Combining Musical and Cultural Features for Intelligent Style Detection

Brian Whitman
MIT Media Lab
Cambridge MA U.S.A.
+1 617 253 0112
bwhitman@media.mit.edu

Paris Smaragdis
MIT Media Lab
Cambridge MA U.S.A.
+1 617 253 0405
paris@media.mit.edu

ABSTRACT

In this paper we present a musical style identification scheme based on simultaneous classification of auditory and textual data. Style identification is a task which often involves cultural aspects not present or easily extracted through auditory processing. The scheme we propose complements any audio driven genre or style detection system with a classifier based on web-extracted data we call 'community metadata.' The addition of these cultural attributes in our feature space aids in proper classification of acoustically dissimilar music within the same style, and similar music belonging to different styles.

1. INTRODUCTION

Musical genres aid in the listening-and-retrieval (L&R) process by allowing a user or consumer a sense of reference. By organizing physical shelves in record stores by genres, shoppers can browse and discover new music by walking down an aisle. But the digitization of musical culture carries an embarrassing problem of how to organize collections: folders full of music recordings, peer-to-peer virtual terabyte lockers and handheld devices all need the same attention to organization as rooftop music stores. As a result, recent work has approached the problem of automatic genre recognition [8] [2], creating top-level clusters of similar music (rock, pop, classical, etc.) from the acoustic content.

While the high level separation of genres is useful, we tend to look more toward styles for discovering new music or for accurate recommendation. Styles usually define subclasses of genres (in the genre Country we can choose from 'No Depression,' 'Contemporary Country,' or 'Urban Cowboy'), but sometimes join together artists across genres. Stores (real or virtual) normally do not partition their space by style to avoid consumer confusion ("*is Bjork in electronic pop or female vocal?*") but they can provide cross-reference data (as in the case of the All Music Guide (`http://www.allmusic.com`); and recommendation engines can utilize styles for high-confidence results.

Style is an imperative class of description for most music retrieval tasks, but is usually considered a 'human' concept and can be hard to model. Some styles evolved with no acoustic underpinnings: a favorite is intelligent-dance-music or 'IDM,' in which the included artists range from the abstract sine-wave noise of Pan Sonic to the calm filtered melodies of Boards of Canada. At first glance, IDM would be an intractable set to model due to its similarity being almost purely cultural. As such, we usually rely on marketing, print publications, recommendations of friends ("*they're like x, only different*") to understand styles on our own.

In this paper we present an automatic style detection system that operates on both the acoustic content of the audio and the very powerful 'cultural representation' of community metadata, using descriptive textual features extracted from automated crawls of the web. The community metadata feature space has previously shown to be effective in a music similarity task on its own [10], and here we augment it with an audio representation. This combined model performs extremely well in identifying a set of previously edited style clusters, and can be used to cluster arbitrarily large new sets of artists.

2. PRIOR WORK

2.1 Genre Classification

Automatic genre classification techniques that explicitly compute clusters from the score or audio level have reported high results in musically or acoustically separable genres such as classical vs. rock, but the hierarchical structure of popular music lends itself to a more finegrained set of divisions.

Using the score level only, (MIDI files, transcribed music or CSound scores) systems can extract style or genre using easily-extractable features (once the music is in a common format, which may require character recognition on a score, or parsing a MIDI file) such as key and frequently used progressions. Systems normally perform genre classification by clustering similar music segments, or performing a one-in-n (where n is the number of genres) classification using some machine learning technique. In [5], various machine learning classifiers are trained on performance characteristics of the score to learn a piece-global 'style,' and in [2] three types of folk music were separated using a Hidden Markov Model.

Approaches that perform genre classification in the audio domain use a combination of spectral features and musically-informed inferred features. Genre identification work undertaken in [8] aims to understand acoustic content enough to classify into a small set of related clusters by studying the spectra along with tempo-sensitive 'timbregrams' with a simple beat detector in place. Similar work treating artists as complete genres (where similar clusters of artist form a 'genre') is studied in [9] and then improved on in [1] with more musical knowledge.

2.2 Cultural Feature Extraction

Cultural features concerning music are not as well-defined and vary with time and interpretation. Any representation that aims to express music as community description is a form of cultural features. The most popular form of cultural features, lists of purchased music, are used in collaborative filtering to recommend music based on their peers' tastes. Cultural features are important to express information about music that cannot be captured by the actual audio content. Many music retrieval tasks cannot do well on audio alone.

A more automatic and autonomous way of collecting cultural features is described in [10]. There, we define 'community metadata' (which is used in this paper) as a vector space of descriptive textual

terms crawled from the web. For example, an artist is represented as their community description from album reviews and fan-created pages. An important asset of community metadata is its attention to time: in our implementation, community metadata vectors are crawled repeatedly and future retrievals take the time of description into account. In a domain where long-scale time is vastly important, this representation allows recommendations and classifications to take the 'buzz factor' into account.

Cultural features for music retrieval are also explored in [4], where web crawls for 'my favorite artists' lists are collated and used in a recommendation agent. The specifics of the community metadata feature vector are described in greater detail below.

3. STYLE CLASSIFICATION

To test our feature space and hypotheses concerning automatic style detection, we chose a small set of artists spanning five separate styles as classified by music editors. In turn, we first make classifications based solely on an audio representation, then a community metadata representation, and lastly show that the combined feature spaces perform the best in separating the styles.

3.1 Data Set

For the results in this paper we operate on a fixed data set of artists chosen from the Minnowmatch music testbed (related work analyzes this database in [9], [1], [10].) The list used contained twenty-five artists, encapsulating five artists each across five music styles. The list is shown in Table 1.

Each artist in represented in the Minnowmatch testbed with one or two albums worth of audio content. The selection of artists in the testbed was defined by the output of a peer-to-peer network robot which computed popularity of songs by watching thousands of users' collections. We have previously crawled for the community metadata for each artist in the Minnowmatch tested in January of 2002.

The 'ground truth' style classification was taken from the All Music Guide at `http://www.allmusic.com` (AMG), a popular edited web resource for music information. We consider AMG our 'best-case' ground truth due to its collective edited nature. Although AMG's decisions are subjective, our intent is to show that a computational metric involving both acoustic and cultural features can approximate an actual labeling from a professional.

The size of our data is intentionally small so as to demonstrate the issues of acoustic versus cultural similarity presented in this paper. This simulation is not meant to represent a fully functioning system due to its scope, but the approach and results propose a viable solution to the problem.

4. AUDIO-BASED STYLE CLASSIFICATION

One obvious feature space for a music style classifier is the audio domain. While we will show that it is not always the best way to discern cultural labels such as styles, we can say it is a very good indicator of the 'sound' of music and perhaps as a higher-level genre classifier.

The audio-based style classifier operates by forming each song into a representation and training a neural network to classify a new song from a test set into one of the five classes. Below, we describe the representation used and the training process.

4.1 Representation

We chose a fairly simple representation for this experiment. For each artist in our set, we chose on average 12 songs randomly from their collection. The audio tracks were downsampled to 11,025Hz, converted to mono, and transformed to zero mean and unit variance. We subsequently extracted the 512-point power spectral density (PSD) of every three seconds of audio and performed dimensionality reduction using principal components analysis (PCA) to the entire training data set to reduce it down to twenty dimensions. The process is described in Figure 1. The series of the reduced PSD features as extracted from all the available audio tracks were used as the representation of every artist.

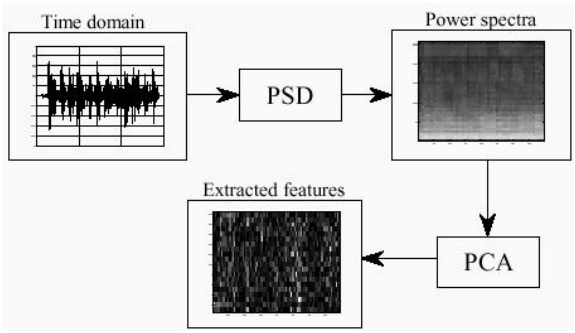

Figure 1: Feature extraction path for audio-based style classification.

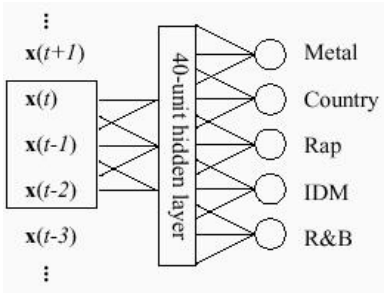

Figure 2: Structure of the learning process. A TDNN, with delays only at the input layer, is used for training. The input is the current, plus the two preceding feature vectors.

4.2 Classification and Learning

Learning for classification on the audio features was done using a feedforward time-delay neural network (TDNN) [3]. This is a structure that allows the incorporation of a short time memory for classification, by providing as inputs samples of previous time points (Figure 2). For training this network we used the resilient backpropagation algorithm [6] and iterated in batch mode (using the entire training set as one batch). The inputs layer has twenty nodes (one for each dimension of the representation) with a memory of three adjacent input frames. We used one hidden layer with forty nodes, and the output layer was five nodes, each corresponding to one of the five styles we wished to recognize. The training targets were a value of 1 for the node corresponding to the style of the input, and values of 0 for all the other nodes.

In the testing phase, the features of the test set were extracted (using the same dimensionality reduction transform derived from the training data), and they were fed to the classification network. Styles were assigned to the output node corresponding to the maximum value.

Table 1: Artist selections for style classification.

Heavy Metal	Contemporary Country	Hardcore Rap	IDM	R&B
Guns n' Roses	Billy Ray Cyrus	DMX	Boards of Canada	Lauryn Hill
AC/DC	Alan Jackson	Ice Cube	Aphex Twin	Aaliyah
Skid Row	Tim McGraw	Wu-Tang Clan	Squarepusher	Debelah Morgan
Led Zeppelin	Garth Brooks	Mystical	Plone	Toni Braxton
Black Sabbath	Kenny Chesney	Outkast	Mouse on Mars	Mya

4.3 Results

We ran a training and testing scheme where each row (collection of five artists across five styles) in turn was selected for testing. The remaining four rows were used for training. This process was executed five times (one for each row as a test,) and the results for each permutation are shown in Figure 3.

As is clearly evident, the results are not particularly good for the IDM style. Most of the artists have been misclassified, and there is little cohesion among that style. This should not be construed as a shortcoming in the training method, as this is a music style that exhibits a huge auditory variance, ranging from aggressive rough beats to abstract and smooth textures. What ties these artists together as a style is not a common sound of their work, but rather a cultural affinity stemming from the use of electronic instruments, and common roots ranging back to electronic dance music. Likewise, we see inconsistent results for Lauryn Hill, classified as a rap artist due to her rap-like production.

Such intra-style auditory inconsistencies are of course hard to overcome using any audio based system, highlighting the need for additional descriptors that factor in additional cultural issues.

5. COMMUNITY METADATA-BASED STYLE CLASSIFICATION

We next describe using cultural features for style classification solely using the community metadata feature vectors described earlier.

The cultural features for the 25 artists in our set were computed during work done on artist similarity early in 2002. Each artist is associated with a set of roughly 10,000 unigram terms, 10,000 bigram terms, 5,000 noun phrases and 100 adjectives. Each term was associated with an artist by it appearing on the same web document as the artists' name– but this alone does not prove a causal relation of description. Associated with each term is a score computed by the software (see Table 2) that considers position from the named referent and a gaussian window around the term frequency of the term divided by its document frequency. (Term frequency is how often a term appears relating to an artist and document frequency is how often the term appears overall.) The gaussian we used is:

$$\frac{f_t e^{-(log(f_d)-\mu)^2}}{2\sigma^2} \quad (1)$$

Here, f_d is the document frequency of a term, f_t the term frequency of a term, and μ and σ are parameters indicating the mean and deviation of the gaussian window.

This method proved well in computing artist similarities (given a known artist similarity list, this metric could predict them adequately) but here we ask the same data to arrange the artists into clusters.

Table 2: Sample adjective terms for the group ABBA with associated weighted scores.

adj Term	Score
nonviolent	0.12
perky	0.1
swedish	0.041
international	0.027
inner	0.025
consistent	0.020
bitter	0.013
classified	0.011
junior	0.009
produced	0.008
romantic	0.008

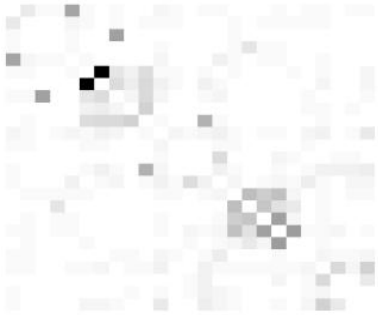

Figure 4: Similarity matrix from cultural similarity. 25 artists are arranged on each axis. Each style is listed one after another, so activity stays mostly on the diagonal.

5.1 Clustering Overlap Scores

The community metadata system computes similarity by a simple 'overlap score.' Each pair of artists is similar with unnormalized *overlap weight* i where i is a additive combination of every shared term's score. These scalars are unimportant on their own, but we can rank their values using each artist in our set as the ground artist to see which artists are more similar each other. Using this method, we compute the similarity matrix M(25,25), using each artist in the five-style set. (See Figure 4.)

This matrix is then used to predict the style of each given artist. For each term type, we take each artist in turn and sort their overlap weight similarities to the other 24 artists in descending order. We then use prior knowledge of the actual styles of the 24 similar artists

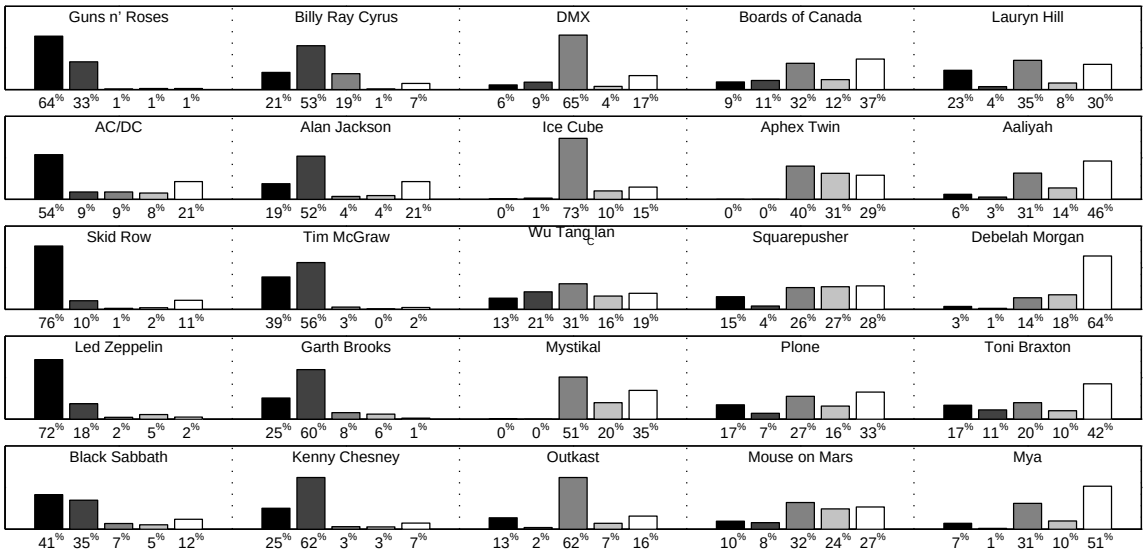

Figure 3: Audio-based style classification. Bars denote the estimated probability of that artist belonging to a particular style.

to find the true style of our target artist: descending the sorted list, once we have counted four other artists in the same cluster, we consider our target artist classified with a normalized score (the amount of cumulated overlap weights the cluster contributed to the total cumulated overlap weights.) The highest cumulated score is deemed the correct classification, and the five style scores are arranged in a probability map. In a larger-scale implementation, this step is akin to using a supervised clustering mechanism which tries to find a fit of an unknown type among already labeled data (by the same algorithm). Because of the small size of the sample set, we found this more manual method more effective.

We do this for each term type in the community metadata feature space and average the returned maps into a generalized probability map. The map defines a confidence value for each style much like the neural network's results above, and the probability approach was crucial in integrating the two methods (which we describe below.)

5.2 Results

In Figure 5 we see that the results for the text-only classifier performs very well for three of the styles and adequately but not perfectly for two of the five styles. There seems to be confusion between the Rap and R&B style sets. However, for the previous problem set (IDM), the cultural classifier works perfectly and with high confidence. We can attribute this to IDM being an almost purely culturally-defined style. One of the issues that plague acoustically-derived classifiers is that often human classifications have little statistical correlation to the actual content being described. This problem also interferes with content-based recommendation agents that attempt to learn a relation model between user preference and audio content: sometimes, the sound of the music has very little to do with how we perceive and attach preference to it.

R&B and Rap's intrinsic crossover (they both appear on the same radio markets and are usually geared toward the same audiences) shows that the cultural classifier can be as confused as humans in the same situation. Here, we present the inverse of the 'description for content' problem: just as often, cultural influences steer us away from treating two almost identical artists as similar entities, or putting them in the same class.

We propose that automated systems that attempt to model listening behavior or provide 'commodity intelligence' to music collections be mindful of both types of influences. Since we can ideally model both behaviors, it perhaps makes the most sense to combine them in some manner.

6. COMBINED CLASSIFICATION

As pointed out in the preceding sections, some features which are crucial for style identification are best exploited in the auditory domain and some are best used in the cultural domain. So far, given our choice of domain, we have produced coherent clusters. Musical style (and even more so musical similarity) requires a complicated definition that can factor in multiple observations ranging from auditory, historical, geographical, ideological, etc. The community metadata is an effort to make up for the latter features, whereas the auditory domain helps on a more staunch judgment on the sound itself. It seems only natural that a combination of these two classifiers can help disambiguate some of the classification problems that we have discussed.

In order to combine the two results we view our classifier data as posterior probabilities and compute their average values. This is a technique that has been shown to be good in practice, when we have a questionable estimate of posterior probabilities [7], as is the case in the cultural-based classification.

6.1 Results

The results of the averaging are shown in Figure 6. It is clear that many of the problems that were present in the previous classification attempts are now resolved. The IDM class, which was problematic in the audio-based classification, is now correctly identified due to strong community metadata coherence. Likewise, the Rap cluster which was not well defined in the metadata classification, was correctly identified using the auditory influence. Overall the combined classification was correct for all samples, bypassing all the problems found in either audio or metadata only classification.

7. FUTURE WORK

One less obvious use of this system is a 'cultural to musical' ratio equation for relations among artists. An application that could know

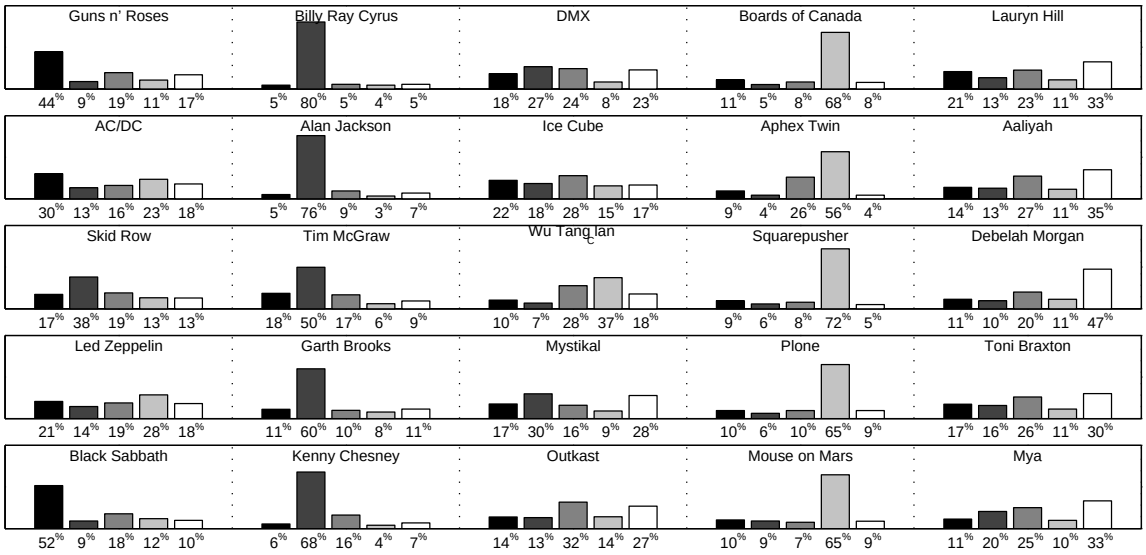

Figure 5: Community Metadata-based style classification. Bars denote the estimated probability of that artist belonging to a particular style.

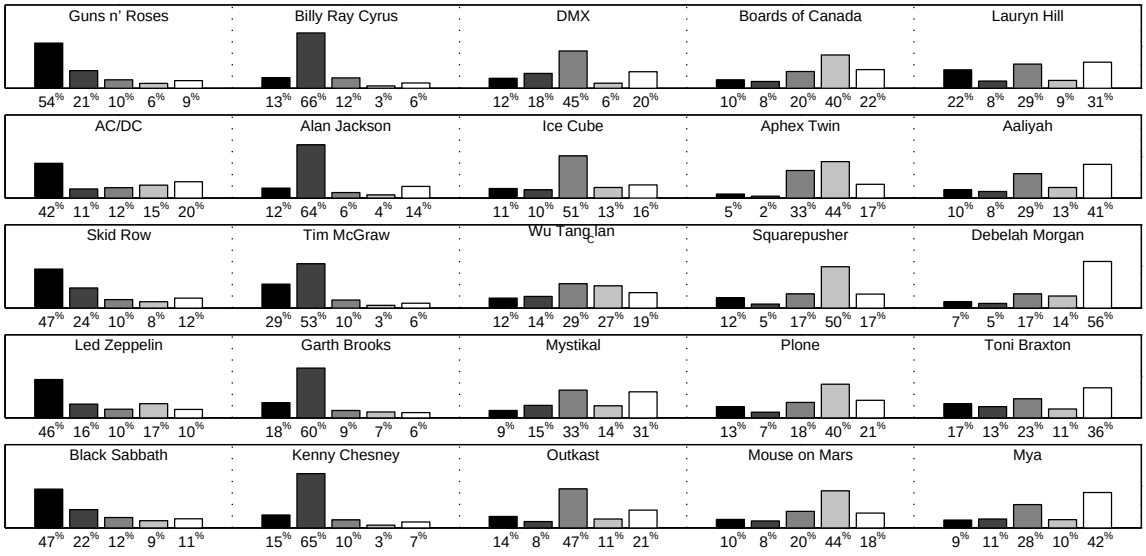

Figure 6: Combined style classification. Bars denote the estimated probability of that artist belonging to a particular style.

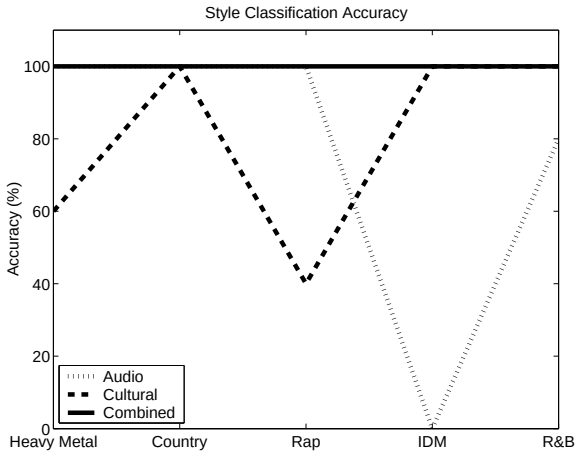

Figure 7: Results for all classification tasks. Accuracy is based on the 1-in-5 style classifier.

in advance how to understand varying types of artist relationships could benefit many music retrieval systems that attempt to inject commodity intelligence into the L&R process.

A good case for such a technology would be a recommendation agent that operates on both acoustic and cultural data. Large scale record shops already compute cultural relationships using sale data fed into a collaborative filtering system, and music-based recommenders such as Moodlogic (`www.moodlogic.com`) operate on spectral features. Both systems have proved successful for different types of music, and a system that could define ahead of time the 'proper' set of features to use would be integral to a combination approach.

We could simply define a 'culture ratio' as

$$C(a,b) = \frac{P(S_c(a,b))}{P(S_a(a,b))} \tag{2}$$

i.e. the probability that artists a and b will be similar using a cultural metric divided by the probability that artists a and b will be similar using an acoustic metric. A high 'culture ratio' would alert a recommender that certain musical relationships (such as almost all in the IDM style) should be treated using a purely cultural feature space. Lower culture ratios would indicate that spectral or musically intelligent features should be used.

8. CONCLUSIONS

We have presented a prominent problem in musical style classification, and proposed a multimodal classification scheme to overcome it. By combining both acoustic and cultural artist information we have achieved classification in styles that exhibit large variance in either domain.

9. ACKNOWLEDGMENTS

This work was supported by the Digital Life Consortium of the MIT Media Lab.

10. REFERENCES

[1] A. Berenzweig, D. Ellis, and S. Lawrence. Using voice segments to improve artist classification of music. 2002. submitted.

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

[3] D. Clouse, C. Giles, B. Horne, and G. Cottrell. Time-delay neural networks: Representation and induction of finite state machines. In *IEEE Trans. on Neural Networks, 8(5)*, page 1065, 1997.

[4] W. W. Cohen and W. Fan. Web-collaborative filtering: recommending music by crawling the web. *WWW9 / Computer Networks*, 33(1-6):685–698, 2000.

[5] R. B. Dannenberg, B. Thom, and D. Watson. A machine learning approach to musical style recognition. In *In Proceedings of the 1997 International Computer Music Conference*, pages 344–347. International Computer Music Association., 1997.

[6] M. Riedmiller and H. Braun. A direct adaptive method for faster backpropagation learning: The RPROP algorithm. In *Proc. of the IEEE Intl. Conf. on Neural Networks*, pages 586–591, San Francisco, CA, 1993.

[7] D. Tax, M. van Breukelen, R. Duin, and J. Kittler. Combining multiple classifiers by averaging or by multiplying? In *Pattern Recognition (33), No. 9*, pages 1475–1485, 2000.

[8] G. Tzanetakis, G. Essl, and P. Cook. Automatic musical genre classification of audio signals, 2001.

[9] B. Whitman, G. Flake, and S. Lawrence. Artist detection in music with minnowmatch. In *Proceedings of the 2001 IEEE Workshop on Neural Networks for Signal Processing*, pages 559–568. Falmouth, Massachusetts, September 10–12 2001.

[10] B. Whitman and S. Lawrence. Inferring descriptions and similarity for music from community metadata. 2002. submitted.