Taylor - *Lady my whole world is you*, [soul]
Kool & the Gang - *Funky stuff*, [funk], (G2)
Kool & the Gang - *Hollywood swinging*, [funk], (G2)
Kool & the Gang - *Jungle boogie*, [funk]
Kool & the Gang - *Spirit of the boogie*, [funk]
Latimore - *Bad risk*, [soul]
Lauryn Hill - *I used to love him*, [rnb]
Lauryn Hill - *Lost ones*, [rnb]
Lauryn Hill - *To Zion*, [rnb], (G1)
Lonnie Green - *I didn't know that funk was loaded (count Funkula)*, [funk], (G1)
Lucy Pearl - *Don't mess with my man*, [rnb], (G2)
Lucy Pearl - *Everyday*, [rnb]
Lucy Pearl - *Lucy Pearl's way*, [rnb]
Manu Dibango - *Big blow*, [funk], (G2)
McKinley Mitchell - *The end of the rainbow*, [soul]
Monica - *For you I will*, [rnb], (G2)
Oslo Gospel Choir - *Nearer my god to thee*, [gospel]
Oslo Gospel Choir - *Open up my heart*, [gospel]
R.Kelly - *I believe I can fly*, [rnb]
Ruth Brown - *Teardrops from my eyes*, [rhythm and blues]
Sade - *Kiss of life*, [soul], (G1)
Sade - *No Ordinary Love*, [soul]
Salt 'n Pepa - *Shoop*, [rnb], (G1)
Salt 'n Pepa feat.En Vogue - *Whatta man*, [rnb]
Staple singers - *Heavy makes you happy*, [soul]

Staple singers - *Long walk to D.C.*, [soul]
Staple singers - *Respect yourself*, [soul]
Stevie Wonder - *For your love*, [soul]
Stevie Wonder - *Superstition*, [soul]
Stevie Wonder - *You are the sunshine of my life*, [soul]
Stevie Wonder - *Master blaster*, [soul]
Take 6 - *Fly away*, [gospel], (G2)
Take 6 - *Mary*, [gospel]
The Rwenzori's - *Handsome boy (e wara)*, [funk]
Toni Braxton - *Let it flow*, [rnb], (G2)
Toni Braxton - *There's no me without you*, [rnb], (G2)
Toni Braxton - *Un-break my heart*, [rnb], (G1)
Vecchio - *Nsambei*, [funk]

World/Folk

Andras Adorjan and Jorge de la Vida - *Jalousie*, [latin american]
Antero Jakoila - *El bandolero*, [latin american]
Antero Jakoila - *El choclo*, [latin american]
Anúna - *The heart's cry*, [folk]
Astrud Gilberto, Stan Getz - *Corcovado*, [latin american]
Brendan Larrissey - *Mist on the mountain/Three little drummers*, [folk]
Clannad - *Coinleach ghlas an fhómhair*, [folk]
Davy Spillane, The riverdance orchestra - *Caoineadh cú chulain*, [folk]
Horacio Salgan and Ubaldo de Lio - *El Choclo*, [latin american]
Joe Derrane with Carl Hession - *Humours of Lissadell/Music in the glenn/ Johnson's*, [folk]
Joni Mitchell - *For free*, [folk]
Joni Mitchell - *Ladies of the canyon*, [folk]
Joni Mitchell - *Rainy night house*, [folk]
Roberto Goyeneche and Nestor Marconi - *Ventanita Florida*, [latin american]
Stan Getz and Joao Gilberto - *Desafinado*, [latin american]
Walter Wanderley - *Ó barquinho*, [latin american]