# AUTOMATIC GENRE CLASSIFICATION USING LARGE HIGH-LEVEL MUSICAL FEATURE SETS

*Cory McKay*
McGill University
Faculty of Music
Montreal, Quebec, Canada
cory.mckay@mail.mcgill.ca

*Ichiro Fujinaga*
McGill University
Faculty of Music
Montreal, Quebec, Canada
ich@music.mcgill.ca

**ABSTRACT**

This paper presents a system that extracts 109 musical features from symbolic recordings (MIDI, in this case) and uses them to classify the recordings by genre. The features used here are based on instrumentation, texture, rhythm, dynamics, pitch statistics, melody and chords.

The classification is performed hierarchically using different sets of features at different levels of the hierarchy. Which features are used at each level, and their relative weightings, are determined using genetic algorithms. Classification is performed using a novel ensemble of feedforward neural networks and k-nearest neighbour classifiers.

Arguments are presented emphasizing the importance of using high-level musical features, something that has been largely neglected in automatic classification systems to date in favour of low-level features.

The effect on classification performance of varying the number of candidate features is examined in order to empirically demonstrate the importance of using a large variety of musically meaningful features. Two differently sized hierarchies are used in order to test the performance of the system under different conditions.

Very encouraging classification success rates of 98% for root genres and 90% for leaf genres are obtained for a hierarchical taxonomy consisting of 9 leaf genres.

**KEYWORDS**

Genre, classification, hierarchical, features, music.

## 1. INTRODUCTION

Musical genre has a particular importance in the field of music information retrieval. It is used by retailers, librarians, musicologists and listeners in general as an important means of organizing music. Anyone who has looked through the discount bins of a music store will have experienced the frustration of searching through music that is not sorted by genre. Furthermore, 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 [13].

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. 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.

There has been a significant amount of research into using low-level features to classify audio recordings into categories based on factors such as genre and style (see Section 2). Although this research is certainly very valuable, the current lack of reliable polyphonic transcription systems makes it difficult to impossible to extract high-level features from audio recordings, as this requires precise knowledge of information such as the pitch and timing of individual notes. Most research to date has therefore made use of primarily low-level, signal-processing based features. Although there have been some very interesting efforts to generate features with musical meaning from audio recordings, the limitations of current signal-processing capabilities has limited these endeavours so far.

The problem of implementing a reliable genre classifier that deals with realistic taxonomies has yet to be solved. It is therefore appropriate to take advantage of whatever resources are available in order to improve performance. There is a large body of existing recordings in symbolic formats. High-level features can be extracted from digital formats such as MIDI, MusicXML, Humdrum and GUIDO, and optical music recognition techniques can also be used to process scores into digital files from which high-level features can be extracted. It is therefore reasonable to pursue research in the use of high-level features extracted from recordings in symbolic formats. This largely untapped source of features could be used to supplement low-level features extracted from audio recordings. If automated transcription does improve, then audio recordings could be translated into symbolic form, and research in the use of high-level features would become even more useful. Furthermore, high-level features make it possible to classify scores, be they paper or digital, when audio

recordings are not available. It therefore seems appropriate to pursue research into classification with high-level features in parallel with further research involving low-level features.

In addition to practical applications, a system that can automatically classify recordings by genre has significant theoretical musicological interest as well. There is currently a relatively limited understanding of how humans construct musical genres, the mechanisms that they use to classify music and the characteristics that are used to perceive the differences between different genres. A system that could automatically classify music and reveal what musical dimensions it is using to do so would therefore be of great interest. Low-level signal processing based features are of little use in this respect, something that further emphasizes the importance of studying the use of high-level features.

This kind of research also has applications beyond the scope of genre classification. The techniques developed for a genre classification system could be adapted for other types of classifications, such as by compositional style or historical period. Once a classification system is implemented, one need only modify the particular training recordings and taxonomy that are used in order to perform arbitrary types of classifications.

The key to the success of all of this is the choice of features. Although effective classifiers are certainly necessary, the performance of even a perfect classifier is limited by its percepts. A realistic genre taxonomy will include tens, and possibly hundreds, of categories. Furthermore, the categories that are actually used in practice are inconsistent and contain a good deal of overlap. It is therefore necessary to have a wide range of features available in order to effectually segment such a difficult feature space.

Unfortunately, the use of too many features can overload classifiers by providing them with too much information. A good feature selection system is therefore essential so that the most relevant features are taken advantage of and the others are eliminated. The use of a hierarchical taxonomy paired with a successful feature selection system can improve performance, as one can first make coarse classifications using certain sets of features, and then use different sets of features to make increasingly finer classifications as one descends down the taxonomical hierarchy.

Section 2 of this paper reviews recent research in automatic genre classification and some musicological research on features in general. Section 3 provides an overview of the features that were used here and the philosophy behind their selection. Section 4 discusses the feature selection and classification techniques that were used and Section 5 explains how these techniques were used to perform hierarchical classifications. Section 6 presents the experiments that were performed to test the effectiveness of the system, and the results. Section 7 provides some final conclusions.

## 2. RELATED RESEARCH

There have been a number of interesting studies on automatic genre classification of audio files. Tzanetakis et al. [19, 20] have published research that used a variety of low-level features to achieve success rates of 61% when classifying between ten genres.

Additional research based on audio recordings has been performed by Pye, who successfully classified music into one of six categories 92% of the time [16]. Deshpande, Nam and Singh constructed a system that correctly classified among three categories 75% of the time [3]. Jiang et al. correctly classified 90.8% of recordings into five genres [8]. Kosina achieved a success rate of 88% with three genres [9]. Grimaldi, Kokaram and Cunningham achieved a success rate of 73.3% when classifying between five categories [6]. Xu et al. achieved a success rate of 93% with four categories [22]. McKinney and Breebaart achieved a success rate of 74% with seven categories [12].

There has been somewhat less research into the classification of symbolic data (e.g. MIDI). Chai and Vercoe were successful in correctly performing three-way classifications 63% of the time [2]. Shan and Kuo achieved success rates between 64% and 84% for two-way classifications [17]. Although these studies are very interesting, they focus more on pattern classification techniques rather than features.

There has also been some important research by Whitman and Smaragdis on combining features derived from audio recordings with “community metadata” that was derived from text data mined from the web [21]. Although beyond the scope of this paper, this line of research holds a great deal of potential, despite the problems related to finding, parsing and interpreting the metadata.

It should be noted that there are number of problematic issues that have been brought up relating to the formation of genre taxonomies in general [1, 14].

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. Lomax and his colleagues in the Cantometrics project [10] have performed the most extensive work to date in this direction. They compared thousands of songs from hundreds of different cultural groups using 37 features. Although there have been a few other efforts to list categories of features, they have tended to be overly broad. Works such as Tag’s “checklist of parameters” [18] are still useful as a general guide, however. A number of musicologists, such as Fabri [4], have also done some very interesting work on musical genre theory.

## 3. FEATURE EXTRACTION

In this study, features were extracted from MIDI files. This format was chosen because a diverse range of files are easily available in this format. Although it is true that

genre classification of MIDI files in particular is not a pressing problem from a practical perspective, the features discussed here could just as easily be extracted from other symbolic formats.

When choosing high-level features to use, it was necessary to keep in mind that it is desirable not only to have features that effectively partition recordings into different categories, but also to have features that are of musicological interest. As an initial step towards arriving at such features, one might look to how humans accomplish this task, 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 an area on which much of the theoretical literature has concentrated. This does not appear to be the case, however. Research has found that humans with little to moderate musical training are able to make genre classifications agreeing with those of record companies 72% of the time (among a total of 10 genres), based on only 300 milliseconds of audio [15]. 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. However, it is probably a better approach to extract features based on simple musical observations rather than on sophisticated theoretical models. Such models tend to have limited applicability beyond the 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 and characteristics of a particular aspect of a recording, as well as how much it varies.

A catalogue of 160 features that can be used to characterize and classify recordings was devised [11], 109 of which were implemented in the system discussed here. Although too numerous to discuss here in detail, these features belong to the following seven categories:

- **Instrumentation** (e.g. whether modern instruments are present)
- **Musical Texture** (e.g. standard deviation of the average melodic leap of different lines)
- **Rhythm** (e.g. average time between attacks)
- **Dynamics** (e.g. average note to note change in loudness)
- **Pitch Statistics** (e.g. fraction of notes in the bass register)
- **Melody** (e.g. fraction of melodic intervals comprising a tritone)
- **Chords** (e.g. prevalence of most common vertical interval)

Two types of features were used: one-dimensional features and multi-dimensional features. One-dimensional features each consist of a single number that represents an aspect of a recording in isolation. Multi-dimensional features consist of sets of related values that have limited significance taken alone, but together may reveal meaningful patterns. The reason for this differentiation is explained in Section 4.

## 4. CLASSIFICATION METHODOLOGY AND FEATURE SELECTION

The first stage in selecting a classification method involves choosing one of the three basic paradigms: expert systems, supervised learning and unsupervised learning. Expert systems involve explicitly implemented sets of rules, something that is not appropriate for genre classification, given the complexity of the task and the limited extent to which the content-based differences between genres are understood.

Unsupervised learning involves simply allowing a system to cluster samples together based on similarities that it perceives in the feature space. Although this is certainly useful for certain types of study, such as grouping anonymous recordings in order to gain insights into possible composers, it is of limited applicability to large-scale genre taxonomies. The genre categories that humans use are often inconsistent and illogical, so the groupings produced by unsupervised learning, although interesting theoretically, would likely bear little resemblance to the actual categories used by humans.

Supervised learning systems appear to be the best option, and were used exclusively here. These systems involve giving classifiers labelled training samples. The classifier then attempts to form (hopefully generalisable) relationships between the features of the training samples and the related categories.

Two well-known types of supervised classification techniques were used in the system described here: feedforward neural networks (NN) and k-nearest neighbour (KNN). NNs have the advantage of being able to simulate logical relationships between features, but can require long training times. KNN classifiers, in contrast, cannot simulate sophisticated logical relationships between features, but require essentially no training time. The use of both techniques allows one to use NNs where the modelling of more sophisticated relationships between features is likely to be most beneficial, while using KNN classifiers elsewhere in order to limit training times.

KNN classifiers operate by treating each sample as a point in feature space and finding the distribution of the categories of the $k$ training points closest to each test sample. Feedforward NNs operate by constructing a network of input units, hidden units and output units connected by weights. Each input to a unit is multiplied by its particular weight, and the sum of the results is fed into an activation function (the sigmoidal function, in this case). Training is performed by iteratively

performing a gradient descent through error space and modifying the weights.

The relative advantages and disadvantages of these two approaches was the motivation behind the one-dimensional and multi-dimensional features discussed in Section 3. For example, the bins of a histogram consisting of the relative frequency of different melodic intervals were treated as a multi-dimensional feature, but the average duration of melodic arcs was treated as a one-dimensional feature. Although it is of course true that all features are potentially interrelated logically, those sub-features grouped into multi-dimensional features were particularly subject to this interdependence.

Each multi-dimensional feature was classified by a separate multi-dimensional neural network, thus increasing the likelihood that appropriate relationships would be learned between the components of each multi-dimensional feature. The one-dimensional features, in contrast, were all processed by a single KNN classifier. This greatly reduced the training time, as the majority of features were one-dimensional, and training a neural network or networks to process them would have been too time consuming.

Feature selection was performed in several stages, all of which used genetic algorithms (GAs). GAs have been used successfully in the past for musical classification [5], and recent research has confirmed their fitness for feature selection and weighting [7].

GAs make use of "chromosomes" that contain bit strings that are iteratively evolved. The "fitness" of each chromosome is evaluated after each generation, and the best performers combine their bit strings to form the next generation. Techniques such as random "mutation" of bits and "cloning" of top performers can also be used. This results in increasingly fit chromosomes whose bit strings represent better solutions to problems.

GAs offer no guarantee of finding optimal solutions, but they often do provide good solutions in a reasonable amount of time. Considering that an exhaustive exploration of the feature selection problem is too computationally expensive when dealing with large numbers of features, GAs are a good alternative.

The first stage of feature selection was performed by using GAs to find the features that provided the best results for the KNN classifier. All other features were then ignored by the KNN classifier, and GAs were applied again to find the best relative feature weightings.

The NNs for each multi-dimensional feature were then trained. The combined classification of the KNN/NN ensemble was found by calculating an average of the classification scores for each category produced by each component of the ensemble (i.e. the KNN and each NN). A final feature selection stage was then performed by applying GAs to each of the components of the ensemble in order to potentially eliminate some. A final weighting was evolved using GAs for each of the surviving members of the ensemble.

The result of all of this after training was a weighted ensemble of classifiers consisting of a single KNN classifier using a weighted subset of all possible one-dimensional features and a set of NNs representing a subset of all possible multi-dimensional features. Such a classifier ensemble could be seen as a black box that took in the entire feature set of a recording as input, ignored the features it had selected out, and output a classification score for each candidate category that it had been trained to recognize. A number of these black boxes were trained to classify recordings hierarchically, as described in Section 5 below.

## 5. HIERARCHICAL CLASSIFICATION

As mentioned previously, classification was performed hierarchically. Recordings were first classified by "root genre" (i.e. the broadest genre categories, such as Jazz, Classical or Popular). Classification then proceeded to the next level of the hierarchy, where only the sub-categories of the winners of the previous stage of classification were considered as candidates. This continued iteratively down the hierarchy of genres until only "leaf genres" (i.e. genres with no sub-categories) remained, and these were chosen as the winning genres.

The classification at each level of the hierarchy involved separately trained specialist classifier ensembles of the type presented in Section 4. Each of these classifier ensembles was trained only on recordings belonging to their candidate categories, and therefore developed feature selections and weightings especially suited to their categories. A Jazz classifier, for example, might be trained only on Swing, Bebop and Funky Jazz recordings, whereas a root classifier would be trained on recordings of all genres. A root classifier would therefore likely be good at making coarse classifications, but a Jazz classifier would likely be better at classifying a recording into specialized sub-genres of jazz once the recording had been labelled as Jazz by the root classifier.

Hierarchical classification has the potential weakness that a mistake made at a broad level of the hierarchy can lead to a decent through an entirely erroneous branch of the hierarchy. Basic flat classification was therefore performed as well as hierarchical classification, in order to provide a reference point that would enable one to determine whether hierarchical classification did indeed improve performance. A more detailed experimental comparison of the application of different classification methodologies to this problem, including hybrid methodologies, will be available in [11].

## 6. THE EXPERIMENT

A total of 950 MIDI recordings were used for training and testing, using a five-fold cross-validation process. Two different taxonomies were used in order to assess the performance of the system under different conditions. The number of candidate features available

for feature selection was also varied, in order to judge the significance of the number of features available.

The first taxonomy that was used consisted of three root genres and nine leaf genres, as shown in Figure 1. This particular taxonomy was chosen because it is comparable in size and diversity to the previous research discussed in Section 2.

| Classical | Jazz | Popular |
|---|---|---|
| Baroque | Bebop | Country |
| Modern | Funky Jazz | Punk |
| Romantic | Swing | Rap |

**Figure 1**. Basic taxonomy used.

The results using hierarchical classification were excellent, as can be seen in Figure 2. On average, the root genre was correctly identified 98% of the time and the leaf genre was correctly identified 90% of the time. These rates are not only much higher than those of the existing symbolic data systems discussed in Section 2, but also as good, and in most cases better, than the existing audio systems. Furthermore, the experiment performed here with flat classification resulted in reduced success rates of 96% and 86% for root and leaf genres respectively. This decrease in performance demonstrates the utility of hierarchical classification.

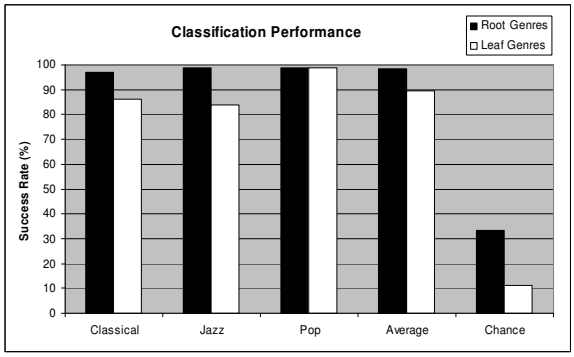


**Figure 2**. Average classification success rate. The leaf genre bars give the average success rates of the leaf genres belonging to the corresponding root genres.

It was found that the particular features selected by each of the classifier ensembles varied significantly. This was as expected, as the types of features that would be useful in differentiating between different types of Classical music, for example, would reasonably be different than those used to differentiate between Country and Rap music. The particular features chosen by different classifier ensembles is a source of great musicological interest, and would be a rich topic for a future paper. Although generalizations are difficult to make due to the variety between classifier ensembles, one interesting observation that can be made is that features related to instrumentation were particularly important, as they were assigned a collective average of 47% of the weightings allocated amongst the features comprising the seven feature groups.

The differences between the weightings evolved by the different classifier ensembles implies that the development of a large catalogue of features than can serve as candidates for feature selection for different specialized classifiers could be of great use. An additional experiment was performed to further investigate this. The experiment described above was repeated three more times, but only a randomly selected subset of the total available features was made visible to the feature selection systems. The results, displayed in Figure 3, demonstrate that performance increased significantly when more candidate features were available for feature selection.

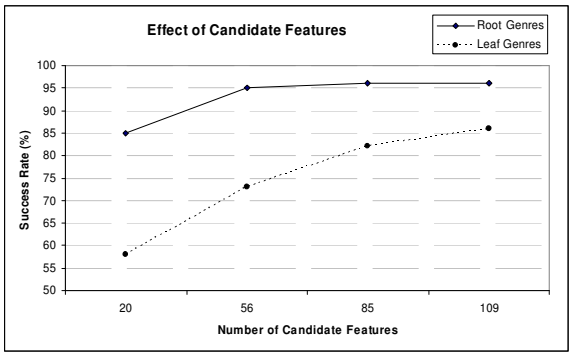


**Figure 3**. Effect of number of features available for feature selection on flat classification success rates.

A taxonomy with nine leaf categories is of course somewhat limited from the perspective of realistic classification problems. In order to investigate the real-world applicability of the system, an experiment was performed using an expanded multi-level taxonomy composed of 9 root categories and 38 leaf categories, a taxonomy much larger than has previously been used to test an automatic genre classification system. The root genres and the leaf genres were respectively successfully identified 81% and 57% of the time. Although not sufficiently high for practical purposes, these rates were significantly higher than chance (11% and 3% respectively), and the results show that there is at least the potential to successfully deal with realistically sized taxonomies, something which has not been done previously.

## 7. CONCLUSIONS

Very encouraging success rates of 98% for root genres and 90% for leaf genres were achieved for a taxonomy consisting of 3 root genres and 9 leaf genres. These rates compare favourably to results achieved in previous research. Further experiments demonstrated that increasing the number of features available to a hierarchical classification system that uses feature selection to train specialist classifiers causes corresponding improvements in performance. A further experiment with a greatly expanded taxonomy showed the potential of the system described here for successfully dealing with a realistic taxonomy, something that is as yet an unsolved problem.

This research demonstrated the effectiveness of large sets of high-level musical features for genre

classification when paired with good feature selection methods and hierarchical classification. Future research into both the use of the hierarchical classification techniques described here and the development of further high-level features is certainly warranted, and could potentially lead to a viable classifier capable of dealing with realistically large genre taxonomies. Further study of which features were selected by which specialist classifier ensembles could also be of great musicological interest.

## 8. ACKNOWLEDGEMENTS

Thanks to the *Fonds Québécois de la recherche sur la société et la culture* for their generous financial support, which has helped to make this research possible.

## 9. REFERENCES

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

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

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

[4] Fabbri, F. 1982. What kind of music? *Popular Music* 2: 131–43.

[5] Fujinaga, I. 1996. Exemplar-based learning in adaptive optical music recognition system. *Proceedings of the International Computer Music Conference*. 55–6.

[6] 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.

[7] Hussein, F., R. Ward, and N. Kharma. 2001. Genetic algorithms for feature selection and weighting, a review and study. *International Conference on Document Analysis and Recognition.* 1240-4.

[8] 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*. 113-6.

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

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

[11] McKay, C. In press. Automatic genre classification of MIDI recordings. *Master's thesis*. McGill University, Canada.

[12] McKinney, M. F., and J. Breebaart. 2003. Features for audio and music classification. *Proceedings of the International Symposium on Music Information Retrieval*. 151–8.

[13] North, A. C., and D. J. Hargreaves. 1997. Liking for musical styles. *Music Scientae* 1: 109–28.

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

[15] 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.

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

[17] Shan, M. K., and F. F. Kuo. 2003. Music style mining and classification by melody. *IEICE Transactions on Information and Systems* E86-D (3): 655–69.

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

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

[20] Tzanetakis, G., G. Essl, and P. Cook. 2001. Automatic musical genre classification of audio signals. *Proceedings of the International Symposium on Music Information Retrieval*. 205–10.

[21] 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.

[22] 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 429–32.