AUTOMATIC SUBGENRE CLASSIFICATION OF HEAVY METAL MUSIC

Valeri Tsatsishvili
Master's Thesis
Music, Mind & Technology
November 2011
University Of Jyväskylä

JYVÄSKYLÄN YLIOPISTO

Tiedekunta – Faculty	Laitos – Department
Humanities	Music
Tekijä – Author	
Valeri Tsatsishvili	
Työn nimi – Title	
Automatic Subgenre Classification Of Heavy Metal Music	
Oppiaine – Subject	Työn laji – Level
Music, Mind and Technology	Master's Thesis
Aika – Month and year	Sivumäärä – Number of pages
November 2011	59

Tiivistelmä – Abstract

Automatic genre classification of music has been of interest for researchers over a decade. Many successful methods and machine learning algorithms have been developed achieving reasonably good results. This thesis explores automatic sub-genre classification problem of one of the most popular meta-genres, heavy metal. To the best of my knowledge this is the first attempt to study the issue. Besides attempting automatic classification, the thesis investigates sub-genre taxonomy of heavy metal music, highlighting the historical origins and the most prominent musical features of its sub-genres.

For classification, an algorithm proposed in (Barbedo & Lopes, 2007) was modified and implemented in MATLAB. The obtained results were compared to other commonly used classifiers such as AdaBoost and K-nearest neighbours. For each classifier two sets of features were employed selected using two strategies: Correlation based feature selection and Wrapper selection.

A dataset consisting of 210 tracks representing seven genres was used for testing the classification algorithms. Implemented algorithm classified 37.1% of test samples correctly, which is significantly better performance than random classification (14.3%). However, it was not the best achieved result among the classifiers tested. The best result with correct classification rate of 45.7% was achieved by AdaBoost algorithm.

Asiasanat – Keywords

Automatic genre classification, heavy metal, subgenre

Säilytyspaikka – Depository

Muita tietoja – Additional information

ACKNOWLEDGMENT

I owe my deepest gratitude to two persons: Tuomas Eerola - who gave me a chance to attend Music, Mind and Technology program and to my wife who gave me strength and stamina to overcome difficult times we have encountered. This thesis would not have been possible to be written without them.

I would like to thank my supervisors Petri Toiviainen and Olivier Latrillot for their time and support. I learnt a lot about MATLAB and coding in general from Olivier. I am grateful to Rafael Ferrer, who showed me how to extract information from Last.fm and opened the world of Python programming language to me. I would also like to thank Pasi Saari for fruitful discussions about feature selection and WEKA software.

TABLE OF CONTENTS

1. INTRODUCTION

During the past decades developments in computer and media technology has been brought about by the dramatic increase of digital music databases in size. This phenomenon resulted in growing attention towards automatic content based organization of digital music databases since it became prohibitively expensive to use human experts for manually indexing such databases. Important developments have been made in music search and recommendation systems; MPEG-7[1] is also a step forward to make multimedia indexing and searching faster and more effective. However, the perceptually subjective nature of many descriptors (such as genres) and the lack of universal models describing genres creates the need for richer metadata. Most of the existing standard search systems remain mostly based on query by metadata paradigm or categorical browsing. Metadata of the most common audio data format nowadays – ID3 tags of MP3 consists of artist name, publisher name, song title, release year and genre of the track, though, it is not guaranteed that information provided in ID3 tags is annotated by an expert and therefore is reliable (e.g. McKay & Fujinaga, 2006). In addition to metadata, content based descriptors are essential for browsing effectively in the sea of audio tracks, especially for discovering new music.

Music Information Retrieval (MIR) is a relatively new field of research which deals with automatic information extraction from music to ease the accessibility of music through information technology. Along with other applications such as music recommendation systems, intelligent search systems, etc. the scope of MIR research involves content based organization of digital music databases.

The first important criterion for dealing with content based organization of music (which, roughly speaking, is clustering similar tracks together) is similarity. Although music similarity is multidimensional, when no specific similarity facet is privileged by the user, genre is the most common descriptor involving virtually all dimensions of similarity. Moreover, genre is the most widely used construct for categorizing music by record labels, record stores, streaming radios, etc. Therefore, genre classification, although quite problematic because of the inherent ambiguity of the genre definition and its subjective character, is and probably will be for a reasonably long time, the most natural paradigm for most of the users browsing music in an audio database. Consequently, when music production growth created the need of automatic

[1] MPEG-7 is an international standard for multimedia content description providing rich metadata including both low-level and high-level description tools for audio.

organization of music databases, automatic genre classification became one of the most popular research topics on content based music organization.

1.1 Motivation

In parallel to growing music production, genres undertake evolution as many new or cross-genres emerge; some genres are merged together or are further divided into subgenres. In addition to constant alteration, complexity of genre hierarchy gradually increases since new layers representing new subgenres appear. This is especially true for heavy metal music. Starting from its emergence in late 1960s the genre has grown dramatically from one small branch of rock music to a big genre consisting of more than 20 subgenres. Nowadays heavy metal is one of the most popular genres spanning from lyrical ballads to the most extreme forms of music. Despite, it is less indulged by attention from scientific community than other more 'traditional' genres of music. This motivated me to explore heavy metal music and partially fill the lack of research on this genre and its subgenres.

Apart from exploring roots and musical characteristics of heavy metal subgenres, it was an interesting challenge to attempt automatic classification on subgenre level where degree of fuzziness in genre definitions increases and boundaries between them blurs extremely. At the same time this attempt is not fully conceptual, but has a solid practical application, since genre remains the most widely used descriptor of music at any level of genre hierarchy. For example, analysis of tags of more than 1000 heavy metal tracks, which were extracted from last.fm[2], showed that the most popular descriptors (tags) applied to this music are genres/subgenres, despite the fact that users are completely free to use any descriptor they prefer (also see Lamere & Pampalk, 2008). Therefore, considering the popularity and diversity of heavy metal music, subgenre classification is as significant as classification on a more general level.

To summarize, the main aims of this thesis are:

1. To automatically classify heavy metal music into its subgenres using machine learning algorithm implemented in this thesis (based on the classification strategy proposed in (Barbedo & Lopes, 2007)).
2. To test if the implemented algorithm is optimal for subgenre classification tasks by comparing its result to two other successfully used pattern classification algorithms tested on the same dataset.

[2] www.last.fm - one of the active social networking and internet radio websites

3. To partly fill the lack of literature exploring musical features, genre hierarchy, and evolution of heavy metal, especially since the 1990s.

1.2 Thesis organization

The thesis consists of seven chapters. Chapter 2 reviews existing research in automatic genre classification. Chapter 3 presents genre taxonomy of heavy metal music along with description of each subgenre. Chapter 4 describes collected audio dataset that was used for system evaluation as well as feature selection and classification algorithms employed. In Chapter 5 results are reported, followed by evaluation and conclusions. Chapter 6 provides outlook for future improvements.

2. BACKGROUND

2.1 Problems with genre taxonomies

As mentioned above, genre is a vaguely defined construct, which makes it inherently subjective. Pachet & Cazaly (2000) addressed inconsistencies in genre taxonomy used by the music industry and by the well-known internet databases such as Amazon[3], Allmusic[4] and Mp3[5]. The authors showed that not only different types of hierarchies are employed in terms of semantics behind relationships between levels, but also different genre labels starting from subgenre level to the most general genres (e.g. rock, pop, etc.). Furthermore, different databases have different sets of artists under the same genres.

Another problem of genre classification is that depending on whether artists, albums or tracks are classified, the results of the taxonomy can be significantly different. Since record labels and music magazines are frequently definers of the genres, taxonomies are album-oriented (Aucouturier & Pachet, 2003; Scaringella, Zoia, & Mlynek 2006), though, such classification may not be effective in many cases as probable diversity of the album content is neglected. For example, many metal albums feature ballads, acoustic instrumentals or short overtures next to the typical heavy metal songs. Although album-level classification could be effective in specific situations, mostly either more general (but less precise) artist-level classification or very specific track-level classification would be more effective option in database browsing.

Problems present in heavy metal genre taxonomies are summarized below with relevant examples from several well-known sources (Table 1) such as online shops Nuclear Blast[6], Relapse records[7] and Amazon.com, musical websites Allmusic.com, Encyclopedia Metallum[8] and from the book 'Metal, The Definitive Guide' by Sharpe-Young (2007).

- Semantics of the taxons are not consistent. Genre labels denote different matters such as historical event (e.g. New Wave of British Heavy Metal), geographical location (e.g. 'Scandinavian Metal', 'Norwegian Black Metal'), instrumentation

[3] www.amazon.com
[4] www.allmusic.com
[5] www.mp3.com
[6] http://www.nuclearblast.de/en/shop/artikel/gruppen/79928.cd.html
[7] http://shop.relapse.com/store/product.aspx
[8] www.metal-archives.com

('Symphonic Metal'), or specific features (e.g. outfit related 'Hair Metal' and lyrics related such as 'Christian' or 'Viking Metal').

- Taxons are used interchangeably. It is not hard to find the same bands or albums listed under different genres in different databases. Especially when band's music spans across more than one genre. A good example would be the band Nevermore. Even though the band has not seen dramatic changes in music during their career, it is classified as 'Progressive Metal' on Nuclear Blast, as 'Power Metal' in (Sharpe-Young, 2007) and on Mp3.com, and as 'Power/Thrash Metal (early) Groove/Progressive Metal (later)' on Encyclopedia Metallum.

- No consensus in taxonomies and their organization exist. Sometimes differences in labels addressed towards the same genre of music can be quite confusing. For instance, in Table 1 'Neo-Metal', 'Rap-Metal', and 'Nu metal' labels used by different sources refer to the same genre (at least they contain mostly the same set of artists), more widely known as Nu Metal. It is worth mentioning a few extremely ambiguous genre labels such as 'Modern Metal' and 'True Metal' found on Nuclear Blast or 'Metal' on Relapse Records (see Table 1).

  In addition to the inconsistencies in taxonomies, their organization is not always clear, or at least not easy to understand. For instance, in his book Sharpe-Young (2007) presumably organizes content by geographical location. For that reason the author adds location markers to the subgenres (e.g. American Death, UK Thrash etc.) or just uses such constructs as 'European Metal' (which is not really a musical genre). However, under such organization it is unclear why there are exceptions such as the section 'Power Metal' which involves bands from different locations. Allmusic.com has inserted 'Heavy Metal' under 'Hard Rock' category and blends subgenres of the two together (e.g. 'Blues Rock' is in the same category as 'Death Metal').

Relapse Records	Nuclear Blast	Encyclopedia Metallum
Ambient Black Metal Death Metal Doom Experimental Goregrind Folk metal Gothic metal Grindcore Industrial Metal Metalcore Noise Power/Traditional Progressive Sludge Thrash	Black Metal Dark Wave/EBM Death Metal Doom/Stoner Rock Gothic Grindcore HardRock Heavy/True/Power metal Industrial Mittelalter/Folk Nu Metal/Metalcore Progressive Rock/Metal Symphonic Metal Thrash Metal Viking/Pagan/Epic Amb/Experimental Modern metal/rock	Black Metal Death Doom Electronic Folk/Viking Gothic Heavy/Traditional Orchestral/Symphonic Power Progressive Speed/Thrash
Allmusic	**Amazon**	**Metal, The Definitive Guide**
Christian Metal Death Metal Grindcore Heavy Metal Speed Metal Hair Metal Alternative Metal British Metal Industrial Metal Rap-Metal Guitar Virtuoso Progressive Metal Neo-Classical Metal Pop-Metal NWOBHM Glitter Punk Metal Stoner Metal Scandinavian Metal Goth Metal Doom Metal Symphonic Black Metall Sludge Metal Power Metal	Compilations Live Albums Alternative Metal British Metal Death Metal Grunge Hard Rock Pop Metal Progressive Metal Thrash & Speed Metal	Heavy NWOBHM American Thrash UK Thrash American Death and Grindcore Norwegian Black Worldwide Black Doom Power Progressive Gothic and Symphonic Metal US Metal German Metal Swedish Metal Finish Metal Japanese Metal South and Central American Metal European Metal Neo-Metal NWOAHM Innovators

Table 1 Genre taxonomies retrieved from Relapse Records, Nuclear Blast, Encyclopedia Metallum, Almusic.com, Amazon.com and (Sharpe-Young, 2007)

- Scalability of genre taxonomies and usage of ‘umbrella terms’. Genres definitions are not constant, but rather they change meaning gradually. A direct consequence of this trend is the existence of the terms ‘Old School’ as in ‘Old School Death Metal’ which refers to the music of early death metal bands in the 1980s. Some commonly used examples of umbrella terms would be ‘Scandinavian Metal’, ‘Melodic Metal’, ‘Extreme Metal’, ‘Experimental Metal’, and so on. All of the listed genres contain a variety of musically diverse styles of music.

The listed problems make existing genre taxonomies rather impractical for training a machine learning algorithm, since it is extremely difficult (if not impossible) to design system which will be able to adapt to such problems like human beings do.

Patchet & Cazaly (2000) attempted to design an objective and very detailed taxonomy which would limit aforementioned problems. However, later, the authors changed their initial idea because they found quite difficult to objectively describe lower level subgenres and to design a taxonomy which would be flexible with respect to the evolution of music. Finally, the authors came up with simpler genre taxonomy for artist-level classification. A natural question here is why would it be reasonable to work on that problem, if it seems impossible to either avoid the aforementioned flaws in the existing genre taxonomies or to design an automatic classification system which would adapt to those problems? It is impossible to design an automatic genre classification system with perfect classification accuracy due to absence of ground truth for training such algorithms. In fact, recently concerns have been raised among researchers about the existence of the ceiling of automatic genre classification algorithm performance using standard features (Aucouturier & Patchet, 2004; Pampalk, Flexer, & Widmer, 2005) and the usefulness of further research in this direction. This issue was addressed in (McKay & Fujinaga, 2006) where the importance of the genre classification was underlined and several ideas to improve existing systems were suggested. These suggestions propose more active involvement of an interdisciplinary approach towards genre including psychological and cultural perspectives of human categorisation, the possibility to assign multiple genres to music pieces where weighting can be used for visualizing relative importance of the assigned genres and labelling individual sections of a recording. The need for such a multi-genre system was mentioned in (Scaringella et al., 2006) as well. However, it is challenging to apply a multi-genre system to real-life data without losing the clarity of the categories. In such a database any set of tracks would be contained within several categories and the clarity of organisation would be poor unless genre weights are solid, objective, and well-defined for the end user. In my opinion, it would be effective if a panel of experts de-

scribed their own ‘templates’ - combination of features for each category, and from all the proposed property combinations for each genre the ones holding the highest degree of agreement could be selected as templates. The weights of genres in the multi-genre system would be consistent if defined by the proportions of different templates present in a piece of music.

The existence of a limit to the accuracy that can be achieved by current automatic classification systems is credible, but an important question is how it can be compared to the results of an average human listener. The lack of empirical evidence on human genre classification ability makes such a question quite difficult to answer. From the few existing researches in human genre classification, the experiment conducted by Perrott and Gjendingen (Gjerdingen & Perrott, 2008) is noteworthy. In the experiment 52 participants were asked to classify 80 pieces of music from 10 genres. The highest accuracy of 70% was reported for three-second excerpts. A more or less similar set of 10 genres was used for testing automatic classifiers in (Tzanetakis & Cook, 2002) and later in (Lee, Shih, Yu, & Lin, 2009) achieving accuracies of 61% and 79% respectively. Nevertheless, it is not possible to directly compare these results since datasets and genre taxonomies employed were not identical. A more valid comparison of human ability to algorithm performance was presented in (Heittola, 2003) where stimuli for a listening experiment were selected from the data that was also used for testing several classification algorithms. Author reported a 10-16% difference between recognition rates, human achieving top of 75% and an automatic classifier 59%. However, the above mentioned work is the only one I was able to find that used the same data for conducting listening experiment on human participants and automatic classification. Therefore, the need for the further research on this problem is evident.

2.2 Automatic genre classification

Since the beginning of 2000s growing interest towards automatic genre classification systems initiated many algorithms using various machine learning methods (an overview can be found in (Scaringella etal., 2006). In general, virtually all supervised pattern classification systems share the same overall structure, which can be divided into three stages: feature extraction, training, and classification.

The idea of feature extraction is to have more compact representation of audio relevant to the specific task. This procedure avoids analyzing any redundant data and results in performing

the task more efficiently and more effectively. For example, by dividing digital audio[9] in short 25ms non-overlapping frames and extracting the average spectral centroid, we will have compact (40 feature values per second) representation characterising dynamics of timbre brightness over the whole audio. Such representation would contain >1000 times less data compared to the raw audio file containing 44100 values per second. However, selecting a compact representation of the audio that contains sufficient information for solving specific problem is a challenging task, since having redundant or irrelevant information can deteriorate performance of the learning algorithm. Therefore features should be selected carefully.

Commonly, features for genre classification are extracted from 20-90ms half overlapping hamming windows, and then often aggregated over the longer segments both to further reduce amount of data representing each file and to capture longer dynamics. Several feature aggregation strategies have been proposed in the literature. One approach is to summarize distributions of feature values over the whole track (e.g. see (Pampalk et al., 2005). Another approach is to summarize feature values over fixed time (less than track length) segments containing several frames, e.g. (Tzanetakis & Cook, 2002). Event based summarization has also been employed where segment size is defined by specific events in the music stream, such as beats or onsets (West & Cox, 2004). Yet another completely different strategy is to directly classify frames of the whole song without any aggregation and derive song genre from a majority vote among its frames. (Xu, Maddage, Shao, Cao, & Tian, 2003)

A systematic study of the effect of segment size on classification accuracy has been done in (Bergsta, Casagrande, Erhan, Eck, & Kegl, 2006) where 16 combinations of four feature sets and four classifiers were tested on seven different segment lengths ranging from 1.8 to 27.9 seconds respectively. For evaluation the authors used the same GTZAN dataset used in (Tzanetakis & Cook, 2002). The experiment showed that the optimal track-level classification rate for three out of four classifiers (namely AdaBoost.Stump, AdaBoost.Tree, and sigmoidal neural network) was achieved for 3.5 second segment sizes, while Support Vector Machines (SVM) required longer segments for optimal performance. West and Cox (2005) also compared a wide range of segmentation methods involving segment sizes from 23ms frames to the length of the whole track. It was reported that the highest classification rate was achieved for event-based segmentation, where segment boundaries were defined by an onset detection function, though, it should be noted that their evaluation dataset was small and inconsistent in genres.

[9] For this example mono PCM wav file with 44.1kHz sampling frequency

Several strategies for feature aggregations have been proposed in the literature. The most common method is the estimation of Gaussian distribution from the mean and covariance matrix of all features e.g. (Li, Ogihara, & Li, 2003) or by mixtures of Gaussian distributions (Pampalk et al., 2005). Another strategy has been proposed in (McKinney & Breebaart, 2003), where a power spectrum from short frame–level feature values was calculated across 740 ms segment and energy was estimated in the following four bands, 0Hz, 1-2Hz, 3-15Hz and 20-43Hz. Meng, Ahrendt, & Larsen (2005) compared the above described methods of feature aggregation to the autoregressive (AR) model. The authors reported superior performance of AR model for Gaussian based and Linear Neural Network classifiers. Yet another method is the long-term modulation spectral analysis (Lee et al., 2009) that captures long-term dynamics from time series of frame-based features and has been reported to improve classification accuracy on two widely used dataset (achieving 86% correct classification rate on GTZAN and 90% - on ISMIR Genre[10]). In this study octave-based spectral contrast (OSC), normalized audio spectral envelope, and MFCC features were used for audio parameterization. K-means algorithm has been also used for feature aggregation (Park, Oh, Yoon, & Lee, 2005). The K-means is an unsupervised learning algorithm where the system automatically forms clusters based solely on the structure of the training data.

In addition to the extraction methods, parametrization plays substantial role in the classification process. Throughout the last decade many descriptors have been used for different genre classification algorithms. The most widely used descriptors in literature, representing timbral, pitch, and rhythmic information of music that proved useful for genre classification tasks are summarized in Table 2. Precise descriptions of these features are available in the literature (e.g. refer to Peeters, 2004; West, 2008; Tzanetakis, 2002), thus, only brief explanations are provided in the table.

There is no theory defining the optimal feature set for music classification, since the results of most of the existing studies are not directly comparable. One of the problems is that there are only few annotated music databases accessible to researchers for evaluating classification systems and most of the algorithms are evaluated on different (in many cases quite small) databases. Nevertheless, a few studies tested different feature sets on the same dataset and classifiers, and interestingly, timbre descriptors (MFCCs or FFT coefficients) have been observed to perform better for genre classification task than pitch or rhythm descriptors alone (e.g. Li et al., 2003; Li & Ogihara, 2006). On the other hand, it was also suggested that

[10] The dataset was used in the ISMIR 2004 Music Genre Classification Contest

Zero Crossing Rate (ZCR) (Burred & Lerch, 2003; McKinney & Breebaart, 2003)	*Number of time domain zero crossings of the signal*
Root Mean Square (RMS) (West, 2008; McKinney & Breebaart, 2003)	*Root Mean Square energy of the signal. Roughly estimates the perceived loudness.*
FFT Spectrum (mostly FFT coefficients) (Bergsta et al., 2006)	*Fourier Transform (FT) of audio signal frame*
Statistical descriptors of spectral shape (Bergsta et al., 2006; Peeters, 2004)	*These include:* *Centroid of Magnitude spectrum of the signal* *Spread of the spectrum around its mean value* *Skewness – Measure of asymmetry of a distribution around its mean value* *Kurtosis - Measures flatness of a distribution around its mean value* *Slope - Represents the amount of spectral amplitude decrease* *Rolloff - Frequency below which 85% of magnitude distribution is located*
Entropy (West, 2008)	*The entropy of the spectrum. High value indicates presence of high amount of noise in the spectrum (flatter spectrum).*
Low energy rate (Tzanetakis & Cook, 2002; Burred & Lerch, 2003)	*Percentage of frames with energy less than average energy over the whole signal*
Spectral Flux (West, 2008)	*Difference between the normalized magnitudes of successive spectra*
Octave Based Spectral Contrast (OSC) (West & Cox, Features and Classifiers for the Automatic Classification of Musical Audio Signals, 2004)	*Octave-scale bandpass filters are applied to the FFT spectrum and in each subband spectral valleys are substracted from spectral peaks*
Roughness (McKinney & Breebaart, 2003)	*Roughness is sensory dissonance perceived as beating when pair of sinusoids is located within the same critical band, corresponding temporal envelope modulations in the range of 20-150 Hz.*
Loudness (Peeters, 2004; Burred & Lerch, 2003)	*Weighting signal spectrum with human ear Frequency response*
Cepstrum	*Fourier transform of the log spectrum of the signal*
Mel Frequency Cepstral Coefficients (MFCC) (Lee et al., 2009)	*Discrete cosine transform of the logarithm of the spectrum computed on mel frequency bands*
Beat Histogram based features (Tzanetakis & Cook, 2002)	*Beat histogram represents beat strength as a function of tempo values*
Beat Spectrum (Foote & Uchihashi, 2001)	*Is calculated by finding periodicities in similarity matrix of the frame based features.*
Pitch Histogram based features (Tzanetakis & Cook, 2002)	*Shows frequency of each pitch (or pitch class) occurrence in audio*
Bandwidth (McKinney & Breebaart, 2003; Barbedo & Lopes, 2007)	*Frequency bandwidth of the signal*
Band energy ratio (McKinney & Breebaart, 2003)	*Ratio of the energy at a certain frequency band to the total energy.*
Linear Predictive Coefficients (LPC) (Bergsta et al., 2006)	*Compressed representation of spectral envelope of the signal*

Table 2. List of commonly used features in genre classification tasks and their brief explanations. Next to each feature papers are referenced where the feature was used.

counting solely on low-level timbre descriptors will lead to a ceiling in classification performance (Aucouturier & Patchet, 2004). McKay and Fujinaga (2006) addressed this problem and underlined the need of major changes in current approach to overcome this problem. The authors reviewed musicological and psychological perspectives of the creation, perception and the evolution of the genre. It was concluded that in addition to commonly used low-level descriptors, higher level musical features as well as cultural features should be considered by the MIR community to extend currently achieved accuracy levels. Only few attempts have yet been made in this direction. For example, Lidy, Rauber, Pertusa, & Iñesta (2007) combined higher-level symbolic descriptors obtained by first transcribing an audio signal such as inter-onset interval (IOI), note pitches, and note durations, with standard low-level features and reported improved genre classification accuracy on the GTZAN dataset, reaching 76.8% of correctly classified samples. Whitman & Smaragdis (2002) tested a combination of low-level and cultural features mined from the web on a fairly small dataset and demonstrated that a combination of those two feature sets performs better than each feature set separately.

Other researchers proposed novel features such as rhythmic cepstral coefficients (RCC) (West, 2008); Daubechies Wavelet Coefficient Histograms (DWCHs), which outperformed other widely used features, achieving highest of 78.5% correct classification on GTZAN dataset (Li et al., 2003). Jang, Jin, & Yoo (2008) also reported improved classification results using a new features and new classification technique obtained by modifying spectral flatness and spectral crest features using modulation spectral analysis.

Despite some pessimistic conclusions about the limitations of the commonly used features for genre classification, most of the reviewed articles are still trying to further improve precision of automatic classifiers by implementing new features or classification algorithms. However, only few authors indirectly questioned usefulness of such competition for real world applications. Perhaps improving the quality of the classification result (i.e. having more acceptable errors) would be more practical and useful for the end user than gaining few more percents in classification accuracy. From my point of view it would be easier to browse a database where there are many low, subgenre-level errors that are easily adaptable for users than one with relatively few but higher level genre misclassifications.

One way to get errors that are more acceptable is to use a hierarchical classification scheme in which classification consists of several stages corresponding to the number of layers in the hierarchy. Depending on the direction of classification, top-down and bottom-up approaches exist. In the former, content is first classified at the highest layer, i.e. broader classes, and the process will go through all the layers to the very bottom sub-genre level, whereas in the

latter, classification starts from the lowest level and ends at the highest level. Burred & Lerch (2003) used a top-down approach for several layer hierarchy consisting of Speech, Music, Background classes at the top and 12 music subgenres at the lowest layer. In the study feature selection algorithm was used for each level of the hierarchy, meaning that for each of the nine split in the hierarchy a respective feature set was obtained. This method is quite interesting. The point is that the same set of features cannot be equally suitable for separating all genres because not all of them have similar distinctive criteria. Thus, employing variable feature sets for each level of hierarchy should perform better than constant parametrization. For comparison, in the study the test data was also directly classified in 17 lowest classes without employing a hierarchical model. Interestingly, both hierarchical and direct classification achieved similar accuracy (58.71% and 59.77% respectively), however, according to the authors, the hierarchical classification produced more acceptable errors.

A bottom-up approach to hierarchical classification was employed in (Barbedo & Lopes, 2007) where the classification procedure was performed on the lowest level and the higher level genres were defined by hierarchy itself. The results were remarkable, achieving 87% for the highest of a four-layer hierarchy consisting of Classical, Dance and Pop/Rock genres and 61% of correctly classified samples for 29 subgenres at the lowest layer.

After parameterization, a classifier is trained with training feature vectors and evaluation data is classified. A wide range of algorithms have been applied to music classification tasks, which can be divided into three categories. A relatively simple category of classifiers is instance-based learners. Probably one of the most widely used classifier in this category is K-Nearest Neighbours (K-NN) (Pampalk et al., 2005; Park et al., 2005). A description of this classifier is provided in section 4.2.3.3

Gaussian classifiers have been successfully used for genre classification as well (Tzanetakis & Cook, 2002; Burred & Lerch, 2003; West & Cox, 2004; McKinney & Breebaart, 2003). A Gaussian learning algorithm assumes that in each class the feature distribution can be modelled using a single Gaussian, or a mixture of several Gaussian distributions (GMM). For example, GMM3 indicates that each class is modelled using a mixture of three Gaussians. Parameters of the distribution (mean and covariance matrix of feature values) are estimated from training data.

Another statistical classifier is the Hidden Markov Model (HMM). A Markov Model (or Markov Chain) is a model describing a system undergoing transitions between N finite number states, provided that the process is random and the next state depends only on current state. For each given time t the system is in a particular state q_t and all the possible transitions to another

state in $t+1$ can be represented by a transition probability matrix $\{a_{ij}\}$ $1<i,j<N$, which contains probabilities for all the possible transitions for time $t+1$. Therefore, using this model the probability of observing the system in any of the states can be calculated. In genre classification setting, the system would be the test data to be classified and the states would be classes. Unlike Markov model, in HMM states themselves are hidden, i.e. not observable.

Another class of classifiers extensively used in the literature is discriminative classifiers. The basic idea of discriminative classifiers is finding a discriminant function resulting in the best separation between classes. Linear Discriminant Analysis (LDA) (West & Cox, 2004) is a simple but fast classifier of this class. The discriminant function for LDA is built by linear combination of feature vectors.

Another successfully used discriminative classifier is Support Vector Machines (SVM) (Mandel, Michael, & Ellis, 2005; Xu et al., 2003; Lidy et al., 2007). For an N-dimensional feature vector space SVM tries to find N-1 dimensional hyperplane which maximizes margin, i.e. the distance between the hyperplane and the nearest data points (Support Vectors).

A comparative study testing performance of the described classifiers was conducted by Li et al. (2003). The authors compared performance of GMM, LDA, KNN and SVM classifiers on several feature sets on GTZAN dataset and found that discriminative classifiers (SVM and LDA achieving the highest classification rate of 78.5% and 71.3% respectively) performed significantly better than statistical GMM (63.5%) and distance-based K-NN (62.1%) learning algorithms. Furthermore, SVM outperformed LDA for all feature sets. Lee et al. (2009) tested K-NN, GMM, and LDA on several feature sets. They used the GTZAN dataset for evaluation and their proposed features. Again, discriminative classifier (LDA) outperformed GMM and K-NN with a 90.6% correct classification rate, which is also the highest result achieved so far on the dataset. In general, SVM is currently one of the best performing classifiers in genre classification; Indeed, SVM was used as a classifier in three of the four studies achieving top results on GTZAN dataset (Table 3).

Authors	***Classifier***	***Achieved accuracy***
(Lee et al., 2009)	*SVM*	*90.6%*
(Bergsta et al., 2006)	*AdaBoost*	*82,5%*
(Li et al., 2003)	*SVM*	*78.5%*
(Lidy et al., 2007)	*SVM*	*76.8%*

Table 3 List of top four results achieved on GTZAN dataset

While performing well in genre classification tasks, SVM classifier is computationally heavy, which makes it impractical to use for large databases. A similarly well performing but more efficient discriminative classifier, AdaBoost, was employed in (Bergsta et al., 2006). In addition to the high results on GTZAN dataset (see table 3), the algorithm also won genre classification task in the 2005 MIREX[11] contest. AdaBoost builds discriminant function by iteratively calling weak learner (in their study decision tree), which votes for or against each class returning a binary vector containing voting results and combining weighted votes of these classifiers. For a given test vector, the class collecting most of the votes is selected.

Classification algorithms recently achieved remarkable accuracy, which arguably is comparable to human ability (such comparison is still impossible due to the lack of empirical evidence). Both direct and hierarchical approaches have been successful. Although there is no evidence that any of the two strategies invariably perform significantly better in terms of achieved accuracy, the hierarchical strategy has been reported to result more acceptable errors (Burred & Lerch, 2003). In addition, the hierarchical approach features few other advantages, namely it is more flexible, new genres can easily be added to the dataset without major changes in the structure. The drawback of the system is that building a consistent genre hierarchy is not an easy task and if not done properly it will significantly degrade classification quality.

[11] Music Information Retrieval evaluation exchange (MIREX) (Downie, 2008)

3. GENRE TAXONOMY

3.1 Introduction

Heavy Metal is an ambiguous term that has two uses. Originally it was coined for the subgenre of Rock music emerged in late 1960s and early of 1970s. As the subgenre evolved and diversified, 'Heavy Metal' remained as an umbrella term for all of its stylistic variations, although these variations became so diverse that they can hardly be merged under one label. Therefore, to avoid confusion, 'Heavy Metal', or 'Traditional Heavy Metal' will be used hereafter for labelling music of mostly 1970s pioneers as well as to their modern followers and 'Metal' will be applied as an umbrella term for all heavy metal subgenres.

Since only a few academic sources have explored subgenres of metal music, an effort has been made to research the topic before organizing the dataset. Existing subgenre taxonomies of metal as well as the historical origins of the subgenres, and their musical features were explored. In Figure 1 outside influences on and interrelationships among the subgenres are presented. For clarity of representation the whole diagram was divided into two parts, therefore some subgenres are present more than once, denoted in shaded boxes. It should be noted that not all the distinguishable features of subgenres are musical, but sometimes the subject of lyrics can be a defining factor of a subgenre (e.g. Christian Metal or Viking Metal), or outfit fashion (e.g. Glam Metal). In this thesis I focused only on genres that are musically distinctive to some degree. In the following sections short historical overview, the influences and main musical features of the subgenres are described.

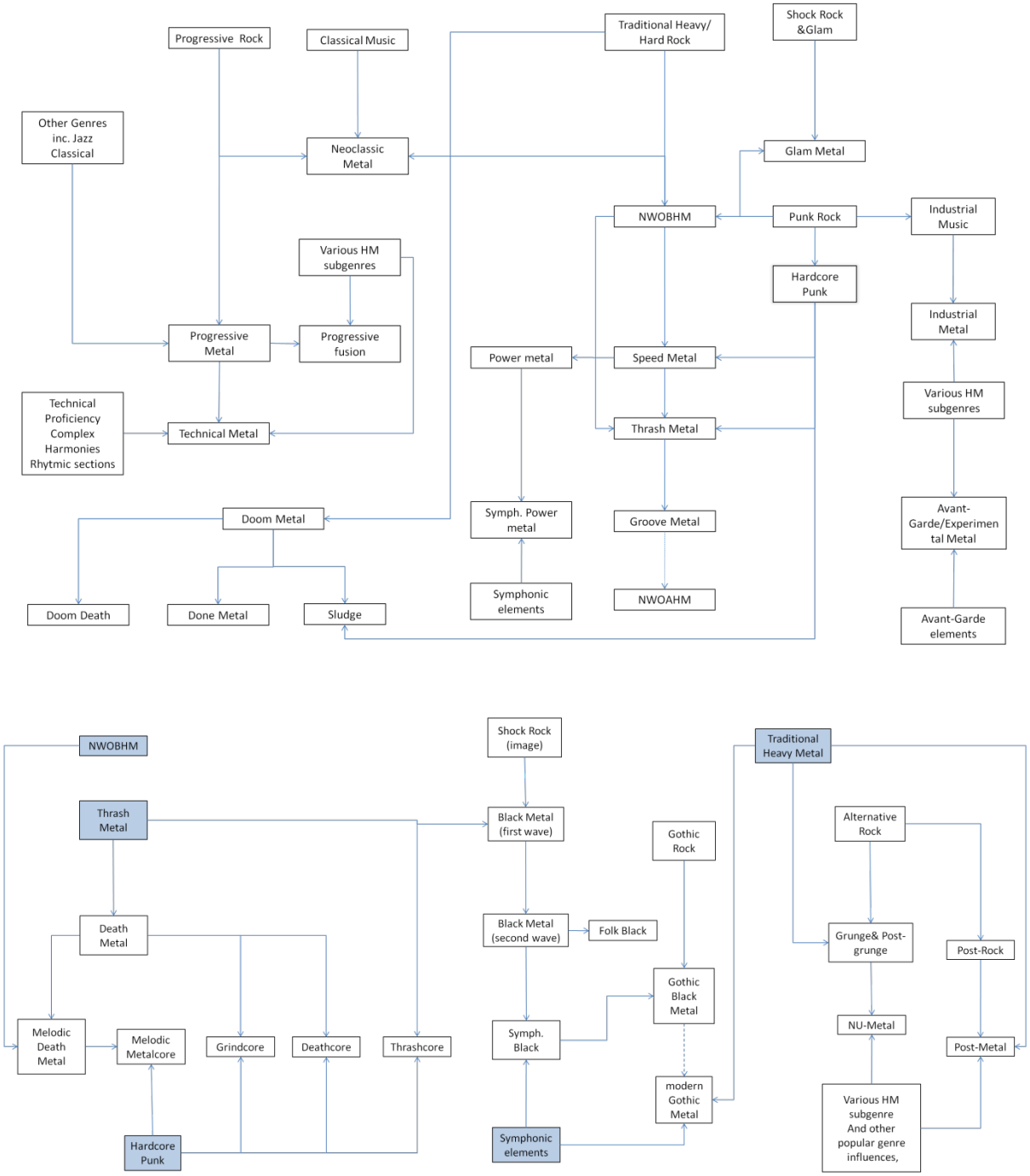

Figure 1 Interrelationships between subgenres of Metal music.

3.2 Subgenres of Metal music

3.2.1 Traditional Heavy Metal

In this thesis traditional heavy metal incorporates three subgenres: early metal, New Wave of British Heavy Metal (NWOBHM), and glam metal. Heavy metal originally was one branch of rock music founded by bands such as Deep Purple, Led Zeppelin, defined by Black Sabbath and later further developed by NWOBHM artists. However, the list of pioneers is heavily debated and somewhat connected to more general confusion between two labels, namely 'Hard Rock' and 'Heavy Metal'. Indeed, whether hard rock and heavy metal are separate genres or just synonyms has been the subject of debates among scholars, listeners, and the music business. Part of the community (e.g. Sharpe-Young, 2007), also cf. (Dunn, McFadyen & Wise, 2005) describes hard rock and heavy metal as two different genres that, despite of an overlap in early history, evolved in different directions, the former remaining close to its blues and British blues roots and the latter drawing more punk and hardcore punk influences. From this point of view Cream, Blue Cheer, Deep Purple and Led Zeppelin were primarily hard rock bands because of apparent presence of strong blues and psychedelic rock influences in their music. However, these bands pioneered features what later would become the key elements of heavy metal, specifically, riff[12]- based music, virtuosic guitar style, falsetto vocal style, double bass drum, more complex rhythmic patterns, usage of power chords[13] and heavily distorted amplification of the guitar sound. Such distortion produced a thicker, 'heavier' guitar sound due to resultant tones of power chords and characteristics of power amplifiers (Walser, 1993).

The opposite opinion is also present in both academic sources and musicians (Charlton, 1998; Dunn, et. al. 2005) and considers the use of the two terms interchangeable. Weinstein (2000) suggests probable political reasons for the coexistence of two different genre labels, specifically, American critics have been avoiding using term 'Heavy Metal' and instead classifying such artists as a minor component of 'Hard Rock', while in Britain term 'Heavy Metal' was widely accepted (as cited in Lilja, 2009, p.24). Esa Lilja (2009) questions the possibility of discriminating between these two by musical features.

It is difficult to find objective reasons to accept or reject either of the two perspectives. Nevertheless, in this thesis bands whose music is associated with hard rock are avoided and traditional heavy metal is considered to have started mostly from NWOBHM movement, the only exception being the band Black Sabbath. Judging from musical features, what nowadays

[12] Riff is repeated melodic figure or chord progression played repeatedly and often forming basis of composition.
[13] The chord consisting of root note and fifth or root and fourth, frequently with octave doublings

is commonly classified as traditional heavy metal is more similar to the music of Black Sabbath and later developed NWOBHM bands than the above mentioned early metal bands (though, their contribution in emerging heavy metal is undeniable).

Although the earliest forms of heavy metal were immensely influenced by blues and rhythm & blues, starting from the late 1970s, the genre underwent further evolution when bands like Saxon, Motörhead and Iron Maiden incorporated elements of then highly popular punk rock while degrading blues influence and consequently producing relatively faster and more aggressive form of heavy metal. This movement, called NWOBHM, gave 'new birth' to heavy metal whose popularity had somewhat declined in the mid 1970s, and became the foundation of several extreme subgenres emerging in the early 1980s. NWOBHM bands adopted a virtuosic approach to the music in contrast to punk's raw simplicity. They almost standardized the concept of two lead guitarists in a band, resulting in more complex arrangements and extended solo sections as well as producing heavier sound. Besides, the music started to separate from pentatonic and blues scale based riffs and melodies by extensively incorporating harmonies based on modal scales, especially Aeolian and Phrygian. Rhythmically NWOBHM music uses common time signatures and consists of 16-th or 8-th note patterns. Guitars tend to play harmonized riffs and solos. The bass often plays in unison with guitar or pedalpoint and employs rhythmic patterns that accent strong beats (to sound more 'heavy'). Keyboards are not common in NWOBHM bands.

In the late 1970s and early 1980s groups such as Mötley Crüe, Twisted Sister and Poison combined shock rock and glam rock image/lyrics with NWOBHM music and created glam metal (also referred as 'Hair Metal' and 'Pop Metal'). The sound of the glam bands was more refined and commercially oriented, drawing pop-rock influences that accounted for their media exposure in 1980s and huge commercial success. Nevertheless, glam metal was more fashion and subculture than a musically distinctive subgenre.

3.2.2 Neoclassical Metal

Neoclassical metal refers to the music of guitar virtuosos who had extensive classical training or who were otherwise influenced by classical music. However, it was not completely new phenomenon developed in metal music. Drawing classical influences in rock music appeared earlier in progressive rock and hard rock acts, notably ELP, Yes and Deep Purple. Consequently, many neoclassical metal artists name those classic bands as their primary influences.

Emerging in the 1980s early neoclassical metal was mostly instrumental music with extremely fast solo sections employing a lot of sweep picking arpeggios and classical harmonies, or even interpretations of virtuosic pieces of Baroque composers (though not restricted to this period and also involving composers of classical and romantic music). Other than employing harmonies and lead guitar phrasing from classical music, neoclassical metal is similar to NWOBHM with its distorted guitar timbre, mid to high tempo, loud drumming and strong bass. The subgenre is mostly guitar-centered, in addition, keyboards (if present) feature extended solo sections, sometimes in a context of the competitor of a guitar. Vocals are mostly high register, inspired by influential hard rock vocalists such as Ian Gillan, but unlike them more frequently deviating from pentationic scale.

Neoclassical metal somewhat declined in popularity in recent years and only several key performers maintained mainstream success. However, its heritage in terms of advanced guitar playing techniques developed by founders of the style, inspired newer generations of guitarists who successfully added this flavour to the sound of their bands.

3.2.3 Speed Metal

Speed metal is a short-lived but historically important subgenre that existed in the beginning of 1980s. Stylistically it was a mixture of NWOBHM with hardcore punk. The latter was developed in the early 1980s by bands such as Black Flag and featured more rhythm than melody oriented songwriting with mostly shouting vocals, quite fast tempos, and specific drum pattern called D-beat. Speed metal became the foundation of two distinct genres; the more aggressive thrash metal and the more NWOBHM inspired power metal.

3.2.4 Power Metal

Power Metal was innovated by NWOBHM inspired bands like Helloween, Gamma Ray and others who played fast and melodic songs with high register vocals, more melodic riffs, but at the same time borrowed rhythmical structure, fast tempo, and extensive usage of two bass drums from Speed Metal. There are two slightly different variations of power metal. One mainly originated in the US and is more influenced by speed metal in terms of more aggressive riffing (e.g. refer to Edguy's song 'Mysteria' from the album Hellfire Club, released in 2004). Unlike its aggressive counterpart, more melodic form of power metal features more NWOBHM characteristics, commonly involving keyboards. Many bands in this genre (especially from Scandinavia) are influenced by neoclassical metal as well, and sometimes even

incorporate folk in their music (e.g. compare to Stratovarius – Black Diamond from the album Visions, released in 1997). Due to its orchestral arrangements and classical flavour, this division is commonly called 'Symphonic Power'. Symphonic power often features operatic vocals, anthemic choruses, a gallop rhythmic pattern (eight-sixteen-sixteen note pattern) on a guitar and a bass or other sixteen note patterns played using palm-muted tremolo picking at high tempos (around 150-200 bpm). Another influence from neoclassical metal is extended solo sections featuring mostly guitar and keyboards.

3.2.5 Thrash Metal

From the early 1980s the metal scene has seen one of the most intensive diversification in the history of its existence. Underground scene bands, notably Metallica, Megadeth, Anthrax and Slayer (referred to as the 'big four' of thrash metal) were creating a new, previously unseen extreme form of metal by mixing the speed and aggression of hardcore punk with more complex song structure and arrangements of NWOBHM. Young musicians wanted to be faster, better players, and more aggressive than others. This tendency, resulting in speed metal, is audible in the early albums of all the bands listed above. Eventually features of their music became clearer and musically distinct from other genres, forming thrash metal, and influencing the emergence of death and black metal later.

Typical thrash metal band consists of two guitars, a bass, drums and vocals. The music features alternating tempos with extensive use of double bass drums. Double time drumming[14] patterns are also common for the genre. Guitars can be tuned down by half or one step and feature technically complex riffs using extensive palm muting[15] in rhythm section and fast solos. Harmonically thrash metal is usually based on modal scales. In thrash metal it is also common to use chromatic notes in a diatonic scale or to use chromatic scale-based riffs lacking any tonal center. Vocals are frequently shouting, - similar to hardcore punk - but singing is also (but not high register as in power metal) often employed.

It should be noted that different bands have different proportions of mixture of their most important influences. For instance Slayer's music and vocals feature more hardcore elements,

[14] A drum rhythm pattern where, in the simplest form, instead of standard beat where the bass drum accents 1st ,3rd beats and snare drum -2nd and 4th, bass drum hits every strong beat and snare hits are on weak beats.
[15] Palm muting is the technique of guitar playing where notes played are muted by placing the picking hand's palm (more precisely the side area below the pinky finger) near the bridge.

whereas later Metallica (from their third album), and Anthrax were biased towards NWOBHM with their refined compositions with melodic solos, riffs and vocals.

In the early 1990s several bands have introduced new features in thrash metal such as slower, midrange tempo and groovy rhythm section. Later, a new generation of bands made more emphasis on rhythm-oriented riffs and lower tempo, while making less focus on fast solos and the exhibition of technical proficiency. This direction was later called 'Groove Metal' or 'Post-Thrash'. Groove Metal is more an evolution of thrash metal rather than a musically separate subgenre (though, traditional or so called 'old-school' thrash is not extinct and still has its followers).

3.2.6 Death Metal

Death metal originated in the mid 1980s as a darker and more extreme form of thrash metal and was inspired by influential thrash acts such as Slayer and Kreator. It started to separate from thrash metal in the late1980s when bands like Death, Morbid Angel, and Obituary released their monumental works featuring a different, heavier or brutal sound than thrash, both in terms of overall timbre and distinctive vocal style similar to that of Chris Barnes (early Cannibal Corpse) and Glen Benton (Deicide).

As a progeny of thrash, early death metal shared many musical features with its ancestor, but eventually, partly due to developments in audio technology, became more extreme and brutal. Almost exclusively death metal band consists of guitars, bass, drums, and vocals. The music is more focused on atonal riffs and triton interval than thrash metal and features abrupt tempo and time signature changes both to achieve an 'evil' sound and to deceive the listener's expectations. Pentatonicism is extremely rare in this subgenre. Guitar solos are not much accentuated usually, but riffs are more intricate and technically demanding to perform. Drummers use double bass drums and as a rule employ extremely fast drum patterns including blast beats[16]. Guitars commonly use low tuning (such as C tuning[17] or lower) and similarly to thrash metal feature fast riffs with extensive usage of palm muting and tremolo picking. The song structure is most frequently different from the standard verse-chorus framework employing sudden key changes, extensively using chromatic scale, diminished seventh arpeggio, and atonal riffs. Probably the most recognizable feature of death metal for the naïve listener is distinctive deep growling vocals - death growls, due to its harsh or 'brutal' nature. When asked

[16] Blast beats are drum pattern featuring bass drum, snare drum and high-hat on eighth note subdivision or alternatively one of the three following sixteenth note subdivision
[17] Guitar is tuned two full steps below standard tuning.

what defined death metal, Paul Ryan, member of one of the prominent band of the genre Origin replied:

"I would say in the beginning of death metal it was unpronounced growling vocals. Then came the blast beats." – Paul Ryan (as cited in Purcell, 2003, p.11)

3.2.7 Melodic Death Metal

Melodic death is a blend of NWOBHM inspired melodic riff-based music and elements of death metal music. In 1991 the Swedish band Dismember released an important album defining Swedish death sound (Purcell, 2003). In contrast to death metal this subgenre sometimes features keyboards in addition to two guitars, a bass, drums, and vocals. Riffs are based more on modal scales, while dissonant intervals in general are less emphasized. Furthermore, melodic death metal songs commonly have a less complex structure and it is not unusual either to include sections or whole tracks featuring acoustic guitar, or clean singing along with growling. From death metal it owes tempo changes, blast beats, growling vocals, and overall timbre.

As in every genre influenced by more than one style of music, the proportions of audible characteristics of its two parent genres in melodic death vary from band to band. This can easily be seen by comparing two albums from influential bands of the genre; the more death metal influenced Slaughter of the Soul (released in 1995) by At The Gates and the more NWOBHM inspired Jester Race (released in 1996) by In Flames. In general, especially modern melodic death is more accessible to a wider audience and is commercially far more successful than death metal.

3.2.8 Traditional Doom Metal

The archetype of this subgenre would be Black Sabbath's self-titled song with its slow tempo and heavy sound combined with pessimistic and grievous themed vocals. The song creates a ponderous atmosphere from which the name of subgenre originated. In the early 1980s several bands such as Saint Virus, The Obsessed and Candlesmass adopted this template and created an absolute opposite of the speed and aggression fashionable at that time. This subgenre usually features slow tempo and long, epic song structures with melancholic lyrics. Although being overshadowed by thrash and glam metal, doom metal survived and even diversified in several directions. Traditional refers to the subgenre created by its pioneers. Musically it is most recognizable by the following features: slow tempo (frequently around 60 bpm), and heavily distorted monotonous riffs mainly in minor key, emphasizing the melody.

Vocals are clean and melancholic vocals. Lyrics are commonly about depression, grief and fear, but also cover such themes as mythology, fantasy or battle. The latter is often referred to as epic doom.

3.2.9 Doom-Death Metal

Doom-Death Metal is practically an extinct subgenre which emerged in the early 1990s and defined acts such as Paradise Lost, My Dying Bride and Anathema. Music from their early albums blends traditional doom music, female vocals, but also deep growls, dissonant riffs, sections with relatively faster tempos borrowed from death metal. Doom-Death in its original form remained underground and declined by the end of 1990s, though, heavily influenced two subgenres, specifically funeral doom and gothic metal.

Funeral Doom takes atmosphere of traditional doom to the extremes. It incorporates both clean and deep growling vocals, frequently unpronounced and with a lot of added reverberation to sound in the background. Commonly, the music features extremely slow (around 40 bpm), long monotonic tracks and ambient keyboards to create dark, funeral atmosphere.

3.2.10 Sludge

Sludge Metal is an underground subgenre which emerged in the 1990s. At that time a new wave of then popular grunge and hardcore punk inspired bands started creating the unique sound but still retained the template of doom metal. Notable musical features include more ‘dirty’ timbre on guitar than typical doom sound, achieved by downtuning guitar, adding a lot of distortion and playing on extremely loud levels on an amplifier. Moreover, feedback among the guitarists of the genre is frequent. Almost universally the songs include sections with fast tempo with double time drumming and shouting vocals adopted from hardcore punk. Notable bands of the genre would be Crowbar, Eyehategod and Down.

3.2.11 Drone

Drone Metal is another underground subgenre originated in the US in the early 1990s. It consists of drone, ambient and noise music elements mixed with a distorted guitar sound and growling or screaming vocals and gloomy atmosphere characterizing to doom metal. Most notable artists of the genre include Earth and Sun o))). Drone metal consists of long (usually more than 10 minutes), mostly instrumental compositions with slow (around 40 bpm) repetitive riffs performed on downtuned and heavily distorted guitars and bass; almost invariable feature

of the music is drones created by low frequency noise. Typical instrumentation consists of a guitar, a bass, drums, and keyboards.

3.2.12 Progressive

Progressive rock extended the compositional complexity and the technical proficiency of rock music by incorporating improvisational approach of jazz and musical structures of classical music in its usually long instrumental sections. Along with progressive rock artists creating concept albums of long songs with several parts (like movements in classical symphonic music), other collectives featured early hard rock and heavy metal elements in their music (e.g. Rush, King Crimson). Inspired by this outlook, in the mid 1980s, heavy metal artists created progressive metal. Similarly to its parent genre, progressive metal music features complex songs both compositionally (long compositions sometimes consisting of few different sections, key changes, complex structures), and rhythmically (frequently employing odd time signatures and polyrhythm). Concept albums are frequently practiced by progressive bands. Another notable characteristic is improvisation when performing live and extended solo sections involving guitar as well as other instruments. The most notable progressive rock influenced artists in the genre would be Dream Theater and Queencrÿche.

In the early 1990s artists from other metal subgenres with virtuosic playing skills and high level musical proficiency started to create various fusions of progressive metal. This made the subgenre one of the most musically diverse of metal subgenres. For instance, Symphony X is power/progressive metal with easily noticeable classical influences (e.g. refer to the album V: the new Mythology Suite released in 2000) while another band Opeth fuses death and heavy metal with progressive elements (e.g. refer to the album Ghost Reveries released in 2006). Tool incorporates alternative and industrial music influences in its still progressive oriented music. Clear jazz music influence is audible in Animals As Leaders' music. In the end of the sonic spectrum is technical metal that takes musical proficiency to extreme levels with its intricate, multilayer rhythmic patterns and elaborated harmonies which require virtuosic skills both to perform and to compose. Excellent representatives of technical metal would be Spiral Architect and Atheist and Meshuggah.

3.2.13 Industrial

Industrial Metal is an umbrella term that refers to merging an experimental approach and production techniques into different metal subgenres. The origin of the term is linked to Industrial Records label created by one of the most prominent industrial artist Throbbing Gristle in the mid 1970s. Early Industrial music artists were experimenting with various noises, tape loops and electronic instruments as well as more traditional instruments.

Industrial metal emerged in the late 1980s when industrial music inspired artists like Ministry, Godflesh and KMFDM started employing distorted guitar sound and heavy metal riffs. Later in the 1990s a new generation of bands such as Rammstein merged electronic music elements into the genre. Due to its versatile nature it is hard to list musical features unifying the entire subgenre. Nevertheless, most frequently the artists of this subgenre employ samples of various non-musical sounds, distorted vocals, synthesizers and rhythmically simple repeating riffs on guitar with straightforward drumming which is accenting strong beats to create 'industrial effect'. Mostly they use common time signatures and consistent tempo.

3.2.14 Avant-garde

Avant-garde or experimental metal is another broad subgenre unifying all artists experimenting with and incorporating any kind of nonconventional elements in music. The subgenre emerged in the 1980s and Celtic Frost is credited to be one of the pioneers of the subgenre. An ideology of avant-garde metal is similar to that of an industrial, as both experiment with non-standard instruments and non-musical sounds, but avant-garde artists push boundaries further by eliminating any limitations in the process of creating music and experimenting with non-standard song structures, chord progressions, or even vocals (singing techniques). In addition, avant-garde artists tend to emphasize abrupt transitions between song sections both rhythmically and harmonically, or, even insert stylistically unrelated segments. Notable artists of the genre would be Atrox, Ephel Duath, and Unexpect.

3.2.15 Metalcore

Metalcore is a fusion genre referring to the mixture of various (mostly thrash, death and melodic death) metal subgenres with hardcore punk. Most notable feature characterizing to all of metalcore subdivisions is the big emphasis on breakdowns – a feature borrowed from hardcore punk. Breakdown in metal refers to the typically slowed down section of a song

(frequently slowed to a half tempo) featuring rhythmically oriented guitar riffs mostly on open lowest strings to achieve heaviest sound, and simple drum patterns to further emphasize heaviness.

The origins of the genre lie in the music of bands such as Integrity and Strife who incorporated elements of thrash metal into hardcore punk. Apart from thrash, several other subgenres were fused with hardcore during the 1990s and created musically diverse directions, such as:

Deathcore - a blend of death and hardcore. Commonly deathcore involves fast tempo, dissonant riffs, blast beat drumming and growling from modern death metal, but features mostly less complex song structures and is less technically complex as well. Another distinctive feature is breakdowns and screaming vocals that are almost universally employed in this subgenre.

Melodic Metalcore - melodic death and hardcore. Pioneered by the mid and late-1990s bands such as Killswitch Engage, Bullet For My Valentine and All That Remains, who imported NWOBHM influences in the genre and created commercially successful subgenre. The border between this subgenre and modern melodic death is quite hard to set. Few distinctive features are more hardcore punk influenced shouting vocals instead of the growling typical to melodic death. Furthermore, melodic metalcore far more frequently involves clean vocals and melodic harmonized guitar solos compared to melodic death music.

Grindcore – mixture of death, industrial, and hardcore punk. One of the pioneers of the genre was the band Napalm Death, whose influential album Scum (released in 1987) featured extremely short songs with incredibly fast tempos and deep grunted vocals, features that became template for the genre.

3.2.16 Black Metal

The label ‘Black Metal’ appeared in the 1980s and referred to the music of bands such as Venom, Bathory and Hellhammer featuring satanic image and lyrics. However, the music, nowadays referred to as the First Wave of Black Metal, was rather close to speed metal of its time in terms of musical features. It was only in the early 1990s that Norwegian bands such as Emperor, Mayhem, Burzum and Darkthrone developed musically distinct form of the genre (called second wave of black metal). Since then the genre further diversified into more commercially oriented form featuring synthesizers by bands such as Dimmu Borgir and Cradle Of Filth. Among synthesized instruments the harpsichord, violin, organ, and choir are most common that give the music an

orchestral feel or a cathedral-like setting. Also occasionally female singing is incorporated. They are generally placed under the Symphonic Black metal label. Other derivations of back metal scene include more experimental approach of bands like Ulver or Arcturus, or influences from ambient music such as in Burzum. Black metal can be typified by its high-pitched 'shrieking' vocals (noticeably different from death growls), extremely fast tempos, blastbeat and D-beat drum patterns. Guitars usually are in standard tuning, riffs are based on modal scales, though, chromatic scales and dissonant intervals are also actively practiced. Guitar riffs usually employ sixteenth note patterns, played using tremolo picking technique. Solos are rarity in the genre.

3.2.17 Gothic Metal

Gothic metal is a diverse genre which refers to a blend of dark, depressive atmosphere of gothic rock, created by multilayer textures of synthesizers, and metal music. Three bands are credited as pioneers of the genre: Anathema, Paradise Lost, and My Dying Bride (e.g. Sharpe-Young, 2007). All three started their careers as doom/doom death bands, doom metal inspired depressive riffs and aggressive growling vocals, but eventually started searching the ways to create darker atmosphere by actively experimenting with keyboards, violins, and even adding female vocals. One of the results of such experiments was so called 'Beauty and the beast' vocals that referred to the duet of clean 'angelic' female vocals and beast like growling usually performed by male. This approach became a standard feature of the genre in the mid 1990s. Symphonic elements practiced by Paradise Lost were adopted by the next generation of the bands such as Tristania, Theatre of Tragedy which founded symphonic gothic branch. Nowadays, the active members of symphonic gothic scene tend to bias towards more polished, commercially oriented music, while almost completely removing growling vocals and other extreme metal influences.

3.2.18 NU-Metal

Nu-Metal is another fusion genre combining the elements of relatively modern popular genres outside of metal such as electronic music, hip hop, post punk, grunge and funk. Emerged in the mid 1990s, Nu-metal was the revival of metal into mainstream, which had been overshadowed by Grunge in the early 1990s. However, some authors argue that the roots of the genre can be traced back in the mid or late 1980s when first attempts to merge rapping with metal music were made (Mclver, 2002, p10). As metal diversifies, it gradually complicates to describe any fusion of it with other genres. Indeed, just listening to several influential and commercially successful acts of the genre, such as Korn, Limp Bizkit, Linkin Park and Slipknot, and suggesting clearly defined com-

mon musical description would be enough to serve as a good argument for this claim. All of the listed bands have different backgrounds and sources of influences. Few recognizable elements associated with NU metal distinguishing from other metal subgenres are incorporation of rapping, samplers and sequencers, and sometimes DJ-s with turntables. Considering its versatility, most probably this genre will not last long as a whole and will quickly divide into at least several new branches.

4. SYSTEM DESCRIPTION

4.1 Overview

In this chapter the methodology and the algorithms used for automatic genre classification are described. The overall structure of the system is depicted in Figure 2. The first stage was pre-processing where audio files were converted into wav format from mp3, down-sampled from 44100Hz to 32000Hz sampling rate and channels were summed to mono. A one minute segment was extracted from each track starting from the middle point. If the duration was less than two minutes, the first minute was selected. Next, the dataset was split into two sets, of which one - the training set - was used for training a classification algorithm and the other - the test set - for evaluation. From both sets commonly used features were extracted. At the next stage a smaller subset of features was selected using feature selection algorithms to reduce the dimensionality of the data and remove irrelevant and redundant features. As shown in Figure 2, feature selection was performed using WEKA[18] data mining software (Hall et al., 2009) and only the training set was employed in this process. The obtained subset of features was used for training and finally the test set was used for evaluating the classifier. Three classification algorithms were tested on the dataset, one of which was implemented in this thesis and is based on the algorithm published in (Barbedo & Lopes, 2007). Other classifiers were selected from a wide range of machine learning algorithms offered by WEKA. In the following sections each stage of the system, including learning algorithms employed, are described in more detail.

[18] WEKA is open source software issued under the GNU General public license. It features a wide variety of machine learning algorithms as well as data transformation and visualization tools which can be accessed via a graphical user interface or command line. More information can be found on its website: http://www.cs.waikato.ac.nz/ml/weka/index.html

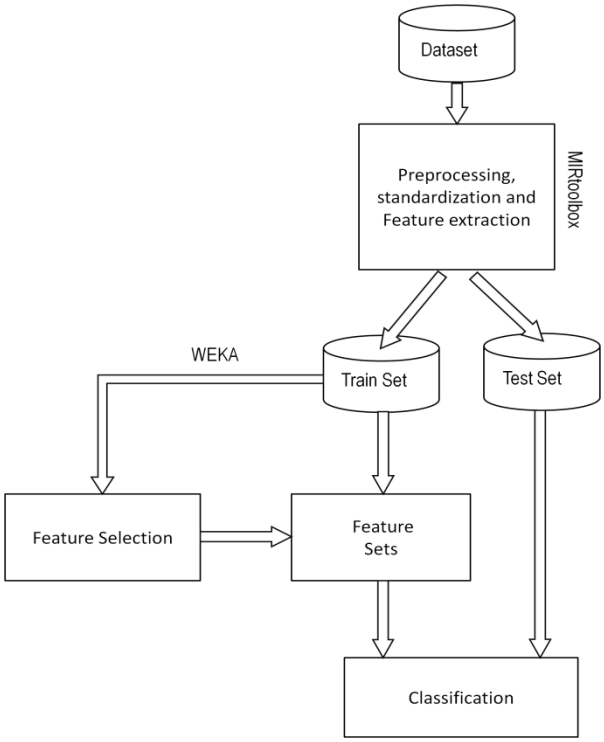

Figure 2 Structure of the system

4.2 Dataset

Since no study has yet been published on automatic genre classification of metal, it was impossible to find any previously used dataset to test the classifier and directly compare the results. Most of the datasets commonly used in genre classification studies (e.g. the above mentioned GTZAN, ISMIR, and Magnatune[19] datasets) are not suitable for our purposes due to being too general in terms of the employed taxonomy. Therefore, an entirely new dataset was collected for this work. It would be forbiddingly expensive to collect the audio material needed to represent all the subgenres provided in chapter 3. Therefore, initially a dataset consisting of 833 tracks representing 17 subgenres was collected.

When designing such dataset, naturally one could keep in mind what features are available for separating the genres and to arrange the dataset with those genres which it will be possible to separate using employed features and achieve a reasonable classification result. It is also possible to include only tracks that clearly exhibit distinctive musical features of the respective classes. However, such a sterile dataset will not be generalizable on real world data, which commonly features (often clearly audible) influences from more than one genre. On the other

[19] www.magnatune.com – Independent record label which kindly allows using its audio database for research purposes.

hand, selecting just any genres and tracks for the dataset can make it impossible to achieve any reasonable results. The initial priority was that the dataset should represent real life data. Therefore almost all popular genres were included in the dataset, but tracks were selected to display strongly the main musical features of their respective classes. The genres in the initial dataset were:

Avant-garde, black, death, doom, drone, gothic, grindcore, traditional heavy metal, industrial, melodic black, melodic death, metalcore, NU-metal, power, progressive, sludge, and thrash metal.

Each class was represented by 49 tracks. The whole set was divided into a training set and a test set, 595 tracks (35 tracks per class) for training and 238 (14 track per class) for testing the classifier. In the dataset most of the artists were represented by several tracks (between 2-5 tracks per artist from different albums, with few exceptions), but it was ensured that the training and the test set contained completely different sets of artists.

Experiments on this fairly complex dataset showed that musical features employed in this work were not sufficient to achieve any reasonable classification results. After a few unsuccessful experiments, one option was to add some more complex features, but expectations that adding a few more features would significantly improve separation between such interwoven genres were quite low. Moreover, the feature extraction and training phases were rather time consuming, therefore it would be more practical to continue experiments on a smaller and also somewhat 'simpler' dataset.

Finally, a subset of 210 tracks representing seven genres including *black, death, melodic death, gothic, heavy, power and progressive metal* was extracted from the original dataset. Each genre was represented by the 30 most relevant tracks from the initial dataset, this time using a 'one track per artist' strategy. The dataset was split into half, leaving 105 tracks for both the training set and the test set. Reasons for selecting the particular set of seven genres were the following:

- These seven genres are fairly popular in the metal community and feature more or less clear distinctive characteristics.
- So called 'umbrella' genres such as avant-garde, industrial, metalcore and NU-metal were excluded from the final dataset, since they feature very diverse musical characteristics and caused most of the misclassifications in the initial experiments.
- It was difficult to collect a sufficient amount of tracks for less well-known or relatively new subgenres such as drone.

4.3 Feature extraction

After organizing and pre-processing the dataset, the set of musical features representing timbral, rhythmic, and dynamic and pitch information of music was extracted from the audio using the MIR toolbox[20] (Latrillot & Toiviainen, 2007). The MIR toolbox is an integrated set of functions written in the MATLAB computing environment, which offers intuitive syntax for extracting a wide range of musical features as well as performing statistical analysis. A set of 35 spectral and temporal descriptors that have previously been employed in genre classification tasks was selected from the available features in the toolbox. Table 4 provides a list of these features and a short description of each. All the features except RMS were extracted from 25ms half-overlapping frames. For RMS longer, 50 ms frames were used. Extracted frame-level features - except Low energy rate and Pulse clarity - were summarized over the whole one minute segment by computing six statistics over all frames including mean, standard deviation (SD), slope, periodicity frequency (PF), periodicity amplitude (PA), and periodicity entropy (PE). Overall, a set of $33 \times 6 + 2 = 200$ dimensional feature vectors were produced, each vector representing one track from training set. This may be the first attempt to use periodicity frequency, periodicity amplitude, and periodicity entropy in a genre classification task. These summarization methods describe periodicities in the time series of frame-level features, more specifically:

- Periodicity amplitude is estimated by calculating the autocorrelation of the feature values along frames and finding the maximal amplitude of the obtained autocorrelation function (obviously zero lag is excluded). This feature indicates the strength of the periodicity in a feature sequence.
- Periodicity frequency is the frequency of the time lag corresponding to the maximum of the autocorrelation.
- Periodicity entropy is obtained by dividing the Shannon entropy of the autocorrelation function *p(x)* by the length of the sequence:

$$H(X) = -\frac{\sum_{i=1}^{n} p(x_i)\, log\, p(x_i)}{log\ (n)}$$

[20] Information about the MIR toolbox, as well as a download link and documentation can be found at the following website: https://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/mirtoolbox

Where x_i, $i=1:n$ is the sequence of feature values. This feature describes the shape of the autocorrelation function curve, one extreme being a flat curve corresponding to the maximal entropy of 1 and on the other extreme - only one predominant peak on the flat background corresponding to the minimal entropy.

Feature	*Description*	*Summarization method*
Spectral Rolloff	*The frequency below which 85% of total energy of the spectrum is contained*	Mean, SD, Slope, PF, PA, PE
Spectral Brightness	*Measures the amount of energy above 1500Hz*	Mean, SD, Slope, PF, PA, PE
Spectral Roughness	*Is estimated by computing the peaks of the spectrum and taking the average of all dissonances between all possible pairs of peaks*	Mean, SD, Slope, PF, PA, PE
MFCC (13 coefficients)	*See table 2 for description. In this work first order differences, or delta MFCC-s are used*	Mean, SD, Slope, PF, PA, PE
Spectral Flux	*The distance between the spectra of successive frames*	Mean, SD, Slope, PF, PA, PE
Regularity	*Measures the degree of variation of the successive peaks of the spectrum*	Mean, SD, Slope, PF, PA, PE
Centroid	*Estimates the geometric center of the spectral distribution*	Mean, SD, Slope, PF, PA, PE
RMS	*Root Mean Square energy of the signal*	Mean, SD, Slope, PF, PA, PE
Low energy	*Mean of the low energy value over all frames*	*Mean*
Pulse clarity	*The strength of the main beat. Estimates rhythmic clarity of the track. Calculated by finding the maximum correlation value from the autocorrelation function of the onset detection curve.* (Latrillot,Eerola,Toiviainen & Fornari, 2008)	*Mean*
Chromagram (12 pitch classes)	*Shows the distribution of spectral energy along the pitch classes.*	Mean, SD, Slope, PF, PA, PE
Zero-crossing rate	*Counts the amount of sign changes of the signal*	Mean, SD, Slope, PF, PA, PE

Table 4 Features and their descriptions (some features were already described in Table 2)

Extracted features were converted into z-scores by the formula:

$$z = \frac{x-\mu}{\sigma}$$

Where x is an input value of a feature in given class to be converted, μ and σ are mean and *SD* of the feature over entire class, respectively.

4.4 Feature selection

It was already mentioned that having redundant or irrelevant features in the feature set will degrade classifier performance. Therefore, the optimal subset of attributes was selected that at least had no less predictive power than the initial set and did not include redundant information. This stage is one of the most important in the classification process.

Various feature selection methods and corresponding algorithms are available for this purpose. In this thesis search-based feature selection algorithms are employed. The concept of the majority of search-based feature selectors is that the algorithm searches through the feature space to find the subset that most likely will predict the class best. Several decisions relating to the search process in this class of algorithms have to be made before starting the actual process. The issue was addressed in (Langley, 1994).

The first decision concerns the starting point in the feature space and the search direction. There are three available options to select: to begin with no features and successively add attributes, known as forward selection; to start with all features and successively remove them, called backward elimination; and a combination of both, to begin somewhere in the middle and go in both directions, known as bi-directional selection.

The second decision concerns search organization. The most trivial search strategy, an exhaustive search, considers all possible subsets of attribute, which is computationally expensive and impractical for large feature sets. A more efficient approach, commonly referred to as 'greedy stepwise', considers local changes, i.e. addition, deletion, or both of a single attribute to the current subset, depending on whether forward selection, backward elimination or a bidirectional search strategy is employed. When a local change improves the merit of the current subset the algorithm selects it and iterates. Another option is to consider all possible local changes and select the best. In both cases the selected changes are not reconsidered later.

The third decision concerns the evaluation strategy. Two common strategies, both employed in this thesis, are filter and wrapper selection. The former evaluates merit and filters out features based on the heuristics defined by the specific model used for selection. Filter algorithms operate independently from the classification algorithm, and rely solely on general characteristics of the training set. Therefore, the filter approach ignores the effects of the selected feature subset on the performance of the induction algorithm, which is its main disadvantage. The advantage of this approach is that it is much faster than wrapper algorithms when dealing with large datasets (Hall M. A., 1999).

Wrapper selection, in contrast, is based on the idea that the selected feature subset will ultimately have high predictive accuracy if the feature selection algorithm will take the biases of the target classification algorithm into account. Wrapper methods generate a feature set and evaluate it by running a classifier on training data and checking the accuracy achieved, which will allow the resulting subset of features to be adjusted to the peculiarities of the classifier. Usually N-fold cross-validation is used for evaluating the performance of the feature subset generated (John, Kohavi, & Pfleger, 1994). N-fold cross-validation consists of splitting the dataset into N partitions in a stratified manner. On each run, an N-1 subset is used for training the algorithm and the remaining single partition is used for testing. Final accuracy is the mean of the results of all N runs. Wrapper selectors repeatedly call the classifier which makes them quite slow on large datasets. (Hall M. A., 1999)

Finally, the fourth decision is the stopping criterion, which defines heuristics for the feature selector to end the searching process. One rule is to stop searching when adding or removing attributes does not improve the merit of the current subset. Another option is to continue adding features as long as merit does not degrade, and finally a more extreme variation is to search the whole feature space and select the best subset.

In this thesis both filter and wrapper selection algorithms were utilized for selecting an optimal attribute subset. For wrapper selection, KNN and decision trees (which will be described later) were used as an evaluator.

The correlation-based feature selection (CFS) algorithm proposed in (Hall M. A., 1999) was selected from the filter selection category. CSF is a filter algorithm that tries to find a subset of attributes which are highly correlated with the class but have the lowest possible inter-correlations. This is achieved by the correlation-based evaluation function:

$$M_S = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}$$

Where M_S represents heuristic 'merit' of a feature subset S containing k features, $\overline{r_{cf}}$ is the mean feature-class correlation and $\overline{r_{ff}}$ is the average feature-feature correlation. The nominator in this equation estimates the predictive power of the feature set and the denominator estimates the degree of redundancy among the features. This function will ignore (or, produce low rank for) irrelevant features, because they will have low correlation with the class (a small value in the numerator). The same is true for redundant features as they will have a high inter-correlation (a large value in the denominator).

The starting point, search organization, and stopping criterion were set similarly for both algorithms; a greedy stepwise algorithm was chosen with a forward selection routine for both CFS and wrapper selection. Feature subsets were evaluated using 5-fold cross-validation. The search ended when adding a new feature did not improve the merit of the subset. The results of feature selection are provided in section 5.2.

4.5 Classification

As mentioned above, in this thesis three learning algorithms were employed to classify the dataset. One of the classifiers was implemented in this thesis and the other two were selected from wide range of available learning algorithms in WEKA. The following sections provide descriptions of the three algorithms.

4.5.1 The classifier implemented in this study

The classifier implemented in this thesis is based on the work of Barbedo & Lopes (2007). The authors performed classification on a four layer genre hierarchy consisting of four broad genres at the top level and 29 subgenres at the lowest, achieving 87% and 61% accuracy for highest and lowest layers respectively. They employed a bottom-up approach, i.e. classification was performed on the lowest layer of the hierarchy, and broader genres were defined by the hierarchy itself. For example, a piece classified as ‘Samba’ belongs to the ‘Latin’ category, which itself is a subcategory of ‘Percussion’ and so on.

The training process of the algorithm consists of selecting six reference vectors for every possible pairs of the genres, three vectors per genre, which resulted in the best separation between the two genres. During the selection all possible six-vector combinations are considered from all potential vectors within each pair of genres. Each of these combinations is used for the classification training data in the pair of genres using a nearest neighbour strategy; the combination achieving the highest accuracy is selected as the reference vector for this pair of genres. Next, the entire process is iterated over all possible pairs of genres. During the training process the algorithm selects reference vectors for every possible pair of the genre.

The classification procedure is partly depicted in Figure 3 using four hypothetical genres: A,B,C,D. The table in the figure represents reference vectors for all possible pairs of genres; each genre in the pair is represented by three reference vectors.

1. The test vector - representing a one-second analysis segment in (Barbedo & Lopes, 2007) - is compared to all the reference vectors and the closest ones

are selected (denoted by shaded boxes) as the local winners for each genre pair.

2. After finishing the voting process, the local victories for each genre are counted and the one with the most victories is selected as the genre of the segment.
3. All the segments are classified and finally the genre of the test track is defined similarly as in the step 2.

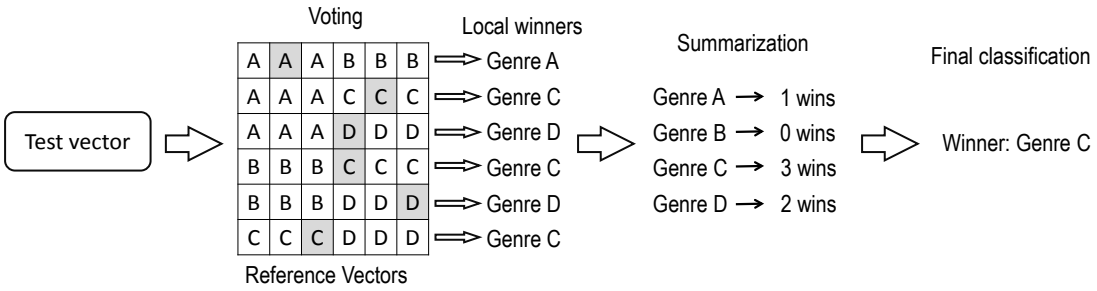

Figure 3 Schematic view of the classification process. The test vector is compared to six reference vectors for each of the possible combinations between A, B, C, D genres. Closest reference vectors are denoted by shaded boxes.

The described strategy was implemented in Matlab. However, it is not an exact implementation of the algorithm and differences are present in the pre-processing step as well as in the implementation. The first difference is in the feature summarization method and feature sets. In this thesis extracted features are summarized over the whole one-minute segment, whereas in (Barbedo & Lopes, 2007) features were summarized over one-second analysis segments and therefore, each test sample consisted of 32 segments.

The training process is also slightly different. For easier illustration, the training stage of implemented algorithm will be described on the example of initial training set consisting of 17 classes where each class contains 35 training vectors.

For each of 136 possible pairs of genres one reference six-vector combination should be selected containing three training vectors from one genre and the rest from another. The reference vector selection starts by producing all possible six-vector combinations for each pair of classes, provided that each genre of the pair contributes with three vectors. One straightfor-

ward way to produce such six-vector combinations is to generate all possible three-vector combinations for each class and then take all possible combinations of obtained sets. However, this will produce a massive $\left(\frac{35!}{3!(35-3)!}\right)^2 = 42\,837\,045$ potential reference vectors for each of the 136 possible pairs. In this implementation, instead of all possible six-vector combinations, potential reference vectors were obtained by generating all possible three-vector combinations in both classes and merging them together. This reduced the number of potential vectors to $\left(\frac{35!}{3!(35-3)!}\right) = 6545$ per pair of genre.

Let us denote the array of all generated potential reference vectors for a given pair of genres with R_i where i=1,2,3...6545. Each R_i consists of $r_a, r_b, r_c, r_d, r_e, r_f$; $a, b, c = 1,2,3, \dots ,35$; $d, e, f = 36,37, \dots ,70$ combinations; T_k, k= 1,2,...,70 denotes the training vectors in a given pair of genres. To select a reference vector, all T_k vectors are classified by each of R_i potential vectors (Figure 4). This process starts with the calculation of distances between all T_k and each r vector from one of the R_i combinations. In each case when the closest T_k to r belongs to the same class as r, the current R_i combination receives one vote. In the Figure 4 a comparison of one potential reference vector R_i to the first two training vectors (T_1 and T_2) is represented. As shown in the figure, R_i gets only one vote from comparison with two training vectors T_1 and T_2, because r_c (closest vector to T_1), belongs to the same A class as T_1, whereas r_f and T_2 belong to different classes. In other words, the given R_i predicted class correctly for only one of the two training vectors. The described voting process is repeated for all 6545 potential vectors

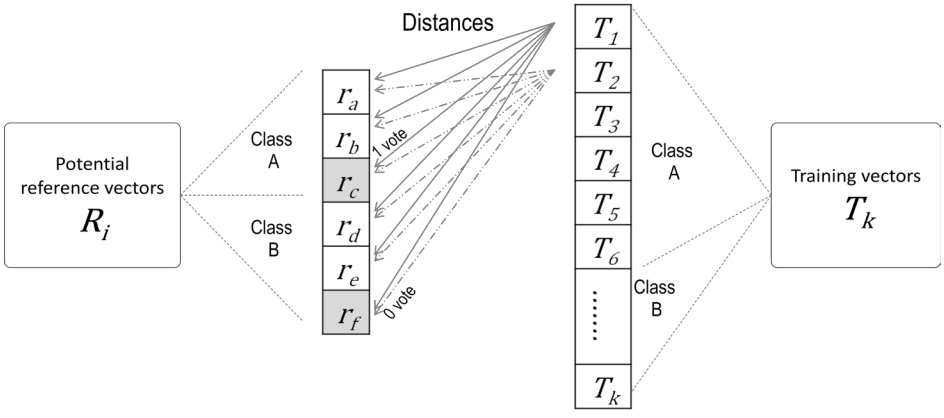

Figure 4 Diagram of the voting process.

and the R_i collecting the most votes, i.e. achieving the best separation between the two classes, is selected as reference vector for a given pair.

The classification procedure is similar to the original algorithm (Figure 3). The only difference is that here the input test sample is not segmented and therefore the third step in the above described classification process is not necessary.

4.5.2 AdaBoost

AdaBoost.M1 is an iterative algorithm that in each iteration calls a predefined classifier to classify the training data. Initially, all instances are weighted equally but after each iteration, the weights for correctly classified instances are reweighted by the formula:

$$W \times \frac{e}{1-e}$$

Where W denotes the weight assigned to the previous step and e is classification error, which is the summed weights of misclassified instances divided by the total weight of all instances. Weights for incorrectly classified instances remain unchanged. The produced weights are subsequently normalized and as an outcome of this process weights are decreased for correctly classified instances and increased for misclassified ones. It should be noted that according to the weighting formula, if a classification error $e = 0$ (perfect classification) then all weights become equal to zero, the algorithm deletes the last result and the boosting process terminates. Similarly the process stops if $e \geq 0.5$; because in this case W, for correctly classified instances, will increase instead of decreasing. The latter condition is equivalent to the requirement that in each iteration a classifier should achieve more than 50% classification accuracy. After several iterations defined by the user, the output of the classifiers is combined using a weighted vote. Weights for classifiers in AdaBoost.M1 are calculated by

$$weight = -\log\frac{e}{1-e}$$

Finally, for a given instance all the predicted classes during iterations are considered and the weights of the classifiers voting for each class are summed and the class with the highest total is selected.

In this thesis the C 4.5 decision tree classifier[21] was selected as a weak learner. In this classifier the learning model is represented as a tree-like structure where each node is a deci-

[21] Hereafter this classifier will be referred as J48, which is Weka implementation name for C4.5.

sion rule testing one attribute, usually by comparing its value to some constant. The leaves of the tree represent classes and therefore any instance reaching a leaf is classified in a corresponding class. A decision tree is built in iterative manner continuously adding nodes until all remaining instances belong to the same class, in which case the current node becomes a leaf and no sub-tree is produced further.

4.5.3 K-Nearest Neighbours

K-NN is one of the simplest classification techniques. It has virtually no learning stage and the classification procedure simply consists of measuring pairwise distances between the input test vector and all training vectors, finding K nearest neighbours to the test vector and choosing the class that the majority of these neighbours belong to. The metric used in this thesis for a distance measurement is Euclidean distance. This classifier, despite being simple, is frequently employed and performs well in various MIR tasks (e.g. Li et al., 2003; Pohle, Pampalk, & Widmer, 2005; Park et al., 2005; Lee et al., 2009)

In spite of having virtually no learning stage, K-NN is not a computationally efficient classifier because it iteratively measures pairwise distances, which becomes tedious for large datasets. However, for such cases algorithms exist for reducing the computational complexity of this classifier (Duda, Hart, & Stork, 2001). In this thesis the dataset was fairly small and K-NN performed sufficiently fast.

5. RESULTS AND CONCLUSIONS

5.1 Overview

In this chapter the results of the three classification schemes employed are presented in three experiments. Each classifier was tested using two feature sets generated by filter and wrapper feature selection strategies. The reason for testing several classification strategies was the difficulty of evaluating the classification results, since no data is available to compare the results obtained in this study. Setting up an experiment on human participants to directly compare accuracy achieved by the algorithm to the listener's ability to classify such data was beyond the scale of this study. Consequently, performance of the implemented algorithm was compared to the results of other more commonly used classifiers in this field. Experiment results are reported in sections 5.3.1 – 5.3.3, discussion and conclusions are provided in the section 5.4.

5.2 Feature sets

As mentioned in the preceding chapter, feature selection was done in WEKA software. Two different feature selection methods, Wrapper and CFS, were employed to select an optimal subset from the initial 33 six dimensional + 2 one dimensional descriptors.

First the CFS algorithm ('CfsSubsetEval') with default settings was applied to the training set. The search method was Greedy stepwise with forward searching strategy. Hereafter, if not otherwise mentioned, all the parameters in the employed algorithms were set to default.

Next, two feature subsets were selected using Wrapper selector ('WrapperSubsetEval') using two different classifiers: decision tree (J48) and K-NN. Similarly to CFS, a greedy stepwise forward searching method was used. Both runs were evaluated via 5-fold cross-validation. Table 5 shows the output of the three selection algorithms. Mostly all three algorithms selected spectral features, among which MFCC, Chromagram, Rolloff, Brightness and Zero-crossing rate were shared by at least two of them. The absence of rhythmic features in the final sets was surprising since it was initially hypothesized that these features would be useful to reliably differentiate between some subgenres.

Correlation based feature selection (CFS)	Wrapper with J48 (W-J48)	Wrapper with KNN (W-KNN)
Spectral Rolloff	*Spectral Rolloff*	*Spectral Rolloff*
Brightness	*Chromagram*	*Brightness*
Centroid	*Mfcc*	*Mfcc*
Mfcc	*RMS*	*Chromagram*
Zerocrossing rate	*Flux*	
Chromagram	*Zerocrossing rate*	
	Regularity	

Table 5 Subsets of features selected by three feature selection algorithms.

5.3 Classification

5.3.1 Classification using implemented algorithm

In this experiment an algorithm similar to the one proposed in (Barbedo & Lopes, 2007) was tested on the subgenre classification of metal music. As described in section 4.5.1, the training stage of this algorithm consists of selecting reference vectors using the nearest neighbour scheme, which is sensitive to irrelevant and redundant features. In their work Barbedo and Lopes (2007) used the following procedure to select a subset of features: each feature in the initial set was ranked according to its ability to differentiate between classes. Next, the classification algorithm was run recursively using the entire initial feature set in the beginning, eliminating one feature with the lowest rank per iteration, until only two features were left. According to the results of the classifications recorded after each iteration, they selected a set of 4 optimal features.

Although the method was successful (as reported by authors) it would be extremely time-consuming to use it for the initial set of 200 descriptors. Instead, a similar strategy was applied to CFS and W-KNN feature sets. First, all 200 descriptors were ranked by feature selection algorithms based on the merit each feature provided (setting ‘generate ranking’ to ‘true’ in WEKA outputs all the features ordered by their respective ranks). For each generated feature set (see table 5), at first the one feature with the lowest rank at a time was excluded until only two features were left. After that, up to five features were added to the same set using a similar scheme, but this time the ones with the highest ranks. Finally, the feature set resulting in the highest classification accuracy was selected.

Output of the algorithm using two optimal parametrization is provided in Table 6 and Table 7 Achieved classification accuracy was 37.14% for CFS subset and 34.29% for W-KNN subset.

	Black	Death	Goth	Heavy	Prog.	Power	MDeath	Recall	Precision	F_measure
Black	**10**	2	0	1	2	0	0	0.667	0.5	0.5714
Death	3	**7**	0	1	2	1	1	0.467	0.4667	0.4667
Goth	1	0	**2**	4	3	4	1	0.133	0.3333	0.1905
Heavy	2	0	0	**8**	1	3	1	0.5333	0.3478	0.4211
Prog.	1	4	1	2	**6**	0	1	0.4	0.3158	0.3529
Power	2	1	1	5	1	**2**	3	0.133	0.1818	0.1538
MDeath	1	1	2	2	4	1	**4**	0.2667	0.3636	0.3077

Table 6. Confusion matrix and detailed accuracy by genre for CFS subset.

	Black	Death	Goth	Heavy	Prog.	Power	MDeath	Recall	Precision	F_measure
Black	**9**	1	0	1	2	2	0	0.6	0.8182	0.6923
Death	0	**10**	0	1	3	1	0	0.6667	0.625	0.6452
Goth	0	0	**5**	2	4	1	3	0.3333	0.625	0.4348
Heavy	1	0	3	**3**	0	6	2	0.2	0.1429	0.1667
Prog.	1	3	0	2	**5**	2	2	0.3333	0.2381	0.125
Power	0	1	0	7	3	**2**	2	0.1333	0.1176	0.1379
MDeath	0	1	0	5	4	3	**2**	0.1333	0.1818	0.1538

Table 7. Confusion matrix and detailed output for W-KNN subset.

5.3.2 Classification using K-NN

In this experiment the K-NN[22] learning algorithm was tested on the CFS and W-KNN feature sets. To select an optimal number of nearest neighbours, the experiment was run several times for each feature set. Similarly, for wrapper selection a 2-NN learning algorithm was used after some experiments with a varying number of the nearest neighbours. In the following the results produced by the optimal combinations are provided.

For the CFS feature set the highest result, 42.8%, was achieved by the 2-nearest neighbours classifier. The output of the algorithm is provided in Table 8. As expected W-KNN feature set achieved a slightly higher 44.8% of correctly classified samples compared to the CFS set. Table 9 provides the output of the algorithm

[22] In all cases K-NN algorithm is used from Weka with default parameters except the number of nearest neighbours.

	Black	Death	Goth	Heavy	Prog.	Power	MDeath	Recall	Precision	F_measure
Black	**11**	1	1	0	0	0	2	0.733	0.688	0.71
Death	3	**9**	0	2	0	0	1	0.6	0.5	0.545
Goth	0	1	**10**	3	1	0	0	0.667	0.385	0.488
Heavy	1	2	7	**4**	1	0	0	0.267	0.1429	0.25
Prog.	0	2	3	2	**6**	1	1	0.4	0.235	0.387
Power	0	1	4	2	3	**3**	2	0.2	0.75	0.316
MDeath	1	2	1	4	5	0	**2**	0.133	0.25	0.174

Table 8 Output of K-NN on CFS feature set.

	Black	Death	Goth	Heavy	Prog.	Power	MDeath	Recall	Precision	F_measure
Black	**8**	2	1	1	0	0	3	0.533	0.8	0.64
Death	1	**8**	0	2	1	1	2	0.533	0.5	0.516
Goth	0	0	**8**	5	2	0	0	0.533	0.571	0.552
Heavy	0	1	2	**7**	3	1	1	0.467	0.269	0.341
Prog.	0	3	0	6	**3**	3	0	0.2	0.25	0.222
Power	0	1	1	4	3	**4**	2	0.267	0.25	0.258
MDeath	1	1	2	1	0	7	**3**	0.2	0.273	0.231

Table 9 Output of K-NN on W-KNN feature set

5.3.3 Classification using AdaBoost

The aim of this experiment was to test the performance of AdaBoost.M1 algorithm with the J48 classifier on the dataset. Only the number of iterations was varied, otherwise all parameters were set to default. Again, the experiment was run two times using CFS and W-J48 feature sets. For the CFS set, 45.7% was achieved for 30 iterations, which is the highest result in this thesis. For W-J48 set, AdaBoost.M1 with 25 iterations classified 42.8% of the test samples correctly. Detailed output of the two experiments is provided in Table 10 and Table 11.

	Black	Death	Goth	Heavy	Prog.	Power	MDeath	Recall	Precision	F_measure
Black	**12**	1	1	0	0	0	1	0.8	0.667	0.727
Death	4	**9**	0	0	0	1	1	0.6	0.529	0.563
Goth	0	1	**5**	6	2	1	0	0.333	0.455	0.385
Heavy	0	4	1	**6**	3	1	0	0.4	0.333	0.364
Prog.	1	0	2	3	**5**	4	0	0.333	0.357	0.345
Power	0	1	1	1	2	**7**	3	0.467	0.389	0.424
MDeath	1	1	1	2	2	4	**4**	0.267	0.444	0.333

Table 10 Output of AdaBoost.M1 on CFS feature set.

	Black	Death	Goth	Heavy	Prog.	Power	MDeath	Recall	Precision	F_measure
Black	**11**	2	0	0	0	0	2	0.733	0.786	0.759
Death	1	**10**	1	0	1	0	2	0.667	0.476	0.556
Goth	0	0	**7**	5	2	0	1	0.467	0.5	0.483
Heavy	2	3	0	**5**	1	2	2	0.333	0.263	0.294
Prog.	0	0	3	2	**4**	3	3	0.267	0.444	0.333
Power	0	3	1	3	0	**4**	4	0.267	0.4	0.32
MDeath	0	3	2	4	1	1	**4**	0.267	0.222	0.242

Table 11 Output of AdaBoost.M1 on W-J48 feature set

5.4 Conclusions

In this thesis three learning algorithms were employed to classify subgenres of metal music. Two different parametrizations were produced by Filter and Wrapper feature selection methods. The results of the classification are summarised in Table 12. All of the results, though not quite high, are reasonably higher than classification by chance (14.28%). Moreover, it is unclear what can be considered a 'good' result for the given problem. It is unknown how accurate an average listener would be in the same task.

	CFS feature set	Wrapper + J48	Wrapper + K-NN
Implemented algorithm	37.1%		34.29%
AdaBoost+J48	45.7%	42.8%	
K-NN	42.8%		44.8%

Table 12 Classification accuracies achieved by different classifiers.

Detailed accuracy by genre is presented in Figure 5. In Figure 6 recall rates are averaged over all experiments and 95% confidence intervals are presented. As shown in the figures black and death metal were the most correctly predicted genres. This was expected because the pair is fairly different in terms of overall timbre from the rest of the genres. Some of the misclassifications are easy to explain. For example, the difference between power and heavy metal or power and progressive metal is mostly in tempo, drum patterns, time signatures and keys, while all of them feature similar instrumentation, vocal style and extended solo sections (often involving keyboard). These differences are difficult to capture using the employed parametrization without higher level rhythmic or harmonic modelling. Melodic death metal was one of the worst predicted genres, mostly misclassified as power, heavy and progressive metal. This can be explained by the fact that the genre is influenced heavily by NWOBHM and power metal, whose features apparently were exhibited in the dataset to a greater extent than death metal features, such as inharmonicity introduced by growling vocals and dissonant riffs.

However, some misclassifications are more difficult to interpret and are in a way unacceptable. As mentioned in chapter two, the quality of the classification result is at least not less important than standard output, such as recall and overall correct classification rate.

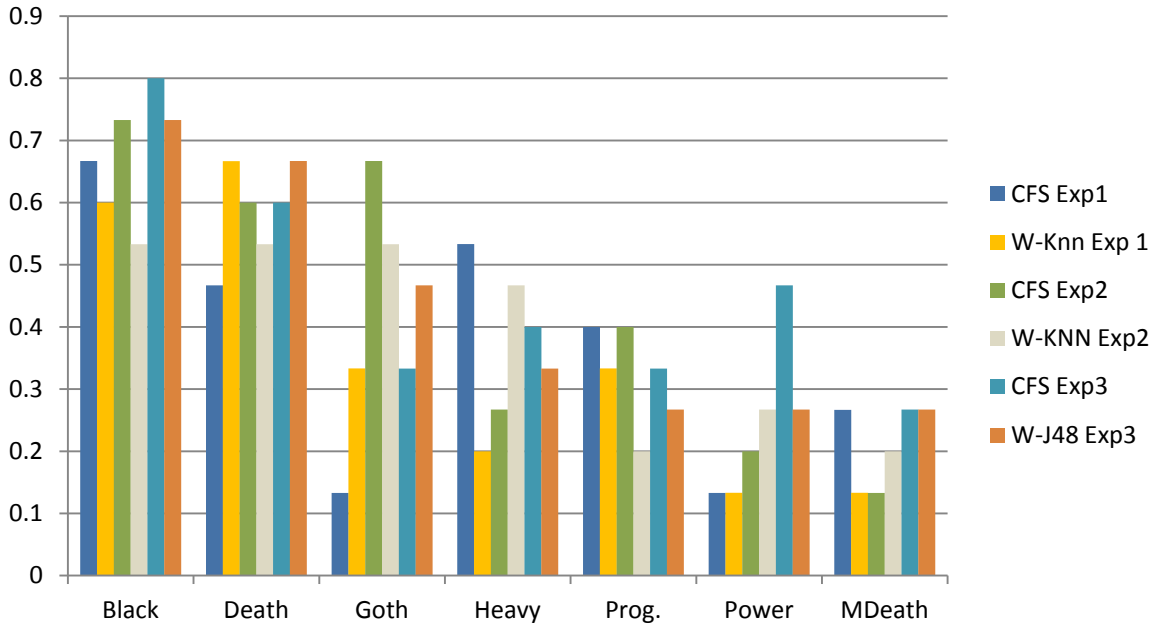

Figure 5 Recall rates for each genre

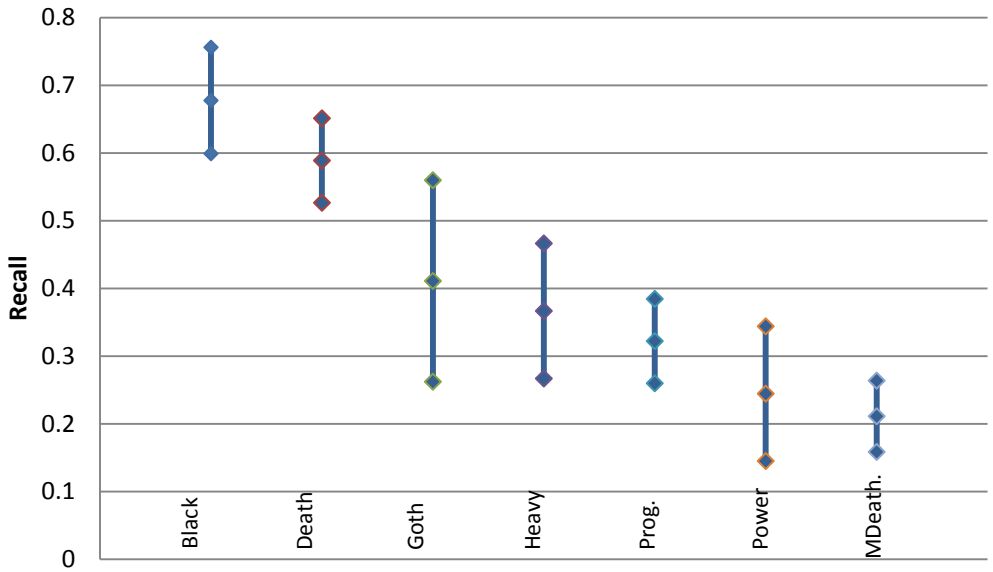

Figure 6 Average recall by genre with 95% confidence intervals

One simple way to characterize the quality of the classification result is to calculate what portion of all classification errors produced is acceptable. In this way the classification quality will be in the range of 0 (all the errors were unacceptable) and 1 (all the errors were acceptable). It is a challenging and subjective task to decide which errors are acceptable or unacceptable, but still possible to some degree. For instance, classifying a heavy metal sample as death or black metal is unacceptable since these genres are quite different both by musical features and overall timbre. On the other hand, erroneously assigning the same heavy metal track to the power metal class can have some explanation, since these genres share some musical features as well as overall timbre. In this thesis a simple acceptability criteria were employed for the dataset with binary classification of errors (i.e. the errors are either acceptable or unacceptable). Table 13 depicts acceptability criteria where acceptable errors are denoted by white cells and unacceptable errors are denoted by grey shading. However, more sophisticated model with weighted acceptability ratings can be created for more complex genre hierarchy involving broader genres.

Comparison of the classification quality produced by the employed learning algorithms (Figure 7) showed that the quality of classification does not explicitly depend on the classification accuracy and can be used as an additional parameter for characterizing the output of a classifier. To illustrate, KNN classifier achieved the highest classification quality, although AdaBoost produced the highest classification accuracy. Figure 8 shows an average classification quality for each genre. Power and heavy metal were expected to induce higher results since they have wider range of acceptable errors than the rest of the genres in the dataset,

however, not all the three classifiers followed such pattern (see Figure 8). Finally, the experiment results confirmed the expectation that both a learning algorithm and parametrization affect the classification quality. However, a larger dataset and more experiments are needed to find out the character of the relationship.

	Black	Death	Goth	Heavy	Prog.	Power	MDeath
Black							
Death							
Goth							
Heavy							
Prog.							
Power							
MDeath							

Table 13 Acceptability criteria. Unacceptable errors are denoted by grey shaded boxes.

One of the aims of this thesis was to test if the concept proposed in (Barbedo & Lopes, 2007), would be more effective for subgenre classification task than other successfully used classifiers in this field. It should be noted that the algorithm is not an exact implementation (see section 4.5.1), but shares the main concept of genre-dependent reference vector selection and classification method with the original. Experiments showed that the implementation produced the lowest result compared to two commonly used classifiers, namely AdaBoost.M1 and K-NN (see table 13), though, McNemar test showed that the performance gain produced by the two classifiers over the implemented algorithm on the same CFS features was not statistically significant (P=0.488, and P=0.302 for K-NN and AdaBoost respectively).

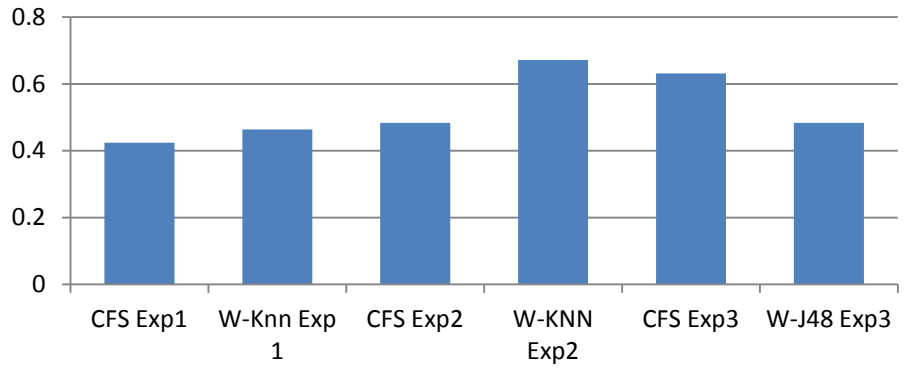

Figure 7 Overall classification quality for each experiment. Exp1 corresponds to the implemented classifier, Exp2 – to K-NN, and Exp3 – to AdaBoost.

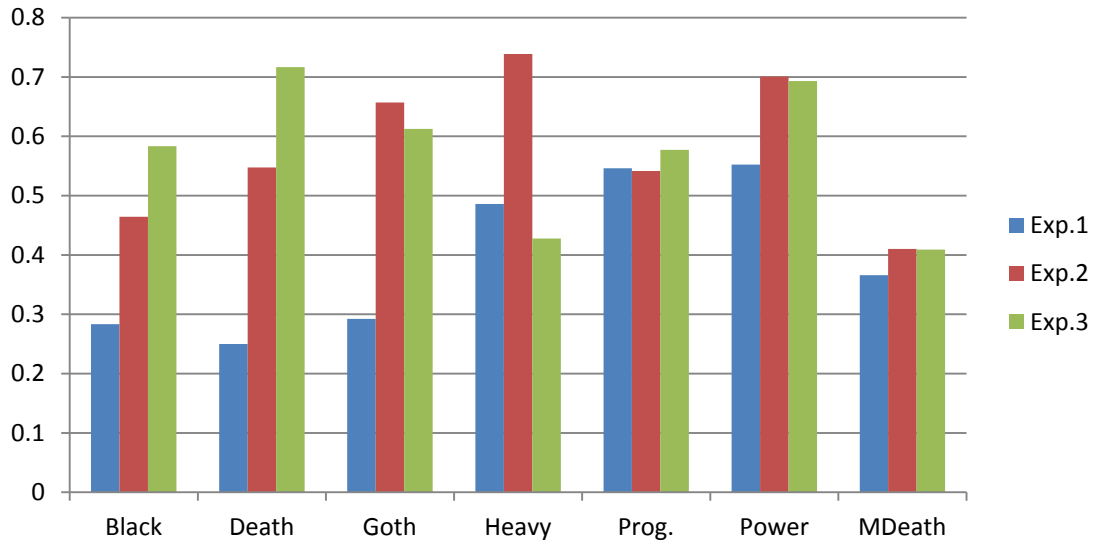

Figure 8 Average classification quality for each genre. Exp1 corresponds to the implemented classifier, Exp2 – to K-NN, and Exp3 – to AdaBoost

It is possible that the original algorithm would achieve a higher accuracy due to employing analysis segments rather than summarizing features over the whole audio sample. Such an approach captures more detailed dynamics of features over the whole piece, which can provide useful information about song structure. Moreover, segmentation is robust to 'outlier' sections in songs (e.g. acoustic guitar part in death metal song), whereas simple statistical summarization is sensitive to such tracks. It should be noted that the summarization method employed in this thesis that captures periodicity in feature sequences was also quite useful. Unfortunately the original algorithm was not available for testing on the same data/feature set; therefore it is hard to tell how classification accuracy was affected by implementation differences only. Still, the aim was to employ a classification strategy rather than an exact algorithm.

Considering the fact that results of different learning algorithms were not significantly different, it would be at least reasonable to propose that it is possibly more or less close to the maximal result that can be achieved by the employed parametrization on a given dataset. Naturally not all the descriptors currently used in the genre classification tasks were tested and there is some room for improvement by adding new features. Nevertheless, credibly most of the existing features will not provide dramatic improvements especially for a larger dataset involving more genres, since higher level features are necessary to detect subtle differences between several subgenres, for example, higher level rhythmic modelling, drum pattern analysis, or even vocal style, whether it is shouting, growling or screaming provides much information about genre. In addition, analysis of guitar riffs for the amount of dissonant intervals present, scales used, and rhythmic patterns would help to separate many genres

reliably, but to achieve this first source separation should be performed, which is in itself a challenging problem.

The need of higher level features was especially clear from the initial experiments on the dataset of 17 subgenres. Feature distributions were quite complex and interwoven making it impossible for the tested classifiers to model. Consequently, even the highest results (about 8-10%) were only slightly above the chance level (about 6%). Besides, the dataset itself had problems as it included several songs from one artist (though not the same artists in train and test sets), which could affect the reference vector selection process and cause overfitting. Nevertheless, due to the lack of results[23] in this field, and the large amount of time needed for the feature selection/training process, it was more practical to perform experiments on a smaller dataset.

[23] The only study about subgenre classification was a master's thesis about subgenre classification of electronic music (Kirss, 2007), but due to the musical differences between the genres, it was impossible to use it for solving problems arising in this study.

6. FUTURE PROSPECTS

I believe that the results of the automatic subgenre classification of metal music achieved in this thesis are not the top limit that cannot be overcome. There is a big room for improvements by fine-tuning the implemented algorithm, better parametrization, and summarisation. First of all, the feature set can be improved by adding more features capturing rhythmic information. Next, a more robust method for extracting a representative one-minute excerpt can be used. The simple method used in this thesis (i.e. extracting a one-minute excerpt from the middle point of a song) showed that some genres have quite steady song structures and the middle point matches to the same section of the song. For example, for most of the power metal songs the extracted section consisted of, or included guitar solo, which is not as helpful for representing the genre as a section with vocals. This is especially true if most of the features represent overall timbre of the song. Alternatively, more than one minute excerpt could be used (even might be necessary) to involve several sections of the song. The combination of longer sections with feature summarization over an analysis segment would effectively capture long term dynamics of the song.

Another option that would be especially effective for a subgenre setting is to give the possibility to the system to predict more than one genre for one song. Coupled with the segment classification strategy, as proposed in (Barbedo & Lopes, 2007), would make it easier to reflect the different influences that songs usually incorporate, resulting in more sensible errors and therefore a higher classification quality. Such an approach would also alleviate the 'ground truth' problem for a dataset as well as make it easier to classify music of fusion genres.

For ground truth extraction social network data can be used quite effectively. In fact, it was attempted to use Last.fm genre tags as ground truth when arranging the dataset in this thesis by ranking genre tags applied to a specific track depending on how many users labelled the track using given word. However, the idea was later dismissed because extracting the number of votes for tags was impossible from API (2009, personal communication with staff).

Finally, as a matter of fact the automatic classification will have the highest quality when errors made by machine will be identical by nature to the errors made by human. To achieve this, as suggested in (McKay & Fujinaga, 2006), the existing musicological and psychological knowledgebase explaining underlying processes of human music classification should be more actively incorporated in the form of high level features even outside of musical content.

BIBLIOGRAPHY

Allmusic. (n.d.). Retrieved August 24, 2009, from http://www.allmusic.com/explore/style/scandinavian-metal-d11954

Allmusic . (n.d.). Retrieved March 9, 2009, from http://www.allmusic.com/explore/style/power-metal-d11959

Anglade, A., Ramirez, R., & Dixon, S. (2009). Genre Classification Using Harmony Rules Induced from Automatic Chord Transcription. *10th International Society for Music Information Retrieval Conference (ISMIR 2009)*, (pp. 669-674).

Aucouturier, J. J., & Patchet, F. (2004). Improving timbre similarity: How high is the sky? *Journal of Negative Results in Speech and Audio Sciences* (1(1)).

Aucouturier, J., & Pachet, F. (2003). Representing Musical Genre: A State of the Art. *Journal of New Music Research , 32*, 83-93.

Barbedo, J. G., & Lopes, A. (2007). Automatic Genre Classification of Musical Signals. *EURASIP Journal on Advances in Signal Processing , Vol. 2007, Article ID 64960*, 1-12.

Bergsta, J., Casagrande, N., Erhan, D., Eck, D., & Kegl, B. (2006). Aggregate features and AdaBoost for music classification. *Machine Learning , 65*, 473-484.

Burred, J. J., & Lerch, A. (2003). A Hierarchical Aprroach to Automatic Musical Genre Classification. *Proceedings of the 6-th international conference on Digital Audio Effects (DAFx-03)*, (pp. 1-4). London.

Charlton, K. (1998). *Rock Music Styles A History.* McGraw-Hill.

Chen, L., Wright, P., & Nejdl, W. (2009). Improving Music Genre Classification Using Collaborative Tagging Data. *ACM International Conference on Web Search and Data Mining* , (pp. 84-93). Barcelona.

Downie, J. S. (2008). The Music Information Retrieval Evaluation Exchange (2005-2007): A window into music information retrieval research. *Acoustical Science and Technology , 29* (4), 247-255.

Duda, R., Hart, P., & Stork, D. G. (2001). *Pattern Classification* (2nd Edition ed.). New York: John Wiley&Sons INC.

Dunn, S., McFadyen, S., & Wise, J. (Directors). (2005). *Metal: A Headbanger's Journey* [Motion Picture].

Foote, J., & Uchihashi, S. (2001). The Beat Spectrum: A new Approach to Rhythm Analysis. *Proc. International Conference on Multimedia and Expo (ICME)*, (pp. 1088-1091).

Gjerdingen, R. O., & Perrott, D. (2008). Scanning the Dial: The Rapid Recognition of Music Genres. *Journal of New Music Research , 37* (2), 93-100.

Hall, M. A. (1999). *Correlation-based Feature Selection for Machine Learning.* The University of Waikato, Hamilton, NewZealand.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & I.H, W. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations , 11* (1), pp. 10-18.

Heittola, T. (2003). Automatic Classification of Music Signals. *MSc Thesis* . Tampere University Of Technology.

Jang, D., Jin, M., & Yoo, C. D. (2008). Music genre classification using novel features and a weighted voting method. *IEEE international conference on multi medi a and expo ICME* , (pp. 1377-1380).

John, G. H., Kohavi, R., & Pfleger, P. (1994). Irrelevant features and the subset selection problem. *Machine Learning: Proceedings of the Eleventh International Conference.* Morgan Kaufmann.

Kahn-Harris, K. (2007). *Extreme Metal: Music and Culture on the Edge.* Berg.

Kirss, P. (2007). *Audio Based Genre Classification of Electronic Music.* Master's Thesis, Universitu of Jyväskylä, Jyväskylä.

Lamere, P., & Pampalk, E. (2008). Social Tags and Music Information Retrieval. *International Society of Music Information Retrieval (Tutorials Session).*

Langley, P. (1994). Selection of relevant features in machine learning. *Proceedings of the AAAI Fall Symposium on Relevance.* AAAI Press.

Latrillot, O., & Toiviainen, P. (2007). MIR in Matlab: A toolbox for musical feature extraction. *Proceedings of the International Conference on Music Information Retrieval.* Wien.

Latrillot, O., Eerola, T., Toiviainen, P., & Fornari, J. (2008). Multi-feature modeling of pulse clarity: Design, validation, and optimization. *International Conference of Music Information Retrieval.* Philadelphia.

Lee, C.-H., Shih, J.-L., Yu, K.-M., & Lin, H.-S. (2009). Automatic Music Genre Classification Based on Modulation Spectral Analysis of Spectral and Cepstral Features. *IEEE Transactions on Multimedia , 11* (4), 670-682.

Li, T., & Ogihara, M. (2006). Towards Intelligent Music Information Retrieval. *IEEE Transactions on Multimdia , 8* (3), 564-574.

Li, T., Ogihara, M., & Li, Q. (2003). A Comparative Study on Content-Based Music Genre Classification. *26th Annual international ACM SIGIR Conference on Research and Development in Information Retrieval*, (pp. 282-289). Toronto.

Lidy, T., Rauber, A., Pertusa, A., & Iñesta, J. (2007). *Improving Genre Classification by Combination of Audio and Symbolic Descriptors Using a Transcription System.* Austrian Computer Society (OCG).

Lilja, E. (2004, May). Characteristics of Heavy Metal Chord Structures Their Acoustic and Modal Construction. Their Relation to Modal and Tonal Context. Licentiate Thesis. University of Helsinki.

Lilja, E. (2009). *Theory and Analysis of Classic Heavy Metal Harmony.* Helsinki: IAML Finland.

Mandel, M., Michael, I. M., & Ellis, D. (2005). Song-Level Features And Support Vector Machines For Music Classification. *International Conference on Music Information Retrieval ISMIR 2005 Proceedings* , (pp. 594-599). London.

Marsicano, D. (n.d.). Retrieved March 3, 2009, from About.com: http://heavymetal.about.com/od/heavymetal101/p/progressivemeta.htm

McKay, C., & Fujinaga, I. (2006). *Musical Genre Classification: Is it Worth Pursuing and How Can it be Improved?* McGill University, Music Technology. Quebec Canada: University of Victoria.

McKinney, M., & Breebaart, J. (2003). Features for Audio and Music Classification. *Proceedings of the International Symposium on Music Information Retrieval*, (pp. 151-158). Washington.

Mclver, J. (2002). *Nu-Metal: The Next Generation of Rock & Punk.* London: Omnibus Press.

Meng, A., Ahrendt, P., & Larsen, J. (2005). Improving Music Genre Classification by Short-time Feature Integration. *IEEE ICASSP*, (pp. 497-500).

Pachet, F., & Cazaly, D. (2000). A Taxonomy of Musical Genres. *Content-Based Information Access Conference (RIAO).* Paris.

Pampalk, E., Flexer, A., & Widmer, G. (2005). *Improvement of Audio-Based Music Similarity and Genre Classification.* Qeen Mary, University of London.

Park, K.-S., Oh, S.-H., Yoon, W.-J., & Lee, K.-K. (2005). A Robust Approach to Content-Based Musical Genre Classification and Retrieval Using Multi-feature Clustering. In M. Maher, *Advances in Computer Science - ASIAN 2004. Higher-Level Decision Making* (pp. 212-222). Springer Berlin / Heidelberg.

Peeters, G. (2004). *A Large Set of Audio Features for Sound Description (Similarity and Classification in the CUIDADO Project.*

Pohle, T. (2005). *Extraction of Audio Descriptors and Their Evaluation in Music Classification Tasks.* .

Pohle, T., Pampalk, E., & Widmer, G. (2005). Evaluation of Frequently Used Audio Features for Classification of Music into Perceptual Categories. *Proceedings of the Fourth International Workshop on Content-Based Multimedia Indexing.*

ProgArchives. (n.d.). Retrieved March 3, 2009, from ProgArchives: http://www.progarchives.com/Progressive-rock.asp

Purcell, N. J. (2003). *Death Metal Music: The Passion and Politics of a Subculture.* McFarland&Company,Inc. USA.

Russ, B. (n.d.). Retrieved April 5, 2009, from BNR Metal pages: http://www.bnrmetal.com/v2/genre.php?ID=M

Russ, B. (n.d.). *Bnr Metal Pages*. Retrieved June 28, 2009, from http://www.bnrmetal.com/v2/genre.php?ID=O

Scaringella, N., Zoia, G., & Mlynek, D. (2006, March). Automatic genre classification of music content: a survey. *Signal Processing Magazine, IEEE , 23* (2), pp. 133-141.

Sharpe-Young, G. (2007). *Metal The Definitive Guide.* Jawbone Press.

Slone, J. (2008, July 12). Retrieved July 2009, from Avantgarde metal: http://www.avantgarde-metal.com/content/stories2.php?id=67

Stetina, T., & Burton, T. (1991). *Speed and Thrash Guitar Method.* Milwaukee: Hal Leonard Publishing Corporation.

Tzanetakis, G. (2002). *Manipulation, Analysis and Retrieval Systems for Audio Signals.* Doctoral Dissertation.

Tzanetakis, G., & Cook, P. (2002). Automatic Genre Classification. *IEEE Transactions on Speech and Audio Processing , 10* (5), 293-302.

Walser, R. (1993). *Running With the Devil: Power, Gender and Madness in Heavy Metal Music.* Wesleyan.

West, K. (2008). *Novel Techniques For Audio Music Classification and Search.* PhD Thesis, School of Computer Sciences, University of East Anglia.

West, K., & Cox, S. (2004). Features and Classifiers for the Automatic Classification of Musical Audio Signals. *Conference on Music information Retrieval (ISMIR).*

West, K., & Cox, S. (2005). Finding an Optimal Segmentation for Audio Genre Classification. *ISMIR 2005, 6th International Conference on Music Information*, (pp. 680-685). London.

Whitman, B., & Smaragdis, P. (2002). Combining Musical and Cultural Features for Intelligent Style Detection. *Proceedings of the International Conference on Music Information Retrieval ISMIR 2002*, (pp. 47-52).

Wikipedia, The Free Encyclopedia. (n.d.). Retrieved June 11, 2009, from http://en.wikipedia.org/wiki/ Heavy_metal_music

Xu, C., Maddage, N. C., Shao, X., Cao, F., & Tian, Q. (2003). Musical genre classification using support vector machines. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, (pp. 429-432).

APPENDIX

Track list of the dataset (Artist Name_Album name_track name)

BLACK METAL

Aeternus_Burning the Shroud_Burning the Shroud
Ajattara_Kuolema_Kituvan Kiitos
Astrofaes_The Attraction_Heavens and Earth I am Immersed to Mistery
Black Horizons_Suicide Symphonies_The Battles Of The Godless Souls
Dimmu Borgir_Enthrone Darkness Triumph_Master Of Disharmony
Emperor_IX Equilibrium_Sworn
Mysticum_In the Streams of Inferno_Wintermass
Nazgul_Awaiting the Battle Ravens_Awaiting the Battle Ravens
Satyricon_The Shadowthrone_The King Of The Shadowthrone
Setherial_Endtime Divine_Crimson Manifestation
Taake_Nekro_Voldtekt
The Black_The Priest Of Satan_Black Blood
Tsjuder_Demonic Possession_Deathwish
Ulver_Aatte Hymne Til Ulven I Manden_Wolf and Passion
War_Total War_Satan
ABYSS_Summon The Beast_The Arrival
Bathory_The Return of Datkness and Evil_Son Of The Damned
Carpathian Forest_Morbid Fascination Of Death_Warlord Of Misanthropy
Dark Funeral_Vobiscum Satanas_The Black Winged Horde
Darkthrone_Under A Funeral Moon_Inn I De Dype Skogens Fabn
Gaahlskagg_Erotic Funeral_I Am Sin
Gloomy Grim_Life__Revelation 666
Gorgoroth_Ad Majorem Sathanas Gloriam_Untamed Forces
Graveland_The Celtic Winter_Hordes Of Empire
Immortal_Sons Of Northern Darkness_One By One
Impaled Nazarene_Ugra-Karma_Sadhu Satana
Marduk_Panzer Division Marduk_502
Mayhem_Chimera_Slaughter Of Dreams
Nargaroth_Black Metal Is Krieg_Far Beyond The Stars
Ragnarok _Blackdoor Miracle_Journey From Life

DEATH METAL

Kataklysm_Serenity In Fire_Blood On The Swans
Lust Of Decay_Purity Through Dismemberment_When Anesthesia Fails
Man Must Die_...Start Killing_War on creation
Misery Index_Discordia_Breathing Pestilence
Monstrosity_Rise to power_Wave of Annihilation
Morbid Angel_Gateways To Annihilation_To The Victor The Spoils
Necrophagist_Onset Of Putrefaction_To Breathe In A Casket
Nile_Ithyphallic_What May Safely Be Written
Origin_Echoes of Decimation_The Burner
Proteus_Personal Narrative Of Cognitive Dreams-capes_Reptilian Matrix
Pungent Stench_Ampeauty_Invisible Empire
Sceptic_Pathetic Being_Only Lies
Spawn Of Possession_Cabinet_Swarm Of The Formless
Vader_Blood_We wait
Vital Remains_Dechristianize_Rush Of Deliverance
Aborted_Goremageddon_Clinical Colostomy
Aeon_Bleeding the False_Morbid Desire To Burn
Atheist_Piece of Time_Beyond
Atrocious Abnormality_Echoes Of The Rotting_The Birth Of Violence
Brain Drill_Apocalyptic Feasting_Apocalyptic Feasting
Cannibal Corpse_Gore Obsessed_Pit Of Zombies
Death_The Sounds Of Perseverance_Spirit Crusher
Gojira_From Mars To Sirius_Backbone
Goreinhaled_Brutal Death Metal Compilation Vol.3-2009_Taste A Putrid Amputation
Gortuary_Manic Thoughts of Perverse Mutilation_Skull Fragments
Grave_Into the Grave_In Love
Guttural Secrete_ Brutal Death Metal Compilation Vol.3-2009_Razorized Ball Gag
Immolation_Unholy Cult_A Kingdom Divided
Macabre_Gloom_Cremator
Malevolent Creation_Warkult [Bonus Track]_Captured

GOTHIC METAL

Darkseed_Astral Adventures_Dying Land
Draconian_Arcane Rain Fell_Daylight Misery
For My Pain._Metal Museum Vol.5 Gothic Metal_Queen Misery
Leave's Eyes_Elegy_A Winter Poem (Non Album Track)
Morgart_Metal Museum Vol.4 Symphonic Met-al_Sinfonie 5 - In A-Dur
Mortal Love_I Have Lost_Adoration
Sentenced_Buried Alive_Brief is the Light
The Gathering_How To Measure A Planet_My Electrici-ty
Theatre Of Tragedy_Aegis_Venus
Therion_Secret Of The Runes_Midgard
Tiamat_Judas Christ_I Am In Love With Myself
Tristania_Beyond The Veil_A Sequel Of Decay
Tystnaden_Metal Museum Vol.5 Gothic Metal_Hamlet
Within Temptation_The Heart Of Everything_Hand Of Sorrow
Xandria_Kill The Sun_Casablanca
Artrosis_Metal Museum Vol.5 Gothic Metal_Nazguls
Crematory_Awake_Lords Of Lies
Entwine_Gone_Thru the Darkness
Epica_The Phantom Agony_Sensorium
Evereve_Metal Museum Vol.5 Gothic Metal_Dies Irae (Grave New World)

Lacrimas Profundere_Metal Museum Vol.5 Gothic Metal_Sarah Lou
Lacuna Coil_Metal Museum Vol.5 Gothic Metal_Swamped
Macbeth_Malae Artes_Lifelong Hope
Mandrake_Metal Museum Vol.5 Gothic Metal_The Necklace
Moonspell_Darkness and Hope_Devilred
Silentium_Metal Museum Vol.5 Gothic Metal_Serpentized
Sirenia_An Elixir For Existence_A Mental Symphony
Sirenia_Nine destinies and a downfall_The Other Side
Sunterra_Metal Museum Vol.5 Gothic Metal_Veil Of Darkness
Trail Of Tears_Profoundemonium_Driven Through The Ruins

TRADITIONAL HEAVY METAL

Astral Doors_Metal Museum Vol.7 Heavy Metal_From Satan With Love
Black Axe_Metal Museum Vol.9 NWOBHM_Red Lights
Blitzkrieg_Metal Museum Vol.9 NWOBHM_Blitzkrieg
Cloven Hoof_Metal Museum Vol.9 NWOBHM_Laying Down The Law
Danzig_Metal Museum Vol.7 Heavy Metal_Do You Wear The Mark
Gaskin_Metal Museum Vol.9 NWOBHM_I'm No Fool
Heaven & Hell_The Devil You Know_08 Follow The Tears
Jaguar_Metal Museum Vol.9 NWOBHM_Back Street Woman
No Vacancy_School Of Rock_Fight
Ozzy Osbourne___Perry Mason
Savatage_Metal Museum Vol.7 Heavy Metal_Damien
Shakra_Infected_Higher Love
Son Of A Bitch_Metal Museum Vol.7 Heavy Metal_18 - Son Of A Bitch - Victim You
U.D.O._Metal Museum Vol.7 Heavy Metal_24-7
Whitesnake_1987_Bad Boys
Accept_Balls To The Wall_Losing More Than You've Ever Had
Black Sabbath_Cross Purposes_Cross Of Thorns
Black Sabbath_Vol.4_Snowblind
Bon Jovi___Its My Life
Bruce Dickinson_Best Of_Broken
Def Leppard_Metal Museum Vol.9 NWOBHM_Wasted
Diamond Head_Metal Museum Vol.9 NWOBHM_Am I Evil
Dream Evil_Metal Museum Vol.7 Heavy Metal_The Sledge
Iron Maiden_Brave New World_Ghost of the Navigator
Judas Priest_The Very Best Of Judas Priest_Living After Midnight
King Diamond_Metal Museum Vol.7 Heavy Metal_Arrival
Manowar_Metal Museum Vol.7 Heavy Metal_Warriors Of The World United
Mercyful Fate_9_Church Of Saint Anne
Motley Crue_Greatest Hits_Girls. Girls. Girls
Motorhead_Hellraiser Best Of The Epic Y_angel city

MELODIC DEATH METAL

Dimension Zero- This Is Hell- Di'i Minores
Immortal Souls- Ice Upon The Night- You
In Flames_A Sense of Purpose- Drenched in Fear
Kill the Romance_Take Another Life- Friend
Lords Of Decadence_Bound To Fall- Point Of No Return
Darkane_Layers Of Lies_Secondary Effects
Ebony Tears_A Handful of Nothing_Harvester Of Pain
Gardenian_Sindustries_Sonic Death Monkey
In Mourning_Shrouded Divine_In the Failing Hour
Insomnium_Above the Weeping World_The Killjoy
Kalmah_They Will Return_My Nation
Misery Speaks_Catalogue of Carnage_To My Enemies
Naildown_World Domination_Eyes Wide Open
Norther_Till Death Unites Us_The End of Our Lives
The Duskfall_A Liftetime Supply of Guild_A Stubborn Soul
Arch Enemy- Rise Of The Tyrant- In This Shallow Grave
At The Gates- Terminal Spirit Disease- The Beautiful Wound
Children Of Bodom- Are You Dead Yet- Next In Line
Dark Tranquility- _- Void Of Tranquillity
Diablo_Eternium_Black Swan
Diablo_Eternium_Symbol Of Eternity
Dimension zero- Silent night fever- The murder-inn
Night in Gales- Necrodynamic- Doomdrugged
Scar Symmetry_Pitch Black Progress_Calculate the Apocalypse
Septic Flesh_Esoptron_Ice Castle
Serpent_Cradle Of Insanity- Sea Of The Silence
Shadow_Shadow_Breath of Awakening
Soilwork- Sworn To A Great Divide_Exile
The Forsaken_Manifest Of Hate_Dehumanized Perspective
Naildown_World Domination_World Domination

POWER METAL

Avantasia_The Metal Opera Pt. I_Breaking Away
Brainstorm_17 X Dynamit Volume 29 November 2001_Blind Suffering
Chinchilla_Madness_Fight
Cydonia_The Dark Flower_Midnight Man
Dark Moor_Dark Moor [Bonus Track]_A Life For Revenge
Dionysus_Anima Mundi_Heart Is Crying
Domain_The Sixth Dimension_Warpath
Dreamtale_Beyond Reality_Dreamland
Duke_Escape From Reality_Friends
Elegy_Principles Of Pain_No Code No Honour
Impellitteri_The Very Best Of Impellitteri Faster Than The Speed Of Light_Beware Of The Devil
Pwr-AXENSTAR_Far from Heaven_Children forlorn
Pwr-Blind Guardian_A Twist In The Myth_Fly
Pwr-Dragonforce_Sonic Firestorm_Fury Of The Storm
Pwr-Hammerfall_Hearts on fire_Hearts On Fire
Arachnes_Primary Fear_The Warning
Arthemis_Golden Dawn_Master Of The Souls
Firewind_Between Heaven And Hell_Between Heaven And Hell
Freedom Call_Eternity_The Eyes Of The World
Freternia_A Nightmare Story_Grimbor The Great
Manigance_Ange Ou Demon_Ange Ou Demon
Metalium_Hero Nation - Chapter Three_Rasputin
Supreme Majesty_Danger_Save Me
Voice_Soulhunter_Devilish Temptation
Helloween_Keeper of the Seven Keys I_Future World
Kamelot_The Forth Legacy_Until Kingdom Come
Kotipelto_Waiting For The Dawn _Battle Of The Gods

Nightwish_Century Child_End Of All Hope
Sonata Arctica_Unia_Paid in Full
Stratovarius_Intermission_Why Are We Here

PROGRESSIVE METAL

Dreamscape_5th Season_Farewell
Fates Warning_Metal Museum Vol.11 - Progressive Metal_Through Different Eyes
James LaBrie_Elements Of Persuasion_Lost
Liquid Tension Experiment_Liquid Tension Experiment_Three Minute Warning
Meshuggah_Destroy Erase Improve_Transfixion
Nightingale_Metal Museum Vol.11 - Progressive Metal_To The End
Nova Art_Metal Museum Vol.11 - Progressive Metal_My Beloved Hate
Opeth_Metal Museum Vol.11 - Progressive Metal_Harvest
Pain Of Salvation_The Painful Chronicles_Oblivion Ocean
Queensryche_Metal Museum Vol.11 - Progressive Metal_Empire
Redemption_The Origins Of Ruin_Bleed Me Dry
Richard Andersson's Space Odyssey_Metal Museum Vol.11 - Progressive Metal_Embrace The Galaxy
Royal Hunt_Metal Museum Vol.11 - Progressive Metal_Ten To Life
Sphere Of Souls_Metal Museum Vol.11 - Progressive Metal_Sweet Sorrow
Spiral Architect_A Sceptic's Universe_Insect
Age Of Nemesis_Terra Incognita_Forgive Me My Foolish Crime
Blotted Science_ The Machinations Of Dementia_ Adenosine Breakdown
Andromeda_The Immunity Zone_Shadow of a lucent moon
Ayreon_Metal Museum Vol.11 - Progressive Metal_Eyes Of Time
Circus Maximus_Isolate_A Darkened Mind
Cloudscape_Metal Museum Vol.11 - Progressive Metal_Out Of The Shadows
Communic_Metal Museum Vol.11 - Progressive Metal_History Reversed
Dali's Dilemma_Manifesto For Futurism_Miracles In Yesteryear
Dream Theater_Metal Museum Vol.11 - Progressive Metal_Just Let Me Breathe
Mayadome_Paranormal Activity_Mindache
Mindflow_Metal Museum Vol.11 - Progressive Metal_Another Point Of View
Stream Of Passion_Embrace The Storm_Embrace The Storm
Symphony X_The Odyssey_King Of Terrors
Time Requiem_Time Requiem_Watching The Tower Of Skies
Tool_Opiate_Part of Me