EVALUATION OF CONTENT-BASED ACOUSTIC FEATURES FOR MUSICAL GENRE CLASSIFICATION

BY

YUHUI LAI

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2016

Urbana, Illinois

Adviser:

Professor Mark Hasegawa-Johnson

ABSTRACT

In this thesis, we evaluate content-based acoustic features for musical genre classification. Effectiveness of various acoustic features are compared using a k-nearest neighbor (KNN) classifier. By utilizing the combinations of acoustic features, an average classification accuracy of 89% for GTZAN database is achieved, which is comparable to prior work. A statistical test, McNemar's test, is applied to support the idea that musical genre is intrinsically related to content-based acoustic features. Especially for some genres, we are able to identify the particular associated acoustic property. In addition, by comparing our KNN results to a psychoacoustic listening experiment, we associate various human perceptual dimensions with low-level acoustic features.

To my parents, for their love and support.

ACKNOWLEDGMENTS

This thesis would not have been completed without the benefit of the courses I took and the assistance of my fellow study-mates. Especially, my group mate Raymond Yeh inspired and enlightened me many times when I had difficulty coming up with strategies for musical genre classification. Also, I would like to thank Michael Friedman and David Jun for lending their support and experience to help me finalize my experimental setups. Their insights and perspectives are very valuable and their contributions are gratefully acknowledged.

My utmost gratitude goes to my advisor, Prof. Mark Hasegawa-Johnson, who provided me plenty of guidance and encouragement in carrying out this thesis work. He motivates me a lot by responding to my doubts promptly and enthusiastically, thereby helping me to improve my work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

EM	Expectation Maximization
GMM	Gaussian Mixture Model
KNN	K-Nearest Neighbor
LMD	Latin Music Dataset
MFCC	Mel-Frequency Spectral Coefficients
MGR	Music Genre Recognition
MIR	Music Information Retrieval
RMS	Root-Mean-Square
STFT	Short Time Fourier Transform
SVM	Support Vector Machine

CHAPTER 1

INTRODUCTION

1.1 Motivation

In recent years, with the rapid development of the Internet, advancement in network bandwidth and tremendous increase in personal computer storage, people are gaining countless opportunities to enlarge their digital audio collections. Music, which can be traced back to the stone age, is one of the most important creations by humans for entertainment. Nowadays, while people are finding their music of interest to become more and more accessible, the Internet is glutted with an overwhelming number of digital copies of music. Hence, the organization of music pieces becomes crucial and a symbolic description of music is necessary.

Musical genres serve as a fundamental criterion for categorizing large digital music databases. In fact, individuals almost always refer to genre label when they describe their musical taste. Therefore, genre is an important metadata element for the description of music content. Since 2002, with the aim of having a robust and accurate autotagging system for electronic copies of music, music record retailers and music researchers in the music information retrieval (MIR) community have paid much more attention to music genre recognition (MGR).

As a matter of fact, genre labels are probably the most widely used descriptors for music [1–3]. Under implicit agreement, we cluster and categorize the vast universe of music into different musical genres such as classical, jazz, hip-hop, rock, and country, etc. However, there are no strict definitions and boundaries for musical genres, as they tend to be easily affected by historical, cultural, public and marketing factors [2]. Most commonly, the distributor of the song or the artist who is responsible for the song creation gets to assign the music to a genre.

Due to the ambiguous nature of musical genre, some [1] argue that genre is intrinsically ill-defined and arriving at a realistic and useful musical taxonomy can be extremely difficult. The boundaries of musical genres are vague, and definitions of musical genre labels are subject to social and historical influences. Though no explicit definition of musical genre exists, the work on automatic genre classification still achieves promising results by employing different feature extraction strategies and various types of classifiers. Most feature extraction frameworks are based on psychoacoustic experiments; that is, genre taxonomy is dictated by human perception of acoustic features such as melody, instrumentation, harmony and rhythm, etc.

The concept of musical genre can be interpreted both intentionally and extensionally [1]. By the intentional definition, the concept of genre is a generalization agreed on by a group of people, which is similar to how we come up with a phrase to necessarily and sufficiently define a level of abstraction. Hence, genre belongs to a linguistic category in this regard and it is used to associate itself to music titles: Yesterday by the Beatles is a Brit-pop title because it is by the Beatles, and we all share cultural knowledge about this group, the 60s, etc. On the other hand, musical genre can be extensional, which is a more analytical approach in which a music track is described in terms of musical elements such as timbre, pitch, tempo, energy distribution, rhythm, or other characteristics. Again, Yesterday by the Beatles is a mellow pop song because it has a cheesy medium tempo melody, string backup and it is sung with a melancholic voice.

The fundamental problem of categorizing music according to genre is that music serves as a way for the artist to express his/her feelings about the world. Feeling or emotion itself is complicated and hard to describe. Sometimes, we indeed do have a clear definition of angry, happy, sad, and relaxed. However, most of the time, people are sentimental and their feelings are complicated and tend to be a mixture of those well-defined emotions. Music is an outcome of those feelings and a means to an end for the creator to vent their emotional state. It is obvious that if description of one's feeling suffers from ambiguity then this kind of categorization of music genre surely endures ambiguity, originating in our dualist view of the world. In other words, genre is intrinsically ill-defined and attempts at defining genre precisely have a strong tendency to end up in circular, ungrounded projections of fantasies [1]. According to a psychoacoustic listening experiment by Seyerlehner et al. [4], not only are there huge variations between individuals in differentiating genres, but also people tend to categorize an unfamiliar music excerpt into folk or country due to vagueness of the categories. Moreover, they find that the majority of the participants strongly agree on just one or two possible genre assignments for most of the songs.

With that being said, genre is intrinsically related to classification: it is useful to assign a genre to a music excerpt, as it contains some musical features that are recognized by humans and can be used to describe the similarity within a genre and the difference across genres. This high level of abstraction of a descriptor for a musical excerpt can be found in our natures and in our irrepressible tendency to classify. Though some genres are ill-defined, some genres can be significantly differentiated from others. There is strong consensus for genres like blues, country and classical as they are aggregates of individual perceptions of one particular kind of music. The main contribution of this thesis is to unravel the intrinsic relationships between different categories of music genre and content based psychoacoustic features.

This rest of this thesis is organized as follows:

- Chapter 2 provides an overview of the current state-of-art music information retrial (MIR) system. In addition, the development of various techniques and approaches for musical feature extraction for music genre classification are discussed.

- Chapter 3 describes our method of musical feature extraction and compares various classifiers. Furthermore, detailed experimental methods are elaborated.

- Chapter 4 presents the numerical results obtained from the experiment conducted in Chapter 3. The KNN classification rate for different choices of k and confusion pattern are demonstrated.

- Chapter 5 discusses the results from Chapter 4 and demonstrates their implications based on statistical significance analysis. In the meantime, the intrinsic and extrinsic properties of each genre are explained. Moreover, Chapter 5 addresses future applications and concerns.

- Chapter 6 summarizes the main idea of this thesis and reinstates the main conclusion based on the experimental results.

1.2 Thesis Statement

What defines a musical genre? Is genre mainly a term dictated by the user's experience and taste? Or is genre intrinsically related to human perception or extrinsically associated with the acoustic features of the song clip?

In this thesis, we answer the above questions by performing a thorough evaluation of different acoustic feature selections. The relation between genres and acoustical features is examined using a KNN classification framework. The experimental results show that genre is actually intrinsically related to content-based acoustic features of various types. There are underlying psychoacoustic properties that are shared within one genre and they are quite distinctive from those of other genres. Furthermore, we are able to link acoustic properties with human perception of genres by analyzing a listening experiment.

CHAPTER 2

BACKGROUND AND LITERATURE REVIEW

2.1 Music Genre Classification

Music genre recognition (MGR) was introduced in the work by Tzanetakis and Cook [5]. The increasing need to autotag music motivates the development of this kind of work. As of 2007, nearly 70% of music autotagging consisted of genre labels [6]. This topic is studied by many researchers, and various frameworks for automatic genre recognition have been proposed. Following is a summary of related work on feature extraction strategies and classifiers:

- Costa et al. [7] extract visual features from spectrograms of music excerpts. Similar to Tzanetakis [5], who extract timbral texture features for classification, they treat time-frequency representation as texture image. A recognition rate of 67.2% is obtained by their proposed framework.

- A Gabor filtering and LPQ texture descriptors approach is utilized by Costa et al. [8] to perform automatic genre classification on a Latin Music Database (LMD). 80% recognition rate is achieved by using a support vector machine (SVM) classifier.

- The texture features are based on Local Binary Pattern, a structural texture operator that has been successful in recent image classification research. Experiments are performed with two well-known datasets: the LMD, and the ISMIR 2004 dataset [9]. The performance of their approach reaches about 82.33%.

- Panagakis et al. [10] presented a novel framework that utilizes joint spare low-rank classification for music genre classification. Their result is comparable or slightly superior to state-of-the-art music genre classification models.

- On-line Dictionary Learning model is proposed by Srinivas et al. [11], and they achieve accuracy of 99.41% on LMD dataset which outperforms all other existing work.

- In the work by Schindler and Rauber [12], the authors combine audio-visual features for music genre classification. Results show that visual features can provide a boost to non-timbral and rhythmic features.

- Lee et al. [13] build a music recommendation system based on users' favorite songs. The genre is categorized based on usage history and a distance metric learning algorithm is applied in order to reduce the dimensionality of feature vector with a little performance degradation.

- In Lykartsis and Lerch [14], a beat histogram is constructed for rhythm-based musical genre classification based on beat analysis [15]. Accuracy of 76.6% is obtained using proposed the approach.

- Burred and Lerch [16] construct a system in a hierarchical way such that the feature selection and the classification are carried out systematically. They achieve 94.59% for differentiating speech and background music.

- Pikrakis [17] proposes a deep-learning architecture which is capable of modelling signatures that represent the rhythm of music recordings. The paper provides supporting evidence that deep-learning networks can be adopted to discriminate between genres based on extracted rhythmic signatures.

- Peeters [18] and Papadapoulos [19] propose a probabilistic framework in which the time of the beats and their associated positions inside a measure, hence the downbeats, are considered as hidden states and are estimated simultaneously using signal observations.

- In Seyerlehner et al. [20], the authors use a set of block-level features for three different tasks: genre classification, tag classification and music similarity estimation. Accuracy of 85.49% is achieved on GTZAN dataset.

- Li et al. [21] propose a new feature extraction method for music genre classification, DWCHs, which captures the local and global information of music signals simultaneously by computing histograms on their Daubechies wavelet coefficients. The result shows that the classification rate of given framework for GTZAN dataset is 76.8%.

These studies employed different features ranging from acoustic to visual elements. Various types of supervised and unsupervised learning classifiers are tested based on all kinds of extracted features.

2.2 Musical Features

The concept of genre can be associated with various acoustic features [2, 22]; one generally attempts to include features from all possibly relevant musical categories for better classification rate. In traditional music theory, we have defined various broad musical dimensions associated with different low-level features of music [23]. In addition, these low-level features are the key elements in our high-level perceptual space. According to McKay and Fujinaga [24], they accumulate 109 different features devised from a catalog of 160 features [25]. 109 different features can be divided into 7 high-level perceptual spaces: instrumentation, timbral texture, rhythm, dynamics, pitch statistics, melody and chord. Detailed description of low-level features that are associated with music dynamics, timbre, rhythm and pitch are presented as follows:

2.2.1 Dynamic Feature

The dynamics of a musical excerpt can be best characterized by a song-level feature such as the root-mean-square (RMS) of the music clip or the energy distribution of the whole piece. Dynamics characterize the general shape and trend of certain musical properties. Song dynamics serve as a high-level descriptor that describes the global shape of the song rather than local spectral shape. RMS, s-lope, attack, and low energy are the most widely used descriptors for representing song dynamics.

RMS

RMS is calculated based on a texture window. It can be calculated as follows:

$$RMS = \sqrt{\frac{\sum_n^N M[n]^2}{N}} \tag{2.1}$$

where $M[n]$ is the discrete sampled signal in time domain and N is the total number of samples. RMS is used to characterize the amplitude of the signal within a texture window.

Slope

Slope provides the general trend and shape of the musical excerpt in time domain.

Attack

Attacks are associated with the onset detection curve as the mark the hit phase of each onset. We retrieve the average of attack phase and attack magnitude for our experiment.

Low-Energy

Low-energy feature: Low energy is the only feature that is based on the texture window rather than the analysis window. Texture window provides information about the whole clip while analysis window segments the music excerpt into a number of frames. Low energy is defined as the percentage of analysis windows that have less RMS energy than the average RMS energy across the texture window. As an example, vocal music with silences will have large low-energy value while continuous strings will have small low-energy value [5].

2.2.2 Rhythm Feature

Rhythm of music is a well-studied topic. Formal researchers have achieved promising results by taking advantage of rhythmic features. Rhythmic content features characterize the movement of music signals over time and contain such information as the regularity of the rhythm, the beat, the tempo, and the time signature. The feature set for representing rhythmic structure is based on detecting the most salient periodicities of the signal [5].

Tempo

Tempo feature is achieved by detecting periodicities from the onset detection curve. Traditionally, the paradigm for tempo estimation is based on detecting periodicities, and choosing the maximum periodicity score for each frame separately [26].

Fluctuation Peak

One way of estimating the rhythm is based on spectrogram computation transformed by auditory modeling, followed by spectrum estimation in each band [27]. Fluctuation peak captures the maximum magnitude of spectrum estimation within each band.

2.2.3 Timbral Texture Feature

Timbral texture feature describes the difference between consecutive frequency spectrals of the song. Timbral texture feature originated from music-speech discrimination [28], and speech recognition [29] as it is a powerful descriptor that can help us to differentiate the mixture of a song within common rhythmic and pitch patterns [30]. In order to extract the timbral texture feature, one applies short time Fourier transform (STFT) to the song excerpt and calculates various attributes within the spectral domain.

Spectral Flux

Spectral flux proposes to calculate the difference between spectral frames. It is defined as the squared difference between the normalized magnitudes of successive spectral distributions [5].

$$F_t = \sum_{n=1}^{N} (N_t[n] - N_{t-1}[n])^2 \tag{2.2}$$

where $N_t[n]$ and $N_{t-1}[n]$ are the normalized magnitude of the Fourier transform at the current frame t, and the previous frame $t-1$, respectively. As such, it is a measure of the change in spectral shape of consecutive frames.

MFCCs

In the area of speech and music classifiers, one of the most commonly used descriptors for timbre texture feature is mel-frequency cepstral coefficients (MFCCs) [31]. MFCCs preserve the stability of the signal by averaging the spectrogram using a mel scale that is logarithmic at high frequencies. The signal becomes invariant to deformation. At the same time, temporal structures such as transients and time-varying characteristics of the signal will be lost due to averaging the spectrogram at a mel scale. In order to minimize the loss, a window size of 23 ms is introduced to prevent losing locally non-stationary signals. There are three steps to calculate MFCCs for speech and music. The first step is to divide the signal into frames. The second step involves taking the amplitude spectrum of the signal. A short-time Fourier transform of signal $x(t)$ is computed as follows:

$$x_{t,T}(u) = x(u)\omega_T(u-t) \tag{2.3}$$

$$\tilde{x}_{t,T}(\omega) = \int x_{t,T}(u)e^{-i\omega u}du \tag{2.4}$$

where ω_T denotes a window of length 23 ms and $\tilde{x}_{t,T}$ is the Fourier transform of $x_{t,T}$. MFCCs are cosine transforms of MFSCs, which average the spectrogram along the frequency axis, giving

$$M_{T,x}(t,j) = \frac{1}{2\pi}\int |\tilde{x}_{t,T}(\omega)|^2 |\tilde{\psi}_j(\omega)|^2 d\omega \tag{2.5}$$

These intervals have a constant frequency bandwidth below 1000 Hz and a constant octave bandwidth above 1000 Hz, where each $\tilde{\psi}_j(\omega)$ covers a mel-frequency interval indexed by j. Then we take the log of $M_{T,x}(t,j)$ at each of the mel frequencies. The last step is to take a cosine transform of $log(M_{T,x}(t,j))$. The MFCCs are the amplitudes of the resulting spectrum. Typically, 13 MFCC coefficients are used to characterize the signal.

DMFCCs and DDMFCCs

The MFCC feature vector describes only the power spectral envelope of an analysis window, but it seems like speech would also have information in the dynamics, prompting one to ask: What are the trends of the MFCC coefficients over time?

Delta-MFCCs (DMFCCs) calculates the trajectories based on the static MFCC:

$$d_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2} \tag{2.6}$$

where d_t is a delta coefficient, from frame t computed in terms of the static coefficients c_{t+N} to c_{t-N}. Delta-delta (acceleration) coefficients are calculated in the same way, but they are calculated from the deltas, not the static coefficients [31].

Zero Crossing Rate

Zero crossing rate represents the number of zero crossings of the time domain signal. Hence, it can provide a reliable measure of the noisiness of the signal [32].

Spectral Rolloff

Spectral rolloff is another descriptor for the measurement of local change within the frequency domain. It is defined as the frequency below R_t where 85% of the magnitude distribution is concentrated [33].

$$\sum_{n=1}^{R_t} M_t[n] = 0.85 * \sum_{n=1}^{N} M_t[n] \tag{2.7}$$

Spectral Centroid

The spectral centroid is the barycentre point of the spectral distribution within a frame [5].

$$SC = \frac{\sum_n k * S[n]}{\sum_n S[n]} \tag{2.8}$$

where S is the magnitude spectrum of a frame. Spectral centroid delivers a fine sense of the spectral shape.

Spectral Irregularity

The irregularity of a spectrum is the degree of variation of the successive peaks of the spectrum. Spectral irregularity is defined as the sum of the squares of the difference in amplitude between adjoining frames [5].

$$Irr = \sum_{n=1}^{N} (M[n] - M[n+1])^2 / \sum_{n=1}^{N} M[n]^2 \tag{2.9}$$

Brightness

Spectral brightness is the amount of energy above the frequency of interest. This term is intrinsically associated with loudness in our perceptual space. Hence, brightness can differentiate sharpness from softness for the timbral structure of the song.

Spectral Spread

Spectral spread computes the standard deviation of the data. Being the squared deviation of the random variable from its mean value, the variance is always positive and is a measure of the dispersion or spread of the distribution. The square root of the variance is called the standard deviation, and is more useful in describing the nature of the distribution since it has the same units as the random variable [5].

Skewness Feature

Spectral skewness calculates the third-order statistical moments of the given spectrum. Skewness is an efficient descriptor for the shape of spectrum based on the texture window [34].

Spectral Flux

Spectral flux is defined as the spectral correlation between adjacent windows [35]. It reflects the degree of change in the spectrum between consecutive texture windows.

Kurtosis

Spectral kurtosis calculates the fourth-order statistical moments of the given spectrum. Similar to skewness, kurtosis also measures the shape of spectrum based on the texture window.

Flatness

Flatness indicates whether the distribution is smooth or spiky, and results from the simple ratio between the geometric mean and the arithmetic mean [32]:

$$\frac{\sqrt[N]{\prod_{n=0}^{N-1} x(m)}}{\frac{\sum_{n=0}^{N-1} x(n)}{N}} \tag{2.10}$$

Roughness

Compared to spectral flatness, roughness provides the opposite information regarding the energy distribution for sequence of spectrum.

Entropy

Entropy computes the relative entropy of one frame spectrum with respect to other frames. The Shannon entropy, used in information theory, is based on the following equation [36]:

$$H(x) = -\sum_{i=1}^{n} p(x_i) log_b p(x_i) \tag{2.11}$$

where b is the base of the logarithm.

2.2.4 Pitch

Pitch serves as a primary measure for harmonics and melody information about musical signals. Using the pitch detection algorithm, we are able to identify the dominant peaks of the autocorrelation function, calculated via the summation of envelopes for each frequency band obtained by decomposing the signal. The envelopes are accumulated into pitch histograms, and the pitch content features are then extracted from the pitch histograms. The common low-level features associated with pitch content features are: the amplitudes and periods of maximum peaks in the histogram, pitch intervals between the two most prominent peaks and the overall sums of the histograms [33].

Chromagram Peak

The spectrum is converted from the frequency domain to the pitch domain by applying a log-frequency transformation. The distribution of the energy along the pitches is called the chromagram [37]. The chromagram is then wrapped by fusing the pitches belonging to same pitch classes. The wrapped chromagram therefore shows a distribution of the energy with respect to the 12 possible pitch classes [38]. Chromagram peak provides the magnitude of each tonality peak for all twelve possible pitches.

Chromagram Centroid

The chromagram centroid is defined as the center of gravity of the magnitude spectrum of the chromagram [5].

Key Clarity

Key information provides a broad estimation of tonal center positions and their respective strengths. Key clarity indicates the key strength associated with the best key(s), i.e., the peak ordinate(s) [32].

Key Mode

Key mode estimates the modality, i.e. major vs. minor, returned as a numerical value between -1 and +1: the closer it is to +1, the more major the given excerpt is predicted to be, the closer the value is to -1, the more minor the excerpt might be [32].

HCDF

The harmonic change detection function (HCDF) is the flux of the tonal centroid [39]. Tonal centroid corresponds to a projection of the chords along circles of fifths, of minor thirds, and of major thirds.

CHAPTER 3

EXPERIMENTAL METHODS

3.1 Algorithms

There are three commonly used classifiers for the task of musical genre classification: support vector machine (SVM) [40], Gaussian mixture model (GMM) [41], and K-nearest neighbor (KNN) [42]. Although SVM has proven to be the most effective in music genre classification [43], it has extensive training time with high computational cost. For the purpose of our experiment, in which we are trying to determine the intrinsic properties of a genre instead of build up a novel system for better classification rate, we restrained ourselves to time-efficient and computationally cheap classifiers. GMM assumes the multidimensional Gaussian distribution of the parameters [44] and estimates the parameters from training data using Expectation Maximization (EM) algorithm [45]. The GMM parameters are mean vectors, covariance matrices and mixture weights from all component densities. Lastly, KNN classifier is an example of a nonparametric classifier where the testing sample is labeled according to the distance measurement from its nearest neighbors [42]. In other words, no explicit probability distribution function is expressed for training model and it is approximated locally using the training set.

For the purpose of our experiment, we choose KNN classifier because it requires relatively no training time and it is easy to compare the classification rate across all different feature selections. From human perception perspective, we tend to associate a new song with the known genre label, and KNN classifier is suitable for this kind of assignment. In addition, compared to other classifiers, KNN enables us to take account of all the low-level features vectors, while other classifiers such as GMM and SVM require us to do dimension reduction for a large chunk of feature vectors, which could result in potential loss of information.

3.2 Multidimensional Scaling

Multidimensional scaling (or MDS) is a set of mathematical operations that explores the hidden structure of the dataset [46]. In our experiment, multidimensional scaling is used to determine the theoretical meaning of the spatial representation of different kinds of genre. We try to represent 10 different genre recognition scores is based on the result of a human listening experiment [4] geometrically by 5 dimensions, such that the interpoint distances correspond in some sense to experimental dissimilarities between genres.

For example, given a confusion matrix $D = [d_{ij}]$, where d_{ij} is the classification score which shows the percentage of the ith genre being classified as the jth genre, we can now fully characterize the coordinates of each genre in an n dimensional space, where n is the number of different genres in the confusion matrix.

$$D := \begin{bmatrix} d_{1,1} & d_{1,2} & d_{1,3} & \dots & d_{1,n} \\ d_{2,1} & d_{2,2} & d_{2,3} & \dots & d_{2,n} \\ \dots & \dots & \dots & \dots & \dots \\ d_{n,1} & d_{n,2} & d_{n,3} & \dots & d_{n,n} \end{bmatrix}$$

The goal of MDS is, given D, is to find n vectors $d_1, ..., d_n \in \Re^N$ such that

$$\|x_i - x_j\| \approx d_{i,j} \tag{3.1}$$

This can be formulated as a optimization problem,

$$\min_{x_1,...x_n} \sum_{i<j} (\|x_i - x_j\| - d_{i,j})^2 \tag{3.2}$$

We named x_i as the semantic feature vector for the ith genre. In our work, we map the coordinates of each genre into a 5-dimensional space with 75.5% energy conserved according to eigenvalue decomposition [47]. In other words, instead of taking $x_1, ..., x_n$, we only use $x_1, ..., x_5$, where x_i is in the descending order associated with its eigenvalue.

3.3 Experimental Setup

3.3.1 Training and Testing dataset

In our experiments, we used a well-known dataset: GTZAN, which consists of 10 genres (blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, and rock), each with 100 music clips. All clips are 16-bit and are sampled at 22,050 Hz monotonically. Due to mechanical limitation and because the goal of this experiment is to evaluate the relationship between content-based acoustic features and musical genre labels, we only take 10 music clips of each genre. Hence, our dataset consists of 100 song excerpts from 10 different genres. In each run, we select one excerpt out of 100 as testing data and leave the remaining 99 as training data.

3.3.2 Feature Extraction

We retrieve all the acoustic features provided in section 2.2 with MIR toolbox, [32] shown in table 3.1. MIR toolbox is resourceful and powerful as it has numerous built-in functions that enable us to obtain the low-level features conveniently. For the purpose of our experiment, we only used 4 out of 7 categories based on popularity of the feature usage in recent years for music genre taxonomy. The music perceptual spaces we utilized are dynamics, timbral texture, rhythm and pitch. Similar to Song et al. [48], we calculate the mean and standard deviation of each low-level feature to form perceptual or acoustic feature vector.

Means and standard deviations of individual low-level features are stacked into a column to represent the overall musical structure of each song excerpt. For example, dynamic feature vector is column vector which stacks the mean and standard deviation of RMS, slope, attack and low energy. In order to make sense of our classification result, normalization of the feature vector is necessary, which ensures equal weight of each feature space. We normalize each feature vector on a 0 to 1 scale:

$$f_{new} = \frac{f_{old} - f_{min}}{f_{max} - f_{min}} \tag{3.3}$$

Table 3.1: List of extracted low-level features and their corresponding perceptual feature spaces

Perceptual Feature	Low-level Features
Dynamic	RMS
	Slope
	Attack
	Low energy
Rhythm	Tempo
	Fluctuation peak
Timbre	Spectrum centroid
	Brightness
	Spread
	Skewness
	Kurtosis
	Rolloff95
	Rolloff85
	Spectral Entropy
	Flatness
	Roughness
	Irregularity
	Spectral flux
	MFCC
	DMFCC
	DDMFCC
Pitch	Chromagram peak
	Chromagram centroid
	Key clarity
	Key mode
	HCDF

where f_{old} is the unscaled feature vector, f_{max} and f_{min} represent the maximum and minimum element in that low-level feature vector space respectively, and f_{new} is the normalized low-level feature vector after scaling. As a result, there are altogether 4 acoustic/perceptual feature vectors being constructed: dynamic, rhythm, timbre and pitch perceptual vector.

3.3.3 Classification

All possible combinations of perceptual vectors are tested on a KNN classifier with $k = 1,2,3,4,5,6,7$. We use Euclidean distance as the metric for evaluating the similarity between testing and training dataset. Owing to the fact that there is basically no training required for KNN classifier, we use leave-one-out cross validation for classification. Results are presented in Chapter 4.

3.4 McNemar's Test

In order to quantitatively evaluate and verify the correlation between content-based acoustic feature and musical genre, a statistical test, McNemar's test, is employed for validation. The McNemar test is used to determine if there are d-ifferences in a dichotomous dependent variable between two related groups [49]. It can be conceptualized as testing two different properties of a repeated measure dichotomous variable. For instance, the test can be applied to table 3.2, which tabulates the outcomes of two tests on a sample of n subjects.

Table 3.2: An example showing the number of testers that pass and fail the first and second test

	Test 2 pass	Test 2 fail	Row total
Test 1 pass	a	b	a+b
Test 1 fail	c	d	c+d
Column total	a+c	b+d	n

The null hypothesis of marginal homogeneity states that the two marginal probabilities for each outcome are the same [50]: i.e., $p_a + p_b = p_a + p_c$ and $p_c + p_d = p_b + p_d$. Thus the null and alternative hypotheses are

$$H_0 : p_b = p_c \tag{3.4}$$

$$H_1 : p_b \neq p_c \tag{3.5}$$

Here, p_a etc., denote the theoretical probability of occurrences in cells with the corresponding label. The McNemar test statistic is:

$$\chi^2 = \frac{(b-c)^2}{b+c} \tag{3.6}$$

Under the null hypothesis, with a sufficiently large number of discordants (cells b and c), χ^2 has a chi-squared distribution with 1 degree of freedom. If the χ^2 result is significant, this provides sufficient evidence to reject the null hypothesis in favor of the alternative hypothesis that $p_b \neq p_c$, which would mean that the marginal proportions are significantly different from each other.

In our case, we have a well-defined dependent variable that is dichotomous with two mutually exclusive categories (i.e., dynamic feature is associated with jazz and dynamic feature is not associated with jazz). The McNemar test is applied to determine whether some particular acoustic features are relevant in defining one particular genre.

CHAPTER 4

EXPERIMENTAL RESULTS

4.1 Overall KNN Classification Result

All combinations for 4 acoustic feature vectors are tested. Hence, there is a total of 15 different combinations, and we assign a feature selection number for each combination. Table 4.1 indicates the features that have been utilized for a given feature selection choice. The average classification rate over various choices of k with different feature selection is shown in figure 4.1.

Table 4.1: List of acoustic features used and their corresponding feature selection numbers

Feature Selection	Perceptual Feature
1	Dynamic
2	Rhythm
3	Timbre
4	Pitch
5	Dynamic + Rhythm
6	Dynamic + Timbre
7	Dynamic + Pitch
8	Rhythm + Timbre
9	Rhythm + Pitch
10	Timbre + Pitch
11	Dynamic + Rhythm + Timbre
12	Dynamic + Rhythm + Pitch
13	Rhythm + Timbre + Pitch
14	Dynamic + Timbre + Pitch
15	Dynamic + Rhythm + Timbre + Pitch

Notice that with $k = 2$ and with all the acoustic features combined, we achieve the highest recognition rate of 89%. Also, even with different selection of feature, 2NN classifier out performs others by considerable amount. Therefore, we choose

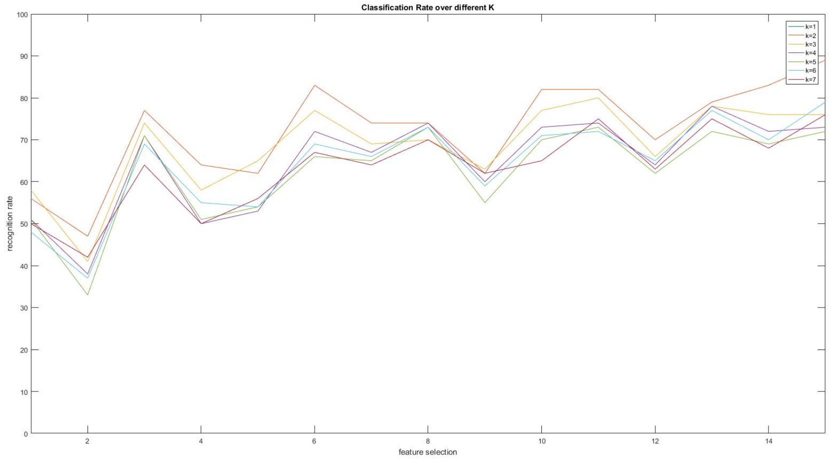


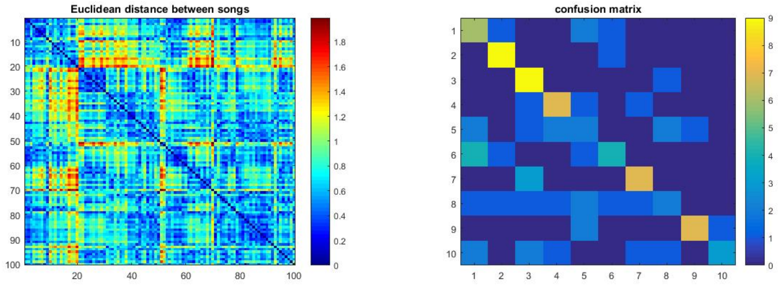
Figure 4.1: Overall classification rate for the choice of $k = 1,2,3,4,5,6,7$

$k = 2$ for the later comparison of classification rate for a single genre across various features and the mapping from acoustic feature to music genre. It is obvious that as the number of features being included increases, so does the classification rate, confirming the validity of our feature selection. One can easily see that the timbre feature plays a major role in deciding the average classification rate across all genres when only one acoustic feature is incorporated. On the other hand, we aim to relate and evaluate various acoustic features for individual genres instead of generalizing one or a few efficient acoustic features for all genres.

4.2 Pairwise Distance and Confusion Matrices

In order to find the correlation between each acoustic feature and the potential corresponding music genre, we compute the pairwise distance between the testing song and all the remaining songs of each feature selection. A confusion matrix is attached for enhanced visualization of how diverse features could affect the genre taxonomy. The results are shown in figures 4.2-4.16.

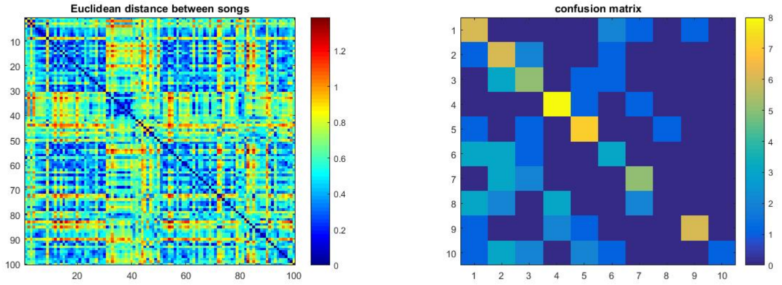
As we can see from figures 4.2, 4.3, 4.5 and 4.6, some songs from one particular genre stand out easily by having a relatively large Euclidean distance from other songs. For example, songs from classical and blues exhibit a high recognition rate regardless of which acoustic feature is selected, while for other genres, the classification rate is closely associated with the chosen acoustic feature.



(a) Pairwise distance for each song in dataset

(b) Confusion matrix. Columns stand for the ground truth label and rows correspond to the predicted label.

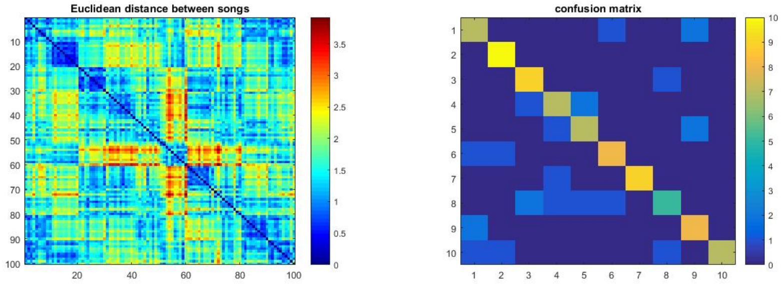
Figure 4.2: Distance measure and confusion matrix for dynamic feature



(a) Pairwise distance for each song in dataset

(b) Confusion matrix. Columns stand for the ground truth label and rows correspond to the predicted label.

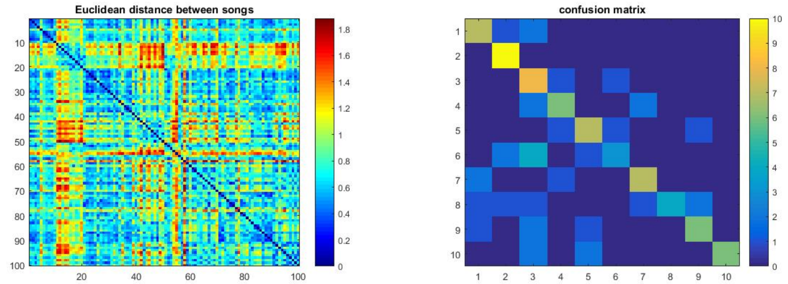
Figure 4.3: Distance measure and confusion matrix for rhythm feature



(a) Pairwise distance for each song in dataset

(b) Confusion matrix. Columns stand for the ground truth label and rows correspond to the predicted label.

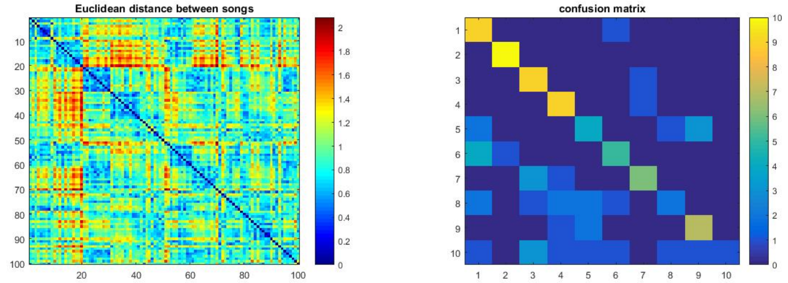
Figure 4.4: Distance measure and confusion matrix for timbre feature



(a) Pairwise distance for each song in dataset

(b) Confusion matrix. Columns stand for the ground truth label and rows correspond to the predicted label.

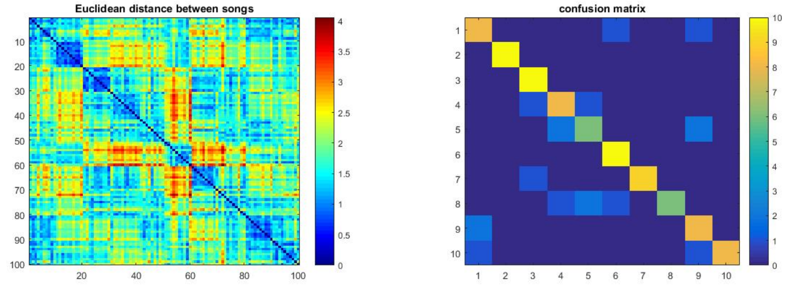
Figure 4.5: Distance measure and confusion matrix for pitch feature



(a) Pairwise distance for each song in dataset

(b) Confusion matrix. Columns stand for the ground truth label and rows correspond to the predicted label.

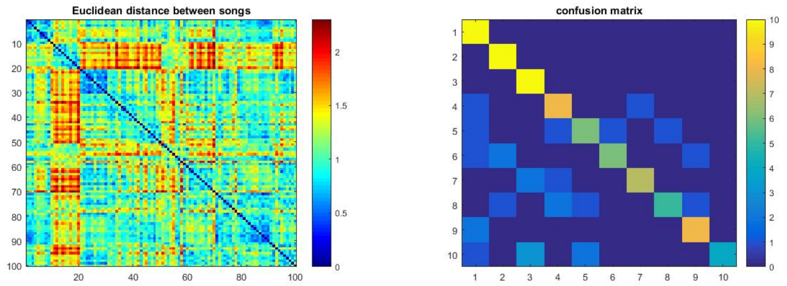
Figure 4.6: Distance measure and confusion matrix for the combination of dynamic and rhythm feature



(a) Pairwise distance for each song in dataset

(b) Confusion matrix. Columns stand for the ground truth label and rows correspond to the predicted label.

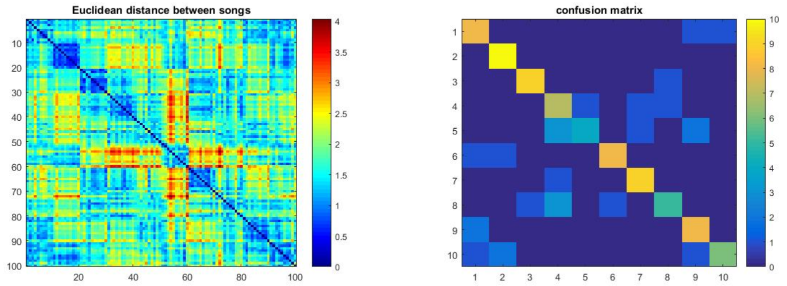
Figure 4.7: Distance measure and confusion matrix for the combination of dynamic and timbre feature



(a) Pairwise distance for each song in dataset

(b) Confusion matrix. Columns stand for the ground truth label and rows correspond to the predicted label.

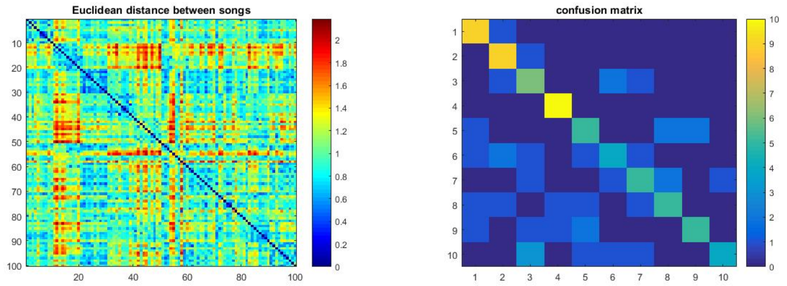
Figure 4.8: Distance measure and confusion matrix for the combination of dynamic and pitch feature



(a) Pairwise distance for each song in dataset

(b) Confusion matrix. Columns stand for the ground truth label and rows correspond to the predicted label.

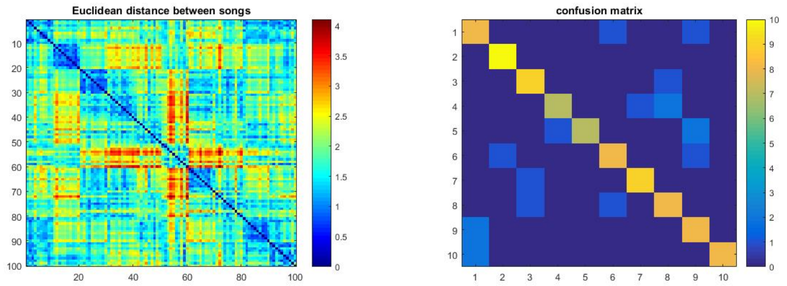
Figure 4.9: Distance measure and confusion matrix for the combination of rhythm and timbre feature



(a) Pairwise distance for each song in dataset

(b) Confusion matrix. Columns stand for the ground truth label and rows correspond to the predicted label.

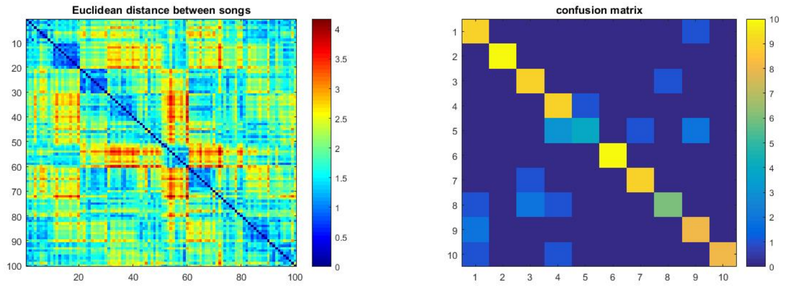
Figure 4.10: Distance measure and confusion matrix for the combination of rhythm and pitch feature



(a) Pairwise distance for each song in dataset

(b) Confusion matrix. Columns stand for the ground truth label and rows correspond to the predicted label.

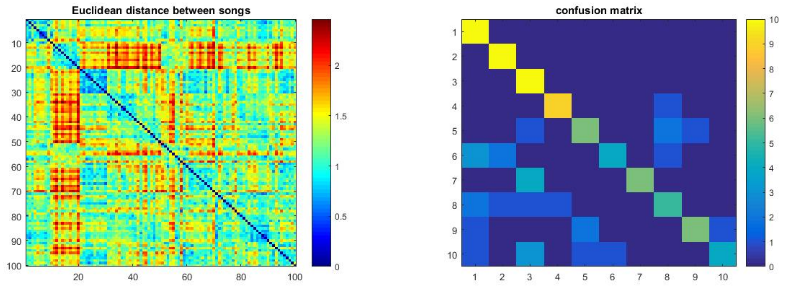
Figure 4.11: Distance measure and confusion matrix for the combination of timbre and pitch feature



(a) Pairwise distance for each song in dataset

(b) Confusion matrix. Columns stand for the ground truth label and rows correspond to the predicted label.

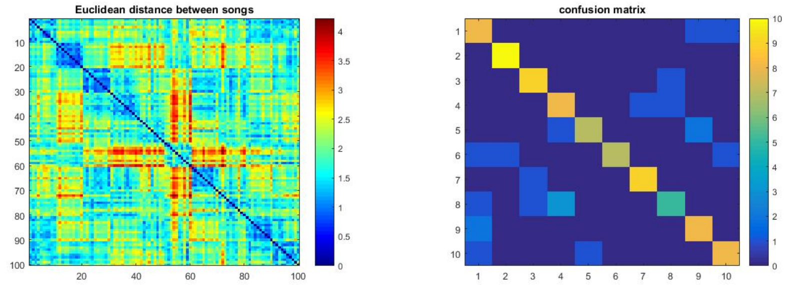
Figure 4.12: Distance measure and confusion matrix for the combination of dynamic, rhythm and timbre feature



(a) Pairwise distance for each song in dataset

(b) Confusion matrix. Columns stand for the ground truth label and rows correspond to the predicted label.

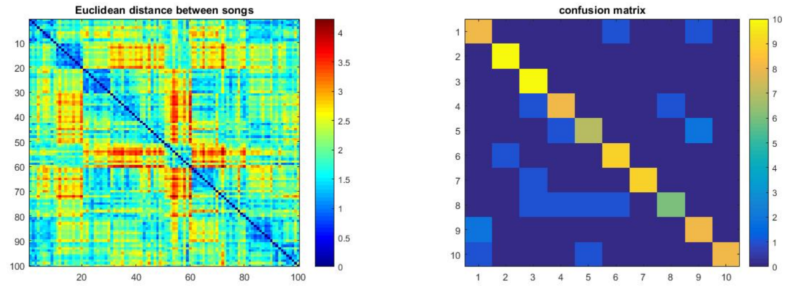
Figure 4.13: Distance measure and confusion matrix for the combination of dynamic, rhythm and pitch feature



(a) Pairwise distance for each song in dataset

(b) Confusion matrix. Columns stand for the ground truth label and rows correspond to the predicted label.

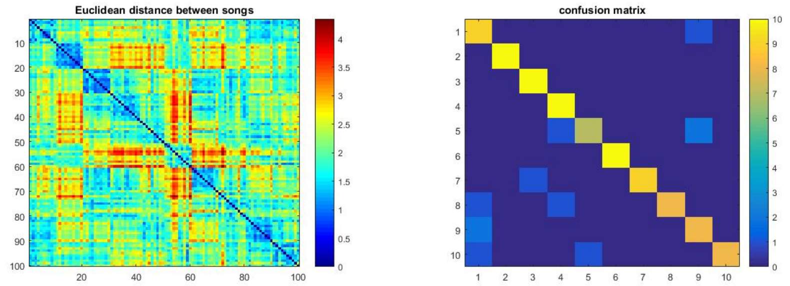
Figure 4.14: Distance measure and confusion matrix for the combination of rhythm, timbre and pitch feature



(a) Pairwise distance for each song in dataset

(b) Confusion matrix. Columns stand for the ground truth label and rows correspond to the predicted label.

Figure 4.15: Distance measure and confusion matrix for the combination of dynamic, timbre and pitch feature



(a) Pairwise distance for each song in dataset

(b) Confusion matrix. Columns stand for the ground truth label and rows correspond to the predicted label.

Figure 4.16: Distance measure and confusion matrix for the combination of dynamic, rhythm, timbre and pitch feature

4.3 Classification Rate for Each Genre

In this section, in order to visualize the relationship between the selection of acoustic feature and individual genre more straightforwardly, we provide the recognition rate of each genre with respect to varied chosen features. Table 4.2 and table 4.3 display the numerical results.

Table 4.2: Classification accuracy in % for each genre for varied chosen feature

Feature Selection	Blues	Classical	Country	Disco	Hiphop
1	60	90	90	70	20
2	60	60	50	80	70
3	70	100	90	70	70
4	70	100	80	60	70
5	90	100	90	90	40
6	80	100	100	80	60
7	100	100	100	80	60
8	80	100	90	70	40
9	90	90	60	100	50
10	80	100	90	70	70
11	90	100	90	90	40
12	100	100	100	90	60
13	80	100	90	80	70
14	80	100	100	80	70
15	90	100	100	100	70

4.4 Semantic Features

In Seyerlehner et al. [4], a listening experiment is done for recording the recognition score of genres by human listeners. A multidimensional scaling approach is applied to the recognition score to represent genres with hidden perceptual subspaces. In other words, we hope to characterize genres in terms of a weighted combination of hidden perceptual subspaces. In order to associate hidden perceptual subspaces with content-based acoustic features, the multidimensional scaling results from a listening experiment are shown in figure 4.17, 4.18, 4.19, 4.20, and 4.21.

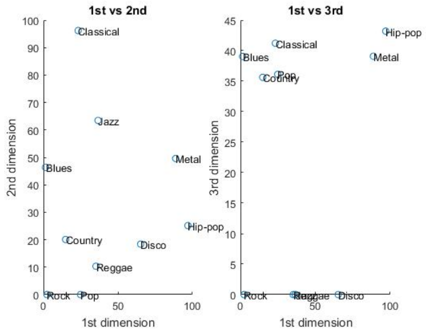


Figure 4.17: Multidimensional scaling results for genres, dimension 1 vs. 2, dimension 1 vs. 3

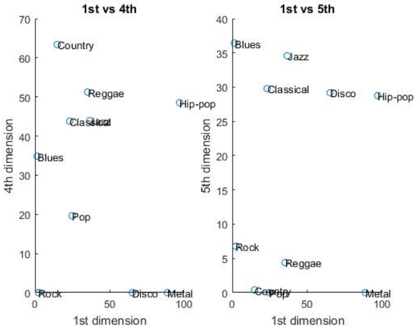


Figure 4.18: Multidimensional scaling results for genres, dimension 1 vs. 4, dimension 1 vs. 5

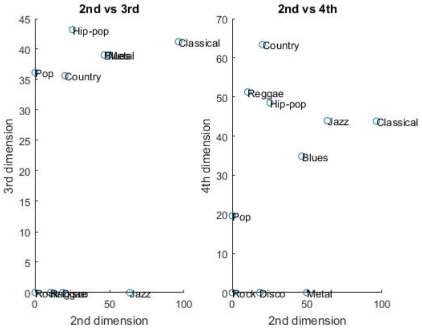


Figure 4.19: Multidimensional scaling results for genres, dimension 2 vs. 3, dimension 2 vs. 4

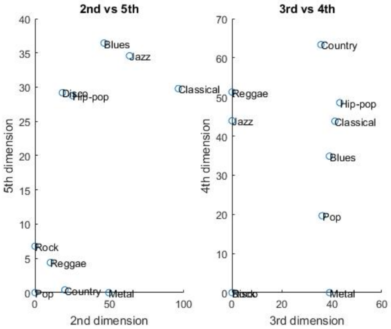


Figure 4.20: Multidimensional scaling results for genres, dimension 2 vs. 5, dimension 3 vs. 4

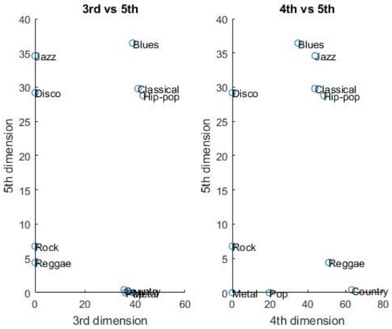


Figure 4.21: Multidimensional scaling results for genres, dimension 3 vs. 5, dimension 4 vs. 5

Table 4.3: Classification accuracy in % for each genre for varied chosen feature

Feature Selection	Jazz	Metal	Pop	Reggae	Rock
1	40	70	20	70	30
2	30	50	0	60	10
3	80	90	50	80	70
4	30	70	40	60	60
5	50	60	20	70	10
6	100	90	60	80	80
7	60	70	50	80	40
8	80	90	50	80	60
9	40	50	50	50	40
10	80	90	80	80	80
11	100	90	60	80	80
12	40	60	50	60	40
13	70	90	50	80	80
14	90	90	60	80	80
15	100	90	80	80	80

To associate each dimension with some given acoustic features, correlation between human perceptual space and acoustic feature is calculated. Correlation factor and associated P value are shown in tables 4.4 and 4.5.

Table 4.4: Correlation factor for perpetual dimension vs. acoustic feature selection

Feature selection/Dimension	1	2	3	4	5
1	-0.1329	0.0478	0.0943	0.2341	-0.8314
2	0.7560	0.3220	0.1194	0.2389	0.4634
3	0.0661	0.6907	0.1422	0.3294	0.049
4	-0.0094	0.4137	0.7366	0.2195	0.0477

4.5 McNemar's Test Result

Although McNemar's test is performed on each pair of combinations chosen above, we only present the tests that are statistically significant in table 4.6. The abbreviations for dynamic, rhythm, timbre and pitch are D, R, T and P respectively. Among all the genres, we observe that for some genres, such as blues, classical, disco and metal, regardless of which perceptual feature is selected for classifica-

Table 4.5: P value for perpetual dimension vs. acoustic feature selection

Feature selection/Dimension	1	2	3	4	5
1	0.7143	0.1944	0.7956	0.5151	0.009
2	0.0118	0.3642	0.7424	0.5063	0.1773
3	0.8668	0.0270	0.6951	0.3527	0.8930
4	0.9794	0.2346	0.0127	0.5423	0.8959

tion, they are fairly recognizable. On the other hand, genres like jazz, pop and rock are heavily influenced by the selection of perceptual vector.

Table 4.6: P value for varied feature selection

Genre	Jazz	Pop	Rock	Country	Hiphop
Feature selection 1	D+R+T+P	D+R+T+P	D+R+T+P	D+R+T+P	D+R+T+P
Feature selection 2	D+R+P	D+R	D+R	R	D
P value	0.0156	0.0156	0.0078	0.0313	0.0313

CHAPTER 5

DISCUSSION

5.1 Result Analysis

Firstly, a promising result of 89% recognition rate is obtained by combining all low-level features with a 2NN classifier. Figure 4.1 confirms the validity of our chosen content-based acoustic feature due to the fact that as the number of acoustic features being involved increases, so does the recognition rate. Also, notice that by only utilizing one acoustic feature for genre taxonomy, the rhythm feature performs the worst while timbre works the best. This is interesting because according to Perrott and Gjerdigen [51], humans can identify musical genre in concurrence with the record companies 71.68% of the time (among a total of 10 genres), based on 300 ms of audio. This indicates that the musical information used to help human perception is short segments rather than whole general musical structure. Our experimental results show otherwise: the dynamic feature is used to characterize the general shape of the music, but it actually works better than rhythm, which corresponds to temporal change within short segments, e.g., tempo and fluctuation.

Regarding individual genres, timbre feature serves as an efficient descriptor for jazz, rock, pop and country, suggesting that these genres are intrinsically related to a certain type of timbral texture. This is especially true for jazz; as shown in table 4.2, by solely including timbre we achieve 80% classification rate. In other words, jazz is a generalization term corresponding to some type of timbre, e.g., we define the mood anger based on certain facial expression and behavior. The same reasoning may apply to rock and pop, but in this case, pop and rock are the nicknames for different combinations of timbre and pitch. Statistical test verifies the idea that jazz is closely related to timbre by showing the statistical significance with a P value of 0.0156, which is less than 0.05. Similarly, the combination of dynamic, timbre and pitch plays a crucial role for country genre classification.

On the other hand, genres like blues, classical, disco, metal and reggae exhibit high recognition rate regardless the feature selection. A possible explanation for this is that genres such as classical contain perceptual features that are so distinctive with respect to other genres that one can easily identify them.

Gouyon et al. [52] address the fact that humans tend to confuse classical music with country and jazz. Our confusion pattern, figure 4.3, which only captures the rhythm pattern, concurs with this. Also, it is clear that the confusion pattern shifts across different choices of acoustic feature, as seen in figure 4.2, 4.3, 4.4 and 4.5.

Strong correlations between various human perceptual dimensions and acoustic spaces are shown in table 4.4 and 4.5. Notice that the first dimension is closely related to rhythmic features by having a correlation factor of 0.756. Similarly, we can associate dimension 2 with timbre, dimension 3 with pitch and dimension 5 with dynamics. The P value analysis agrees with the observation by having P less than 0.05 for all those acoustic features. Interestingly, dimension 4 is almost equally weighted across all acoustic feature selections. In general, the one-to-one correspondence of acoustic feature to perceptual space is found for dimension 1, 2, 3 and 5. These results provide the theoretical meaning of the spatial representation of genres; that is, we are able to interpret the hidden structure of perceptual space according to acoustic features.

Also, new grouping techniques of music genres are possible. Observe that there is literally zero Euclidean distance between rock and blues in dimension 1, which suggests that blues and rock have similar rhythm, and so do metal and hip-hop. According to this grouping technique, we can cluster musical genres in terms of acoustic features. For example, rock, reggae and disco are grouped together due to similar pitch, blues and metal are grouped together owing to similar dynamics, etc. This can lead to a hierarchy modeling of genres categorization.

5.2 Extension

The dataset should be extended from 100 songs to 1000 songs if there is no machine and time limitation. In order to identify the intrinsic attributes that are related to music genres, supplemental feature need to be added, i.e., lyrics and instrumental features. A hierarchy modeling of low-level features can be utilized to enhance the classification rate. In addition, more types of classifiers could be tested for comparison of different feature selections.

CHAPTER 6

CONCLUSION

In this thesis, we provide a comparative study of various feature extraction scenarios and investigate the classification performance of those features based on a k-nearest neighbor (KNN) classifier. By utilizing the combinations of acoustic features, an average classification accuracy of 89% for the GTZAN database is achieved, which is comparable to the state-of-the-art. We can conclude that content-based acoustic features are proven to be useful for automatic music genre classification. Statistical methods were used to determine which features are significant for a specific music genre.

Experimental results showed that music genre is indeed intrinsically related to music attributes. For some genres, we are able to identify the particular associated acoustic property. Though Talupur et al. [53] argued that different genres have different classification criteria, our results demonstrate that using the combination of all low-level features provides the optimal recognition rate. In addition, we are able to associate various human perceptual dimensions with content-based acoustic features.

CHAPTER 7

REFERENCES

[1] J.-J. Aucouturier and F. Pachet, “Representing musical genre: A state of the art,” *Journal of New Music Research*, vol. 32, no. 1, pp. 83–93, 2003.

[2] N. Scaringella, G. Zoia, and D. Mlynek, “Automatic genre classification of music content: a survey,” *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 133–141, 2006.

[3] C. McKay and I. Fujinaga, “Musical genre classification: Is it worth pursuing and how can it be improved?” in *ISMIR*, 2006, pp. 101–106.

[4] K. Seyerlehner, G. Widmer, and P. Knees, “A comparison of human, automatic and collaborative music genre classification and user centric evaluation of genre classification systems,” in *International Workshop on Adaptive Multimedia Retrieval*. Springer, 2010, pp. 118–131.

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

[6] T. Bertin-Mahieux, D. Eck, and M. Mandel, “Automatic tagging of audio: The state-of-the-art,” *Machine Audition: Principles, Algorithms and Systems*, pp. 334–352, 2010.

[7] Y. M. Costa, L. S. Oliveira, A. L. Koericb, and F. Gouyon, “Music genre recognition using spectrograms,” in *Systems, Signals and Image Processing (IWSSIP), 2011 18th International Conference on*. IEEE, 2011, pp. 1–4.

[8] Y. Costa, L. Oliveira, A. Koerich, and F. Gouyon, “Music genre recognition using Gabor filters and LPQ texture descriptors,” in *Iberoamerican Congress on Pattern Recognition*. Springer, 2013, pp. 67–74.

[9] Y. M. Costa, L. Oliveira, A. L. Koerich, F. Gouyon, and J. Martins, “Music genre classification using LBP textural features,” *Signal Processing*, vol. 92, no. 11, pp. 2723–2737, 2012.

[10] Y. Panagakis, C. L. Kotropoulos, and G. R. Arce, “Music genre classification via joint sparse low-rank representation of audio features,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1905–1917, 2014.

[11] M. Srinivas, D. Roy, and C. K. Mohan, "Music genre classification using online dictionary learning," in *2014 International Joint Conference on Neural Networks (IJCNN).* IEEE, 2014, pp. 1937–1941.

[12] A. Schindler and A. Rauber, "An audio-visual approach to music genre classification through affective color features," in *European Conference on Information Retrieval.* Springer, 2015, pp. 61–67.

[13] J. Lee, S. Shin, D. Jang, S.-J. Jang, and K. Yoon, "Music recommendation system based on usage history and automatic genre classification," in *2015 IEEE International Conference on Consumer Electronics (ICCE).* IEEE, 2015, pp. 134–135.

[14] A. Lykartsis and A. Lerch, "Beat histogram features for rhythm-based musical genre classification using multiple novelty functions," *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, 2015.

[15] E. D. Scheirer, "Tempo and beat analysis of acoustic musical signals," *The Journal of the Acoustical Society of America*, vol. 103, no. 1, pp. 588–601, 1998.

[16] J. J. Burred and A. Lerch, "A hierarchical approach to automatic musical genre classification," in *Proceedings of the 6th International Conference on Digital Audio Effects.* Citeseer, 2003, pp. 8–11.

[17] A. Pikrakis, "A deep learning approach to rhythm modeling with applications," in *6th International Workshop on Machine Learning and Music (MML13)*, 2013.

[18] G. Peeters, "Spectral and temporal periodicity representations of rhythm for the automatic classification of music audio signal," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1242–1252, 2011.

[19] G. Peeters and H. Papadopoulos, "Simultaneous beat and downbeat-tracking using a probabilistic framework: theory and large-scale evaluation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1754–1769, 2011.

[20] K. Seyerlehner, M. Schedl, T. Pohle, and P. Knees, "Using block-level features for genre classification, tag classification and music similarity estimation," *Submission to Audio Music Similarity and Retrieval Task of MIREX*, vol. 2010, 2010.

[21] T. Lidy, A. Rauber, A. Pertusa, and J. M. I. Quereda, "Improving genre classification by combination of audio and symbolic descriptors using a transcription systems." in *ISMIR*, 2007, pp. 61–66.

[22] F. Pachet and D. Cazaly, “A taxonomy of musical genres,” in *Content-Based Multimedia Information Access-Volume 2*, 2000, pp. 1238–1245.

[23] D. Temperley, *The Cognition of Basic Musical Structures*. MIT Press, 2004.

[24] C. McKay and I. Fujinaga, “Automatic genre classification using large high-level musical feature sets.” in *ISMIR*, vol. 2004. Citeseer, 2004, pp. 525–530.

[25] C. McKay, “Automatic genre classification of midi recordings,” Ph.D. dissertation, McGill University, 2004.

[26] O. Lartillot, “Mirtempo: Tempo estimation through advanced frame-by-frame peaks tracking,” *Proceedings of the Music Information Retrieval E-valuation eXchange (MIREX), Utrecht, Netherlands*, pp. 1–2, 2010.

[27] C. N. Silla Jr, C. A. Kaestner, and A. L. Koerich, “Automatic genre classification of latin music using ensemble of classifiers,” in *Proc. of the 33rd Integrated Software and Hardware Seminar*, 2006, pp. 47–53.

[28] E. D. Scheirer and M. Slaney, “Multi-feature speech/music discrimination system,” May 27 2003, US Patent 6,570,991.

[29] M. Hariharan, S. Yaacob, M. Hasrul, and O. Q. Wei, “Speech emotion recognition using stationary wavelet transform and timbral texture features,” *ARP-N Journal of Engineering and Applied Sciences*, 2006.

[30] C. L. Krumhansl, “Rhythm and pitch in music cognition.” *Psychological bulletin*, vol. 126, no. 1, p. 159, 2000.

[31] L. Rabiner and B.-H. Juang, “Fundamentals of speech recognition,” 1993.

[32] O. Lartillot and P. Toiviainen, “A Matlab toolbox for musical feature extraction from audio,” in *International Conference on Digital Audio Effects*, 2007, pp. 237–244.

[33] T. Li, M. Ogihara, and Q. Li, “A comparative study on content-based music genre classification,” in *Proceedings of the 26th Annual International ACM SIGIR conference on Research and Development in Informaion Retrieval*. ACM, 2003, pp. 282–289.

[34] I. Fujinaga, “Adaptive optical music recognition,” Ph.D. dissertation, McGill University Montréal, Canada, 1996.

[35] S. McAdams, “Perspectives on the contribution of timbre to musical structure,” *Computer Music Journal*, vol. 23, no. 3, pp. 85–102, 1999.

[36] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 5, no. 1, pp. 3–55, 2001.

[37] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler, "A tutorial on onset detection in music signals," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 1035–1047, 2005.

[38] E. Gómez, "Tonal description of polyphonic audio for music content processing," *INFORMS Journal on Computing*, vol. 18, no. 3, pp. 294–304, 2006.

[39] C. Harte, M. Sandler, and M. Gasser, "Detecting harmonic change in musical audio," in *Proceedings of the 1st ACM Workshop on Audio and Music Computing Multimedia*. ACM, 2006, pp. 21–26.

[40] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. ACM, 1992, pp. 144–152.

[41] G. Lu and T. Hankinson, "A technique towards automatic audio classification and retrieval," in *Proc. IEEE Intl. Conf. on Signal Processing*, vol. 2, 1998, pp. 1142–1145.

[42] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[43] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[44] D. Reynolds, "Gaussian mixture models," *Encyclopedia of Biometrics*, pp. 827–832, 2015.

[45] J. A. Bilmes et al., "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models," *International Computer Science Institute*, vol. 4, no. 510, p. 126, 1998.

[46] J. B. Kruskal, "Nonmetric multidimensional scaling: a numerical method," *Psychometrika*, vol. 29, no. 2, pp. 115–129, 1964.

[47] I. Jolliffe, *Principal component analysis*. Wiley Online Library, 2002.

[48] Y. Song, S. Dixon, and M. Pearce, "Evaluation of musical features for emotion classification." in *ISMIR*. Citeseer, 2012, pp. 523–528.

[49] L. Gillick and S. J. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on*. IEEE, 1989, pp. 532–535.

[50] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947.

[51] D. Perrot and R. Gjerdigen, "Scanning the dial: An exploration of factors in the identification of musical style," in *Proceedings of the 1999 Society for Music Perception and Cognition*, 1999, p. 88.

[52] F. Gouyon, S. Dixon, E. Pampalk, and G. Widmer, "Evaluating rhythmic descriptors for musical genre classification," in *Proceedings of the AES 25th International Conference*, 2004, pp. 196–204.

[53] M. Talupur, S. Nath, and H. Yan, "Classification of music genre," *Project Report for*, vol. 15781, 2001.