Two Systems for Automatic Music Genre Recognition

What Are They Really Recognizing?

Sturm, Bob L.

Published in:
Proceedings of the second international ACM workshop on Music information retrieval with user-centered and multimodal strategies

DOI (link to publication from Publisher):
10.1145/2390848.2390866

Publication date:
2012

Document Version
Early version, also known as pre-print

Two Systems for Automatic Music Genre Recognition: What Are They Really Recognizing?

Bob L. Sturm
Dept. Architecture, Design and Media Technology
Aalborg University Copenhagen, A.C. Meyers Vænge 15, DK-2450
Copenahgen SV, Denmark
bst@create.aau.dk

ABSTRACT

We re-implement and test two state-of-the-art systems for automatic music genre classification; but unlike past works in this area, we look closer than ever before at their behavior. First, we look at specific instances where each system consistently applies the same wrong label across multiple trials of cross-validation. Second, we test the robustness of each system to spectral equalization. Finally, we test how well human subjects recognize the genres of music excerpts composed by each system to be highly genre representative. Our results suggest that neither high-performing system has a capacity to recognize music genre.

Categories and Subject Descriptors

H.3 [**Information Search and Retrieval**]: Content Analysis and Indexing; H.4 [**Information Systems Applications**]: Miscellaneous; J.5 [**Arts and Humanities**]: Music

General Terms

Machine learning, pattern recognition, evaluation

Keywords

Music genre classification, music similarity

1. INTRODUCTION

The problem of automatically recognizing the genre of music in recorded audio remains an unsolved problem, and one that has been superseded in part by the more easily-defined problem of predictive music tagging [8]. Nonetheless, we have seen significant progress in the past decade. Tzanetakis and Cook [13] propose combining short-term signal features (both time- and frequency-domain computed over windows of 23 and 43 ms duration) with long-term features (pitch and beat histograms computed over entire signals), and either modeling these with Gaussian mixture models for parametric classification, or using them in k-nearest neighbor classification. Their best-performing system achieves a mean accuracy of 61% in ten different genres. Since then, the accuracy of such systems have climbed: to 83% [4], and reportedly to 91% [10], on the same dataset used in [13].

In this paper, we re-implement two state-of-the-art systems. The first approach [4] combines weak classifiers trained by a multiclass version of AdaBoost [6, 11] on statistically summarized bags of features. This approach won the 2005 MIREX music genre classification competition, and continues to be one of the best performing approaches since. The second approach [10] uses sparse representation classification of auditory features. This approach, originally presented as achieving over mean accuracies of over 90% [10], has been contested to give no higher than 70% mean accuracy. However, our version in this paper performs on par with [4].

After confirming our implementations work as reported with respect to mean classification accuracies, we inspect their behavior more closely than has ever been done before. In the first experiment, we perform multiple runs of cross-validation tests, and find excerpts that each system consistently and persistently mislabels. In the second experiment, we find that when we apply minor changes to the spectrum of the audio signal, each system confidently classifies the same excerpt of music in radically different genres. In the third experiment, we test how well humans label music excerpts composed by each system to be highly representative of each genre. These experiments provide evidence that each system, while having high accuracies in music genre recognition, are not actually comparing genre. While this may not be controversial to experts in the field, we continue to see the problem of genre recognition addressed without any critical analysis of the validity of the experiments, performance measures, and conclusions for such a complex problem.

2. TWO GENRE RECOGNITION SYSTEMS

We now review the two systems we re-implement, and present an overview of their results. We make available all code used in this work at: `http://removed.edu`

2.1 AdaBoost with decision trees, and bags of frames of features (AdaBFFs) [4]

The multiclass AdaBoost method [6, 11] attempts to create a strong classifier by combining "votes" cast by a growing collection of weak classifiers. Its use for music genre recognition was first proposed in [4]. In our implementation of this system, we first find the features of a given audio signal using Hann windows of length 1,024 samples (46.4 ms), overlapped by 50%. For each window, we compute 40 Mel-frequency cepstral coefficients (MFCCs) as in [12], the number of zero crossings of the waveform, the variance

ACM Special Session '13 Nara, Japan

and mean of the power spectrum, 16 quantiles of the power spectrum (converting the discrete spectrum to a probability mass distribution, we find the highest frequencies at which the cumulative distribution function is less than $m/17$ for $m \in \{1, 2, \ldots, 16\}$), and the error of a least-squares optimized 32-order linear predictor (autoregression). Then, we compute their mean and variances of the features in every 129 consecutive windows (3 seconds), which gives segmental feature vectors having dimension 120.

Given the segmental features of a labeled training set, iteration l of the multiclass AdaBoost [6,11] adds a new decision tree $\mathbf{v}_l(\mathbf{x})$ and weight w_l such that a total prediction error is minimized. For a segmental feature vector $\mathbf{x}$ belonging to one of K classes, the decision tree $\mathbf{v}_l(\mathbf{x})$ produces a length-K vector of ± 1 specifying its prediction. A 1 in element k means it favors class k, whereas -1 means the opposite. We use the "multiboost package" [3], with decision trees as the weak learners, AdaBoost.MH [11] as the strong learner, and all other parameters left to their defaults. After iteration L, we have a trained classifier combining L decision trees, which produces the length-K vector of scores

$$\mathbf{f}(\mathbf{x}) := \sum_{l=1}^{L} w_l \mathbf{v}_l(\mathbf{x}). \tag{1}$$

From this, we can predict the class of $\mathbf{x}$ by finding the row of $\mathbf{f}(\mathbf{x})$ having the largest score.

Considering that we have several features from a single piece of recorded music $\mathcal{X} := \{\mathbf{x}_i\}$, we will have several scores produced by the classifier. Thus, we pick the class for the set of features $\mathcal{X}$ by using the logistic method [9]:

$$P[k|\mathcal{X}] := \gamma_{\mathcal{X}} \left[1 + \exp\left(-2 \sum_{i=1}^{|\mathcal{X}|} [\mathbf{f}(\mathbf{x}_i)]_k \right) \right]^{-1} \tag{2}$$

where we define $\gamma_{\mathcal{X}}$ such that $P[k|\mathcal{X}]$ acts as a probability mass distribution over the K genres.

2.2 Sparse representation classification with auditory modulations (SRCAM) [10]

Given a matrix of N features $\mathbf{D} := [\mathbf{d}_1|\mathbf{d}_2|\cdots|\mathbf{d}_N]$, and the set of class identities $\cup_{k=1}^{K} \mathcal{I}_k = \{1, \ldots, N\}$, where $\mathcal{I}_k$ specifies the columns of $\mathbf{D}$ belonging to class k, sparse represenation classification (SRC) [15] first finds a sparse representation of an unlabeled feature $\mathbf{x}$ by solving

$$\min \|\mathbf{a}\|_1 \text{ subject to } \|\mathbf{x} - \mathbf{D}\mathbf{a}\|_2 < \epsilon \tag{3}$$

for $\epsilon \geq 0$. This problem is known as basis pursuit denoising [5]. We then define a set of weights $\{\mathbf{a}_k\}$ by

$$[\mathbf{a}_k]_n := \begin{cases} a_n, & n \in \mathcal{I}_k \\ 0, & \text{else} \end{cases} \tag{4}$$

where a_n and $[\mathbf{a}_k]_n$ is the nth row of $\mathbf{a}_k$. In this way, $\mathbf{a}_k$ preserves weights in $\mathbf{a}$ specific to class k. Finally, SRC classifies $\mathbf{x}$ simply by

$$\hat{k}(\mathbf{x}) := \arg\min_k \|\mathbf{x} - \mathbf{D}\mathbf{a}_k\|_2^2. \tag{5}$$

We may gauge the confidence of this classifier by comparing the class-dependent errors. To this end, we define the confidence of SRCAM for assigning class k to $\mathbf{x}$ as

$$C(k|\mathbf{x}) := \frac{\max_{k'} J_{k'} - J_k}{\sum_l [\max_{k'} J_{k'} - J_l]} \tag{6}$$

where $J_k := \|\mathbf{x} - \mathbf{D}\mathbf{a}_k\|_2$. In this way, $C(k|\mathbf{x})$ acts as a probability mass distribution over the genres.

The use of SRC for music genre recognition was first proposed in [10], which describes using auditory modulation features of 30 s music excerpts to achieve mean accuracies above 90%. These results have been contested, but our adaptations of the approach give mean accuracies in the neighborhood of state-of-the-art. Here, we use the Lyon Passive Ear Model implemented in [12], and a downsampling factor of 40, to produce auditory spectrograms of audio signals 30 s in duration. For modulation analysis, we pass the zero-meaned signals of each frequency band through a bank of 8 Gabor filters sensitive to modulations rates in $\{2, 4, 8, \ldots, 256\}$ Hz. Finally, we find the squared energy of each Gabor filter, giving a distribution of energy in frequency and modulation rate; and we vectorize the representation to produce the 768-dimensional feature vector $\mathbf{x}$.

To produce the dictionary $\mathbf{D}$ used by SRCAM, we take the set of features $\{\mathbf{x}_i\}$ and standardize them by first mapping all values in each dimension to $[0, 1]$ (subtracting the minimum value observed, and dividing by the largest difference observed), and then making the mapped observations of each dimension have unit variance. Finally, we make all standardized features have unit ℓ_2 norm and compose the dictionary by concatenating the features as its columns. To attempt to solve (3), we use the SGPL1 solver [14] with at most 100 iterations, and $\epsilon := 0.1$. Though we find that the solver converges only about 20% of the time in 100 iterations, its output still appears favorably discriminative. Before the decomposition and classification of an unlabeled feature, we apply the standardization transformation used to create $\mathbf{D}$, and make it have unit ℓ_2 norm.

2.3 Experimental results

To confirm each system is working, we perform stratified cross-validation with the same dataset used in [4, 10]. This dataset (GTZAN), created for the work in [13], has 1000 music excerpts of 30 seconds duration with 100 examples labeled with each of 10 different music genres: Blues, Classical, Country, Disco, Hip Hop, Jazz, Metal, Popular, Reggae, and Rock. As in [4], we use decision trees of 1 node (stumps), and 5-fold stratified cross-validation with AdaBoost run for 2500 iterations. As in [10], we use 10-fold stratified cross-validation. However, unlike these past works, we run 10 independent trials of cross-validation to obtain mean statistics across random distributions of the data into training and testing sets. Figure 1 shows the mean confusions for each system. We see that the overall mean accuracy of AdaBFFs, with a 95% confidence interval, is 0.7755 ± 0.0022. This is about 5% less than that reported on the same dataset of 83% in [4], but this could be due to their use of decision trees of unspecified number of nodes. The overall mean accuracy of SRCAM is 0.8203 ± 0.0019, which is about 10% worse than reported in [10], and is due to a bug in the original study (personal communication with Panagakis et al.).

3. A CLOSER LOOK AT THE BEHAVIOR

We now look closer than ever before at the behavior of these two systems. First, we find specific music excerpts for which the systems persistently label the same wrong genre. Second, we test the sensitivity of the systems to the spectral equalization of a music excerpt. Finally, we test how well humans label excerpts each system composes to be highly representative of particular genres.

(a) AdaBFFs

True Genre	Blues	Classical	Country	Disco	Hip hop	Jazz	Metal	Pop	Reggae	Rock
Blues	84.70 ±1.63	0.00 ±0.00	3.60 ±0.52	1.70 ±0.78	0.30 ±0.30	1.20 ±0.57	0.20 ±0.39	0.90 ±0.20	2.40 ±0.60	6.10 ±0.68
Classical	0.20 ±0.26	95.70 ±0.42	0.00 ±0.00	1.00 ±0.00	0.00 ±0.00	6.00 ±0.00	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00	0.80 ±0.39
Country	4.80 ±0.76	0.80 ±0.26	75.70 ±1.13	2.90 ±0.54	1.10 ±0.46	3.50 ±0.89	0.10 ±0.20	5.90 ±0.85	5.80 ±0.57	13.80 ±1.23
Disco	1.40 ±0.52	0.00 ±0.00	4.40 ±1.28	71.40 ±1.73	3.70 ±0.66	0.90 ±0.46	0.60 ±0.43	2.90 ±0.46	2.90 ±0.35	14.00 ±0.88
Hip hop	0.80 ±0.39	0.00 ±0.00	0.30 ±0.30	2.70 ±0.59	72.90 ±1.38	0.00 ±0.00	0.80 ±0.49	1.80 ±0.76	15.60 ±1.47	1.10 ±0.74
Jazz	2.30 ±0.30	1.40 ±0.43	0.80 ±0.57	0.00 ±0.00	0.00 ±0.00	87.40 ±1.38	0.00 ±0.00	0.10 ±0.20	0.30 ±0.30	1.50 ±0.44
Metal	1.20 ±0.76	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00	2.60 ±0.43	0.00 ±0.00	92.90 ±0.94	0.00 ±0.00	0.60 ±0.32	4.50 ±0.44
Pop	0.00 ±0.00	0.00 ±0.00	2.10 ±0.62	7.40 ±0.98	4.50 ±0.79	0.00 ±0.00	0.10 ±0.20	82.70 ±0.42	4.90 ±0.35	3.00 ±0.51
Reggae	1.10 ±0.46	0.00 ±0.00	3.40 ±0.73	4.20 ±0.64	12.40 ±0.73	0.20 ±0.26	0.00 ±0.00	0.80 ±0.39	62.60 ±1.58	5.70 ±0.88
Rock	3.50 ±0.84	2.10 ±0.46	9.70 ±1.13	8.70 ±0.97	2.50 ±0.60	0.80 ±0.49	5.30 ±0.66	4.90 ±0.46	4.90 ±0.54	49.50 ±2.52

(b) SRCAM

True Genre	Blues	Classical	Country	Disco	Hip hop	Jazz	Metal	Pop	Reggae	Rock
Blues	94.40 ±0.60	0.00 ±0.00	3.30 ±0.66	0.30 ±0.42	0.20 ±0.26	2.10 ±0.46	0.00 ±0.00	0.00 ±0.00	3.90 ±0.46	4.30 ±0.59
Classical	1.50 ±0.33	96.20 ±0.39	4.40 ±0.52	2.20 ±0.26	0.00 ±0.00	3.10 ±0.20	0.00 ±0.00	0.60 ±0.32	1.30 ±0.30	2.10 ±0.35
Country	0.20 ±0.26	0.00 ±0.00	73.50 ±1.14	2.60 ±0.43	1.00 ±0.00	1.10 ±0.46	0.20 ±0.26	0.60 ±0.32	1.70 ±0.30	6.20 ±0.70
Disco	0.80 ±0.39	0.00 ±0.00	1.90 ±0.35	69.50 ±1.44	3.10 ±0.99	0.80 ±0.26	0.10 ±0.20	3.80 ±0.82	5.30 ±1.06	5.40 ±0.67
Hip hop	0.00 ±0.00	0.00 ±0.00	0.50 ±0.33	6.10 ±0.54	83.00 ±1.20	0.20 ±0.26	0.70 ±0.30	2.30 ±0.66	5.70 ±0.78	0.10 ±0.20
Jazz	2.30 ±0.42	0.80 ±0.26	2.60 ±0.60	0.00 ±0.00	0.90 ±0.20	90.40 ±0.78	0.30 ±0.30	0.70 ±0.30	1.20 ±0.26	2.10 ±0.46
Metal	0.30 ±0.30	1.00 ±0.00	0.00 ±0.00	0.50 ±0.33	2.60 ±0.32	1.00 ±0.00	95.50 ±0.53	1.60 ±0.32	1.00 ±0.00	15.00 ±1.01
Pop	0.00 ±0.00	0.00 ±0.00	5.60 ±0.84	6.20 ±0.49	3.40 ±0.60	0.10 ±0.20	0.00 ±0.00	85.90 ±0.99	5.30 ±0.51	1.30 ±0.30
Reggae	0.40 ±0.32	0.00 ±0.00	1.10 ±0.46	7.20 ±0.87	5.80 ±0.76	0.20 ±0.26	0.00 ±0.00	2.60 ±0.52	72.90 ±1.41	4.50 ±0.53
Rock	0.10 ±0.20	2.00 ±0.51	7.10 ±0.74	5.40 ±0.73	0.00 ±0.00	1.00 ±0.00	3.20 ±0.49	1.90 ±0.20	1.70 ±0.30	59.00 ±1.57

Figure 1: Confusion matrices with 95% confidence intervals shown below means.

3.1 Consistent & Persistent Misclassifications

Figure 2 shows for the Disco-labeled excerpts in GTZAN how the two systems labels them in the 10 trials of cross-validation shown in Fig. 1. The darkness of a square represents the frequency of the applied label, with black being 10 times (each excerpt is classified ten times based on different training data). We define a mislabeling of an excerpt as "consistent and persistent" (CPM) if the wrong label applied is the same in all ten trials of cross-validation. We see the same behavior for the other nine genres in both systems, but only show the results for Disco for lack of space. All excerpts we discuss below can be heard in the dataset GTZAN.[1]

Disco is a style of dance music which emerged from Funk and Soul in the early 1970s in the USA, and quickly became a world-wide phenomenon [1]. Disco music typically uses the common time meter at a steady tempo of around 120 beats per minute, a distinctive use of the open hi-hat on the off beats, prominent electric bass lines, and often rich backing textures provided by female vocals, keyboards, synthesizers, strings, and horns [1]. Of the Disco-labeled excerpts in GTZAN, we find AdaBFFs has 10 consistent and persistent mislabelings (CPMs): four as Pop, three as Rock, one as Classical, one as Country, and one as Hip hop. And for SRCAM we find 12 CPMs: three as Hip hop, three as Pop, three as Reggae, two as Classical, and one as Rock.

Both systems share three of the same CPMs, but each of these are in some sense forgivable based on the musical content of the excerpts. First, 23 is persistently labeled Pop; but this excerpt comes from Latoya Jackson's 1988 album "Bad Girl," and the top last.fm tag[2] applied to the artist is "pop."[3]

Second, 29 is persistently labeled Pop as well; but this excerpt is of Evelyn Thomas singing "Heartless" in 1984, and none of the last.fm tags applied to the song include "Disco" (they are, "female vocalists, soul, dance, hi-nrg"). Finally, 47 is persistently labeled Classical; but this 30 second excerpt features Barbra Streisand and Donna Summer singing at a slow tempo and accompanied softly by only piano and strings. Though the top last.fm tag associated with the entire song is "disco," a human listening to the excerpt could focus on the strings and piano, the lack of percussion and bass, and thus not label it as disco.

Individually, each system has other forgivable CPMs. AdaBFFs persistently labels excerpt 27 as Hip hop; but the top last.fm tags applied to this song are "Hip-Hop" and "rap." SRCAM persistently labels excerpt 21 as Pop; but while the source of this excerpt is the song "Never Can Say Goodbye," originally covered in 1974 by Gloria Gaynor, this excerpt is not taken from that recording, but instead is a much later remix with Gaynor singing. SRCAM also persistently labels excerpt 41 as Country, while the top last.fm tag for this song is "soul" — a genre label not present in GTZAN. Finally, SRCAM persistently labels excerpt 85 as Hip hop; and while the top last.fm tags for this song ("new wave, 80s, funk") do not include include Hip hop, this excerpt features a sparse up-tempo electronic drum loop over which a female vocalist educationally raps about words.

Each system, however, has CPMs that are not so forgivable. AdaBFFs persistently labels as Pop excerpts 15 and 28 while each has a top last.fm tag of "disco." AdaBFFs persistently labels excerpts 67, 83, and 84 as Rock. The top 5 last.fm tags of ABBA's "Dancing Queen" (67) and "Disco Duck" (83) include "disco" but not "rock." The identity of excerpt 84 is unknown, but its content — funky bass and guitar, electronic drums and horns — sounds much more Disco than Rock. SRCAM persistently labels as Hip hop excerpts 48 and 79 while the former has the top last.fm tag of "disco," and the artist of the latter has the top last.fm tags "funk,

[1] `http://marsyas.info/download/data_sets`

[2] `http://last.fm` is a crowd-sourcing music business that gathers and makes publicly available (through their API) information on listening habits from a diverse world-wide community. A tag is something a user of last.fm creates to describe a music group or song in their music collection.

[3] There are no tags applied to the specific song.

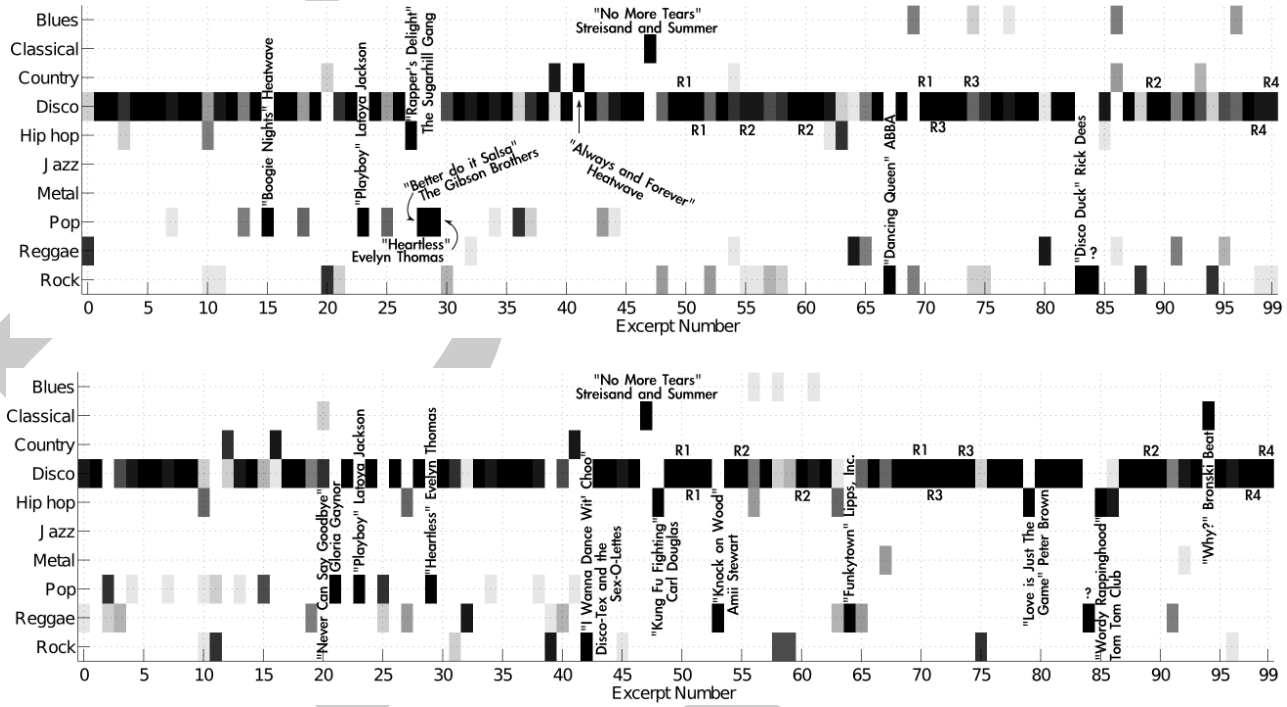

Figure 2: Applied labels (y-axis) of Disco excerpts in GTZAN over 10 trials of cross-validation by AdaBFFs (top) and SRCAM (bottom). Black is an excerpt labeled the same 10 times. R# show replicated excerpts.

disco funk." SRCAM persistently labels as Reggae excerpts 53, 64 and 84. The first two excerpts have the top last.fm tag of "disco;" the last excerpt, discussed above as sounding more Disco than Rock, also sounds more Disco than Reggae because it is up-tempo, lacks spring reverberation, and the instruments are not played as they are in Reggae [1]. SRCAM persistently labels as Rock excerpt 42, while the top last.fm tag applied to Disco-Tex and the Sex-O-Lettes is "disco." And finally, SRCAM persistently labels as Classical excerpt 94, which has top last.fm tags of "80s, new wave, synthpop, dance, pop." Though this excerpt features a long-held string pad, its quick-tempo electronic drum loop and synthesized melody make such a CPM unforgivable.

(a) AdaBFFs

(b) SRCAM

Figure 3: Wordle of last.fm tags applied to Disco excerpts of (top) CPMs of AdaBFFs (15, 28, 67, 83), and (bottom) SRCAM (42, 48, 53, 64, 79).

	Label, No.	Origin
AdaBFFs	Country 39	Wayne Toups "Johnnie Can't Dance"
	Hip hop 00	Afrika Bambaata "Looking for the Perfect Beat"
	Pop 12	Aretha Franklin, Celine Dion, Mariah Carey, et al. "You Make Me Feel Like A Natural Woman"
	Reggae 23	Bob Marley "Sun is Shining"
	Reggae 59	Bob Marley "One Love"
	Rock 27	The Beach Boys "Good Vibrations"
	Rock 31	The Rolling Stones "Honky Tonk Woman"
	Rock 37	The Rolling Stones "Brown Sugar"
	Rock 40	Led Zeppelin "The Crunge"
	Rock 43	Led Zeppelin "The Ocean"
	Rock 57	Sting "If You Love Somebody Set Them Free"
	Rock 81	Survivor "Poor Man's Son"
	Rock 82	Survivor "Burning Heart"
SRCAM	Hip hop 00	Afrika Bambaaya "Looking for the Perfect Beat"
	Pop 63	Diana Ross "Ain't No Mountain High Enough"
	Reggae 01	Bob Marley "No Woman No Cry"
	Rock 31	The Rolling Stones "Honky Tonk Woman"
	Rock 77	Simply Red "Freedom"

Table 1: All GTZAN excerpts consistently and persistently mislabeled Disco by AdaBFFs & SRCAM.

Figure 3 shows a graphic representation of all tags applied by users of last.fm to these excerpts. The font size of each tag is proportional to the "count" supplied by last.fm (100 is the most frequent, and 1 the least frequent). From this we see the most frequent tag of these CPM excerpts is disco. Looking outside of the Disco data, AdaBFFs and SRCAM also consistently and persistently mislabel as Disco other excerpts. In Table 1, we see how both AdaBFFs and SRCAM consistently and persistently label as Disco, "Honky Tonk Woman" by The Rolling Stones, and "Looking for the Perfect Beat" by Afrika Bambaata.

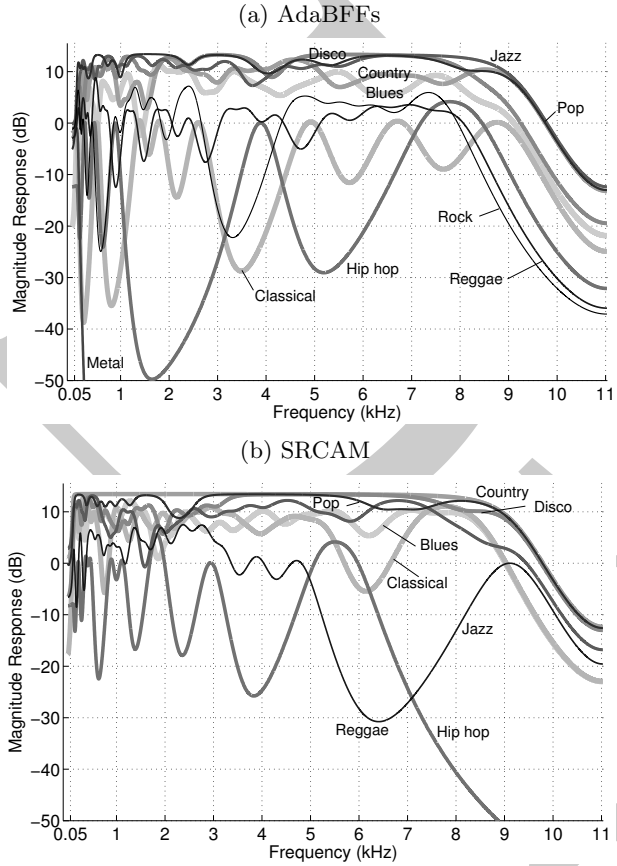

Figure 4: Equalizations for each system to claim a particular genre for Western Swing song "Big Balls in Cowtown" by Bob Wills and the Texas Playboys.

3.2 Genre-shifting Mastering

We now test the robustness of these systems to changes in the spectral characteristics of an excerpt. We train AdaBFFs with the entire 1000 excerpts of GTZAN. For SRCAM, the standardized ATMs of all 1000 excerpts compose the dictionary. In the case of AdaBFFs, we use 2500 training iterations of AdaBoost. We take a musical excerpt, pass the signal through a 90-channel Gammatone filterbank with channels either on or off, and then classify it. If a system is recognizing genre, then it should be quite robust to minor changes in the spectral characteristics of the excerpt. For instance, humans can recognize the genre of a piece of music whether it is heard from AM or FM radio.

We take a 30-second excerpt of the Western Swing song "Big Balls in Cowtown" by Bob Wills and the Texas Playboys. Each classifier labels this as Country. However, often with only minor changes to the sound, AdaBFFs labels the same music with all ten genres, and SRCAM labels it with eight genres. Figure 4 shows the magnitude responses of the equalizations applied to this excerpt for each system. With a majority of the bands turned on, AdaBFFs labels the same music Disco, Pop, Reggae and Rock. Likewise, SRCAM labels it Disco, Jazz, Pop and Reggae. We find this same behavior with many other excerpts, including "Symphony of Destruction" by Megadeth (Metal), "Poison" by Bel Biv Devoe (Hip hop), and even "blues.00001" from GTZAN. In all cases, while the classifiers label correctly the "original" versions, we are able to coax each to apply many of the other labels with only perceptually minor spectral equalization.

3.3 Representative Excerpts

We now test whether human subjects can recognize the genres of excerpts each classifier composes to be highly representative of each genre. To do this, we take the 1,198 sample loops that accompany Apple's GarageBand program.[4] These loops include cover the variety of genres in GTZAN, e.g., drum patterns characteristic of Disco, Hip hop, Jazz, Reggae, and Rock; piano and organ played characteristic to Blues, Classical, Disco, Jazz, Pop, Reggae and Rock; bass played characteristic to Blues, Country, Disco, Jazz, Metal, Pop, Reggae, and Rock; guitar (and banjo) played characteristic to Blues, Country, Jazz, Metal, and Rock; melodies played on recorder, orchestral brass, and strings; and sound effects, like vinyl scratching characteristic to Hip hop. From these, we make random combinations of four sample loops to create excerpts 30 seconds long. We then have each classifier, trained as in the previous experiment on all GTZAN, select an excerpt to best represent each genre. To be sure a randomly composed excerpt is representative of genre k according to AdaBFFs, we select it only if $P[k|\mathcal{X}] > 0.999$. For SRCAM, we select an excerpt of genre k only if $C(k|\mathbf{x}) \geq 1.7C(k'|\mathbf{x}) \; \forall \; k' \neq k$. With these choices, we find that about 1 in 5 random combinations result in a representative excerpt. We find that the most likely labels applied by AdaBFFs to randomly generated excerpts are Country (25%) and Rock (23%), and the least often Pop ($< 1\%$). For SRCAM, the most likely labels are Classical (20%) and Reggae (35%), and the least often Country ($< 1\%$). Finally, we take one excerpt of each genre selected by each classifier (20 in total) for a listening experiment.

We perform each listening test as follows. First, we tell the subject that they will listen to up to 30 musical excerpts of about 10 seconds in length. They are to pick one of the ten genres listed that best describes each one. They can listen to each excerpt as many times as needed, but cannot return to previous excerpts or change previous answers. They must select a genre before advancing an excerpt. Then the subject dons a pair of headphones, and interacts with a GUI we built in MATLAB. In order to screen subjects for their ability to recognize the ten genres, the first ten excerpts are ones that we selected from GTZAN for their representability.[5] The test ends if the subject makes an error in these; otherwise, the subject is then presented the 20 representative excerpts. The presentation order of all excerpts are randomized, with the exception that the first ten are from GTZAN, the second ten are representative according to AdaBFFs, and the final ten are representative according to SRCAM.

With this experimental design we test whether a subject able to recognize real excerpts from the same genres can recognize the same genres among the representative excerpts. The null hypothesis $\mathcal{H}_0$ is thus: those able to recognize the genres of 10 real excerpts are unable to recognize the same genres of the representative excerpts. Twenty

[4]This program is made for people to start making music easily using a sample loop-based sequencing environment.

[5]Blues 5 John Lee Hooker, "Sugar Mama"; Classical 96 Antonio Vivaldi, "Summer"; Country 12 Bobby Bare, "Music City USA"; Disco 66 Peaches & Herb, "Shake Your Groove Thing"; Hip hop 47 A Tribe Called Quest, "Award Tour"; Jazz 19 Joe Lovano, "Birds of Springtimes Gone By"; Metal 4 Dark Tranquillity, "Of Chaos and Eternal Night"; Pop 95 Mandy Moore, "Love You for Always"; Reggae 71 Dennis Brown, "Big Ship"; Rock 40 Led Zeppelin, "The Crunge."

subjects passed the screening, and so we define statistical significance by $\alpha = 0.05$. Assuming independence between each trial, we model the number correct N as a random variable distributed Binomial$(20, 0.1)$. The expectation $E[N] = 20(0.1) = 2$, and $\text{Var}[N] = 20(0.1)(0.9) = 1.8$. In our experiments, the mean no. correct is 1.85 (median is 2; mode is 1); the variance is 1.19. The maximum no. correct is 4 (1 person), and the minimum is 0 (2). Since the probability $P(N > 3) > 0.13$, and $P(N = 0) > 0.12$, we cannot reject $\mathcal{H}_0$ for any subject, let alone all of them. Looking at how each subject performs for excerpts specific to each classifier, we find no behavior statistically significant from chance for either. We also test for a significant difference between the two sets of representative excerpts, i.e., $\mathcal{H}_0$ is that accuracy on the two sets are not significantly different. A two-tailed t-test shows we cannot reject $\mathcal{H}_0$.

It is clear from the data and conversation afterwards that even though all genres are equally represented (10%), subjects most often selected Jazz (28.3%), followed by Disco (15.0%), Pop (12.0%), Rock (11.8%), Reggae (9.5%), Hip hop (7.3%), Blues (6.5%) and Country (6.5%), Classical (2.0%), and Metal (1.3%). The SRCAM-composed Reggae excerpt contains a prominent banjo, and so most subjects (90.0%) classified it as Country. The AdaBFFs-composed Rock excerpt contains a prominent walking acoustic bass, and so most subjects (70.0%) classified it as Jazz. Subjects mentioned that while the first ten real excerpts were easy to classify, the last 20 were very difficult, and many sounded as if elements of more than 1 genre were combined.

4. CONCLUSION

It has been acknowledged several times now, e.g., [2, 7], that low-level features summarized by bags of frames, such as those used in [13], are unsuitable for music similarity tasks, which includes genre recognition. This has motivated the fusion of features over longer time scales, such as those used in AdaBFFs [4] and SRCAM [10], from which we see significant increases in mean classification accuracy by over 20% compared to past low level approaches like [13]. In this paper, we have investigated, more deeply than has ever done before, the behavior of AdaBFFs and SRCAM, and to what extent their high classification accuracies really reflects a capacity to recognize music genre.

First, we inspected the CPMs of each system for the Disco-labeled excerpts of GTZAN. We should not expect perfect performance; but if they have a capacity to recognize a genre, then we should not expect such CPMs of excerpts that clearly meet the stylistic rules of a genre supposedly embodied in GTZAN. Figure 2 shows the CPMs of the Disco excerpts. Figure 3 shows that the frequencies of tags applied by people to the "unforgivable CPMs" are dominated by "disco." And Table 1 lists all non-Disco excerpts of GTZAN that each system consistently and persistently labels as Disco. Second, we find both systems are quite sensitive to even minor changes in the spectral characteristics of signals. While the underlying music does not change, each system labels them in several widely differing genres. This is, of course, not surprising given that these systems heavily rely on spectral characteristics; but that they are so sensitive argues for new methods while at the same time questioning how they perform so well in the first place. Finally, we invert each system such that they output musical material instead of labels. This allows us to "hear" what each system is hearing when it comes to their internal models of genres. We find that we cannot reject the null hypothesis that humans do not recognize the genres supposedly represented by each system. In summary, our deep analysis casts doubt on any conclusion that either system can recognize genres.

5. REFERENCES

[1] C. Ammer. *Dictionary of Music*. The Facts on File, Inc., New York, NY, USA, 4 edition, 2004.

[2] J.-J. Aucouturier, B. Defreville, and F. Pachet. The bag of frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music. *J. Acoust. Soc. America*, 122(2):881–891, Aug. 2007.

[3] D. Benbouzid, R. Busa-Fekete, N. Casagrande, F.-D. Collin, and B. Kégl. Multiboost: a multi-purpose boosting package. *J. Machine Learning Res.*, 13:549–553, Mar. 2012.

[4] J. Bergstra, N. Casagrande, D. Erhan, D. Eck, and B. Kégl. Aggregate features and adaboost for music classification. *Machine Learning*, 65(2-3):473–484, June 2006.

[5] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, 20(1):33–61, Aug. 1998.

[6] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Computer System Sci.*, 55:119–139, 1997.

[7] G. Marques, M. Lopes, M. Sordo, T. Langlois, and F. Gouyon. Additional evidence that common low-level features of individual audio frames are not representative of music genres. In *Proc. Sound and Music Comp.*, Barcelona, Spain, July 2010.

[8] C. McKay and I. Fujinaga. Music genre classification: Is it work pursuing and how can it be improved? In *Proc. Int. Symp. Music Info. Retrieval*, 2006.

[9] A. Niculescu-Mizil and R. Caruana. Obtaining calibrated probabilities from boosting. In *Int. Conf. Uncertainty in Artificial Intell.*, pages 413–420, 2005.

[10] Y. Panagakis, C. Kotropoulos, and G. R. Arce. Music genre classification via sparse representations of auditory temporal modulations. In *Proc. European Signal Process. Conf.*, Glasgow, Scotland, Aug. 2009.

[11] R. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37(3):297–336, 1999.

[12] M. Slaney. Auditory toolbox. Technical report, Interval Research Corporation, 1998.

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

[14] E. van den Berg and M. P. Friedlander. Probing the pareto frontier for basis pursuit solutions. *SIAM J. on Scientific Computing*, 31(2):890–912, Nov. 2008.

[15] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Trans. Pattern Anal. Machine Intell.*, 31(2):210–227, Feb. 2009.