Issues in Automatic Musical Genre Classification

Cory McKay
Faculty of Music, McGill University
cory.mckay@mail.mcgill.ca

ABSTRACT

A novel software system that automatically classifies musical recordings based on genre is presented and discussed. This system is intended as a demonstration of how automated musical feature extraction from MIDI files, machine learning and pattern recognition techniques can be applied to the general tasks of music classification and grouping.

The nebulous definitions and overlapping boundaries of genres makes reliable and consistent genre classification a difficult task for humans and computers alike. Traditional rules-based classification systems are severely limited by these factors as well as by the dynamic nature of genres. The techniques used in this software system are presented as alternative methods that can help to overcome these limitations.

Arriving at a realistic and useful musical taxonomy can also be a difficult task. The problems associated with this task are briefly reviewed and some possible ways in which technology can be applied to improve the process of taxonomy construction are presented.

The highlights of the catalogue of musical features that the software extracts from symbolic musical data are presented in the context of how the features can be used both for automatic classification and for statistical musicological studies. The easy to use and flexible interface of the software is also demonstrated as a resource that could easily be adapted to a variety of areas of musical research. Several automated pattern recognition and classification techniques are also briefly presented in order to demonstrate how they can be applied to musical research.

1. INTRODUCTION

Musical genre is used by retailers, libraries and people in general as a primary means of organizing music. Anyone who has attempted to search through the discount bins of a music store will have experienced the frustration of searching through music that is not sorted by genre. Listeners use genres to find music that they're looking for or to get a rough idea of whether they're likely to like a piece of music before hearing it. The music industry, in contrast, uses genre as a key way of defining and targeting different markets. The importance of genre in the mind of listeners is exemplified by research indicating that the style in which a piece is performed can influence listeners' liking for the piece more than the piece itself (North & Hargreaves 1997).

Unfortunately, consistent musical genre identification is a difficult task, both for humans and for computers. There is often no generally accepted agreement on what the precise characteristics are of a particular genre and there is often not even a clear consensus on precisely which genre categories should be used and how different categories are related to one another.

This brings to light two of the main problems of genre classification. The first of these is which musical features (a term commonly used in pattern recognition that, in this case, refers to characteristic pieces of information that can be extracted from music and used to describe or classify it) to consider for classification and the second is how to devise a taxonomy into which recordings can be classified.

The need for an effective automatic means of classifying music is becoming increasingly pressing as the number of recordings available continues to increase at a rapid rate. It is estimated that 2000 CDs a month are released for wide distribution in Western countries alone (Pachet & Cazaly 2000). Software capable of performing automatic classifications would be particularly useful to the administrators of the rapidly growing networked music archives, as their success is very much linked to the ease with which users can search for types of music on their sites. These sites currently rely on manual genre classifications, a methodology that is slow and unwieldy. An additional problem with manual classification is that different people classify genres differently, leading to many inconsistencies.

Research into automatic genre classification has the side benefit that it can potentially contribute to the theoretical understanding of how humans construct musical genres, the mechanisms they use to classify music and the means that are used to perceive the differences between different genres. The mechanisms used in human genre classification are poorly understood, and constructing an automatic classifier to perform this task could produce valuable insights.

The types of features developed for a classification system could be adapted for other types of analyses by musicologists and music theorists. Taken in conjunction with genre classification results, the features could also

provide valuable insights into the particular attributes of different genres and what characteristics are important in different cases.

Automatic feature extraction and learning / pattern classification techniques have the important benefit of being adaptable to a variety of other content-based (i.e. relating directly to and only to the music itself) musical analysis and classification tasks, such as similarity measurements in general or segmentation. Systems could be constructed that, to give just a few examples, compare or classify pieces based on compositional or performance style, group music based on geographical / cultural origin or historical period, search for unknown music that a user might like based on examples of what he or she is known to like already, sort music based on perception of mood, or classify music based on when a user might want to listen to it (e.g. while driving, while eating dinner, etc.). Music librarians and database administrators could use these systems to classify recordings along whatever lines they wished. Individual users could use such systems to sort their music collections automatically as they grow and automatically generate play lists with certain themes. It would also be possible for them to upload their own classification parameters to search on-line databases equipped with the same classification software.

2. SYMBOLIC AND AUDIO REPRESENTATIONS

Musical data is generally stored digitally as either audio data (e.g. wav, aiff or MP3) or symbolic data (e.g. MIDI, GUIDO or Humdrum). Audio data represents actual sound signals by encoding analog waves as digital samples. Symbolic data, in contrast, stores musical events and parameters themselves. Symbolic data is therefore a high-level representation and audio data is a low-level representation and, in general, symbolic representations store information that includes the pitch, time of attack, duration, instrumentation and, sometimes, dynamics of each note.

Although the classification of audio data is certainly very important from a practical perspective, the emphasis here is placed on symbolic data. Automatic transcription systems have not yet achieved the point where they can accurately transcribe anything other than monophonic melodies. This means that audio classification systems must rely on low-level features related to signal processing rather than direct musical information. This is of limited utility for musicological research that requires knowledge of the parameters of actual notes.

The use of high-level features extracted from symbolic recordings has the additional advantage of making it possible to classify music for which no audio recordings are available. Optical music recognition techniques could be used, for example, to read in paper scores so that they could be classified. Furthermore, future advances in automatic audio transcription could make it possible to make use of both low and high-level features.

MIDI files were used in the particular experiment presented later in this paper because a diverse range of such recordings are widely available. Other symbolic formats, such as Humdrum or GUIDO, could just as easily have been used.

3. CLASSIFICATION TECHNIQUES

There are three main classification paradigms that can be used to perform automated classification:

- **Expert Systems:** Use pre-defined rules to process features and arrive at classifications.
- **Supervised Learning:** Attempt to formulate classification rules by using machine learning techniques to train on model examples. Previously unseen examples are classified into one of the model categories using the patterns learned during training.
- **Unsupervised Learning:** Cluster the data based on similarities that the systems perceive themselves. No model categories are used.

Expert systems are a tempting choice because known rules and characteristics of genres can be implemented directly. A great deal of potentially useful work has been done analyzing and generating theoretical frameworks in regards to classical music, for example. Given this body of research, it might well be feasible to construct a rules-based expert system to classify such types of music. There are, however, many other kinds of music for which this theoretical background does not exist. Many types of Western folk music, a great deal of non-Western music and Western popular music do not, in general, have the body of analytical literature that would be necessary to build an expert system.

There have, of course, been some efforts to at least consider general theoretical frameworks for popular and/or non-Western music, such as in the work of Middleton (1990). Unfortunately, these studies have not been precise or exhaustive enough to be applicable to the task of automatic genre classification, and it is a matter of debate as to whether it is even possible to generate a framework that could be broad enough to encompass every possible genre. Although there are broad rules and guidelines that can be informally expressed about particular genres, it would be very difficult to design an expert system that could process rules that are often ill-defined and inconsistent across genres. A further problem is that new genres are constantly appearing and existing ones often change. Keeping a rules-based system up to date would be a very difficult task.

Systems that rely on pattern recognition and learning techniques hold more potential. Such systems can

analyze musical examples and attempt to learn and recognize patterns and characteristics of genres in much the same way that humans do, although the precise mechanisms used differ. A side benefit of such systems is that they may recognize patterns that have not as of yet consciously occurred to human researchers. These patterns could then be incorporated into theoretical research.

This leaves the options of supervised and unsupervised learning. Although very well suited to automated systems that measure musical similarity in general, unsupervised systems are not well suited to the particular problem of genre classification because the categories produced might not be meaningful to humans. Although unsupervised learning avoids the problems related to defining a set genre hierarchy discussed below, and the categories produced might well be more accurate than human genre categories in terms of objective similarity, a genre classification system that uses its own genre categories would be of limited utility to humans who want to use genres that are meaningful and familiar to them.

Supervised learning is the best option, despite the fact that a manually classified and therefore biased model training set is a necessary but unavoidable drawback. Such systems form their own rules without needing to interact with humans, meaning that the lack of clear genre definitions is not a problem. These systems can also easily be retrained to reflect changes in the genres being classified.

There are a number of particular pattern classification techniques that can be used, including neural networks and k-nearest neighbour. Duda, Hart and Stork's book (2001) is one particularly good reference on such techniques.

4. FORMING GENRE TAXONOMIES

It can be difficult to find clear, consistent and objective definitions of genres, and genres are rarely organized in a consistent or rational manner. The differences between genres are vague at times, rules distinguishing genres are often ambiguous or inconsistent, classification judgments are subjective and genres can change with time. The categories that come to be are a result of complex interactions of cultural factors, marketing strategies, historical conventions, choices made by music librarians, critics and retailers and the interactions of groups of musicians and composers.

In order to train an automatic classification system using supervised learning it is first necessary to have a set of genre categories that the training examples can be partitioned into. The lack of a commonly accepted set of clearly defined genres makes it tempting to simply devise one's own artificial labels for the purposes of making an automatic classification system. These labels could be designed using reasonable, independent and consistent categories, a logical structure and objective similarity measures. One could even use unsupervised learning techniques to help accomplish this if desired. The genre labels in common use are often haphazard, inconsistent and illogical, and one would certainly wish to devise a system that does not suffer from these problems.

It is argued here that this would be a mistake, however. One must use the labels that are meaningful to real people in order for the labels to be useful to them, which is to say that genre categories must be consistent with how a person with moderate musical knowledge would perform categorizations. Furthermore, genre labels are constantly being created, forgotten and modified by musicians, retailers, music executives, DJs, VJs, critics and audiences as musics develop, so a static, ideal system is not sustainable. Genre is not defined using strictly objective and unchanging qualities, but is rather the result of a dynamic cultural process. One must therefore be careful to avoid thinking of genres in terms of immutable snapshot, as both their membership and their definitions change with time.

Another approach to finding an appropriate labelling structure is to look at the categories used by music sales charts such as Billboard, or by awards shows such as the Grammies. Unfortunately, there are also a number of problems with this approach. Charts such as those used by Billboard often only reflect the current trends in music to the exclusion of older genres. A proper system should include old genres as well as new. Furthermore, these categories tend to reflect the labelling system that the music industry would ideally like to see, not the one which is actually used by the public. Charts and award categories therefore often have labels based on marketing schemes more than common perceptions, and do not even offer the advantages of being consistent or well thought out from a taxonomical perspective.

Specialty shows on radio or television do offer a somewhat better source of labels, as they often reflect categories that attract listeners interested in specific genres, both new and old. They do still suffer from the influence of commercial biases, however, as the contents of shows tend to be influenced at least as much by the preferences of advertisers relating to age, income and political demographics as by the musical preferences of listeners. Although university radio stations do not suffer from this problem in the same way, they are often limited in scope and by the variable expertise and knowledge of their DJs.

Retailers, particularly on the Internet, may perhaps be the best source of labels. They use categories that are likely the closest to those used by most people, as their main goal is to use a taxonomy that makes it easy for customers to find music that they are looking for. Al-

though retailers can sometimes be a little slow to respond to changes in genre, they nonetheless do respond faster than some of the alternatives discussed above, as responding to new genres and keeping existing genres up to date allows them to draw potential buyers into areas that contain other music that they may wish to buy, therefore increasing sales.

Although one might argue that it would be preferable to base labels on the views of concert goers, clubbers, musicians, DJs, VJs, music reporters and others who are on the front line of genre development, doing so would be disadvantageous in that genres at this stage of development may be unstable. Additionally, favouring the genre labels used by specialists may result in some confusion for non-specialists. Waiting for retailers to recognize a genre and thus make it "official" is perhaps a good compromise in that one keeps somewhat abreast of new developments, while at the same time avoiding contradictions and excess overhead in terms of data collection and computerized training.

The problem of inconsistency remains, unfortunately, even with the taxonomies used by retailers. Not only do record companies, distributors and retailers use different labelling systems, but the categories and classification judgements between different retailers can also be inconsistent. This is, unfortunately, an avoidable problem, as there are no widely accepted labelling standards or classification criteria. Employees of different organizations may not only classify the same recording differently, but may also make selections from entirely different genre labels, or may emphasize different identifying features. One must simply accept that it is impossible to find a perfect taxonomy, and one must make do with what is available.

An important part of constructing a genre taxonomy is determining how different categories are interrelated. This is, unfortunately, a far from trivial problem. Attempts to this point to implement an automatic classification system have sidestepped these issues by limiting their testing to only a few simple genres. Although this is acceptable in the early stages of development, the problem of taxonomical structures needs to be carefully considered if one wishes to construct a system that is scalable to real-world applications.

This problem is discussed in a paper by Pachet and Cazaly (2000). The authours observe that retailers tend to use a four-level hierarchy: global music categories (e.g. classical, jazz, rock), sub-categories (e.g. operas, Dixieland, heavy metal), artists and albums. Although this taxonomy is effective when navigating a physical record store, the authours argue that this taxonomy is inappropriate from the viewpoint of establishing a major musical database, since different levels represent different dimensions. In other words, a genre like "classical" is fundamentally different from the name of an artist.

Pachet and Cazaly continue on to note that Internet companies, such as Amazon.com, tend to build tree-like classification systems, with broad categories near the root level and specialized categories at the leaves. The authours argue that, although this is not in itself necessarily a bad approach, there are some problems with it. To begin with, the level that a category appears at in the hierarchy can vary from taxonomy to taxonomy. Reggae, for example, is sometimes treated as root-level genre and is sometimes considered a sub-genre of world music.

A further problem is that there is a lack of consistency in the type of relation between a parent and a child. Sometimes it is genealogical (e.g. rock -> hard rock), sometimes it is geographical (e.g. Africa -> Algeria), sometimes it is based on historical periods (e.g. Baroque -> Baroque Opera), etc. Although these inconsistencies are not significant for people manually browsing through catalogues, they could be problematic for automatic classification systems that are attempting to define genres using content-based features, as musics from the same country or same historical period can be very different musically.

An additional problem to consider is that different tracks in an album or even different albums by an artist could belong to different genres. Many musicians, such as Neil Young and Miles Davis, write music in different genres throughout their careers. Even a single album by such a musician can contain music from several different genres. It seems clear that attempting to classify by musicians rather than individual recordings is problematic.

Pachet and Cazaly argue that it therefore seems apparent that, ignoring potential problems related to size, it would be preferable to base taxonomies on individual recordings, rather than on artists or albums. In a later paper, however, Aucouturier and Pachet (2003) argue that one should in fact use taxonomies based on artist rather than title, as taxonomies based on title involve many more entries and result in categories that are overly narrow and have contrived boundaries.

Pachet and Cazaly argue that it is necessary to build an entirely new taxonomy to meet the needs of any large scale musical database. They emphasize the goals of producing a taxonomy that is objective, consistent, independent from other metadatabase descriptors and that supports searches by similarity. They suggest the use of a tree-based system organized based on genealogical relationships, where only leaves would contain musical examples. Each node would contain its parent genre and the differences between its own genre and that of its parent.

The concerns with existing taxonomies expressed by Pachet and Cazaly are certainly valid, but their proposed solution unfortunately has some problems of its

own. To begin with, defining an objective classification system is much easier said than done, and getting universal agreement on a standardized taxonomy is most probably an intractable task. Furthermore, their system does not deal with the reality that a single recording can sometimes reasonably be said to belong to more than one genre, nor does it deal with the potential problem of multiple genealogical parents that can compromise the tree structure.

It seems apparent that some modifications are needed to Pachet and Cazaly's system, but some sort of hierarchal tree-based taxonomy nonetheless appears to be a convenient and realistic genre structure. Franco Fabbri (1982) suggests that, when faced with describing a genre to a person who is unfamiliar with it, most individuals do so by defining the genre as an intersection of other similar genres with labels known to both parties, by using a broader label under which the genre in question might fall or by explaining the genre using familiar terms such as definitions and emotive meanings. The former two methodologies are certainly consistent with a hierarchal structure with visible parents and siblings.

A further issue to consider is the variable degree to which different genres branch out into sub-genres. Considered from a hierarchal tree-based perspective, this variability applies to both the depth and breadth of various branches. Some genres have many very specialized sub-genres, such as electronic dance music (e.g. techno, jungle, rave, etc.). Others, such as pop-rock, tend to have fewer, broader and less specified sub-genres. For the purposes of creating a genre hierarchy, one must accept these inconsistencies rather than imposing unrealistically broad or narrow categories in order to avoid dissymmetry in the genre structure.

Aucouturier and Pachet (2003) divide methods of genre classification into three categories: manual, prescriptive and emergent. The manual approach involves humans performing the classification task by hand, while the prescriptive and emergent approaches involve automatic systems.

Aucouturier and Pachet define the prescriptive approach as an automatic process that involves a two-step procedure: feature extraction followed by machine learning / classification. The prescriptive approach assumes a pre-existing taxonomy that a system can learn. Aucouturier and Pachet argue, reasonably enough, that prescriptive systems tend to be based on contrived taxonomies and that a truly useful system would need to be able to deal with much larger taxonomies than can successfully be modelled and kept up to date. A further problem is that it can be difficult to find training samples that are unambiguously representative enough to train a classifier properly.

Aucouturier and Pachet argue that the emergent approach is the best alternative. Rather than using existing taxonomies, an emergent system attempts to emerge labels according to some measure of similarity. The authors suggest using similarity measurements based on audio signals as well as on cultural similarity gleaned from the application of data mining techniques to text documents. They propose the use of collaborative filtering to search for similarities in the taste profiles of different individuals and of co-occurrence analysis on the play lists of radio programs and the track listings of CD compilation albums.

The emergent approach is untested, however, and it is difficult to predict how effective it would be in real life. Implementing the data mining techniques required would be quite a difficult task. Furthermore, there is no guarantee that the recordings that get clustered together would be consistent with groupings that humans use in reality or would find convenient to use, nor is there any obvious provision for defining the types of genre structures and interrelations that humans find useful when browsing through categories. Nonetheless, the emergent approach holds more promise than naive unsupervised learning, although Aurcouturier and Pachet's argument that it is superior to the prescriptive approach is not entirely convincing.

In any case, the notion of developing modules that collect and consider non-content-based sociocultural data is intriguing. Whether it is prescriptive or emergent systems that end up being more effective, the idea of automatically exploiting text documents to gather sociocultural data should be explored in future research.

5. FEATURE EXTRACTION

In order to train a computer classifier, it is first necessary to extract features from musical recordings that can be given to the classifier as percepts. Simply giving the recording directly to a classifier would create an excess of information that would make the classification very slow and, quite likely, impossible. Extracting features from recordings and providing these to classifiers reduces the amount of information that must be processed and emphasizes aspects of recordings that are, hopefully, salient to the process of category discrimination.

Choosing which features to use is, unfortunately, a difficult problem. Although there has been a great deal of work on analyzing and describing particular types of music, there has been relatively little research on deriving features from music in general. Alan Lomax and his colleagues in the Cantometrics project (Lomax 1968) have performed the most extensive work, by comparing several thousand songs from hundreds of different cultural groups using thirty-seven features. These features provide a good starting point for developing a library of high-level features. Although there have been a few other efforts to list categories of features, they have

tended to be overly broad. Work such as Phillip Tag's "checklist of parameters" (1982) are still useful as a general guide, however.

As an initial step, one might look to how humans accomplish this task for inspiration, as we are able to successfully perform genre classifications, so we do provide one, albeit not the only, viable model.

One might imagine that high-level musical structure and form play an important role, given that this is the area on which much of the theoretical literature has concentrated. This does not appear to be the case, however. Research by Perrott and Gjerdingen (1999) found that humans with little to moderate musical training are able to make genre classifications agreeing with those of record companies 71.68% of the time (among a total of 10 genres), based on only 300 milliseconds of audio. This is far too little time to perceive musical form or structure. This suggests that there must be a sufficient amount of information available in very short segments of music to successfully perform classifications. This does not mean that one should ignore musical form and structure, as these are likely useful as well, but it does mean that they are not strictly necessary.

This is an indication that it is probably a better approach to extract features based on simple musical observations rather than using sophisticated theoretical models. Such models tend to have limited applicability beyond those limited spheres which they were designed to analyze, and sophisticated automatic musical analysis remains an unsolved problem in many cases.

Ideally, one would like to use features consisting of simple numbers. This makes storing and processing features both simpler and faster. Features that represent an overall aspect of a recording are particularly appropriate in this respect. Features based on averages and standard deviations allow one to see the overall behaviour of a particular aspect of a recording, as well as how much it varies.

The development of a large set of features is necessary to perceive the differences between any individual arbitrary pair of genres coming from the large superset of genres in general. Although it is not feasible from a classification standpoint to deploy all of these features during a single classification operation, the use of a hierarchical taxonomy makes it possible to perform multiple classifications on different sub-trees of the hierarchy, each using specialized features. In other words, one could first make a coarse classifications with a certain sets of features, and then use different sets of features to make finer classifications.

A catalogue of 160 features that can be used to characterize and classify recordings was constructed. Although too numerous to discuss here in detail, these features belong to the following seven categories:

- Instrumentation
- Texture
- Rhythm
- Dynamics
- Pitch Statistics
- Melody
- Chords

These features could be used for any classification task, such as composer or performer style, not just genre classification. Future studies of which features are discriminating in which contexts could be of musicological interest.

It should be noted that many people use features beyond those that can be derived from the actual musical content of a recording or a performance. Genre is very much linked to the social, economic and cultural placement of both musicians and listeners. One need only see a photo or watch an interview with a musician, without ever having heard his or her music, to be almost certain whether the musician plays rap, heavy metal or classical music, for example. The style of album art, web pages and music videos are all features that humans can use to identify genre. Similarly, a performer's appearance and actions on stage (facial expressions, ritual gestures, types of dancing, etc.) provide clues towards genre, as do an audience's demographics, dress and behaviour (clapping, shouting, sitting quietly, dancing, etc.). The fine distinction between some sub-genres may well be related to such sociological features more than musical content.

Although the current study is only concerned with content-based features, future research that uses data mining techniques to gather sociological features to supplement content-based features could be highly useful. There has been some initial research in this direction (Whitman & Smaragdis 2002) that has had encouraging results.

6. OTHER GENRE CLASSIFICATION SYSTEMS

There have been a number of previous studies on automatic genre classification of audio files. The work of George Tzanetakis and his colleagues (Tzanetakis, Essl & Cook 2001; Tzanetakis & Cook 2002) is particularly widely cited. The authours used a variety of low-level features to achieve success rates of 61% when classifying between ten genres.

Additional research has been performed by Grimaldi, Kokaram and Cunningham (2003), who achieved a success rate of 73.3% when classifying between five categories. Kosina (2002) achieved a success rate of 88% with three genres. Xu et al. (2003) achieved a success rate of 93% with four categories. Deshpande, Nam and Singh (2001) constructed a system that correctly

classified among three categories 75% of the time. McKinney and Breebaart (2003) achieved a success rate of 74% with seven categories. Jiang et al. (2002) correctly classified 90.8% of recordings into five genres.

There has been somewhat less research into the classification of symbolic data. Shan and Kuo (2003) achieved success rates between 64% and 84% for two-way classifications. Chai and Vercoe (2001) were successful in correctly performing three-way classifications 63% of the time. Although these studies are very interesting, they focus more on pattern classification techniques rather than on features.

There has also been a significant amount of work on using unsupervised learning techniques to group recordings by similarity. Although most of these are not directly relevant to genre classification in particular, one exception is the work of Ponce de Leon and Inesta (2002), who constructed a system that correctly grouped 77% of their MIDI recordings into groups roughly corresponding to either classical or jazz music.

7. THE EXPERIMENT

An initial experiment was performed to evaluate the feasibility of automatic genre classification using symbolic musical representations and determine the potential of future research in this direction. Given that this was only an initial investigation, only twenty features were implemented for this experiment. A very limited taxonomy was used for the same reasons.

The training and testing data consisted of 225 MIDI files hand classified hierarchically into three parent genres (Classical, Jazz and Pop) and nine sub-genres (Baroque, Romantic, Modern Classical, Swing, Funky Jazz, Cool Jazz, Rap, Country and Punk). The particular files were selected to represent each category as broadly as possible (e.g. the Baroque category included operas, violin concertos, harpsichord sonatas, etc. not just organ fugues). This significantly increased the difficulty of the task, as each sub-genre only had 20 training recordings (five recordings were reserved for testing in each run) to learn a broad range of music. This was done in order to truly test the viability of the system and its features.

The features were classified using an array of eight feed-forward neural networks that consisted of four networks for identifying parent genres and four networks for identifying sub-genres. This division into two groups made it possible to classify parent genres independently from sub-genres. A coordination system considered the certainty score output by the networks for each sub-genre in combination with the certainty for each parent genre, and produced a final classification using weighted averages. This particular classification system was used because it allowed the independent comparison of different groups of features as well as a comparison of how well parent genres were classified relative to sub-genres.

A five-fold cross-validation was used to test the performance of the system. This means that five testing runs were performed. 80% of the data was used for training and 20% for testing during each of these runs. The result was that every piece was used for training during four runs and for testing during one run. The results are shown in **Tables 1 and 2**.

	Set 1	Set 2	Set 3	Set 4	Set 5	Average
Classical	93	80	100	93	100	**93.2**
Jazz	73	80	60	53	40	**61.2**
Pop	100	100	100	100	100	**100.0**
Average	88.7	86.7	86.7	82.0	80.0	**84.8**

Table 1: Classification success rates (in percentages) for parent genres for all five cross-validation testing runs.

	Set 1	Set 2	Set 3	Set 4	Set 5	Average
Baroque	80	40	80	80	80	**72.0**
Romantic	0	40	0	20	40	**20.0**
Modern	100	40	100	40	80	**72.0**
Swing	40	80	20	40	20	**40.0**
Funky Jz.	60	40	60	40	0	**40.0**
Cool Jz.	40	20	20	20	0	**20.0**
Rap	80	60	80	60	20	**60.0**
Country	80	100	100	100	100	**96.0**
Punk	100	100	100	100	100	**100.0**
Average	64.4	57.8	62.2	55.6	48.9	**57.8**

Table 2: Classification success rates (in percentages) for sub-genres for all five cross-validation testing runs.

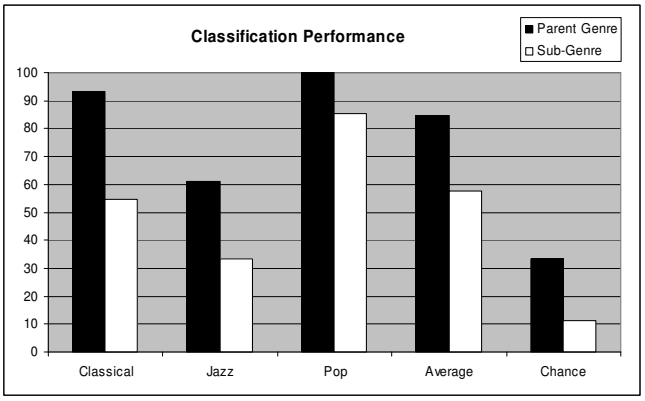

Figure 1: Average classification success rates on test sets. The sub-genre bars give the average success rates of the sub-genres belonging to the corresponding parent genre.

Overall success rates of 84.8% were achieved for parent genres and 57.8% for sub-genres across all five training runs. These results were fairly consistent across training runs. There was also a consistent difference in which categories were successfully classified, with

Punk and Country performing very well and Cool Jazz and Romantic performing very poorly.

As can be seen from **Figure 1**, the test set was classified at a success rate significantly higher than chance in all cases. Furthermore, the system achieved success rates comparable to existing audio classification systems using similar numbers of categories and better than existing systems using symbolic data. This is particularly encouraging, given the limited feature set, small training sample and broad categories used here. This appears to provide a strong argument that further research in this area is justified.

8. SOFTWARE INTERFACE

A user-friendly interfaced is being developed that will be ported to the classification system. This will allow the user to input arbitrary taxonomies and lists of recordings, choose which features to extract, extract the selected features from recordings, evaluate the usefulness of particular features in different contexts and perform actual classifications.

An emphasis has been put on making the interface easy to use and flexible so that it can be used for a variety of research and applied purposes by people with little technical expertise. The software has been built so that the taxonomies and other lists of recordings can be altered directly within the program's GUI, without having to exit it or edit arcane configuration files. **Figure 2** gives two brief sample of what the interface looks like.

The software has also been designed so that additional features can be designed and added to the software easily and painlessly by anyone with some basic Java programming skills. This makes the software expandable for a variety of research purposes.

9. CONCLUSIONS AND FUTURE RESEARCH

The experiment discussed in **Section 7** gave classification results that indicate that there is certainly significant potential for further research in the area of automatic genre classification. Future research will initially concentrate on the implementation of the full feature library, the use of a greatly expanded list of recordings and more realistic taxonomy, feature selection methods and a more sophisticated classification methodology.

More long-term research will examine the use of data-mining techniques to automatically tap text resources that can be used to refine taxonomies and provide features. Musical similarity in general will also be studied.

This research is interesting from both applied and theoretical musicological perspectives. The relative effectiveness of different features in distinguishing between categories has theoretical interest. The way in which the system makes genre classifications based entirely on "objective" content-based grounds, without any of the cultural context which humans can never entirely ignore, could also provide inspiration into research on how humans form genre categories and the extent to which content-based features are important in this. Research in automatic genre classification could potentially contribute to the theoretical understanding of how humans construct musical genres, the mechanisms they use to classify music and the means that are used to perceive the differences between different genres.

The software interface has been designed to allow a broad range of people use it. The potential of this system extends well beyond genres, as it can be used to perform classifications of any type, such as composer style. The incorporation of an unsupervised clustering module in the future will also expand the range of tasks and research goals to which the system could be applied.

10. ACKNOWLEDGEMENTS

Thanks to Ichiro Fujinaga for his invaluable advice. Thanks also to the *Fonds Québécois de la recherche sur la société et la culture* for their generous support, which has helped to make this research possible.

11. REFERENCES

Aucouturier, J. J., and F. Pachet. 2003. Representing musical genre: A state of the art. *Journal of New Music Research* 32 (1): 1–12.

Chai, W. and B. Vercoe. 2001. Folk music classification using hidden Markov models. *Proceedings of the International Conference on Artificial Intelligence*.

Deshpande, H., U. Nam, and R. Singh. 2001. Classification of music signals in the visual domain. *Proceedings of the Digital Audio Effects Workshop*.

Duda, R.O., P.E. Hart, and D.G. Stork. 2001. *Pattern classification*. New York: John Wiley & Sons Inc.

Fabbri, F. 1982. What kind of music? *Popular Music* 2: 131–143.

Grimaldi, M., A. Kokaram, and P. Cunningham. 2003. Classifying music by genre using a discrete wavelet transform and a round-robin ensemble. *Work Report*. Trinity College, University of Dublin, Ireland.

Jiang, D.N., L. Lu, H.J. Zhang, J.H. Tao, and L. H. Cai. 2002. Music type classification by spectral contrast feature. *Proceedings of Intelligent Computation in Manufacturing Engineering*.

Kosina, K. 2002. Music genre recognition. *Diploma thesis*. Technical College of Hagenberg, Austria.

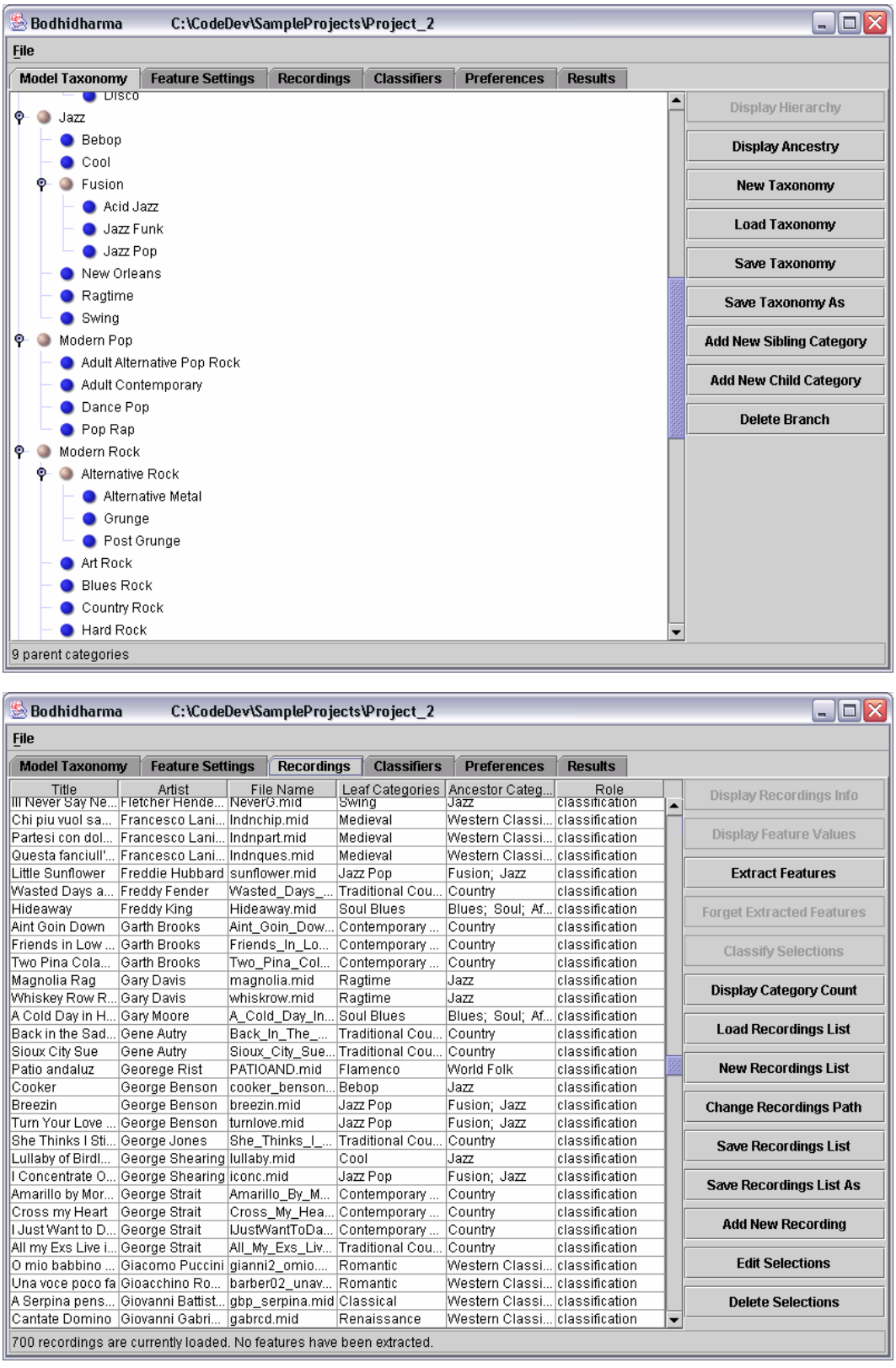

Figure 2: Two sample screens from the classification software interface. The top screen shows a sample taxonomy and the bottom screen shot shows a sample list of recordings that have been classified.

Lomax, A. 1968. *Folk song style and culture*. Washington, D.C.: American Association for the Advancement of Science.

McKinney, M. F., and J. Breebaart. 2003. Features for audio and music classification. *Proceedings of the International Symposium on Music Information Retrieval*. 151–158.

Middleton, R. 1990. *Studying popular music*. Philadelphia: Open University Press.

North, A. C., and D. J. Hargreaves. 1997. Liking for musical styles. *Music Scientae* 1: 109–128.

Pachet, F., and D. Cazaly. 2000. A taxonomy of musical genres. *Proceedings of the Content-Based Multimedia Information Access Conference*.

Perrott, D., and R. O. Gjerdingen. 1999. Scanning the dial: An exploration of factors in the identification of musical style. *Research Notes*. Department of Music, Northwestern University, Illinois, USA.

Ponce de Leon, P. J., and J. M. Inesta. 2002. Musical style identification using self-organising maps. *Proceedings of the International Conference on Web Delivery of Music*. 82–89.

Shan, M. K., and F. F. Kuo. 2003. Music style mining and classification by melody. *IEICE Transactions on Information and Systems* E86-D (3): 655–659.

Tagg, P. 1982. Analysing popular music: Theory, method and practice. *Popular Music* 2: 37–67.

Tzanetakis, G., and P. Cook. 2002. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing* 10 (5): 293–302.

Tzanetakis, G., G. Essl, and P. Cook. 2001. Automatic musical genre classification of audio signals. *Proceedings of the International Symposium on Music Information Retrieval*. 205–210.

Whitman, B., and P. Smaragdis. 2002. Combining musical and cultural features for intelligent style detection. *Proceedings of the International Symposium on Music Information Retrieval*. 47–52.

Xu, C., N. C. Maddage, X. Shao, F. Cao, and Q. Tian. 2003. Musical genre classification using support vector machines. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*. V_429–V_432.