COMBINING FEATURES EXTRACTED FROM AUDIO, SYMBOLIC AND CULTURAL SOURCES

Cory McKay **Ichiro Fujinaga**
Music Technology Area and CIRMMT, Schulich School of Music, McGill University
Montreal, Quebec, Canada
cory.mckay@mail.mcgill.ca, ich@music.mcgill.ca

ABSTRACT

This paper experimentally investigates the classification utility of combining features extracted from separate audio, symbolic and cultural sources of musical information. This was done via a series of genre classification experiments performed using all seven possible combinations and subsets of the three corresponding types of features.

These experiments were performed using jMIR, a software suite designed for use both as a toolset for performing MIR research and as a platform for developing and sharing new algorithms.

The experimental results indicate that combining feature types can indeed substantively improve classification accuracy. Accuracies of 96.8% and 78.8% were attained respectively on 5 and 10-class genre taxonomies when all three feature types were combined, compared to average respective accuracies of 85.5% and 65.1% when features extracted from only one of the three sources of data were used. It was also found that combining feature types decreased the seriousness of those misclassifications that were made, on average, particularly when cultural features were included.

1. INTRODUCTION

Music is a multi-faceted area of inquiry, and there are many factors that influence any individual's experience of music. Given the variety of possible approaches to studying music, researchers can have a tendency to focus their attentions on limited types of music or on particular specialized approaches to studying music. One of the most exciting aspects of the contemporary music information retrieval (MIR) field is that it attempts to break down these barriers using a multi-disciplinary approach to music.

In the past, MIR research has tended to focus on studying either audio (e.g., MP3), symbolic (e.g., MIDI) or cultural (e.g., web data, user tags, surveys, etc.) sources of information about music. Traditionally, research using each of these sources has been relatively segregated based on whether a researcher had a corresponding background in signal processing, music theory or library sciences and data mining. In recent years, however, MIR researchers have increasingly begun to study these sources of information in combination, with a particular emphasis on research combining audio and cultural sources of data.

Features extracted from all three types of data can provide valuable information for use in music classification and similarity research. Audio is clearly useful because it is the essential way in which music is consumed, and cultural data external to musical content is well-known to have a large influence on our understanding of music [8].

Symbolic data has recently been receiving less attention from MIR researchers than it did in the past. The value of symbolic data should not be overlooked, however, as much of the information associated with high-level musical abstractions that can be relatively easily extracted from symbolic formats is currently poorly encapsulated by the types of features that are typically extracted from audio, which tend to focus primarily on timbral information

Symbolic formats can thus, at the very least, be a powerful representational tool in automatic music classification. This characteristic will become increasingly valuable as polyphonic audio to symbolic transcription algorithms continue to improve. Even though such technologies are still error-prone, it has been found that classification systems can be relatively robust to such errors [6].

This paper focuses on an experimental investigation of the extent to which combining features extracted from the three types of data can be advantageous in automatic music classification. If the orthogonal independence of the feature types is high, then performance boosts can potentially be attained in a variety of applications by combining features extracted from the different data types.

This investigation was performed via sets of automatic genre classification experiments. Genre classification in particular was chosen because it is a complex and difficult task that combines diverse musical variables. The experiments could just as easily have been performed using other types of classification, however, such as mood or artist classification. The essential issue being investigated remains the potential performance improvements attained by combining features extracted from the three types of data.

2. RELATED RESEARCH

There has been a significant amount of research on combining audio and cultural data. Whitman and Smaragdis

[14] performed particularly important early work on combining audio features with cultural data mined from the web, and achieved substantial performance gains when doing so. Dhanaraj and Logan [3] took a more content-based approach by combining information extracted from lyrics and audio. Others have combined audio and cultural data for the purpose of generating music browsing spaces (e.g., [5]). Aucouturier and Pachet [2] used a hybrid training approach based on acoustic information and boolean metadata tags. Research has also been done on using audio data to make correlations with cultural labels, which can in turn improve other kinds of classification (e.g., [12]).

There has been much less work on combining symbolic data with audio data. Lidy et al. [6] did, however, find that combining audio and symbolic data can result in improved performance compared to when only audio data is used.

To the best knowledge of the authors, no previous research has been performed on combining symbolic and cultural features or on combining all three feature types.

Far too many papers have been published on automatic genre classification in general to cite with any completeness here. One influential work that bears particular mention, however, is that of Tzanetakis and Cook [13].

3. THE JMIR SOFTWARE SUITE

jMIR is a suite of software tools developed for use in MIR research. It was used to perform the experiments described in this paper. jMIR includes the following components:

- **jAudio** [7]: An audio feature extractor that includes implementations of 26 core features. jAudio also includes implementations of "metafeatures" and "aggregators" that can be used to automatically generate many more features from these core features (e.g., standard deviation, derivative, etc.).
- **jSymbolic** [10]: A symbolic feature extractor for processing MIDI files. jSymbolic is packaged with 111 mostly original features [8].
- **jWebMiner** [11]: A cultural feature extractor that extracts features from the web based on search engine co-occurrence page counts. Many user options are available to improve results, including search synonyms, filter strings and site weightings.
- **ACE** [9]: A meta-learning classification system that can automatically experiment with a variety of different dimensionality reduction and machine learning algorithms in order to evaluate which are best suited to particular problems. ACE can also be used as a simple automatic classification system.

The jMIR components can be used either independently or as an integrated suite. Although the components can read and write to common file formats such as Weka ARFF, jMIR also uses its own ACE XML file formats that offer a number of significant advantages over alternative data mining formats [9].

The components of jMIR were designed with the following goals in mind:

- Provide a flexible set of tools that can easily be applied to a wide variety of MIR-oriented research tasks.
- Provide a platform that can be used to combine research on symbolic, audio and/or cultural data.
- Provide easy-to-use and accessible software with a minimal learning curve that can be used by researchers with little or no technological training.
- Provide a modular and extensible framework for iteratively developing and sharing new feature extraction and classification technologies.
- Provide software that encourages collaboration between different research centers by facilitating the sharing of research data using powerful and flexible file formats.

In order to improve accessibility, each of the jMIR components is, with the temporary exception of ACE, packaged with an easy-to-use GUI. The jMIR components also include user manuals and help systems.

The jMIR components are all implemented in Java, in order to make them as platform-independent as possible. They are open-source and distributed free of charge.[1]

4. THE SAC DATASET

The SAC (Symbolic, Audio and Cultural) dataset was assembled by the authors in order to provide matching symbolic, audio and cultural data for use in the experiments described in Section 5. SAC consists of 250 MIDI files and 250 matching MP3s, as well as accompanying metadata (e.g., title, artist, etc.). This metadata is stored in an iTunes XML file,[2] which can be parsed by jWebMiner in order to extract cultural features from the web.

It was decided to acquire the matching MIDI and audio recordings separately, rather than simply synthesizing the audio from the MIDI. Although this made acquiring the dataset significantly more difficult and time consuming, it was considered necessary in order to truly test the value of combining symbolic and audio data. This is because audio generated from MIDI by its nature does not include any additional information other then the very limited data encapsulated by the synthesis algorithms.

SAC is divided into 10 different genres, with 25 pieces of music per genre. These 10 genres consist of 5 pairs of similar genres, as shown in Figure 1. This arrangement makes it possible to perform 5-class genre classification experiments as well as 10-class experiments simply by combining each pair of related genres into one class. An additional advantage is that it becomes possible to meas-

[1] jmir.sourceforge.net

[2] Contact cory.mckay@mail.mcgill.ca for access to the file.

ure an indication of how serious misclassification errors are in 10-class experiments by examining how many misclassifications are in an instance's partner genre rather than one of the other 8 genres.

Blues: Modern Blues *and* Traditional Blues
Classical: Baroque *and* Romantic
Jazz: Bop *and* Swing
Rap: Hardcore Rap *and* Pop Rap
Rock: Alternative Rock *and* Metal

Figure 1: The ten genres found in the SAC dataset.

5. EXPERIMENTAL PROCEDURE

The first step of the experiment was to use jMIR's three feature extractors to acquire features from each matched audio recording, MIDI recording and cultural metadata. Details on the particular features extracted are available elsewhere ([7], [8], [10] and [11]).

To provide a clarifying example, features might be extracted from a Duke Ellington MP3 recording of *Perdido,* from an independently acquired MIDI encoding of the same piece, and from automated search engine queries using metadata such as artist and title. Three types of feature sets were therefore extracted for each piece, one corresponding to each of the three data types.

These three types of features were then grouped into all seven possible subset combinations. This was done once for each of the two genre taxonomies, for a total of fourteen sets of features (as shown in Table 1), in preparation for fourteen corresponding classification experiments designed to measure how well the different feature sets performed relative to one another.

Feature Type	5-Genre Code	10-Genre Code
Symbolic	S-5	S-10
Audio	A-5	A-10
Cultural	C-5	C-10
Symbolic + Audio	SA-5	SA-10
Audio + Cultural	AC-5	AC-10
Symbolic + Cultural	SC-5	SC-10
Symbolic + Audio + Cultural	SAC-5	SAC-10

Table 1: The identifying codes for the 14 experiments.

ACE was then trained on and used to classify each of the fourteen feature sets in fourteen independent 10-fold cross-validation experiments. This resulted in two classification accuracy rates for each of the seven feature type combinations, one for each of the two genre taxonomies. As a side note, ACE includes dimensionality reduction functionality, so training was actually performed with automatically chosen subsets of the available features.

It is desirable not only to determine how effective each of the feature type combinations are at achieving correct classifications, but also how serious those misclassifications that do arise are. Two classifiers with similar raw classification accuracy rates can in fact be of very different value if the misclassifications made by one classifier result in classes that are more similar to the "correct" class.

A normalized weighted classification accuracy rate was therefore calculated for each of the 10-genre experiments in order to provide insight on error types. This was calculated by weighting a misclassification within a genre pair (e.g., Alternative Rock instead of Metal) as 0.5 of an error, and by weighting a misclassification outside of a pair (e.g., Swing instead of Metal) as 1.5 of an error.

6. RESULTS AND DISCUSSION

6.1. Results

The average classification accuracy rates across cross-validation folds for each of the fourteen experiments outlined in Table 1 are shown in Table 2, including weighted and unweighted results. Figures 2 and 3 illustrate the unweighted results for the 5-genre and 10-genre experiments respectively. These results are summarized in Table 3 and Figure 4, which show the average results for all experiments using one feature type, all experiments using two feature types and all experiments using three feature types.

6.2. Effects of Combining Feature Types on Accuracy

As can be seen in Figures 2 and 3, all combinations of two or three feature types performed substantially better than all single feature types classified independently. Furthermore, combining all three feature types resulted in better performance than most pairs of feature types.

This is illustrated in Figure 4, which shows important average increases in performance when feature types are combined. Combining all three feature types resulted in increases in performance of 11.3% on the 5-genre taxonomy and 13.7% in the 10-genre taxonomy, compared to the average performances of each of the single feature types classified individually. Considered in terms of percentage reduction in error rate, this corresponds to impressive improvements of 78.0% and 39.3% for the 5 and 10-genre taxonomies, respectively.

A Wilcoxon signed-rank test indicates that, with a significance level of 0.125, the improvements in performance of two or three feature types over one type were statistically significant in all cases. However, the improvements when three feature types were used instead of two were not statistically significant. The corresponding average increases in performance were only 2.3% and 2.7% for the 5 and 10-genre taxonomies, respectively.

Overall, these results indicate that, at least to some extent, the three different types of features contain orthogonally independent information, and can therefore be profitably combined for a variety of purposes.

	S	A	C	SA	AC	SC	SAC
5-UW	86.4	82.8	87.2	92.4	95.2	94	96.8
10-UW	66.4	67.6	61.2	75.6	78.8	75.2	78.8
10-W	66.4	67.4	66.6	78.6	84.6	81.2	84.2

Table 2: The unweighted classification accuracy rates for the 5-genre (5-UW) experiments and both the unweighted (10-UW) and weighted (10-W) accuracy rates for the 10-genre experiments. Results are reported for each feature type combination, as described in Table 1. All values are average percentages calculated over cross-validation folds.

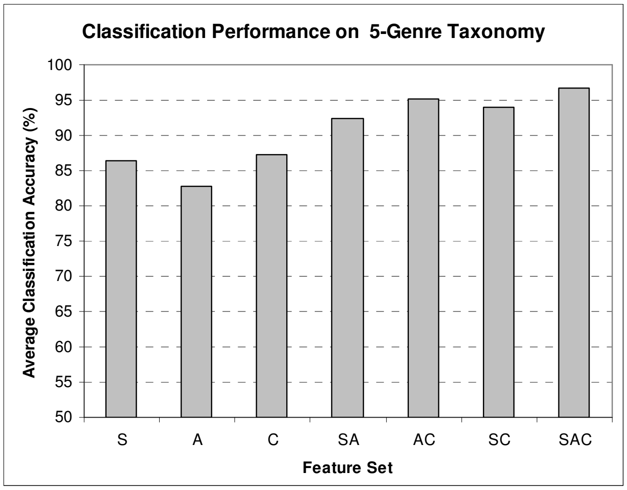

Figure 2: The classification accuracy rates for the 5- genre taxonomy, as described in Table 1.

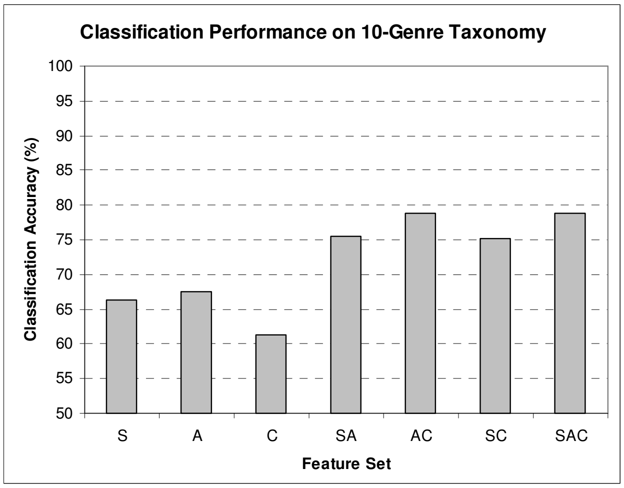

Figure 3: The unweighted classification accuracy rates for the 10-genre taxonomy, as described in Table 1.

	1 Feature Type	2 Feature Types	3 Feature Types
5-UW	85.5	93.9	96.8
10-UW	65.1	76.5	78.8
10-W	66.8	81.5	84.2

Table 3: The average classification accuracy rates for all experiments employing just one type of feature (S, A and C), two types of features (SA, AC and SC) or all three types of features (SAC). Results are specified for the 5-genre taxonomy (5-UW), the unweighted 10-genre taxonomy (10-UW) and the weighted 10-genre taxonomy (10-W). All values are percentages, and are calculated as simple mean averages from Table 2.

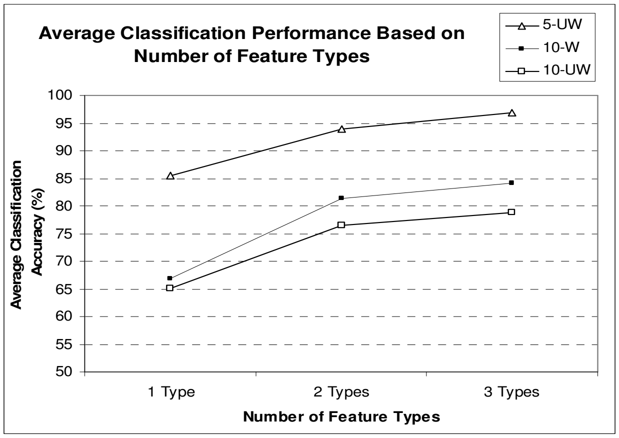

Figure 4: The average classification accuracy rates for all experiments employing just one type of feature (S, A and C), two types of features (SA, AC and SC) or all three types of features (SAC). The three trend lines refer to the 5-genre taxonomy (5-UW), the unweighted 10-genre taxonomy (10-UW) and the weighted 10-genre taxonomy (10-W). This data corresponds to the values in Table 3.

6.3. Types of Misclassifications

As described in Section 5, normalized weighted classification accuracy rates were calculated for the experiments on the 10-genre taxonomy in order to evaluate the seriousness of the particular misclassifications that were made. The results, and how they compare to the unweighted classification accuracies, are shown in Table 2 and Figure 5.

The weighted and unweighted accuracies were not significantly different when the audio and symbolic features were processed individually. However, the weighted performance was 3% higher than the unweighted performance when these two feature types were combined. Although this is not a dramatic increase, it is an indication that combining these feature types can make those misclassifications that do occur result in classes closer to the model classes in addition to increasing classification accuracy itself, as discussed in Section 6.2.

The differences between the weighted and unweighted classification accuracies were greater in all feature sets that included cultural features. These weighted rates were higher than the unweighted rates by an average of 5.7%, a difference that, based on Student's paired t-test, is statistically significant with a significance level of 0.005.

Overall, these results indicate that the types of misclassifications that occur when cultural features are used are less serious than when audio or symbolic features are used alone. Quite encouragingly, it also appears that this improvement in error quality carries through when cultural features are combined with audio and symbolic features.

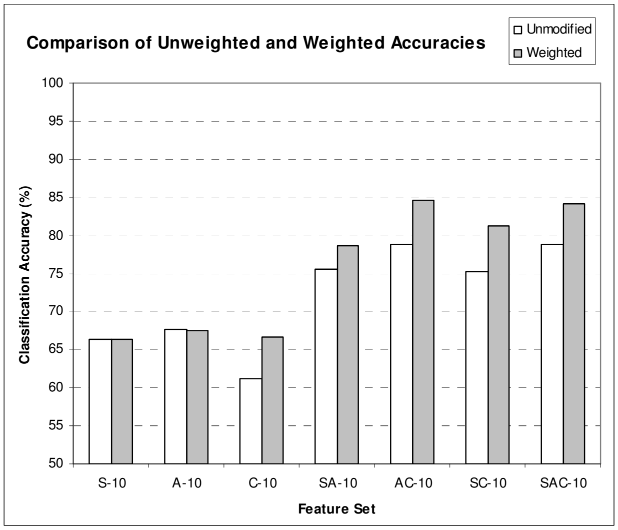

Figure 5: The differences between the unweighted and weighted classification accuracies on the 10-genre taxonomy for each of the seven feature type combinations. This data corresponds to the last two rows of Table 2.

6.4. General Genre Classification Performance

In order to put the experimental results described here in context, it is appropriate to compare them with classification accuracies achieved by high-performing existing specialized genre classification systems. It is important, however, to keep in mind the essential caveat that different classification systems can perform dramatically differently on different datasets, so direct comparisons of classification accuracies calculated on different datasets can give only a very rough indication of comparative performance.

The MIREX (Music Information Retrieval Evaluation eXchange) evaluations offer the best benchmarking reference points available. Although no evaluations of genre classification based on cultural data have been carried out yet at MIREX, both symbolic and audio genre classification evaluations have been held, most recently in 2005 and 2007, respectively. The highest accuracy for symbolic classification was 84.4%, attained on a 9-genre taxonomy by McKay and Fujinaga's Bodhidharma system [15]. The highest classification accuracy attained in audio classification was 68.3%, achieved by the University of Illinois's International Music Information Retrieval Systems Evaluation Laboratory (IMIRSEL) team, on a 10-genre taxonomy [16].

The experiments described in this paper achieved classification accuracies of 67.6% using only features extracted from audio and 66.4% using only features extracted from symbolic data. This is comparable to the best MIREX audio result of 68.3%, but significantly lower than the best MIREX symbolic result of 84.4%, which was achieved on a taxonomy only smaller by one class.

This latter result is intriguing, as jSymbolic uses the same features and feature implementations as Bodhidharma. The difference is likely due at least in part to the specialized and sophisticated hierarchical and round-robin learning classification ensemble algorithms used by Bodhidharma [8], whereas ACE only experiments with general-purpose machine learning algorithms.

When all three feature types were combined, the jMIR experiments described in this paper achieved a success rate of 78.8% which was still lower than Bodhidharma's performance, but significantly better than the best audio MIREX results to date.

Taken in the context of the particular difficulty of the SAC dataset, and when it is considered that the accuracy on the 10-genre taxonomy improves to 84.2% when weighted, the results attained here are encouraging, and may be an indication that the ultimate ceiling on performance might not be as low as some have worried [1]. It may well be that the use of more sophisticated machine learning approaches, such as those used by Bodhidharma or by DeCoro et al. [4], combined with the development of new features, could significantly improve performance further.

7. CONCLUSIONS AND FUTURE RESEARCH

The experimental results indicate that it is indeed substantively beneficial to combine features extracted from audio, symbolic and cultural data sources, at least in the case of automatic genre classification. Further research remains to be performed investigating whether these benefits generalize to other areas of music classification and similarity research.

All feature groups consisting of two feature types performed significantly better than any single feature types classified alone. Combining all three feature types resulted in small further improvement over the feature type pairs on average, but these additional improvements were not as uniform nor were they statistically significant.

The results also indicate that combining feature types tends to cause those misclassifications that do occur to be less serious, as the misclassifications are more likely to be to a class that is more similar to the model class. Such improvements were particularly pronounced when cultural features were involved.

Encouragingly high genre classification accuracy rates were attained. The results of the experiments as a whole provide hope that any ultimate ceiling on genre classification performance might not be as low as has been worried.

The jMIR software suite was demonstrated to be an effective and convenient tool for performing feature extraction and classification research. The SAC dataset was also found to provide a good basis for performing combined audio, symbolic and cultural experiments.

An important next step is to repeat the experiments performed here with MIDI files transcribed from audio, in order to investigate more practically significant use cases where MIDI files do not have to be manually harvested. This would also enable experiments on larger datasets.

There are also plans to combine feature types in more sophisticated ways, such as by segregating them among different weighted specialist classifiers collected into blackboard ensembles. Sophisticated classification techniques that take advantage of ontological structuring, such as those utilized in Bodhidharma, also bear further investigation, and could fruitfully be incorporated into ACE. There are also plans to expand SAC as well as to apply jMIR to a wider variety of MIR research applications, such as mood and artist classification.

8. ACKNOWLEDGEMENTS

The authors would like to thank the *Centre for Interdisciplinary Research in Music Media and Technology* (CIRMMT), the *Social Sciences and Humanities Research Council* (SSHRC) and the *Canadian Foundation for Innovation* (CFI) for their generous financial support.

9. REFERENCES

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

[2] Aucouturier, J. J., and F. Pachet. 2007. Signal + context = better. *Proceedings of the International Conference on Music Information Retrieval.* 425–30.

[3] Dhanaraj, R., and B. Logan. 2005. Automatic prediction of hit songs. *Proceedings of the International Conference on Music Information Retrieval.* 488–91.

[4] DeCoro, C., Z. Barutcuoglu, and R. Fiebrink. 2007. Bayesian aggregation for hierarchical genre classification. *Proceedings of the International Conference on Music Information Retrieval.* 77–80.

[5] Knees, P., M. Schedl, T. Pohle, and G. Widmer. 2006. An innovative three-dimensional user interface for exploring music collections enriched with meta-information from the web. *Proceedings of the ACM International Conference on Multimedia.* 17–24.

[6] Lidy, T., A. Rauber, A. Pertusa, and J. M. Iñesta. 2007. Improving genre classification by combination of audio and symbolic descriptors using a transcription system. *Proceedings of the International Conference on Music Information Retrieval.* 61–6.

[7] McEnnis, D., C. McKay, and I. Fujinaga. 2006. jAudio: Additions and improvements. *Proceedings of the International Conference on Music Information Retrieval.* 385–6.

[8] McKay, C. 2004. Automatic genre classification of MIDI recordings. *M.A. Thesis.* McGill University, Canada.

[9] McKay, C., R. Fiebrink, D. McEnnis, B. Li, and I. Fujinaga. 2005. ACE: A framework for optimizing music classification. *Proceedings of the International Conference on Music Information Retrieval.* 42–9.

[10] McKay, C., and I. Fujinaga. 2006. jSymbolic: A feature extractor for MIDI files. *Proceedings of the International Computer Music Conference.* 302–5.

[11] McKay, C., and I. Fujinaga. 2007. jWebMiner: A web-based feature extractor. *Proceedings of the International Conference on Music Information Retrieval.* 113–4.

[12] Reed, J., and C. H. Lee. 2007. A study on attribute-based taxonomy for music information retrieval. *Proceedings of the International Conference on Music Information Retrieval.* 485–90.

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

[14] 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.

[15] *2005 MIREX contest results – symbolic genre classification.* Retrieved April 1, 2008, from http://www.music-ir.org/evaluation/mirex-results/sym-genre/index.html.

[16] *Audio genre classification results - MIREX 2007.* Retrieved April 1, 2008, from http://www.music-ir.org/mirex/2007/index.php/Audio_Genre_Classification_Results.