Comparing Features for the Automatic Classification of Vocal Music

Nav Chue

HAIT Master Thesis series nr. 11-011

Thesis submitted in partial fulfillment of the requirements for the degree of Master of Arts in Communication and Information Sciences, Master Track Human Aspects of Information Technology, at the Faculty of Humanities of Tilburg University

Thesis committee:

Dr. J. J. Pajmans
Dr. M. M. van Zaanen

Tilburg University
Faculty of Humanities
Department of Communication and Information Sciences
Tilburg center for Cognition and Communication (TiCC)
Tilburg, The Netherlands
September 2011
Abstract

Traditional approaches to the task of automatic classification of music by genre generally focus on the use of note-duration symbols to represent musical content. As well, most studies generally extract information from a combination of instruments. This paper compares the traditional method of using note-duration symbols with another method based on time slice representations of the music to see which performs better. It differentiates the task for the correct classification of Western Classical and Modern Pop music based on the vocal part of a musical piece. The TiMBL memory-based learning software was used to process the data and to obtain accuracy results. The research in this thesis demonstrates that using a combination of high-level features symbols is more effective at classifying music by genre than looking at the music based on time slice measurements.

To build the framework for the experiments, a description of musicology focusing on musical genres and concepts and the process of machine learning are discussed. The method for and design of the classification task is given, as well as an analysis of the impact of feature extraction on the classification task. Finally, the results of the experiments are outlined and discussed.
Acknowledgments

I would like to thank my family (Mom, Papa, Jan², Jason, and Niels) and Mareks for their invaluable support throughout this difficult year. I would also like to extend gratitude to my supervisor, Professor J. J. Paijmans (Paai), for guiding me through this research.
# Contents

Abstract .......................................................................................................................... i
Acknowledgments ........................................................................................................... ii
Figures and Tables .......................................................................................................... v

Chapter 1. Introduction................................................................................................... 1
  1.1. Thesis Outline ........................................................................................................ 2
  1.2. State of the Art ...................................................................................................... 2
  1.3. Experimental Strategies ...................................................................................... 4

Chapter 2. Exploration of Musicology ......................................................................... 6
  2.1. Music Theory ....................................................................................................... 6
    2.1.1. Notes ............................................................................................................. 6
    2.1.2. Keys ............................................................................................................. 7
    2.1.3. Chords ......................................................................................................... 7
    2.1.4. Pitch ........................................................................................................... 7
    2.1.5. Melody ........................................................................................................ 9
    2.1.6. Harmony .................................................................................................... 9
    2.1.7. Rhythm ....................................................................................................... 10
    2.1.8. Intervals ..................................................................................................... 10
    2.1.9. Octaves ..................................................................................................... 10
    2.1.10. Beats ....................................................................................................... 11
    2.1.11. Vocal Range ........................................................................................... 12
  2.2. Music Genre ........................................................................................................ 13
    2.2.1. Western Classical Music ............................................................................ 13
      2.2.1.1. Renaissance ......................................................................................... 14
      2.2.1.2. Baroque ............................................................................................. 15
    2.2.1.3. Classicism ............................................................................................. 15
    2.2.1.4. Romanticism ......................................................................................... 15
    2.2.2. Modern Pop Music .................................................................................... 15
      2.2.2.1. Dance ................................................................................................. 15
      2.2.2.2. Jazz and Blues ................................................................................... 16
      2.2.2.3. Pop/Rock ........................................................................................... 16
      2.2.2.4. Soul/R&B/Funk ............................................................................... 16
    2.2.3. Classical Music versus Pop Music ............................................................... 16

Chapter 3. Machine Learning ...................................................................................... 19
  3.1. Automatic Classification of Music by Genre ....................................................... 19
    3.1.1. Classification Algorithms .......................................................................... 19
      3.1.1.1. Unsupervised Classification .............................................................. 19
      3.1.1.2. Supervised Classification ................................................................ 20
    3.2. Classification Problems .................................................................................... 21
  3.3. TiMBL .................................................................................................................. 21

Chapter 4. Research Problem ..................................................................................... 23
  4.1. Research Question ............................................................................................... 23
  4.2. Research Design ................................................................................................ 23
Figures and Tables

Tables
Table 1. Highest accuracy results of MIREX 2010 mixed genre classification..................................................3
Table 2. Notes per measure ........................................................................................................................................11
Table 3. Results of 10-fold cross-validation for 6 variations in window-length.........................................................33
Table 4. Information gain on the first fold for feature-sets .......................................................................................33
Table 5. Excerpt from Bach chorale “Hilf, Herr Jesu, lass gelingen” (Greentree, 2011) ........................................34
Table 6. Excerpt from Bach chorale “Hilf, Herr Jesu, lass gelingen” (Greentree, 2011), showing the first line of Figure 13 in the first stave ........................................................................................................35
Table 7. Time slice results .............................................................................................................................................36

Figures
Fig. 1. High and low notes (“Reading Music”) ..............................................................................................................6
Fig. 2. Notes on the treble (top) and bass clefs (“A beginner's guide,” n.d.) .................................................................7
Fig. 3. Commonly used notes on keyboard position and their pitch position (“Reading Music”) ..........................8
Fig. 4. Basic pitch class representation (Cochrane, n.d.) ..........................................................................................8
Fig. 5. Stave representation of vocal ranges with the clef symbol which is used on all note lies to indicate the position of middle C (“C clef”, n.d.) ..........................................................................................12
Fig. 6. Time periods for Western classical music (Naxos, no date) .........................................................................14
Fig. 7. Basic pop vocal arrangement (Smith, 2005) ..................................................................................................17
Fig. 8. Memory based machine learning (Paijmans, 2010) ......................................................................................23
Fig. 9. Whole note middle C (Good, 2002) ..................................................................................................................26
Fig. 10. Portion of XML representation of middle C note in XML (“MusicXML”, n.d.) ..................................................26
Fig. 11. Symbolic representation of the first stave of Table 5 ..................................................................................26
Fig. 12. Slice representation of the first stave of Table 5 .........................................................................................34
Chapter 1. Introduction

The focus of this research is the comparison of two different strategies for extracting features from musical scores based on their usefulness for the classification of music by genre. In the experiments for this thesis, the two experimental strategies explained in Section 1.3 (note-duration and time slice analysis) will be used to test if one performs significantly better than the other. Notes, tones, chords, intervals and pitch have all been found to be significant for classifying musical pieces by genre. As such, the experimental setup will analyze which of these features will be extracted from the musical data for use in creating the classification model. The method will involve the extraction and analysis of features both in isolation and in combination with each other. A system of automatic classification will be developed based on a dataset of MusicXML files covering two genres (Western Classical and Modern Pop), and will include a variety of sub-categories of these genres. The data will then be refined and processed to discover which vocal features or combination of these features are most reliable for differentiating classical music files from other genre. In essence, this thesis will be checking for distinct patterns.

The experiments will test the anecdotal hypothesis that classical vocalization displays different, explicit features as compared to vocalizations from different genres. Previous research (McKay and Fujinaga, 2006; Alluri, 2007) has demonstrated that it can be difficult for automatic music genre classifiers to differentiate classical pieces from other genres based on the analysis of musical instruments and combination of instruments. (For example, the chords and the chord progressions in classical and pop music are very similar.) Therefore, the vocal line will provide the data for testing the two experimental methods, excluding other instruments. By isolating and analyzing some of its parameters, it will be possible to see if the vocal line can be defining for music classification. As well, the voice line is monophonic, which makes extraction less difficult in a classification task (section 1.3). This thesis will focus specifically on the differences between contemporary pop vocal music and classical vocal music. This will involve a the review of classical music in general, followed by an extrapolation of theories related to vocal music. The aim is to discover which attributes in classical vocal music can be correlated to deduce the genre of a classical piece, using one or the other of the experimental approaches.

The researcher's interest in this topic emerged from a personal interest in classical music. The study of classical music, combined with education in the theories and methods involved in information retrieval resulted in more interest in the classification of music by genre. Specifically, the researcher's curiosity stemmed from a interest in whether or not classical vocalists tend to sing different notes from signers from other genres.

The theoretical approach for this paper is the analysis of classical musical pieces (sheet music scores) in MusicXML format. The two genres that will be used for classification in this study are Western Classical music and Modern Pop music and their taxonomy is similar to that found on popular online music websites that employ genre categories. The musical pieces include a variety of styles within each broad category, ensuring that the data pool is representative and realistic.

This research will include the following primary tasks:

- Collect XML files for analysis – The XML files will be retrieved from free databases on the Internet such as Musescore.com and Wikifonia.org.
• Feature extraction – The features to be extracted conform to either of two experimental strategies:
  a) note-duration symbols
  b) time “slices” through music scores
• Manually determine (hand label) the genre of dataset files – Because genre classification is a subjective process, each file in the test set of will need to be assigned a genre before they are analyzed by the classification system. This can be done by hand-labeling the test files according to a reputable Internet classification guide/resource.
• Automatically process the data – The data will be run through TiMBL to determine performance on the classification task using both experimental methods.
• Evaluate results – The performance results from both experiments will be compared to check the validity of the methods for use in automatic genre classification research.

1.1. Thesis Outline

The thesis consists of the following sections:

• Chapter 2 – A description of musicology focusing on the genres and musical concepts (and their importance) relevant to the thesis. The stylistic and musical differences between the two genres is analyzed.

• Chapter 3 – The process of machine learning is discussed, along with a detailed description of the proposed method of automatic genre classification, including and analysis of the impact of feature extraction on the classification task.

• Chapter 4 – A general description of the research to be undertaken, including the research problem and the research design.

• Chapter 5 – The specifics of the classification task are outlined, focusing on the experimental process, including the dataset used, and the features selected.

• Chapter 6 – This chapter reveals the results of the experiments.

• Chapter 7 – The discussion of the experiment results, as well as suggestions for future work will be covered here.

1.2. State of the Art

The MIREX 2010 competition had a similar contest task focused on the classification of audio musical pieces by genre based on a set of ten mixed genre categories (MIREX, 2010). As illustrated in Table 1, the best performance achieved was approximately 89% on the Rap/Hip-Hop genre, closely followed by the Baroque and Country genres.
Table 1. Highest accuracy results of MIREX 2010 mixed genre classification

<table>
<thead>
<tr>
<th>CLASS</th>
<th>CLASSIFICATION RESULTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>baroque</td>
<td>81.2857</td>
</tr>
<tr>
<td>blues</td>
<td>78.8571</td>
</tr>
<tr>
<td>classical</td>
<td>71.8571</td>
</tr>
<tr>
<td>country</td>
<td>82.8571</td>
</tr>
<tr>
<td>edance</td>
<td>74.5714</td>
</tr>
<tr>
<td>jazz</td>
<td>79.5714</td>
</tr>
<tr>
<td>metal</td>
<td>79.8571</td>
</tr>
<tr>
<td>rap/hiphop</td>
<td>88.8571</td>
</tr>
<tr>
<td>rock-roll</td>
<td>61.1429</td>
</tr>
<tr>
<td>romantic</td>
<td>63.8571</td>
</tr>
</tbody>
</table>

The precursor to the MIREX competition, ISMIR, ran a competition in 2004 to compare the performance of audio description algorithms on a genre classification task (Polotti and Rocchesso, 2008). The best performance was a classification accuracy of 84.07% correct answers on a test set of 729 pieces with six genre categories. For the 2005 MIREX competition, the contest focused on the accuracy of different classification methods for symbolic recordings into appropriate music genres, and the best results were 65.28% for the raw classification accuracy, and 65.28% for mean hierarchical accuracy (Karydis, 2006). Since that first competition, MIREX results have increased. The best results for the symbolic genre classification contest is a performance of 84% based on a taxonomy of nine genres (McKay, Burgoyne, Hockman, Smith, Vigliensoni, and Fujinaga, 2010). As at the 2010 competition, the contest focused on the performance of audio genre classification tasks specifically, and no entry with more than six genre categories has performed successfully above an 80% rank.

Using data in a format similar to MIDI, Karydis (2006) selected features expressed in repeating pitch patterns, other pitch attributes, and histograms of pitch durations. Working with five sub-genres of Classical music, a performance accuracy of approximately 90% was achieved.

Pitch histograms were used by Tzanetakis, Ermolinskyi, and Cook (2005) to demonstrate symbolic and acoustic representations of pitch in music signals. The results presented pitch histograms as a way to represent the pitch content of music signals both in symbolic and acoustic form. The features selected resulted in a histogram showing the accurate categorization of 50% of the data from 5 different genres. Lidy and Rauber (2005) used audio data separated into three corpora comprised of 698 files from eight genres; 1458 from 10 genres; and 1000 files from 10 genres. Classification performances achieved accuracies for each corpus at 70.4%, 84.2%, and 74.9%, respectively (Roos and Manaris, 2007).

Basili et al. (2004) used approximately 300 MIDI files based on six genres (Classical, Jazz, Blues, Disco, Pop, and Rock) to extract features such as pitch range, time changes and melodic intervals. A classification accuracy of 70% was achieved. McKay and Fujinaga (2004) worked with 950 MIDI files from three genres (Classical, Jazz, Popular), selecting 109 features – including pitch statistics, melody, and chords – for their experiments, achieving an accuracy rate of 98%. Also in 2004,
Dixon et al. (2004) selected rhythmic features from audio to classify 698 files from eight Ballroom Dance sub-genres and had a performance accuracy of 96%. This high level of accuracy is tempered by the research finding that “rhythm classification is much easier than general genre classification” (Roos and Manaris, 2007). Working with the Dance genre, Gouyon et al. (2004) used 698 files from eight sub-genres, selecting rhythmic features, and achieved a 78.9% accuracy. Uhle and Dittmar (2004, in Polotti and Rocchesso, 2008) worked with musical pieces from eleven genres, selected rhythmic features and achieved a performance of 67.6%.

Using audio features based on amplitude diversity, Li et al. (2003) used 1000 music works across 10 genres to attain an accuracy of 78.5%.

Working with audio data on pitch, timbral, and rhythmic features, and a test set of 1000 files equally split across ten genres (Classical, Blues, Classical, Disco, Hip-Hop, Jazz, Pop, Rock, Country, Metal, and Reggae), Tzanetakis and Cook (2002) were able to achieve a 61% performance accuracy.

Finally, in doing experiments, researchers have found that there is a high level of accuracy achieved in classification tasks for distinguishing Pop and Classical genres. Polotti and Rocchesso (2008) identify previous research findings: For example, Costa et al. (2004) achieve up to 90.3% classification accuracy, and Mierswa and Morik (2005) report even 100% on 200 pieces. In both cases, the baseline is one half. Although Xu et al. (2003) report a classification accuracy of 93% for four genres, in general the classification accuracy decreases when the number of genres grows.

The goal of the experiments in this thesis is not to outperform any of these systems, but rather to compare the two different approaches to feature extraction that are described in Section 1.3.

1.3. Experimental Strategies

In traditional methods of music classification, the central features for experiments are notes, note durations, or a combination of the two. This research proposes a different method of automatic classification, whereby a feature is created based on the traditional parameters for the voice or melody line in the music. Using the voice in this way as a classification feature focuses specifically on the staves with the melody. The decision to focus on the vocal line is based on the monophonic nature of voice and the fact that it represents the melody of a musical piece. Monophonic means that the music is represented as one note at a time rather than a simultaneity of notes, and this feature of the vocal/melody line makes it easier to process. Previous research has identified that detection of the singing voice is a monophonic problem, and that pitch estimation for monophonic signals is much simpler than polyphonic musical data (Akan, Pande, and Limaye, 2010). One major obstacle is that with polyphonic data it is difficult to put all the notes in the music together because they are of different lengths.

In the proposed, non-traditional method, instead of using the voice plus duration representation of musical data, the music will be sampled to extract “slices” of predetermined sizes at different time periods in the music. Each measure in the music is divided into slices, and slices are sampled at a certain rate, primarily in fractions of the 256 slices in a measure. Each slice is then processed to compute what note is played at that particular point in time. This process transforms the data into learning representations based on the time slices for each musical piece which are then input into TiMBL as a training set for analysis. Essentially, there is a vote - based on the slices - for the category/class of the musical piece. As an example, the traditional method of feature use would
examine a note of say E, and then add the duration of that note, such as five beats, or would use a combination of both features (E_5). In the alternative method, the process will involve one slice of every nth second, and for every slice, the note played at that specific time is extracted. The E note with a duration of five beats is no longer considered to be one feature, but it can be sliced into two or more features, making the feature equivalent to a specific point in musical time that is being described.

The key parameters of each time slice (adapted from Whitman, Flake, and Lawrence, 2001) are:

- **Slice jump** - The length in seconds of each slice extracted from the musical piece. Slice length is the length in seconds of each individual slice of a song. This determines the value of the length of each part that the music is divided into for the representation. Slices of about two beats in length (approximately 0.8 to 1 second or 10 to 40 milliseconds) are most effective because that is how long it takes for a change that defines music to be identified. It is the number of points (256 in a measure) between each slice.

- **Slice start** - The offset where the program starts taking slices at different points within the musical piece. This indicates the frequency/distribution of slices within a song in the slice representation.

The next chapter will introduce the musical concepts that form the basis for this research in so much as the importance of the musical features that can be used as data inputs and outputs for the research experiments are highlighted.
Chapter 2. Exploration of Musicology

In order to be able to extract and interpret any high level features from MusicXML files it is essential to understand musical terms and how they relate to musical pieces. In this chapter, an overview of the musical theory that is most important for this research will be outlined. These theoretical concepts are essential for the creation of optimal experiment features and to describe how the features are to be implemented in the MusicXML protocol.

2.1. Music Theory

The following section will explore the basic structure of the musical data gleaned from musical scores that will be used for the experiments.

2.1.1. Notes

Sometimes called “tones”, notes are a single sound of a given musical auditory frequency (pitch), duration (rhythm), and velocity. The pressing of a piano key or a single movement of a bow on a violin represents a note being struck (Ke, 2004). The musical scale has eight notes - C, D, E, F, G, A, B, C - in repetitive ascending and descending patterns comprising an octave relationship (section 2.1.9 and Figures 1 and 2). On music scores, notes and other musical symbols are set on a stave consisting of five horizontal lines to indicate the pitch and time of the music. These symbols are read left to right and the higher the position on the stave, the higher the note or pitch.

![Figure 1. High and low notes (“Reading Music”)](image)

Figure 1 also shows the treble and clef symbols (left corner at the top and bottom) which span the stave lines at the top and bottom, respectively. The treble clef is used to indicate notes that have a higher pitch than those on the bass clef, and there bass clef denotes those notes with a lower pitch than those the treble staves. The notes, according to the treble and stave representations are shown in Figure 2.
Note duration is important because of its relation to rhythm. Since the number of notes in a bar is an explicit representation, “...smaller duration values lead to more notes within a bar (see “Beats”, section 2.1.10), thus making the rhythm faster. The effect of the note durations on rhythm is rather important, as the music genres, generally, abide to rhythmic patterns” (Karydis, 2006).

2.1.2. Keys

The key of a musical piece (or a portion of it) is the dominant pitch in the work (Turton, 2008). While the key of a note is sometimes confused with its pitch of the entire musical piece, it is important to note that a note in a single key indicates all pitches in a pitch class (section 2.1.4). A key is the central note of each piece of music or part of a musical piece. Every key is based on a unique scale which starts and ends on the note that demarcates the next scale's octave. On a piano, “...the key of C-major uses a scale that starts on C and uses only the white keys of the piano. In a piece composed in the key of C, the music is likely to end on the note C, and certain combinations of notes based on C will predominate” (Gehrkens, 1914). In the Western world, keys can be major or minor keys, and these two concepts are tied to the mood of the music, either indicating a happy or sad mood of the music (Turton, 2008).

2.1.3. Chords

When any number of notes (commonly three or four) are sounded together, the combination is called a chord. Playing notes simultaneously conveys a specific harmonic feature for a musical piece, and, as such, chords can either be harmonious or dissonant (Turton, 2008). Chord dissonance and harmony are subjective measures that are dependent on the correspondence between the pitches and the tonal properties in different sections of music (Klapuri and Davy, 2006).

Chords usually accompany by melodies, a combination that imbues the harmony of a musical piece while also enhancing the phrasing (Turton, 2008).

2.1.4. Pitch

Pitch refers to the highness or lowness of a musical sound. It can be measured on a musical
scale from low to high based on the frequency of the sound and is based on the auditory perception of the listener, also making it a subjective musical feature (Klapuri and Davy, 2006). Each pitch level is mapped to the musical scale according the vibration frequency of the auditory tone that the listener hears (Plack, Oxenham, and Fay 2005). In music notation, each pitch is assigned a name based on the first seven letters of the alphabet (A-G), indicating the position of each note relevant to the next note on the scale (Figure 3).

In Western music, a “semitone” is the smallest pitch difference in music notation, such as G to G sharp (Ke, 2004), and is the distance from one pitch class to an adjacent pitch class (Takeuchi and Hulse, 1993). For this research, the notion of pitch will be restricted to the sense defined by Plack, Oxenham, and Fay (2005), that is, “...that attribute of sensation whose variation is associated with musical melodies”.

There are twelve unique pitches starting from the “C” key, as indicated in Figure 3. Each of these unique pitches is known as a pitch class, and is a group of notes that have the same pitch name. Pitch names have two parts, the first being the pitch class (such as C, C# (C sharp)), and the second is the octave representation. In the octave representation, “...the octave runs from C up to B, and middle C (261.6 Hz) is designated C4. The B next to middle C (246.9 Hz) is B3. The C an octave above middle C (523.2 Hz) is C5, the C an octave below middle C (130.8 Hz) is C3, and so on. The pitch A4, used to tune orchestras, corresponds to 440.0 Hz by modern convention” (Takeuchi and Hulse, 1993). For example, all the C notes belong to the same pitch class (C2, C4, and C7 all belong to pitch class "C"), while all the C#s (C sharp notes) belong to another class. Each pitch class groups all pitches in an octave relationship (Thul, 2006). A pitch class set with notes C, E, and G can be represented by the set (0, 4, 7). As seen in Figure 4, the twelve pitches are numbered from 0 to 11 and numbering restarts at 0 when an octave is reached because both the notes at the start and end are C notes, that is, they have the same pitch class and pitch number. The pitch class/variety is a measure of the diversity of the pitch class set used in writing the melody (Gutierrez, 2002).

![Fig. 3. Commonly used notes on keyboard position and their pitch position (“Reading Music”)](image)

![Fig. 4. Basic pitch class representation (Cochrane, n.d.)](image)
“Pitch class equivalence” refers to two pitches that are separated by one or more octaves (Dimond, 2007).

In terms of vocal music, pitch refers to the key in which the singer has to sing. Essentially, the melody of a piece is a sequence of pitches. Pitches are the primary way in which sung musical content is conveyed (Pollastri, 2002; Stegemoller, Skoe, Nicol, Warrier, and Kraus, 2008). In any music piece, vocalization and instrumentation are interrelated in the sense that the higher the notes/keys that are being played in a musical piece, the higher the notes that the vocalist has to sing. That is, vocalists have to match the pitch of instruments. Unlike other instruments, the human voice is not percussive and can deviate from the precise notes (that is, pitch values). Klapuri and Davy (2006) explain that “when...singing..., for example, both intentional and unintentional deviations take place from the nominal note pitches”.

Pitch is one of the most useful feature attributes for classifying musical genres. For example, “...jazz music tends to have more chord changes, classical music has a large variability of harmonic properties and in rock music high-pitched notes are mainly absent and they seldom exhibit a high degree of harmonic variation” (Heittola, 2003).

2.1.5. Melody

The melody of the music (also known as the tune) is a complete musical phrase in a musical piece, represented by a sequence of meaningful notes (such as for understanding the mood of the piece) (Chakraborty et al., 2010). This is in contrast with a musical segment which is also a sequence of notes, but which is not a complete phrase/entity and is not meaningful for understanding a musical piece. Melody is the harmonious progression of a number of pitches that are consistent, cohesive and audibly defined enough to be considered as musical and not just as noise (for example, discordant sounds) (Randel, 2003). The melody of a musical piece is usually intended to be sung and as such, the vocal range of the human voice is an important, related concept (discussed in Section 2.1.11).

2.1.6. Harmony

Harmony is the most emphasized and highly developed element in Western music. The harmony of a musical piece refers to notes played simultaneously to imbue the piece with a certain artistic coherence. It is a group of notes that are played behind, beneath, and around the melody and most harmony is based on chords (Schmidt-Jones, 2011). Harmony supports the melody and gives the music texture or mood. The sound of a musical piece can be changed by changing the harmony alone. There are different types of harmonic textures, two of them being monophony (most relevant to this thesis) and polyphony. In monophony, a melody all by itself can have an implied harmony, even if no other notes are sounding at the same time. In other words, the melody can be constructed so that it strongly suggests a harmony that could accompany it. Some melodies are not meant to be played with harmony such as certain types of modern art music and Native American music. By contrast, polyphony refers to a texture of music in which there is more than one independent melodic line at the same time.
2.1.7. Rhythm

Rhythm indicates the duration of each note used in a musical piece. It is based on frequent occurrences of beats using a system of measures. Rhythm is based on the number of notes played at a specific tempo in a section of music (Karydis, 2006). “Since rhythm implies continuity, there must usually be at least two...measure groups in order to make musical rhythm possible” (Gehrken, 1914). There are two different levels of rhythm within music, with a higher level comprised of bars, measures, and beats, and a lower level comprised of tatum (Heittola, 2003). Rhythm is useful because the duration of notes can be used to capture changes in pitch. As well, Heittola (2003) notes that it is especially useful for genre classification:

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

2.1.8. Intervals

An interval is the correlation between two tones based on their pitch. Intervals are measured by the number of pitches between each semitone and is represented by ordinals (for example, 2 = second, 5 = fifth, et cetera) (Turton, 2008). An octave is a special type of interval, based on twelve semitones. Intervals come in many sizes within an octave:

- Same Note (count = 1) Unison
- 1 step (count = 2) Second
- 2 steps (count = 3) Third
- 3 steps (count = 4) Fourth
- 4 steps (count = 5) Fifth
- 5 steps (count = 6) Sixth
- 6 steps (count = 7) Seventh
- 7 steps (count = 8) Octave (Back to the same note name)

Two tones sounding together are usually termed a harmonic interval, and as a melodic interval if the tones are sounded consecutively (Gehrken, 1914). The melodic interval is the distance between the separate notes on the melodic line. Perfect intervals can be found in Gregorian vocal music which features monks singing in “...two voices or more moving in perfect musical intervals which are the Unisons, 4ths, 5ths and octaves” (Gehrken, 1914).

The relationship between chords and intervals is that chords are combinations of three or more tones/notes sounded simultaneously for which the distances (intervals) between the tones are based on a particular formula/pattern.

2.1.9. Octaves

An octave is defined as “a series of eight notes occupying the interval between (and including) two notes, one having twice or half the frequency of vibration of the other” (“Octave”, n.d.). To the human ear, octaves are two notes that sound the same though one note sounds like it has a higher pitch.
Intervals exist between a specific musical pitch and another pitch with half or double its frequency, a ratio of 2:1 (Pitkow, 2000).

Music theory from the Western world identifies octave equivalency as a musical gold standard in which notes corresponding by one or more octaves work the same way in the harmonic or melodic structure of a musical piece. Pitches that are written in ascending order “…use a circular scale in which notes separated by octaves are given the same name: DO-re-mi-fa-so-la-ti-DO in the West, SA-re-ga-ma-pa-dha-ni-SA in south Asia. Octaves just seem to sound the same” (Pitkow, 2000).

Musical melodies are often designed to be sung and the typical human vocal range is about two octaves (Ke, 2002). Banda (2000) highlights the usefulness of octaves in differentiating vocals with an accessible example, noting that the American national anthem, “The Star-Spangled Banner” has a high technical requirement for singers: “[It] has a vocal range of an octave plus a 5th; that is, an octave plus the fifth note of the next, higher octave. Even professional singers have problems singing this song. Different voices sing comfortably in different ranges”.

### 2.1.10. Beats

The beats in any musical piece comprise a constant rhythmic pattern that is used to count time as the piece progresses from beginning to end (Gehrken, 1914). As such, beats are a set measurement of time and are used to describe the length of notes. Groups of beats are called a bar or a measure, while a meter is used to refer to groups of strong and weak beats.

With regard to the rhythmic aspects of beats, a crochet is a note held for one beat (quarter note), a dotted note is a note that is held for one and half times its original length, while a minim is note held for two beats (also known as a half note) (Turton, 2008). A quaver note is held for half a beat (also known as an eighth note).

The time (or meter) signature is found at the start of musical pieces in the form of two numbers after the clef mark, with the top number representing the amount of beats per measure and the bottom showing which note the beat belongs to. Two half notes or four quarter notes equal a whole note and, “for the most popular time signature, 4/4, there are 4 quarter notes per measure” (Ke, 2002), as indicated in Table 2.

<table>
<thead>
<tr>
<th>Bottom Number</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Whole note</td>
</tr>
<tr>
<td>2</td>
<td>Half note</td>
</tr>
<tr>
<td>4</td>
<td>Quarter note</td>
</tr>
<tr>
<td>8</td>
<td>Eighth note</td>
</tr>
<tr>
<td>16</td>
<td>Sixteenth note</td>
</tr>
</tbody>
</table>

**Table 2.** Notes per measure

Of the different kinds of time signatures, the most widely used are 4/4 meter which uses four beats per
measure (called “common time”), and the 3/4 and 6/8 meters – mostly found in classical music – which use three beats and six beats per measure, respectively.

Beats can exist in simple or compound time based on how the beat is divided. Simple time divides the beat in two equal sections, while compound divides it into three. “All meters where the lower number is both larger than, and divisible by 3 are compound (6, 9, 12, etc.). All meters where the upper number is an even number that is not divisible by 3 (2, 4, 8, etc.) are simple” (“Music”, n.d.).

A duple refers to two beats in a measure, while a triple is three beats in the measure. Different beat patterns also exists such as quadruples and quintuples, with quadruples being a common beat pattern. A quadruple is two duples with the importance on the first measure (“Music Theory Basics”, n.d.). For example, with a quadruple there is an emphasis on both the first and third beat, but beat one is stronger than beat three.

McKay (2002) has found that beat patterns are different for different genres of music:

...Clear periodicities that are often multiples of each other is typical of Punk music...
Other types of music demonstrated very different patterns in their beat histograms. Techno, for example, often had very clearly defined beats...[while] Modern Classical music, to cite another example, often had much less clearly defined beats.

2.1.11. Vocal Range

One hypothesis of this thesis is that there are differences in the vocalizations amongst different genres of music. Of specific relevance here is the researcher's anecdotal observation that the vocal range of Classical singers is generally higher than that of vocalists in other genres of music, such as Pop and Rock.

Vocal range is the distance between the highest and lowest notes that the human voice can match. Figure 5 illustrates the general vocal ranges associated with different voice type with the use pitch notation where middle C is equivalent to C4.

![Fig. 5. Stave representation of vocal ranges with the clef symbol which is used on all note lies to indicate the position of middle C (“C clef”, n.d.)](image)

Some singers within these voice types may be able to sing somewhat higher or lower. However, it should be noted that within some classical forms of vocalization such as opera, vocalists do not generally and are often not required to sing certain notes. This is because opera singers have to perform without a microphone (Appelman, 1986), so they only sing notes that are audibly harmonious...
to audiences when sung in accompaniment with an orchestra. Appelman notes that “in contrast, a pop artist could include notes that could be heard with the aid of a microphone.”

The normal range of the human voice is from lowest or bass voice to highest or soprano voice. This range corresponds to an octave range of about two octaves (Ke, 2004). The human voice reaches its depth at the note F4 and peaks at the highest or soprano voice at note C6. There are various differences between classical and non-classical vocal music which have been identified. In some cases the differences are hard to determine since some singers of non-classical music may sing in the same range as classical vocalists: “...Mariah Carey's vocal range spans approximately 123-830 Hz (B2-G#5). Typical pop singers cover a far less impressive range than that” (Izhaki, 2007). Therefore, the features that will be selected from the MusicXML representation needs to be sufficiently representative in order to compensate for this ambiguity. The similarity between vocalization features in different genres is not the only complexity in classifying music by genre. A careful examination of the different musical features that can be extracted and examined to determine the genre of a musical piece reveals ambiguities related to pitch, melody, and tempo (Chapter 5). As an example, it has been found that there is a high correlation between pitch and vocal range.

2.2. Music Genre

The genre of a musical piece refers to the category of the piece based on parameters such as its tempo, place of origin, and chords. Music can be classified according to the timbral, spectral and harmonic information available for music pieces, but the decision about which combination of these parameters are indicative of a specific genre is still very much a subjective measure (Alluri, 2007).

McKay and Fujinaga (2006) differentiate musical “genre” from “style” by identifying that the former is more directly related to cultural aspects/differences that are either generally acknowledged by a community as representative, while the latter can be differentiated by the features that are distinguished in the music according to the agents involved in its creation such as the composer or the musicians involved in the creation of the piece of music. Specifically, “genre” can be defined as “a kind of music, as it is acknowledged by a community for any reason or purpose or criteria, i.e., a set of musical events whose course is governed by rules (of any kind) accepted by a community” (McKay and Fujinaga, 2006).

The two genres that will be used for classification in this study are Western Classical and Modern Pop.

2.2.1. Western Classical Music

Music has developed over the ages in tandem with changes in societies, including economic, social, and religious changes. Western Classical music can be defined as music that has evolved from specific eras and composers throughout specific historic times in Europe and, if one is to be liberal, North America. Very rigid historical definitions of classical music only include music composed circa 1770-1830 (the Classical period), but this thesis will be more expansive in defining this genre. Western Classical music derives from a particular system of European musical education and tradition and is commonly precisely created for each specific instrument that is part of the musical piece (Heittola, 2003). Classical music is distinct in that it is “.... usually performed by an orchestra, a large group of musicians playing a variety of different instruments, or by a chamber group, a smaller group of
musicians” (Heittola, 2003) and majority of any given musical piece is strictly adhered to, with only a few elements allowed for interpretation by the performers.

This category of music can be divided into several historical eras, varying widely according to researcher. Though there exists some overlap in the definition of the different periods of music to be included in the genre of Western Classical, major sub-categories include Renaissance, Baroque, Classicism, and Romanticism (Naxos, no date). The categories, as well as the major musical figures within each era are illustrated in Figure 6.

Fig. 6. Time periods for Western classical music (Naxos, no date)

2.2.1. Renaissance

The Renaissance (also called “Early”) music category dates from the ninth century and before and focuses on Christian musical pieces (Alluri, 2007). The features of this music include the singing of Latin religious texts without the accompaniment of instruments, with no fixed rhythm. These musical pieces were sung almost exclusively by young males (female voices were introduced towards the end of the period), a style popular referred to as the Gregorian chant. Named for Pope Gregory the First, these chants were compiled and adapted by him for the religious services of the time. A single note or key was the anchoring element in the musical pieces of this period. “The keys used during that time were known as ‘modes’ corresponding roughly to the scales starting on the white keys of the piano” (Alluri, 2007). The songs/chants during the Renaissance period employed four modes, as
opposed to the two modes used in modern music. The more important and ostentatious the religious occasion (such as during Christmas), the more elaborate the melodic style would be. The lack of instrumentation and the preponderance of a single note within Early Music pieces make this style very distinctive.

2.2.1.2. Baroque

The Baroque period spans music created from approximately 1600 to 1750 (Alluri, 2007). Baroque composers were able to deviate from the standards of the Renaissance period, primarily due to the waning influence of the Church. Instrumental music was radically developed and the 12-interval (that is, octave) system was introduced, leading to structural changes in harmony. A major vocal technique used during this period was counterpoint, which is the “... superposition of voices that are independent in rhythm, harmony and contour” (Alluri, 2007). Within Baroque musical pieces there was a limited dynamic range which restricted the level of the music. Once a level was chosen at the start for the pitch and other musical elements, there would be little deviation for the duration of the piece. This music featured a continuous progression of notes used to aggrandize the piece, and a discrete playing style. Unlike Renaissance music, rigid rhythmic elements and meters were adhered to. The opera and cantata musical styles were developed in the Baroque era, and this heralded a change in vocal music.

2.2.1.3. Classicism

The Classical period of music lasted from approximately 1750 to 1820. Musical elements such as rhythmic variations, musical tonality, and a wider dynamic range dominated classical pieces, giving rise to greater flexibility, diversity, and contrasts in the style (Alluri, 2007). Musical form became rigidly defined, giving rise to musical forms such as sonatas, symphonies, and concertos. In contrast with the previous eras, compositions became simpler with more emphasis on melody and a lighter feel, and were also characterized by the use of “...modulation to create tension and release, and harmonic rhythm to define large-scale forms” (Alluri, 2007).

2.2.1.4. Romanticism

Spanning the time period from approximately 1820 to 1910, the Romantic period saw composers using their artistry for the love of music rather than for the adherence to rigid musical standards based on form and purpose (Alluri, 2007). This genre is characterized by “complex harmonies, uneven changes in rhythm and meter, a large increase in dynamic range to enhance spontaneity” (Alluri, 2007). Short melodies with recurring themes (leitmotif) and very long pieces with fancy instrumentation were developed, and diverse tones were created by playing instruments differently.

2.2.2. Modern Pop Music

Modern Pop music is quite a large category of music and includes many sub-categories such as Dance/Electronic music, Hip-Hop, Jazz/Blues, and Rock/Pop. The Modern Pop music used in this study will include pieces from all the following categories.

2.2.2.1. Dance

Dance music evolved from experiments with electronic instruments by musicians such as
Kraftwerk (Heittola, 2003). Their experiments and the introduction of the synthesizer (which made
music creation more accessible) and other instruments such as samplers led to the widespread use of
these instruments in popular music. The music in this sub-category is characterized by strong beats and
a strong rhythmic structure.

2.2.2.2. Jazz and Blues

Jazz and Blues music originated in Afro-American communities as traditional folk music in the
early 1900's. It is a mix traditional African and European folk music forms. Blues music is known
for its “blues notes” which have a uniform structure, the most common being the the twelve-bar blues
form “...where three four-measure long phrases are repeated” (Heittola, 2003). Jazz evolved from
Blues and is characterized by unique music styling by the performer (improvisation), and “...a
propulsive rhythm, melodic freedom, and improvisation solos” (Heittola, 2003).

2.2.2.3. Pop/Rock

The category of Pop/Rock is a wide one which is generally used to described modern popular
music. The genre label is still useful, however, for capturing a class of music with unique
characteristics from other sub-genres. Pop music is characterized by “...short and simple melodic
pieces having some catching tune to make them easily memorable” (Heittola, 2003). Rock music has
specific provenance, having originated from Blues and Country music. It is characterized by strong
beats and the dominant use of the electric guitars.

2.2.2.4. Soul/R&B/Funk

The final sub-category of Soul/R&B/Funk music encompasses musical pieces that evolved
when Blues music was popularized in urban locations. R&B music produced in the 1960's and 1970's
can almost exclusively be categorized as “soul” music, and has distinct vocal intensity. Soul music
grew into Funk when Rock music influences and heavier grooves and beats were added, while more
modern R&B is now more similar to Pop and Hip Hop music. “However, there are still some
characteristic features like smooth vocals with a bouncy beat” (Heittola, 2003).

2.2.3. Classical Music versus Pop Music

Modern Popular music is not as restricted to the strict technical requirements that are imposed
on other genres of music such as Classical (Browne and Skinner, 2006). The fact that most modern
vocalists use only about two octaves of their vocal range is undeniably related to the developments in
the music/sound engineering industry that have led to changes in the way that musical pieces are
processed before release to the public. Smith (2005) notes that “creative developments and advanced
technology have led to a very formulated song structure, constructed from layering vocals, backing
vocals and harmonies to create an arrangement that is constantly moving”. Songs within the Modern
Pop music genre have become over-produced, and it is often the case that a pop song is “enhanced”
through a process of of layering different kinds of vocals throughout the song. Vocal samples can be
modified to match time changes/signatures, pitch, loudness and other musical features (Lau, 2002). As
demonstrated in Figure 7, these layers consist of “…doubled chorus sections and a high vocal double
brought in at the end” (Smith, 2005).
These techniques are used to create more complex and challenging vocal pieces with the ultimate goal of providing a greater impact on the listener. They are not the product of the vocal precision, but can be created solely with software such as Pro Tools and Autotune (Koldestam, 2004; Tyrangiel, 2009; Smith, 2005). Tyrangiel (2009) points out one notable example:

If you haven't been listening to pop radio in the past few months, you've missed the rise of two seemingly opposing trends. In a medium in which mediocre singing has never been a bar to entry, a lot of pop vocals suddenly sound great. Better than great: note- and pitch-perfect, as if there's been an unspoken tightening of standards at record labels or an evolutionary leap in the development of vocal cords. At the other extreme are a few hip-hop singers who also hit their notes but with a precision so exaggerated that on first listen, their songs sound comically artificial, like a chorus of ‘50s robots singing Motown. The force behind both trends is an ingenious plug-in called Auto-Tune, a downloadable studio trick that can take a vocal and instantly nudge it onto the proper note or move it to the correct pitch. …Auto-Tune doesn't make it possible for just anyone to sing like a pro, but used as its creator intended, it can transform a wavering performance into something technically flawless. "Right now, if you listen to pop, everything is in perfect pitch, perfect time and perfect tune," says producer Rick Rubin. "That's how ubiquitous Auto-Tune is".

Examples of the adjustments that can be made to vocal tracks include:

- Increasing the vocal range
- Adjustments to tone and timbre, “...to help shape and mold the recorded sound into a desirable vocal sound” (Smith, 2005).
- Pitch manipulation - Since “one of the most crucial elements of a vocal performance is the pitch of the sound...” (Smith, 2005), pitch correction systems such as Antares Autotune are used to track the pitch of the singer's voice and adjust it to the correct note.
Techniques such as pitch, tone and timbre correction all aim to hide the singer's inabilities, and the technological advances have created tools that are faster and more accurate at this job. Many Pop/Rock singers have difficulty achieving a certain level of musical performance, a problem often characterized by pitchiness (Koldestam, 2004). Tyrangiel (2009) concurs, noting that “...many beloved songs are actually off-pitch or out of tune. There's Ringo Starr on "With a Little Help from My Friends," of course, and just about every blues song slides into notes as opposed to hitting them dead on”.

With Classical music, there is a more rigid standard for technical execution. This emanates from different agents within the Classical music scene such as composers who want to achieve and display technical achievement in their compositions and classical performers with the same goals. Classical singing is still regarded as a gold standard of vocal performance though it is its popularity has waned since its heyday (Wilson, 2002). As Miller (2002) notes, “The ideal in the 'classical' singing voice includes an extended pitch range with a 'perfect scale, or the ability to move through a series of pitches with no audible discontinuity in voice quality” (Miller, 2002). One way in which this gold standard is manifest is in the duration, breadth and eventual scope specificity required of Classical musicians (Wilson, 2002; Kovačić, Boersma, and Domitrović, 2003; Stegemoller et al., 2008). Their education involves the technical mastery of complex areas of music including “...proficiency in sight-reading and ensemble playing, thorough understanding of tonal and harmonic principles, knowledge of performance practice, and a familiarity with the style/musical idiom inherent to a given period, composer or musical work” (Wilson, 2002; “Classical Music”, n.d.). Whereas with Modern Pop music where much of the complexity is not being evoked or created by the singer but rather by artificial tools, Western Classical pieces demonstrate complexity by applying an appropriate range of musical attributes such as “...thematic development, phrasing, harmonization, modulation (change of key), texture, and, of course, musical form itself” (“Classical Music: Reference”, n.d.).

Now that the musical concepts have been explained in detail, relevant theories surrounding machine learning will be explained, since the experiments will involve the analysis of musical parameters through the application of machine learning techniques.
Chapter 3. Machine Learning

Machine learning involves a computer system using different methods to learn how to perform a distinct task (such as music classification) on its own. This automatic process employs algorithms in conjunction with other techniques to obtain information on a “...statistical and computational basis” (Kanters, 2009). A system generally engages in learning by obtaining information from a training cycle and by applying that information to different instances/events during the repetition of the same cycle in order to improve task performance. Extrapolating information about known instances (training set) to new instances is done by checking for patterns/similarities between the training set and the test set. By combining these patterns with existing taxonomies or by using them to create new taxonomies, meanings can be ascribed and the resulting genre categories can be used to classify new music pieces (McKay and Fujinaga, 2006). This method is akin to the machine learning technique of instance/concept learning whereby a pre-established space of possible hypotheses is searched for the hypothesis/pattern/class that best suits the training set (Mitchell, 1997 in Kanters, 2009). In learning from hypotheses, the inductive learning approach “...assumes that any hypothesis found to approximate the target function well over a sufficiently large set of training examples will also approximate the target function well over other unobserved examples” (Kanters, 2009). This theory supports the idea that the genre of a musical piece can be assigned/inferred by extrapolating the patterns found in known/training examples to new data.

The taxonomy for machine learning algorithms is based on the desired outcome of the algorithm. Two important elements of this taxonomy are supervised and unsupervised learning, and they are discussed in the next sections of this Chapter.

3.1. Automatic Classification of Music by Genre

While the automatic classification of music by genre can be a difficult process, the use of advanced machine learning and pattern recognition techniques greatly improves the accuracy of classification systems. Techniques used can involve the high- or low-level features as input to the process, after which they are processed by classification algorithms.

3.1.1. Classification Algorithms

Features or combinations of features such as pitch and melody can be classified by a variety of data mining techniques and algorithms. These techniques, which can be either supervised or unsupervised, are used to classify music by genre in a high dimensional feature space.

3.1.1.1. Unsupervised Classification

Unsupervised methods use feature data to determine a genre without any prior information about the genre itself, and only observes where the data lies in high-dimensional feature space (Chu, 2010). As new musical data is introduced, the same method is used each time in order to see if data forms clusters in the feature space, again without being able to identify the genre. These clusters find commonalities between musical pieces, but are unable to classify them into a specific genre. Unsupervised methods group items together based on a statistical or arithmetic model that is generally
different from a human system of understanding (such as the concept of genres, which is a human, not digital concept). Agglomerative hierarchical classification and K-Means are two unsupervised methods used to classify music by genre.

When K-Means is applied, each feature spanning a high-dimensional feature space is regarded as a feature set. To determine the feature sets, first, an initialized, \( k \) number of centroids are given as central/cluster points amongst features scattered in the high-dimensional feature space. Individual, scattered features are then grouped as belonging to specific centroids based on a calculation that reduces the distances between centroids and individual points through the use of the mean squared distance. New centroids for these feature sets/clusters are then calculated, and the entire process is repeated until the centroids of feature sets no longer change. The final clusters will contain features that display similar patterns, but there is no information about the category that they belong to.

In agglomerative hierarchical classification, clusters are created using a tree structure. Each point in the high-dimensional feature space is considered to be a cluster to which each cluster/point is compared to see which has the shortest distance to the initial point/cluster. An initial cluster and the clusters with the shortest distance from it are grouped to create a new cluster/point. The process is repeated until “until the distance between clusters is smaller than the threshold or the clusters number is enough” (Chu, 2010). Points are more similar based on their depth within the tree structure, and this is very useful for applications like music recommendation.

While the agglomerative algorithm may seem to yield more fruitful results than K-Means, all unsupervised methods possess similar disadvantages. Similarity measures are problematic for genre classification because some genres (such as Rock and Pop) will yield very similar classification results. The “…resulting clusters may be arbitrary and unable to define their genres” (Chu, 2010). Therefore, unsupervised approaches are rarely effective and seldom used.

3.1.1.2. Supervised Classification

Unlike unsupervised approaches to music genre classification, supervised approaches do not rely on clustering based on similarity, and start out with some level of information about the music genres. Supervised approaches are trained on data that has been hand-labeled according to genre (Chu, 2010). New, unlabeled data is introduced and the data can be assigned to a genre that the system knows, based on the parameters of the labeled data. Three prominent supervised approaches are support vector machines, K-nearest neighbor, and Naïve Bayes.

The Naïve Bayes algorithm takes the training data for the classifier and tests it statistically. The algorithm outputs data outlining the estimated maximum likelihood that an item belongs to different categories and “…maximizes conditional probabilities on the observed feature values as decision criteria” (Basili, Serafini, and Stellato, 2004).

Support vector machines try to design a strategy whereby the separation/margin between different types of data is maximal. A model of the data in the feature space is created to optimize the high-dimensionality of the space and, by extension, the results (Chu, 2010). The kernel which does this work of increasing dimensionality calculates the distance between points in the feature space, with points with the least distance between them belonging to the same class/category.

For the K-nearest neighbor (KNN) algorithm, data is classified according to the number of \( K \)
neighbors closest to it. It is a rule-based algorithm (Basili, Serafini, and Stellato, 2004). Points representing data that is labeled by category/class are scattered in the high-dimensional feature space (Chu, 2010). A point that will be classified is marked within this space, and, as with K-Means, a value for the $K$ nearest neighbors is initialized to indicate how many neighbors will be classified. Measurements of the distance between all the points are taken, most commonly the Euclidean distance. The points of the five closest neighbors are linked to the data point to be classified. Finally, the point is assigned a genre corresponding to the nearest neighbors with the most votes, that is, the class that has a higher representation amongst the points (Chu, 2010; Kanters, 2009). Therefore, KNN employs instance/concept learning.

3.2. Classification Problems

Both supervised and unsupervised approaches are susceptible to some problems which may make music genre classification difficult. The two methods assign a specific genre to the musical piece, but this may not be a reliable classification because of the similarities and ambiguities amongst genres (Chu, 2010). Called a hard-decision problem, this situation can be remedied by the use of fuzzy c-means. Instead of specifying a specific genre, for each musical track, this method assigns percentages of similarity to a variety of genres such as “...Metal: 50.1%, Classical: 49.9%”, so that the intersections/overlaps are revealed (Chu, 2010). However, this revelation does not simplify the problem of music genre classification and is rarely used.

Overfitting is another problem that can occur with both supervised and unsupervised classification methods. It occurs when the system used for classification is modeled too closely to training data, making it difficult for the system to accurately classify new data/examples. In essence, overfitting occurs because the existing system cannot find a fit for new data that matches the parameters of existing training data. The solution to this problem is to design a system that “…find[s] the right trade-off for the generalization: having a system that is complex enough to capture the differences in the implied models, but simple enough to avoid overfitting” (Chu, 2010).

In addition to solving problems such as overfitting, there are methods that can be used to determine how well a supervised or unsupervised system is able to classify unseen/unknown data. Cross-validation is one such method and it works by partitioning data into $k$ folds, facilitating the classification of data on a single subset/fold - which is designated as the test set - instead of the entire data pool. In this way, for each cycle of classification tasks, a fold can be reserved for the test set, while all other data subsets will be used as the training set (Kanters, 2009).

3.3. TiMBL

TiMBL, the Tilburg Memory Based Learner, is an automatic classification software. The software is an example of supervised learning/classification. It works with a representative training set that is stored in computer memory and classifications are made by finding patterns in unseen data that is similar to the training set (Kanters, 2009; Le, 2003). Using “...discrete data...new classifications are made by extrapolating from the most similar stored cases” (Kanters, 2009). In this way, TiMBL's classification algorithms work similarly to the k-Nearest Neighbor algorithm, and is, therefore, an example of both instance learning and the inductive learning hypothesis at work.

In addition to the KNN classification kernel that is the basis of the software, Le (2003) notes
that TiMBL includes several other features, including:

- Feature weighting for dealing with features of differing importance: information gain, gain ratio, chi-squared, shared variance metric (MVDM) for making graded guesses of the match between symbolic values
- Class voting in the k-NN kernel according to distance of nearest neighbors (inversed, inverse linear, exponentially decayed)
- Optimizations for fast classification: conversion of case base to tree, inverted indexing
- Even faster classification using heuristic approximations: the IGTREE decision tree algorithm and the TRIBL hybrid
- Fast leave-one-out testing and internal \( n \)-fold cross-validation

The experiments in this thesis will use TiMBL, but with the default settings.

Now that the musical and machine learning concepts have been explained, the discussion will now move to the experiment components that will be analyzed within the context of these concepts.
Chapter 4. Research Problem

The discussion will now move to the experimental setup, focusing on how the hypothesis of this thesis will be explored, that is, the way in which testing will be executed and results obtained and evaluated. For experiments involving the automatic classification of music by genre, the major tasks are the preparation of the dataset for analysis, the selection of features to be used, an evaluation of the work of the classification algorithm used by the classifier, the work of the classifier on unseen data, and the results.

4.1. Research Question

In this research, the goal is to differentiate the task for the correct classification of Western Classical and Modern Pop music based on the vocal part of a musical piece. The central question is whether the note-duration symbols perform better than time slice representations.

4.2. Research Design

As demonstrated in Figure 8 below (which describes the detection of chronological expressions), the classification system employed in this research will involve the setup and training of the system and the subsequent use of the resulting model for the categorization of unseen data.

![Diagram](Fig. 8. Memory based machine learning (Paijmans, 2010))
4.2.1. Dataset Preparation

The major step in this stage is the collection of music scores for the different genres. Once obtained, the scores will be tagged according to their appropriate genre category. The two genres for classification are Western Classical and Modern Pop. Western Classical music will include the sub-categories of Renaissance, Baroque, Classicism, and Romanticism, while Modern Pop music will include music that fits into the Dance, Jazz and Blues, Pop/Rock, and Soul/R&B/Funk sub-categories.

Some important aspects of the dataset include:

• Dataset type – The data is in MusicXML format.
• Dataset size – Many research studies employ larger datasets, but even datasets with few elements (50 files or less) can be representative (Guaus, 2009). Larger datasets can produce better results, but that is not always the case. Guaus (2009) explains that smaller file sets “may better represent the genre space than a large number of ambiguous files [so] maybe it is enough to train a system with a few representative...excerpts per class”.
• Genre specificity – The work of the classifier can be affected by the specificity of the genres included in the dataset. The use of more general genres produce better results than more narrow categories. Differentiating between Pop, Rock and Folk music is more difficult than doing the same for Classical, Jazz, and Rock categories.
• Dataset variation – Maximal variability can be achieved by ensuring that each specific genre used in the study includes a good cross-section and sub-genres for the genre category. The best way of doing so is to only include one file per performer per album for each genre. This prevents a classic overfitting problem know as “the producer effect”, a problem in which a classifier can become biased to specific features which are representative of the album but not of the genre category (Guaus, 2009). For this thesis, this standard was achieved for the Modern Pop data used in the experiments by using only one composer per song. The same results were not achieved for the Western Classical pieces due to the difficulty in obtaining vocal pieces in MusicXML format.
• Dataset balance – Each genre category should have a similar number of songs because unbalanced datasets bias the classifier regardless of the strength/suitability of the features extracted and the classification system used.

4.2.2. Feature Extraction

The features that are selected for input to an automatic classification system are important because the classification task is dependent on these features. A classifier that has a gold-standard design will not perform optimally if appropriate features are not selected for its classification (McKay and Fujinaga, 2005). A poor selection of features can result in musical pieces being assigned to incorrect genre classes. The features selected and implemented in this study will be based on original combinations as well as those gleaned from a survey of previous research.

During selection, a good range of features must be allocated to provide the most useful input to the classifier. They must be filtered to omit irrelevant or redundant information that will hamper the classification process. The features must have a general scope to ensure that they are applicable to
various genres, and even styles, of music. “Simple and descriptive measurements of overall characteristics and variances in a piece of music can be particularly effective in this respect, and tend to be more general than features derived from sophisticated theoretical analyses” (McKay and Fujinaga, 2005). Features that suit this description can be broadly grouped into categories related to the melody, instrumentation, texture, rhythm, dynamics, and pitch attributes of a musical piece (McKay and Fujinaga, 2005; Basili, Serafini, and Stellato, 2004). To test the hypothesis of this thesis, the focus will be on the use of simple features which have been proven to provide a first level for eliciting information that is useful for music genre classification, that is the note-duration symbols and time slices.

Feature selection affects the work of a classification system by influencing its ability to select the musical properties that are relevant to the task (that is, the task based on either note/pitch duration symbols or the time slices method) while at the same time retaining information that is important for genre identification. The features selected all display the attributes discussed in this section and have also been used by other researches for symbolic genre classification.

4.2.2.1. Feature Types

Features can be selected from audio formats such as audio files on compact discs or MPEG and MP3 files. They can also be selected from symbolic and visual formats (utilizing optical character recognition) and represented as such in different ways such as MIDI and MusicXML files. Features can be categorized as metadata, such as composer, and year of publication), making them high-level features in representations that are more meaningful to humans and include features such as tempo and metre (McKay and Fujinaga, 2005). Finally low-level features are in their most raw, signal processing format and their low-level attributes focus on representations of music in base forms such as spectral frequency ratios and zero crossings. High-level, symbolic features are extremely hard to obtain from audio recordings, while it is difficult to use low-level features to extrapolate musical information that is theoretically meaningful. Therefore, in this thesis, high-level features will be extracted from the symbolic MusicXML format in order to have an optimal selection of high-level features for use in a classification task.

For their research into the automatic classification of genre, McKay and Fujinaga (2005) used classification software to analyze high-level music features. Numerous music files in MIDI format and from different musical genres were manually categorized and their high-level musical features were extracted and, using an algorithm, weighted to assign different levels of importance to each one. The features were then used to train and test a classifier. This is the same method that will be used for the research described in this thesis with the exception that the feature files will be in MusicXML and not MIDI format.

4.2.2.2. MusicXML

XML (eXtensible Markup Language) is a programming language structures and formats text by using annotations (McKay, 2010). Most often associated with the creation of web pages, XML allows users to structure data in specific ways for optimal processing by computers.

The representation of music in MusicXML format is a method designed to make musical scores machine-readable (Simoes, Louenco, and Joao Almeida, 2007), as demonstrated in Figures 9 and 10 below. This format offers the advantage of “...strict hierarchical organization, declarative approach,
clear distinction between syntax and semantics, intelligibility, modularity, editability, extensibility, implementability, reliability” (Ludovico, 2006). XML uses tags to differentiate different music elements using music syntax.

![Fig. 9. Whole note middle C (Good, 2002)](image)

```xml
<key>
  <fifths>0</fifths>
</key>
<time>
  <beats>4</beats>
  <beat-type>4</beat-type>
</time>
<clef>
  <sign>G</sign>
  <line>2</line>
</clef>
</attributes>
<note>
  <pitch>
    <step>C</step>
    <octave>4</octave>
  </pitch>
  <duration>4</duration>
  <type>whole</type>
</note>
```

![Fig. 10. Portion of XML representation of middle C note in XML (“MusicXML”, n.d.)](image)

The scope of this thesis does not include a complete explanation of MusicXML notation for music. What is important to note is that this and other representative forms extract various features from music scores. Features such as the notes, pitches, beats, and time signatures, are demonstrated, respectively, by the XML tags <note>, <pitch>, <beat>, and <time> (Figure 10). Notes are defined using letter notation A, B, C, D, E, F, G, and are separated by spaces (Alethis.net, n.d.). Relative lengths of notes are indicated by adding a numerical value after the note letter, for example:

- 1 = whole note
- 2 = half note
- 4 = quarter note
- 8 = eighth note
- 16 = sixteenth note

The <key> tag defines the key signature: C major, A minor, et cetera. (“C clef”, n.d.).

### 4.2.2.3. Feature Dimensionality

Features will all be analyzed on two levels. First, one-dimensionally, taking each feature in isolation to take advantage of individual feature characteristics.
The approach begins with sliding windows of variable sizes. The first series of experiments will use note-duration symbols, both combined and separate.

- A#4_0.5 – combined
- A#4 0.5 – separated

The second series involves series of time slices of varying “thickness”. For example, a sample of 128 pulses/beats can be used (one measure is separated to 256 parts, explained in Chapter 6), so:

- A#4 A#4 C-3 C-3

The research task/problem has now been fleshed out, so the next chapter will describe the musical data that can potentially be used in the classification task.
Chapter 5. Classification Task

After close analysis of the research considerations within the context of the research problem, the focus now turns to the design of the experiments to be performed. The specifics of the classification task in relation to the setup of the experiment (including the dataset used) are discussed, as well as the range of features that can be extracted from the MusicXML data.

5.1. Experimental Setup

In order to compare the features, the length of each feature record in the comparison pair needs to be approximately the same length. Each experiment requires the researcher to select a feature database of type 1, compute its accuracy, then take a feature of type 2 of approximately the same length, compute its accuracy, and finally, check to see if there are any significant differences. The following section describes the type of musical data that can potentially be used in these experiments.

5.2. Data

Two datasets will be used in the experiments. The first is comprised of files belonging to the Western Classical genre, and the second will have files from the Modern Pop category. The files were obtained from online databases of MusicXML files such as Wikifonia.org and MuseScore.com. Both genres are divided into four sub-genres each (although the sub-genres are not treated as separate classes).

5.2.1. Western Classical

The Western Classical genre is composed of 400 full-length music pieces distributed across 4 genres. The four sub-genres are Renaissance, Baroque, Classicism, and Romanticism. In Renaissance music there is a lack of instrumentation other than the voice and a single note dominates the piece. Baroque pieces have a limited dynamic range for pitch, other musical elements and feature a continuous progression of notes used to aggrandize the piece, and there is an adherence to strict rhythmic elements and meters. The Classicism genre has musical elements such as rhythmic variations and tonality, a wider dynamic range, and more emphasis on melody. Music from the Romantic period feature complex harmonies, uneven changes in rhythm and meter, a large increase in dynamic range, and short melodies.

5.2.2. Modern Pop

The Modern Pop category is composed of 400 full-length music pieces, uniformly distributed over 4 genres. The sub-genres are Dance, Jazz and Blues, Pop/Rock, and Soul/R&B/Funk. Dance music is characterized by strong beats and a strong rhythmic structure, while Jazz/Blues pieces have a propulsive rhythm and loose melody. The Pop/Rock genre includes music that has short and simple melodies and strong beats. Finally, Soul/R&B/Funk musical piece have distinct melodic intensity, heavy beats, and bouncy beats.
5.2.3. Data Sources

McKay (201) identifies three main types of music data that can be used in automatic music classification:

- **Audio data** – Physical audio signals are made into digital representations such as the MP3 and WAV formats. This data is in a digital format with spectral information in the form of spectral waves.
- **Symbolic music representations** – Sound is represented using abstract symbols that are musically meaningful, including representations that directly correlate to the music notation in music scores. Formats such as MIDI, MusicXML and Humdrum store symbolic data.
- **Cultural data** – This is data that does not directly represent the actual sound that makes up the music, but which is still related to the musical piece. Cultural data can be extracted from sources such as album reviews by experts, music sale statistics, and album art images. It is also frequently gleaned from the Internet in the form of edited metadata repositories, unedited listener tags, playlists and searchable web pages.

This thesis will employ the XML symbolic representation because it is easier to extract features from this representation: “Instruments, notes and durations are given by the data itself. From this data, many statistics can be computed (...note distribution, most common intervals, etc.)” (Guarda, 2009).

5.3. Potential Features

When the theoretical music knowledge discussed in Chapter 2 is examined, specific relevant features and combination of features can be identified for use in automatic classification experiments. The sections of this chapter describe features that can potentially be used for the TiMBL feature catalogue. These features included rhythmic elements, including the note duration on the vocal line; pitch features, including the range of the piece and the variety of the pitch; the melody, specifically the level of melodic variation, as well as the types of phrases and the occurrence of repetition of these phrases. The potential features are categorized according to broad headings and are identical or similar to feature sets used in previous research (McKay, 2010; Turton, 2008; Karydis, 2006; Thul, 2006; Tymoczko, 2006; Heittola, 2003; Gutierez, 2002; Ke, 2002; McKay, 2002; Dworak, n.d.).

5.3.1. Pitch

Pitch statistics highlight musical attributes of features such as the commonality pitches and pitch classes are in a piece, its tonality, and the range that it covers (McKay, 2010). The following pitch statistics can be calculated from MusicXML data:

- **Pitch class** – The pitch class that each note in the piece belongs to, which can be expressed by the formula:

  \[
  \text{Note Number} - ((\text{INT} \ (\text{Note Number} / 12)) \times 12) \text{ OR } \text{n} \% 12
  \]

  (where \( n \) is the note, \( \% \) is the mod function and 12 is the argument)
• Pitch class variety (common) - Number of pitch classes that represent at least 9% of the notes
• Pitch class variety (rare) - Number of pitch classes played at least once
• Pitch Distance - The absolute value of the difference between the two pitches $p$ and $q$ (interval).
  \[ |q - p| \] (where p, q are pitches)
• Pitch class distance - Distance between two given pitch classes
  \[ \text{Pitch Class Distance} = ||a-b||_{12}Z \]
  (where a,b are pitch classes. “The smallest non-negative real number $x$ such that, if $p$ is a pitch belonging to pitch class $a$, then either $p + x$ or $p - x$ belongs to pitch class $b$” (Tymoczko, 2006.).)
• Pitch Range - Difference between highest and lowest pitches
  Highest pitch - Lowest pitch
• Most common pitch class - All pitches existing in an octave relationships.
• Pitch/register variety - Number of pitches used at least once.
  Number of Distinct Pitches / Number of Pitches
• Number of Common Pitches - Number of pitches that account individually for at least 9% of all notes.
• Repeated pitch patterns of several notes - Number of repeated patterns of $n$ notes (n-grams)
  Number of repeated note pitch sequences / (Number of notes - n - 1)
• Most Common Pitch Prevalence - Fraction of notes corresponding to the most common pitch.
  Number of notes of pitch type $x$ / Total number of notes in piece
• Number of notes - Average number of notes per musical piece
• Notes per measure - Average number of notes per measure. The formula is:
  Total number of notes in all measures / Total number of measures in piece
• Note durations - Relative lengths of notes.
  \[ \frac{1000}{\text{Note Duration of specific note}} \]
• Key - The dominant key signature of the musical piece
• Weighted Pitch and Duration - The last feature is a modified feature consisting of a “...weighting scheme that allows the prediction of the genre to be more or less affected by one of the two features, in order to determine their contribution”(Karydis, 2006).

5.3.2. Rhythm
• Rhythm can be defined by the beats per measure.
  \[ \frac{(\text{Number of beats / Time-duration of audio sample}) \times 60}{\text{Number of beats in one measure}} \]
  OR

30
The beats per minute is based on the calculations of Ke (2002) for relative and absolute measure values:

1) Half note = 120 / BPM
2) Quarter note = 60 / BPM
3) Eighth note = 30 / BPM
4) Sixteenth note = 15 / BPM
5) Dotted-quarter note = 90 / BPM
6) Dotted-eighth note = 45 / BPM
7) Dotted-sixteenth note = 22.5 / BPM
8) Triplet-quarter note = 40 / BPM
9) Triplet-eighth note = 20 / BPM
10) Triplet-sixteenth note = 10 / BPM

5.3.3. Interval
- Main tonal interval relation - Whether a piece has simple (fifth or fourth interval) or complex harmonic structure (Heittola, 2003)
- Dominant interval - Number of semitones between the two most common pitch classes

5.3.4. Octave
- Octave range - Dominant octave range
- Key octave - The octave of each key can be determined from the formula:

\[
\text{Octave} = \text{INT} \left( \frac{\text{Note Number}}{12} \right) + 1
\]

(\text{where INT represents the integer result of the division by 12})

5.3.5. Melody
- Number of Common Melodic Intervals - Number of melodic intervals that represent at least 9% of all melodic intervals.
- Melodic Octaves - Fraction of melodic intervals that are octaves.
- Melodic Pitch Variety - Average number of notes that go by before a note is repeated.

The research task/problem has now been elucidated, so the next step is to apply the machine learning techniques to the processing of the musical data explained in this chapter in order to make some conclusions about the two experimental methods explained in section 4.1.
Chapter 6. Experiment Results

As described in section 5.2, four hundred Western Classical and four hundred Modern Pop files were hand-classified as “C” and “M”, respectively. Two different methods were used to test the dataset. First, a system of ten-fold cross-validation was used to test the consecutive notes in the two genres of music (section 5.3.1). Next, samples of the music were taken and processed by the system; a sample for every $n$ seconds (section 5.3.2). The two experiments involved analyzing a sequence of features comprising five notes (the default number of features is five) in combination with the duration, versus a slice/section of music that has a tone. The goal of the the experiments was to find out if the “slicing” method (section 6.1) performs as well or better than the note-duration method (section 6.1).

In the TiMBL database, every record gives the same information about the time slice. In the traditional method, every record has widely varying information about the music. The format that the records are in is $<\text{note}>-<\text{octave}>_<\text{duration}>$ (for example, an F sharp in the fourth octave with a duration of 4 beats). There is an equal number of records for every class and the duration of the record and time slice records are the same length, or rather, as close as possible. In this way, time slices will cover the same average time as the regular files. The results of both tests are discussed in this chapter.

6.1. Note-Duration/n-gram Results

From the music files in MusicXML format, a TiMBL database table was created with a record for every line of $n$ consecutive notes in each musical piece from the two genres. Ten-fold cross validation was used to evaluate the margin of error of the classifier's learning algorithm. In the process, TiMBL uses every position in the score (derived from the MusicXML representation), selects a feature, then describes the music based on the feature as a number of n-grams (Section 5.3.1).

In this first experiment the input to TiMBL is a music file with examples for the training set, and another test file with unseen data that has not been classified by genre. Every line in the training and test files will have the same amount of fields. When the command `Timbl -f <trainfile> -t <testfile>` is executed, a file based on the original test file is created which has a class added by TiMBL.

Ten-fold cross validation was used to evaluate the margin of error of the classifier's learning algorithm. For the execution of the 10-fold classification task, the table is split into ten parts, and the “words” (sequences of features within the musical line) of the resultant TiMBL table are counted to calculate how many lines each part will have. In the final step, a file is created with the list of the ten parts, and then TiMBL is launched. Each distinct fold was used once as the test set, in combination with the remaining folds that comprise the training set. Basically, each case it taken in turn as a test file and the other $x-1$ (9) parts are added to be used as the training file. The results were an average of each experiment result done on each fold during training and testing (McKay and Fujinaga, 2005). The classifier created ten folds and a list with names that were used for the 10-fold classification option of TiMBL.

A cross-section of the averages of the overall accuracy results for a ten feature folds are illustrated in Table 3.
The five features are five consecutive symbols for $pitch + length$. Initially, all the features (pitch and note duration) were analyzed, then combinations based on these features were considered.

The best performance was achieved on a combination of the first four feature sets ($pitch + length_{1,2,3,4}$) with a 78% accuracy. The worst performance was exhibited on a single feature test, feature five ($pitch + length_5$), which yielded a result of 73%. Overall, single features did not exhibit very different performance, with the best result being 73% on feature set number one ($pitch + length_1$). The conclusion here is that single features in the case are statistically equivalent.

TiMBL also outputs an information gain table for each fold to show the relative importance of each feature to the success of the feature-set for the classification task. An initial, secondary aim in the experiments was to select features with the highest information gain. Table 4 shows the information gain for the first fold, “aa”, of each feature-set highlighted in Table 3. For each feature-set, the highest and lowest levels of information gain are highlighted, where applicable. The largest information gain was 0.025 which was achieved on feature three in both cases, on combination tests of features 1, 2, 3 and four ($pitch + length_{1,2,3,4}$), and features 1, 2, and 3 ($pitch + length_{1,2,3}$).

<table>
<thead>
<tr>
<th>Feature</th>
<th>All fields</th>
<th>$pitch + length_{1,2,3}$</th>
<th>$pitch + length_{1,2}$</th>
<th>$pitch + length_1$</th>
<th>$pitch + length_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature 1</td>
<td>0.02466301</td>
<td>0.02466301</td>
<td>0.02466301</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Feature 2</td>
<td>0.02472215</td>
<td>0.02453585</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Feature 3</td>
<td>0.02481807</td>
<td>0.024818073</td>
<td>0.024818073</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Feature 4</td>
<td>0.02434881</td>
<td>0.024348812</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Feature 5</td>
<td>0.02395456</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Results of 10-fold cross-validation for 6 variations in window-length

Table 4. Information gain on the first fold for feature-sets
However, because the features in each case/test are equivalent to each other, the data on information gain is not useful.

6.2. **Time Slice Results**

In this second experiment, “slices” were taken at different points throughout the musical piece, a sample for every \( n \) seconds.

In Table 5, a Bach chorale is shown in its stave representation.

| Table 5. Excerpt from Bach chorale “Hilf, Herr Jesu, lass gelingen” (Greentree, 2011) |
|nThe musical information on these four staves can be broken down into the note letter or key, the pitch of the note and the duration of each note. The format is \(<\text{note}>_{\text{key}}-\text{<note duration>}\). A “G” note with in the fourth key and a duration of two beats would be represented as G\(_{4-2}\). Using the information from Figure 2 about the placement of notes on staves, the symbolic representation (with the first four cases omitted) begins as demonstrated in Figure 11. |
| Fig. 11. Symbolic representation of the first stave of Table 5 |

Starting with the first note of the soprano line (and recalling the musical theory provided in Chapter 2 and Figure 2 specifically), we see that the first two notes before the vertical line on the stave (indicating a bar) are “G” notes. Hence, the first two representations read “G\(_{4-2}\)  G\(_{4-1}\)”. In the second bar, the
two first notes are “D” notes. The “C” at the end of each line denotes that the musical representation belongs to the Western Classical genre. The time slice representation for a distance of 64 using the first stave of the Bach score (Table 5) is shown in Figure 12.

\[
\begin{array}{cccccc}
G_4 & G_4 & G_4 & G_4 & D_5 & C \\
G_4 & G_4 & G_4 & D_5 & D_5 & C \\
G_4 & G_4 & D_5 & D_5 & D_5 & C \\
G_4 & D_5 & D_5 & D_5 & D_5 & C \\
D_5 & D_5 & D_5 & D_5 & C_5 & C \\
\end{array}
\]

Fig. 12. Slice representation of the first stave of Table 5

When analyzing only one slice, the information provided for that slice is a small sample of the musical piece as a whole. If the number of slices used in the record is increased, this translates to more of the music being captured for the experiment. Thus, more cases (slices) mean a more representative look at the music. On average, the time on every line of music in the representation is one or two measures. The traditional representation has twice the amount of data. While this can be doubled in the time slices, the resulting representation will not have any greater information. The feature data is gathered and computed at every slice. At every predetermined time slice the classifier summarizes the values collected during the current slice and saves its relevant statistics. The output of the slice analysis is statistics with a single row for each time slice.

Each row contains the features (five by default) that come from different measures. This statistic is the input for the training and classification process. In the experiment, the fifth feature is the first slice of the next measure. That is, if the experiment starts analysis on the first measure, the program then shifts the window/slice over to the first note of the second measure. When it gets to the second measure, the first slice of second measure is processed, then the second slice of that measure, and so forth. Table 6 illustrates how the first measure of each bar in the Bach piece is divided into four time slices.

Table 6. Excerpt from Bach chorale “Hilf, Herr Jesu, lass gelingen” (Greentree, 2011), showing the first line of Figure 13 in the first stave

35
For the slice experiment, 10,000 Modern Pop and 10,000 Classical segments were observed. The time slice was created by taking the value of a slice at every fixed number of points in the measure. For each piece of music, every measure is sampled in both 256 or 64 samples/parts, so when the music was sliced at every 64\textsuperscript{th} slice, every measure was sampled 4 times per measure. With slices at every 64 points, there is less information in each time slice/span, but there is more fine-grained information within each slice.

In the first run, samples were taken at every 256 points within the measure, with every measure being sampled once. The result was 57\% (rounded from 0.568193). When 10 cases were processed using 256 slices, the accuracy increased to 78\%, similar to the performance within the traditional method which has an approximate accuracy of 73\%. With only 1 slice is analyzed, the accuracy was 40\% only. Next, the tracks were sampled at every 64\textsuperscript{th} point, yielding four times as many samples per measure. The results of the experiments on the 64-slice is a worse performance; 46\%. The time slice does not work as well in the traditional approach. A snapshot of the experiment results is shown in Table 7.

<table>
<thead>
<tr>
<th>Slices/Features</th>
<th>256 Slices</th>
<th>64 Slices</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.40</td>
<td>0.28</td>
</tr>
<tr>
<td>5</td>
<td>0.57</td>
<td>0.49</td>
</tr>
</tbody>
</table>

\textbf{Table 7.} Time slice results
Chapter 7. Discussion and Conclusion

This work describes two methods of analyzing music information for the extraction of genre categories. Information about notes, note durations, the key of those notes, and the measures within which notes appear were all used for the experiments. The research in this thesis demonstrates that using a combination of high-level features symbols is more effective at classifying music by genre than looking at the music based on time slice measurements. A genre classification experiment using the traditional method of note-duration symbol analysis successfully classified MusicXML scores among eight categories 78% of the time. This result compares favorably to previous research (section 1.2) in this area. By contrast, the best performance achieved using the time slice method is 57%, well below the standard results achieved in recent MIREX competitions.

As observed in sections 5.3.1 and 5.3.2, the time slice method does not work as well as the traditional approach. The classifier performed better on data sets analyzed using the regular method than on data analyzed using time slices broken into 256 parts per measure. When the data is split into even smaller slices (such as the 64-slice), the performance dropped significantly. This decrease in accuracy can be attributed to the fact that even though the information in slices is more fine-grained, the information is more detailed for the slice, but not for the musical piece as a whole. Essentially, because each time slices is more specific, having more time slices is more representative per slice, but there is no information that is specific for the entire piece.

Looking at the results, the next step is to think about improvements that could be made to the current models for improving accuracy. For the note-duration symbol method, the best results were generated by using multiple features. As such, it can be hypothesized that using more features can improve the results. For the time slice experiments, since more fine-grained information (obtained by using fewer slices) results in a bigger improvement, it is possible that using larger slices (for example, slices of 512 parts) would improve the classification results for the musical piece. Second, the accuracy measure only takes into account harmonization exactly as Bach composed it. Because of the way that accuracy is measured in the experiments, features that exhibit low performance may actually still be musically valid.

There are a number of directions in which this work can continue to develop. For the most part, this thesis only dealt with the vocal line of the music. By expanding on the instrument component of the music that is used for the experiments, different feature sets can be created (some possibly unique), and this could translate to greater accuracy on the classification task. Another direction in which this thesis could expand is with the application of the same techniques expounded in chapter 5 to spectral data, instead of representative data in MusicXML format. In this way, it could be possible to extract further information not only from the notes, but also from spectral parameters surrounding the notes. As well, there are other, more contextual features that can be used in these experiments, such as “...generated chords...as features... [the] harmonization of previous notes ...to predict chords...the location of a note in a bar and the location of a bar in a piece” (Chilvers and van Zaanen, 2009).

As well, the number of features selected from the MusicXML scores can be increased in order to provide greater dimensionality to the experiments and also to observe the relationship between different features and specific genres.
Works Cited


Chu, M-L. (2010). Automatic music genre classification. In National Taiwan University digital


41


