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

Bob L. Sturm
Department of 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 two state-of-the-art systems for music genre recognition, and closely examine their behavior. First, we find specific excerpts each system consistently and persistently mislabels. Second, we test the robustness of each system to spectral adjustments to audio signals. Finally, we expose the internal genre models of each system by testing if human can recognize the genres of music excerpts composed by each system to be highly genre-representative. Our results suggest that, though they have high mean classification accuracies, neither system is recognizing music genre.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Content Analysis and Indexing; J.5 [**Arts and Humanities**]: Music

General Terms

Algorithms, Experimentation, Performance

Keywords

Music genre recognition, classification performance

1. INTRODUCTION

The problem of automatically recognizing music genre of remains an unsolved problem, and one that has been superseded in part by the problem of tag prediction [8]. Nonetheless, we have seen significant progress over the past decade. Tzanetakis and Cook [16] combine 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 long durations), and either model these by Gaussian mixture models for parametric classification, or use them for k-nearest neighbor classification. For the benchmark dataset GTZAN [14, 16],[1] their best-performing system achieves a mean classification accuracy (MCA) of 61%. Since then, the MCAs of such systems trained and tested with GTZAN have climbed to 83% [4], and reportedly above 91%, e.g., [10].

In this paper, we examine two state-of-the-art systems in ways that, to our knowledge, have never been done.[2] The first approach [4] combines weak classifiers trained by multiclass AdaBoost [6,11] with bags of frames of features. In the 2005 MIREX audio genre classification competition,[3] this approach performed the best with respect to MCA across two different datasets (both different from GTZAN). The second approach, developed from [10], uses sparse representation classification of auditory features. Originally presented as achieving MCAs greater than 90% in GTZAN [10], this result has yet to be reproduced [15]. The version we develop in this paper, however, performs slightly better than that of [4] with respect to MCA in GTZAN.

After we discuss the particulars of our re-implementations, we take a closer look at their behaviors. We discover that each system consistently and persistently mislabels particular excerpts across multiple runs of stratified cross-validation (SCV). We then find that we can make each system confidently classify the same excerpt of music in radically different genres by minor filtering of the audio signal. Finally, we conduct an experiment revealing humans cannot recognize the genres of music excerpts composed by each system to be highly representative of each genre. We conclude with a discussion about the implications of our observations.

2. TWO GENRE RECOGNITION SYSTEMS

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

Multiclass AdaBoost [6, 11] creates a "strong" classifier by combining "weak" classifiers, e.g., decision stumps. Its use for music genre recognition is first proposed in [4]. Given the feature vectors (fvs) 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 fv $\mathbf{x}$ extracted from a signal in one of K classes, the lth decision tree $\mathbf{v}_l(\mathbf{x})$ produces a length-K vector with its kth element equal to 1 if it prefers class k, or -1 if not. After iteration L, we have a classifier combining L decision trees that produces the length-K vector of scores

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

with which we classify $\mathbf{x}$ by the row of the largest element

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

[2] We make available all our code: `http://imi.aau.dk/~bst`

[3] `http://www.music-ir.org/mirex/wiki/2005`

MIRUM'12, November 2, 2012, Nara, Japan.

in $\mathbf{f}(\mathbf{x})$. For a set of several fvs $\mathcal{X} := \{\mathbf{x}_i\}$, we have the set of score vectors $\{\mathbf{f}(\mathbf{x}_i)\}$. We thus pick the class for the set of features $\mathcal{X}$ by maximizing the logistic [9]:

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

where $[\mathbf{b}]_k$ denotes the kth element of $\mathbf{b}$, and we set $\gamma_{\mathcal{X}}$ so that $P[k|\mathcal{X}]$ is a probability distribution.

In our re-implementation, 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. To extract the features of an audio signal, we first segment it by 50%-overlapped Hann windows of duration 46.4 ms. We compute for each segment: the number of zero crossings; the variance 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 $m \in \{1, 2, \ldots, 16\}$); the error of a least-squares optimized 32-order linear predictor (autoregression) and the 40 Mel-frequency cepstral coefficients (MFCCs) as in [13]. For every 129 consecutive windows (3 seconds duration), we compute their mean and variances of these features to give a fv $\mathbf{x}_i$ with 120 dimensions.

2.2 Sparse representation classification with auditory modulations (SRCAM)

Sparse representation classification (SRC) [18] is motivated by the idea that sparse approximations can be class-discriminative. The use of SRC for music genre recognition is first proposed in [10]. Given a matrix of N fvs $\mathbf{D} := [\mathbf{d}_1|\mathbf{d}_2|\cdots|\mathbf{d}_N]$, and the set $\cup_{k=1}^{K} \mathcal{I}_k = \{1, \ldots, N\}$, where $\mathcal{I}_k$ specifies the columns of $\mathbf{D}$ belonging to class k, SRC first finds the sparse approximation of an unlabeled fv $\mathbf{x}$ by basis pursuit denoising for an $\epsilon^2 \geq 0$ [5]:

$$\mathbf{a}_{\mathbf{x}} = \arg\min_{\mathbf{a} \in \mathbb{R}^N} \|\mathbf{a}\|_1 \text{ subject to } \|\mathbf{x} - \mathbf{D}\mathbf{a}\|_2^2 < \epsilon^2. \quad (3)$$

SRC then defines the set of class-restricted weights $\{\mathbf{s}_k \in \mathbb{R}^N : k \in \{1, \ldots, K\}\}$ by

$$[\mathbf{s}_k]_n := \begin{cases} [\mathbf{a}_{\mathbf{x}}]_n, & n \in \mathcal{I}_k \\ 0, & \text{else.} \end{cases} \quad (4)$$

Finally, SRC selects the class simply by

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

We measure the "confidence" of SRC by comparing the errors of the class-dependent approximations, defined for class k by $J_k := \|\mathbf{x} - \mathbf{D}\mathbf{s}_k\|_2$. The confidence of class k for $\mathbf{x}$ is thus

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

The auditory modulation features as described in [10] are irreproducible [15]; but we are able to create similar features with help from the authors of [10]. We use the Lyon Passive Ear Model implemented in [13], and a downsampling factor of 40, to produce auditory spectrograms of signals 30 s in duration. To create the 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 as described in [15]. Finally, we find the squared energy of each Gabor filter output, which gives a distribution of energy in frequency and modulation rate. and then vectorize this to produce a 768-dimensional fv $\mathbf{x}$.

To produce the dictionary $\mathbf{D}$, we take the set of fvs $\{\mathbf{x}_j : j \in \{1, \ldots, N\}\}$ 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 scaling each dimension to have unit variance. Finally, we make each standardized fv have unit ℓ_2 norm and compose the dictionary by concatenating them all as columns. We apply the same standardization transformation used to create $\mathbf{D}$, to each fv to be classified. To solve (3), we use the SGPL1 solver [17] with at most 100 iterations, and $\epsilon^2 := 0.01$. Though we find that the solver converges only about 20% of the time, its output still appears favorably discriminative.

2.3 Experimental results

We train and test each system with SCV in GTZAN [16] — a dataset with 1,000 music recording excerpts of 30 s duration with 100 labeled examples in each of 10 different music genres: Blues, Classical, Country, Disco, Hip Hop, Jazz, Metal, Popular, Reggae, and Rock. Though GTZAN is a problematic dataset [14], we use it here for two reasons: to confirm that our re-implementations perform as is reported for GTZAN [4,10]; and because we are specifically interested in finding what makes them perform so well for GTZAN.

To train and test AdaBFFs, we use decision trees of one node (stumps), and 5-fold SCV with AdaBoost.MH run for 2500 iterations (as in [4]); and to test SRCAM, we use 10-fold SCV (as in [10]). Here, however, we run 10 SCV trials for the mean statistics of each system over training and testing dataset distributions. Each excerpt in GTZAN is thus classified ten times using different training data each time. Figure 1 shows the mean confusions of each system, and their 95% confidence intervals. We see that the overall MCA of AdaBFFs is 0.7755 ± 0.0022, which is significantly less than 83% reported in [4]. This could be due to their use of decision trees with an unspecified number of nodes for testing on GTZAN (though their MIREX 2005 submission did use decision stumps).[4] The MCA of SRCAM is 0.8203 ± 0.0019, which is about 9% worse than that reported in [10], but more than 10% higher than what has been achieved with previous attempts at reproducing their results in GTZAN [15].

3. A CLOSER LOOK AT BEHAVIORS

In the literature, the performance analysis of genre recognition systems often consists of only MCA and confusion matrices. We move beyond this to examine the behaviors underlying these systems. First, we look at specific excerpts that each system consistently and persistently mislabels. Second, we test the sensitivity of the systems to the spectral equalization of a music excerpt. Finally, we attempt to survey the internal genre models of each system by testing how well humans recognize the genres of excerpts it composes to be highly representative of a particular genre.

3.1 Consistent & Persistent Misclassifications

We define a *consistent and persistent mislabeling* (CPM) to be when an excerpt is mislabeled the same way in all 10 SCV trials. For lack of space, we only discuss CPMs of Disco-labeled excerpts in GTZAN. Disco is a style of dance

[4]Personal communication with James Bergstra.

(a) AdaBFFs

	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

True Genre

(b) SRCAM

	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

True Genre

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

music that emerged in the early 1970s in the UK and USA, and quickly became a popular style world-wide [12]. Disco blends, among others, Funk and Soul styles, typically uses a common time meter at a steady tempo of around 120 beats per minute. typically has a distinctive use of the open hi-hat on the off beats, as well as prominent and bouncy electric bass lines, and harmonized rich accompaniment by strings, brass, keyboards, female vocals, and synthesizers [1].

Figure 2 shows how each system, in the 10 trials of SCV used for Fig. 1, labels the Disco-labeled excerpts of GTZAN. The darkness of a tile represents the frequency of the applied label. We find AdaBFFs has 10 CPMs: four as Pop, three as Rock, and one each as Classical, Country, and Hip hop. For SRCAM, we find 12 CPMs: three each as Hip hop, Pop, and Reggae, two as Classical, and one as Rock. Of these, both systems share three of the same CPMs, but each of these are "forgivable" for the following reasons. First, (excerpt) 23 comes from Latoya Jackson's album "Bad Girl" from 1991, more than a decade after the climax of Disco in the USA [12]. The song does not have characteristics particular to Disco, and the top last.fm tag[5] applied to this artist is "pop" (from here on all tags come from last.fm). Second, 29 comes from Evelyn Thomas's "Heartless" from 1984, which is not tagged "disco." Finally, 47 is only Barbra Streisand and Donna Summer singing at a slow tempo and softly accompanied by piano and strings. Though the song from which it comes is exemplary Disco, the excerpt lacks drums and bass lines so distinctive of Disco that we consider this CPM as forgivable. Individually, each system has other forgivable CPMs: excerpt 27 for AdaBFFs (top tags are "Hip-Hop" and "rap"); and for SRCAM excerpts 21 (its source is the Disco song "Never Can Say Goodbye," but this excerpt comes from a "modern pop"-sounding version), and 85, "Wordy Rappinghood," featuring a sparse sequenced drum loop and a female vocalist rapping about words.

Each system, however, has CPMs that are not so forgivable. Excerpt 84 is certainly Rock for AdaBFFs, and Reggae for SRCAM. The identity this excerpt is currently unknown [14], but, to us, its content (up-tempo sequenced drums, bass and horns, funk/disco rhythm guitar, cowbells and handclaps, no spring reverberation) has little in common with the Rock- or Reggae-labeled GTZAN excerpts. "Boogie Nights" — a classic in the discography of Disco [12] — and excerpt 28 both have the top last.fm tag "disco," but AdaBFFs is adamant that they are Pop. Similarly, AdaBFFs insists that ABBA's "Dancing Queen" and Dee's "Disco Duck" are both Rock though their top 5 last.fm tags include "disco" and not "rock"; and Heatwave's "Always and Forever", with a top last.fm tag of "soul" is Country. For SRCAM, Disco classic "Kung Fu Fighting" [12], and excerpt 79 are undoubtedly Hip hop. The former has the top tag "disco," and the artist of the latter has the top tags "funk, disco funk." Also for SRCAM, though they all have the top tag "disco," "Funkytown" and excerpt 53 are always Reggae, and the excerpt by Disco-Tex and the Sex-O-Lettes is Rock. SRCAM is strangely insistent that "Why?" by Bronski Beat (top tags: "80s, new wave, synthpop") is Classical.

Table 1 shows for each system the CPMs as Disco for excerpts in other categories of GTZAN. Some of these are forgivable, e.g., Pop 12 and 63 having Aretha Franklin and Diana Ross, respectively; but other CPMs are not so forgivable. Both systems are unanimous in their Disco classification of "Honky Tonk Woman" from 1969 by The Rolling Stones, and the Hip hop "Looking for the Perfect Beat" by Afrika Bambaata. And AdaBFFs is resolute that The Beach Boys's 1966 hit "Good Vibrations" is as Disco as Chic.

3.2 Genre-shifting Mastering

We now test the robustness of each system to changes in the spectral characteristics of an excerpt. If a it is capable of recognizing genre, then it should be robust to spectral

[5] `http://last.fm` is an on-line service gathering information about music from listeners around the world. A tag is a word or phrase applied by a listener describing an artist or song.

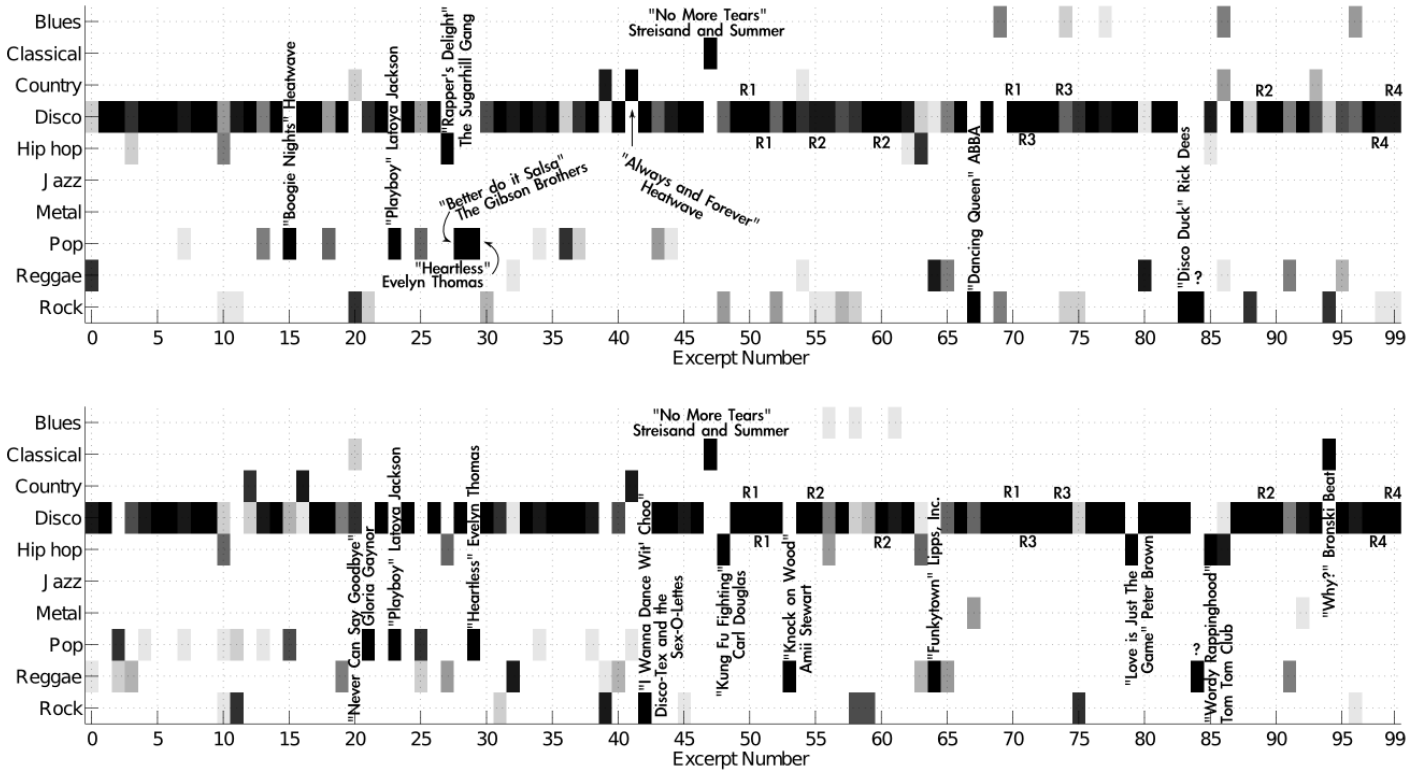


Figure 2: Assignments (y-axis) of Disco-labeled excerpts in GTZAN in 10 SCV trials by AdaBFFs (top) and SRCAM (bottom). Black is an excerpt labeled the same 10 times. R# show exact replicas [15].

	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 Bambaata "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.

changes of an excerpt. For instance, humans can recognize any of the 10 genres in GTZAN whether such music is played on AM or FM radio. To the end, we train AdaBFFs with the all of GTZAN using 2500 training iterations of AdaBoost.MH. For SRCAM, the standardized ATMs of all excerpts compose the dictionary. We take a 30 s excerpt of recorded music (sampled at 22.050 kHz), pass it through a 94-channel Gammatone filterbank with channels either on or off (center frequencies from 110 Hz to about 9 kHz), and sum the output with that of a lowpass filter preserving the frequency content below 110 Hz (Fig. 3). Each system then classifies the "equalized" signal.

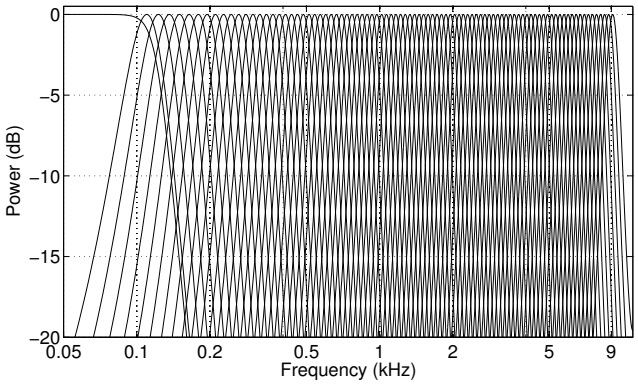
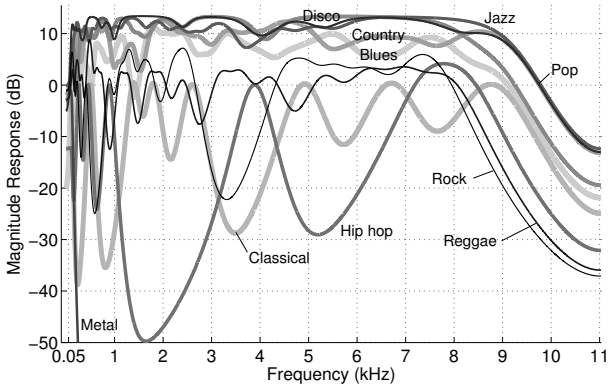


Figure 3: Magnitude response of 95-band filterbank.

To begin, we take a 30 s excerpt of the Western Swing song "Big Balls in Cowtown" by Bob Wills and the Texas Playboys. Each system labels this excerpt as Country. However, often with only minor changes to the sound, we are able to make AdaBFFs label this music with all ten genres of GTZAN, and SRCAM with eight genres. For the particular label applied to the excerpt by each system, Fig. 4 shows the magnitude responses of the equalizations. We see that with a majority of the filters on, AdaBFFs applies Disco, Pop, Reggae and Rock, and SRCAM applies Disco, Jazz, Pop and Reggae. Listening to these filtered excerpts shows that they are not changed so significantly that a human would make such confusions. Furthermore, we find the two systems behavior similarly with other excerpts, including "Symphony of Destruction" by Megadeth, "Poison" by Bel Biv Devoe, and even "blues.00001" from GTZAN. In all cases, while these systems correctly label the "original" ver-

(a) AdaBFFs



(b) SRCAM

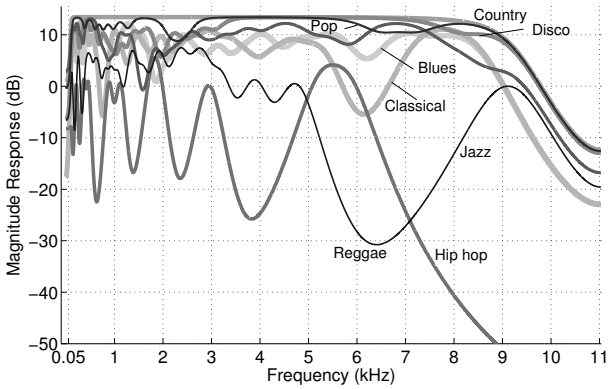


Figure 4: Equalizations for each system to claim a particular genre for Western Swing song "Big Balls in Cowtown" by Bob Wills and the Texas Playboys.

sions, we can coax each to label the same music with widely different genres by spectral equalization that is often only perceptually minor.

3.3 Representative Excerpts

We now test whether human subjects can recognize the genres of excerpts each system (trained on all excerpts of GTZAN) composes to be highly representative of each genre. To do this, we take the 1,198 sample loops that accompany Apple's GarageBand program.[6] These loops 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 randomly combine four sample loops, each one repeated to last 30 seconds, and have each system select an excerpt to best represent each genre. AdaBFFs selects a randomly composed excerpt to be representative of genre k only if $P[k|\mathcal{X}] > 0.999$. Likewise, SRCAM selects it 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 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 ($< 1\%$).

Taking one representative excerpt produced by each system for each of the ten genres in GTZAN, we perform listening tests as follows. First, we tell the subject that they will listen to up to 30 musical excerpts of about 10 seconds in length. For each one, they are to pick one of the ten genres listed that best describes it. 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. Then the subject dons a pair of headphones, and interacts with a GUI built in MATLAB. In order to screen subjects for their ability to recognize the ten genres, the first ten excerpts are ones we selected from GTZAN for their genre representability.[7] 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 from AdaBFFs, and the final ten are from 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 subjects passed the screening, and so we define statistical significance by $\alpha = 0.05$. Assuming independence between each trial, we model the number (no.) correct N as a random variable distributed Binomial$(20, 0.1)$. The expectation $E[N] = 20(0.1) = 2$, and variance $\text{Var}[N] = 20(0.1)(0.9) = 1.8$. 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 system, 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 $\mathcal{H}_0$. In summary, subjects are unable to recognize the genres of the representative excerpts.

It is clear from the data that even though all genres are equally represented, test 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%), Country (6.5%), Classical (2.0%), and Metal (1.3%). The SRCAM-composed Reggae excerpt contains a loud 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 several genres 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

[6]This program is made for people to easily make music within a loop-based sequencing environment.

[7]Artists and titles in [14]: Blues 5, Classical 96, Country 12, Disco 66, Hip hop 47, Jazz 19, Metal 4, Pop 95, Reggae 71, Rock 40.

as those used in [16], 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 MCA by over 20% compared to past low-level approaches like [16]. In this paper, we have examined whether the high classification accuracies of both AdaBFFs and SRCAM reflects a capacity to *recognize* genre.

First, we have found that each system commits CPMs, and have looked specifically at the Disco-labeled excerpts of GTZAN. We do not expect human performance; but we should not expect from a system with a capacity to recognize a genre such CPMs of excerpts that, for the most part, clearly meet the stylistic rules of a genre supposedly learned. This type of analysis might be argued unfair, and that we must look at the instances of correct classification — seen for Disco in Fig. 2. Looking at CPMs, however, illuminates the deficiencies in the genre schema of a system, and seeks to answer the question, "How does it recognize Disco?" instead of, "How often does it recognize Disco?"

Second, we have found that 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 raises the question of how they are performing so well in the first place. It might be argued that if we include in the training data the original excerpts and filtered versions, then these systems might perform better. In addition to the problems of having to define what is "unfiltered" when everything is filtered, and specifying what filters to use and how many permutations, it is not clear how this could help either system learn music genre in a more complete way.

Finally, we have attempted to tease out the internal genre models of each system by having them "compose" highly genre-representative music from sample loops. Through formal listening tests and statistical analyses, we find humans do not recognize the genres supposedly represented. It might be argued that the composed excerpts are not "real music" when compared to the data with which the systems are trained. Thus, it would be better to use "relatively unknown but real songs." Such an experiment, however, does not help reveal the inner models. Consider a horse that can correctly clop the sum of two integers spoken by its trainer. Asking exhaustively for the sums of every pair of digits does not test its *understanding* of arithmetic. We wish to ask instead, e.g., "What sums make 5?"

Acknowledgments

Thanks to Yannis Panagakis and Costas Kotropoulos for their help in building features for SRCAM; to James Bergstra and Norman Casagrande for discussions on AdaBoost; and to Mark Plumbley, Geraint Wiggins, Nick Collins and Fabien Gouyon for several helpful discussions. This work supported in part by: Independent Postdoc Grant 11-105218 from Det Frie Forskningsråd; the Danish Council for Strategic Research of the Danish Agency for Science Technology and Innovation in project CoSound, case no. 11-115328; EPSRC Platform Grant EP/E045235/1 at the Centre for Digital Music of Queen Mary University of London.

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 worth pursuing and how can it be improved? In *Proc. Int. Soc. Music Info. Retrieval*, Victoria, Canada, Oct. 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] P. Shapiro. *Turn the Beat Around: The Secret History of Disco*. Faber & Faber, London, U.K., 2005.

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

[14] B. L. Sturm. An analysis of the GTZAN music genre dataset. In *Proc. ACM Workshop on Music Information Retrieval with User-Centered and Multimodal Strategies*, Nara, Japan, Nov. 2012.

[15] B. L. Sturm and P. Noorzad. On automatic music genre recognition by sparse representation classification using auditory temporal modulations. In *Proc. Int. Symp. Computer Music Modeling and Retrieval*, London, U.K., June 2012.

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

[17] 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.

[18] 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.