On Automatic Music Genre Recognition by Sparse Representation Classification using Auditory Temporal Modulations

Bob L. Sturm[1] and Pardis Noorzad[2]

[1] Department of Architecture, Design and Media Technology
Aalborg University Copenhagen
Lautrupvang 15, 2750 Ballerup, Denmark
[2] Department of Computer Engineering and IT
Amirkabir University of Technology
424 Hafez Ave., Tehran, Iran
bst@create.aau.dk, pardis@aut.ac.ir

Abstract. A recent system combining sparse representation classification (SRC) and a perceptually-based acoustic feature (ATM) [31, 30, 29], is reported to outperform by a significant margin the state of the art in music genre recognition, e.g., [3]. With genre so difficult to define, this remarkable result motivates investigation into, among other things, why it works and what it means for how humans organize music. In this paper, we review the application of SRC and ATM to recognizing genre, and attempt to reproduce the results of [31] where they report 91% accuracy for a 10-class dataset. We find that only when we pose the sparse representation problem with inequality constraints, and, more significantly, reduce the number of classes by half, do we begin see accuracies near those reported. In addition, we find evidence that this approach to classification does not benefit significantly from the features being based on a perceptual analysis.

1 Introduction

Simply because we lack clearly definitive examples, and any utilitarian definitions, the automatic recognition of music genre is different from other tasks in music information retrieval. The human categorization of music seems natural, yet appears fluid and often arbitrary by the way it appears motivated by more than measurable characteristics of audible changes in pressure [9, 26, 17]. Extra-musical information, such as artist fashion, rivalries and the fan-base, associated dance styles, lyrical subjects, societal and political factors, religious beliefs, and origins in time and location, can position a particular piece of music into one category or another, not often without debate [38]. With the changing fashions of communities and the needs of companies, new genres are born [17]. And genres become irrelevant and lost, though we might still hear the recorded music and classify it as something entirely different.

It seems daunting then to make a computer recognize genre with any success. Yet, in developments between 2002 and 2006 [23], we have seen the accuracy of such algorithms progress from about 60% — using parametric models created from bags of features [43] — to above 80% — aggregating features over long time scales and boosting weak classifiers [3]. The majority of approaches developed

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

so far use features derived only from the waveform, and/or its symbolic form. Some work has also explored mining user tags [27] and written reviews [1], or analyzing song lyrics [22]. Since humans have been measured to have accuracies around 70% after listening to 3 seconds of music — which surprisingly drops only down to about 60% for only half a second of listening [13] — the results of the past decade show that the human categorization of music appears grounded to a large extent in acoustic features, at least at some coarse granularity.

Recently, we have seen a large leap in genre classification accuracy. In [31, 30, 29], the authors claim that with a perceptually-motivated acoustic feature, and a framework of sparse representation classification (SRC) [46], we move from 82.5% accuracy [3] to up to 93.7%. SRC, which has produced very promising results in computer vision [46, 47] and speech recognition [11, 35], can be thought of as a generalization of k-nearest neighbors (kNN) for multiclass classification with many important advantages. It is a global method, in the sense that it classifies based on the entire training set; and it does not rely only on local similarity information as does kNN. SRC can prevent overcounting of neighborhood information by virtue of its emphasis on sparsity in the representation. Additionally, SRC assigns a weight to each training set sample, thus quantifying the degree of its importance. All of these points make SRC a strong classification method.

The massive improvement in genre recognition that accompanies this approach motivates many questions, not only about what is working and why it is working so well, but also about how we perceive rich acoustic scenes, and the way we think and talk about music. For instance, are these purely acoustic features so discriminative because they are modeled on the auditory system of humans? Since the features in [31] are computed from segments of very long duration (30 s), how robust is the method to shorter observations, e.g., can it reach 60% for 500 ms? Do its misclassifications make sense, and to some extent forgivable? Do the features cluster in a sensible way, and do subgenres appear as smaller clusters within larger clusters? Can we compute high-level descriptors from these features, such as rhythm, harmony, or tempo?

In this work, we review the approach proposed in [31], and describe our attempt to reproduce the results, making explicit the many decisions we have had to make to produce the features, and to build the classifier. The accuracies we observe, however, are 30–40% inferior to those in [31], even with the improvement we observe when posing the sparse representation problem using inequality constraints rather than the equality constraints specified in [31, 30, 29]. Only when we reduce by half the number of classes tested in [31] do we see the reported high accuracies. In addition, we find evidence that the perceptual nature of the features has no significant impact on the classifier accuracy. We make available our MATLAB code, both classification and feature extraction, with which all results and figures in this article can be reproduced: `http://imi.aau.dk/~bst/software/`.

2 Background

We now review SRC from a general perspective, and then we review modulation analysis for feature extraction, and its application specifically to music genre recognition. Throughout, we work in a real Hilbert space with inner product $\langle \mathbf{x}, \mathbf{y} \rangle := \mathbf{y}^T \mathbf{x}$, and p-norm $\|\mathbf{x}\|_p^p := \sum_i |[\mathbf{x}]_i|^p$, for $p \geq 1$, where $[\mathbf{x}]_i$ is the ith component of the column vector $\mathbf{x}$.

2.1 Classification via sparse representation in labeled features

Define a set of N labeled features, each belonging to one of C enumerated classes

$$\mathcal{D} := \{(\mathbf{x}_n, c_n) : \mathbf{x}_n \in \mathbb{R}^m, c_n \in \{1, \ldots, C\}\}_{n \in \{1, \ldots, N\}} . \tag{1}$$

And define $\mathcal{I}_c \subset \{1, \ldots, N\}$ as the indices of the features in $\mathcal{D}$ that belong to class c. Given an unlabeled feature $\mathbf{y} \in \mathbb{R}^m$, we want to determine its class using $\mathcal{D}$. In kNN, we assume that the neighborhood of $\mathbf{y}$ carries class information, and so we classify it by a majority vote of its k-nearest neighbors in $\mathcal{D}$. Instead of iteratively seeking the best reconstruction of $\mathbf{y}$ by a single training sample (i.e., its ith nearest neighbor), we find a reconstruction of $\mathbf{y}$ by all training samples. Then we choose the class whose samples contribute the most to the reconstruction. We have SRC when we enforce a sparse reconstruction of $\mathbf{y}$.

SRC essentially entails finding nearest to an unlabeled feature its linear approximation by class-restricted features. To classify an unlabeled feature $\mathbf{y}$, we first find the linear combination of features in $\mathcal{D}$ that constructs $\mathbf{y}$ with the fewest number of non-zero weights, regardless of class membership, posed as

$$\min_{\mathbf{a} \in \mathbb{R}^N} \|\mathbf{a}\|_0 \ \text{ subject to } \ \mathbf{y} = \mathbf{D}\mathbf{a} \tag{2}$$

where we define the $m \times N$ matrix $\mathbf{D} := [\mathbf{x}_1 | \mathbf{x}_2 | \ldots | \mathbf{x}_N]$, and the pseudonorm $\|\mathbf{a}\|_0$ is defined as the number of non-zero weights in $\mathbf{a} := [a_1, a_2, \ldots, a_N]^T$. We might not want to enforce equality constraints, and so we can instead pose this

$$\min_{\mathbf{a} \in \mathbb{R}^N} \|\mathbf{a}\|_0 \ \text{ subject to } \ \|\mathbf{y} - \mathbf{D}\mathbf{a}\|_2^2 \leq \epsilon^2 \tag{3}$$

where $\epsilon^2 > 0$ is a maximum allowed error in the approximation. All of this, of course, assumes that we are using features that are additive. We can extend this to non-linear combinations of features by adding such combinations to $\mathcal{D}$ [35], which can substantially increase the size of the dictionary.

We now define the set of class-restricted weights $\{\mathbf{a}_c\}_{c \in \{1,2,\ldots,C\}}$

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

The non-zero weights in $\mathbf{a}_c$ are thus only those specific to class c. From these, we construct the set of C approximations and their labels $\mathcal{Y}(\mathbf{a}) := \{\hat{\mathbf{y}}_c(\mathbf{a}) := \mathbf{D}\mathbf{a}_c\}_{c \in \{1,2,\ldots,C\}}$, and we assign a label to $\mathbf{y}$ simply by a nearest neighbor criterion

$$\hat{c} := \arg \min_{c \in \{1,\ldots,C\}} \|\mathbf{y} - \hat{\mathbf{y}}_c(\mathbf{a})\|_2^2. \tag{5}$$

Thus, SRC picks the class of the nearest approximation of $\mathbf{y}$ in $\mathcal{Y}(\mathbf{a})$.

We cannot, in general, efficiently solve the sparse approximation problems above [8], but there exist several strategies to solve them. We briefly review the convex optimization approaches, but [42] provides a good overview of many more; and [48] is a large study of SRC using many approaches. Basis pursuit (BP) [6] proposes relaxing strict sparsity with the convex ℓ_1-norm

$$\min_{\mathbf{a}\in\mathbb{R}^N} \|\mathbf{a}\|_1 \text{ subject to } \mathbf{y} = \mathbf{D}\mathbf{a}. \tag{6}$$

And without equality constraints, BP denoising (BPDN) [6] poses this as

$$\min_{\mathbf{a}\in\mathbb{R}^N} \|\mathbf{a}\|_1 \text{ subject to } \|\mathbf{y} - \mathbf{D}\mathbf{a}\|_2^2 \leq \epsilon^2. \tag{7}$$

One could also change the ℓ_2 error to ℓ_1 to promote sparsity in the error [47, 12]. We have the LASSO [41] when we switch the objective and constraint of BPDN

$$\min_{\mathbf{a}\in\mathbb{R}^N} \|\mathbf{y} - \mathbf{D}\mathbf{a}\|_2^2 \text{ subject to } \|\mathbf{a}\|_1 \leq \rho \tag{8}$$

where $\rho > 0$. Furthermore, we can pose the problem in a joint fashion

$$\min_{\mathbf{a}\in\mathbb{R}^N} \frac{1}{2}\|\mathbf{y} - \mathbf{D}\mathbf{a}\|_2^2 + \lambda\|\mathbf{a}\|_1 \tag{9}$$

where $\lambda > 0$ tunes our preference for sparse solutions versus small error.

Along with using the ℓ_1 norm, we can reduce the dimensionality of the problem in the feature space [47]. For instance, the BPDN principle (7) becomes

$$\min_{\mathbf{a}\in\mathbb{R}^N} \|\mathbf{a}\|_1 \text{ subject to } \|\mathbf{\Phi}\mathbf{y} - \mathbf{\Phi}\mathbf{D}\mathbf{a}\|_2^2 \leq \epsilon^2 \tag{10}$$

where $\mathbf{\Phi}$ is a fat full-rank matrix mapping the features into some subspace. To design $\mathbf{\Phi}$ such that the mapping might benefit classification, we can compute it using information from $\mathbf{D}$, e.g., by principal component analysis (PCA) or non-negative matrix factorization (NMF), or we can compute it non-adaptively by random projection. With PCA, we obtain an orthonormal basis describing the directions of variation in the features, from which we define $\mathbf{\Phi}^T$ as the $d \leq m$ significant directions, i.e., those having the d largest principal components.

Given $d \leq m$, NMF finds a positive full rank $m \times d$ matrix $\mathbf{U}$ such that

$$\min_{\mathbf{U}\in\mathbb{R}_+^{m\times d}} \frac{1}{N}\sum_{n=1}^{N} \|\mathbf{x}_n - \mathbf{U}\mathbf{v}_n\|_2^2 \text{ subject to } \mathbf{v}_n \succeq 0. \tag{11}$$

The full-rank matrix $\mathbf{U}$ contains d templates that approximate each feature in $\mathbf{D}$ by an additive combination. Thus the range space of $\mathbf{U}$ provides a good approximation of the features in $\mathbf{D}$, with respect to the mean ℓ_2-norm of their errors. In this case, we make $\mathbf{\Phi}^T := (\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T$.

Finally, we can reduce feature dimensionality by random projection [7, 4, 21], where we form the entries of $\mathbf{\Phi}$ by sampling from a random variable, e.g., Normal, and without regard to $\mathbf{D}$. We normalize the columns to have unit ℓ_2-norm, and ensure $\mathbf{\Phi}$ has full rank. While this approach is computationally simple, its non-adaptivity can hurt classifier performance [21].

2.2 Modulation Analysis

Modulation representations of acoustic signals describe the variation of spectral power in scale, rate, time and frequency. This approach has been motivated by the human auditory and visual systems [44, 15, 36, 40, 24]. In the literature, we find two types of modulation representations of acoustic signals, which seemingly have been developed independently. One might see these approaches as a form of feature integration, which aggregate a collection of small scale features.

In [44, 24], the authors model the output of the human primary auditory system as a multiscale spectro-temporal modulation analysis, which [44] terms a "reduced cortical representation" (RCR). To generate an RCR, one first produces an "auditory spectrogram" (AS) approximating the time-frequency distribution of power at the output of the early stage of the auditory system [49]. This involves filtering the signal with bandpass filters modeling the frequency responses of the hair cells along the basilar membrane, then calculating activations of the nerve cells in each band, and finally extracting a spectral power estimate from the activation patterns [49, 44, 24]. In the next step, which models the central auditory system, one performs a "ripple analysis" of the AS, giving the local magnitudes and phases of modulations in scale and modulation rate over time and frequency [44, 24]. This procedure uses 2-D time-frequency modulation-selective filters, equivalent to a multiresolution affine wavelet analysis sensitive to fast and slow upward and downward changes in frequency [44]. To obtain spectro-temporal modulation content [24], one integrates this four-dimensional representation over time and/or frequency.

A similar representation is proposed in [15], where the authors extract modulation information by applying a Fourier transform to the output of a set of bandpass filters modeling the basilar membrane. The magnitude output of this gives a time-varying modulation spectrogram. One could instead apply a wavelet transform to each row of a magnitude spectrogram, and then integrate the power at each scale of each band along the time axis. This produces a modulation rate-scale representation [40].

Motivated by its perceptual foundation [49, 44], and success in automatic sound discrimination [45, 24], the work of [28] appears to be the first to use modulation analysis features for music genre recognition, which they further refine in [31, 30, 32]. In [28], the authors use the toolbox by the Neural Systems Laboratory (NSL) [33, 34] to derive an RCR and perform a ripple analysis. They then average this across time to produce tensors of power distributed in modulation rate, scale, and acoustic frequency. While the features in that work are built from the RCR [44, 24], the features used in [31, 30] are a joint scale-frequency analysis [40] of an AS created from the model in [49]. This feature, which they call an "auditory temporal modulation" (ATM), describes power variation over modulation scale in each primary auditory cortex channel.

3 Recreating the Features and Classifier of [31]

In this section, we first describe how we generate ATM features, which are described in part in [31, 32]; and then we describe the approach to classify an ATM using SRC presented in part in [31].

3.1 Building Auditory Temporal Modulations

The authors take 30 seconds of music, downsample it to 16 kHz, then make it zero mean and unit variance. They then compute an AS following the model of the primary auditory system of [49], except they use a constant-Q transform of 96 bandpass filters covering a 4-octave range (24 filters per octave), whereas [49] uses an affine wavelet transform of 64 scales covering 5 octaves from about 173 Hz to 5.9 kHz. Finally, they pass each channel of the AS through a Gabor filterbank sensitive to particular modulation rates, and form the ATM by integrating the energy output at each filter.

To create ATMs, we have tried to follow as closely as possible the description in [31, 32]. We first generate a constant-Q filter bank with 97 bands spaced over a little more than four octaves, with $N_f = 24$ filters per octave. We center the first filter at 200 Hz because that is specified in [49]. The last filter is thus centered on 3200 Hz. Since in [49] the model of the final stage of the primary audio cortex computes first-order derivates across adjacent frequency bands, we end up with a 96 band AS as specified in [31, 32].

We create our constant-Q filter bank as a set of finite impulse response filters designed by the windowing method [25]. Since it is not mentioned in [49, 31, 32], we make all filters independent, and to have the same gain. To generate the impulse responses of our filterbank, we modulate a prototype lowpass window to logarithmically spaced frequencies. Because of its good low passband characteristic, we use a Hamming window, which for the kth filter ($k \geq 1$) produces the impulse response sampled at F_s Hz

$$h_k(n) := \gamma_k \left[0.54 - 0.46 \cos\left(\frac{2\pi n}{l_k}\right)\right] e^{j2\pi\omega_k n/F_s},\ 0 \leq n < l_k \tag{12}$$

with a modulation frequency $\omega_k := f_{\min} 2^{(k-1)/N_f}$ Hz, and length in samples

$$l_k := \left\lceil \frac{q}{2^{k/N_f} - 2^{(k-1)/N_f}} \frac{F_s}{f_{\min}} \right\rceil. \tag{13}$$

We set the gain γ_k such that there is no attenuation at the kth center frequency, i.e., $|\mathcal{F}\{h_k(n)\}(\omega_k)| = 2$, where $\mathcal{F}\{x(n)\}(\omega)$ is the Fourier transform of $x(n)$ evaluated at frequency ω. The factor $q > 0$ tunes the width of the main lobe. We choose $q \approx 1.316$ such that adjacent filters overlap at their -3 dB stopband.

This model of the basilar membrane is simplified considering its non-adaptive and uniform nature, e.g., it does not take into account masking and equal loudness curves. An alternative model of the cochlea is given by Lyon [20], which involves a filterbank with center frequencies spread uniformly below a certain frequency, and logarithmically above [37]. Figure 1 shows that the Lyon model attenuates single sinusoids at frequencies tuned to the center frequencies of its filterbank. Our filterbank uniformly passes these frequencies, albeit over a smaller four octave range [31, 32] assumed to begin at 200 Hz. Figure 1 also shows that the filterbank of the NSL model [34] by and large has a uniform attenuation.

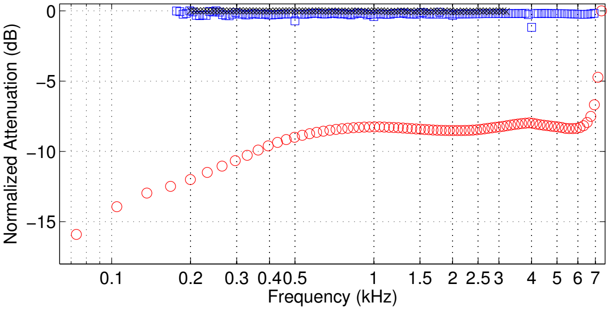


Fig. 1. Attenuations of single sinusoids with the same power, at frequencies identical to center frequencies in the filterbanks. (x) Our constant-Q filter bank. (o) Lyon passive ear model [20, 37]. (□) NSL ear model [34].

We pass though our constant-Q filter bank a sampled, zero-mean and unit-variance acoustic signal $y(n)$ [31, 32], which produces for the kth filter the output

$$y_k(n) := \sum_{m=0}^{l_k-1} h_k(m)y(n-m-\Delta_k) \tag{14}$$

where $\Delta_k > 0$ is the group delay of the kth filter at ω_k. This delay correction is necessary because the filters we use to model the basilar membrane have different lengths. This correction is unnecessary in the implementation of the Lyon [37] or NSL models [34], since they use second-order sections with identical delays.

As in [31, 32], we next take the sample wise difference in each band

$$y'_k(n) := y_k(n) - y_k(n-1). \tag{15}$$

which models the action potential of the hair cell [49]. This now goes through a non-linear compression, followed by a low pass filter modeling leakage in the hair cell membrane. Referring to [49], we see the compression can be modeled as a sigmoidal function, and that the output of the kth channel is

$$g_k(n) := \frac{1}{1+e^{-\gamma y'_k(n)}} - \frac{1}{2} \tag{16}$$

where $\gamma > 0$ is depends on sound pressure level [49]. Furthermore, "... saturation in a given fiber is limited to 30-40 dB" [49], implying γ is somehow set adaptively. In reality, we cannot equate the values of the digital samples in $y'_k(n)$ with the physical pressure embodied in this compression. However, working naively, we might absorb into γ such a conversion, and find some value that actually compresses. Figure 2 shows the cumulative distribution of amplitudes input to the compressor (15) with a 30 second music signal having unit energy [31, 32]. For $\gamma = 1$, we see that this distribution is compressed, whereas setting $\gamma = 10$ results in an expansion. Thus, we set $\gamma = 1$ independent of the input, and assume it compresses $y'_k(n)$ from any 30 second music signal scaled to have unit energy.

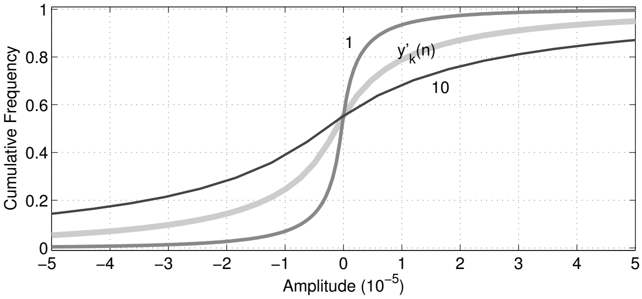


Fig. 2. Cumulative distributions of amplitude input to compressor ($y'_k(n)$), and output as a function of γ (labeled).

The compressor output $g_k(n)$ is then smoothed by the hair cell membrane and attendant leakage [49, 32], which passes frequencies only up to 4 − 5 kHz [49]. Thus, we pass each $g_k(n)$ through a 6th-order Butterworth filter having a cutoff frequency of 4 kHz, producing $f_k(n)$. This is then processed by a "lateral inhibitory network," described in [49], which detects discontinuities in the response. This entails a spatial derivative across channels with smoothing, a half-wave rectifier, and then integration; but [31, 32] does not specify smoothing, and states the process can be approximated by a first order derivative across logarithmic frequency. Thus, we compute for channel $s \in \{1, \ldots, 96\}$

$$v_s(n) := [f_{s+1}(n) - f_s(n)]\mu[f_{s+1}(n) - f_s(n)] \tag{17}$$

where $\mu(u) = 1$ if $u \geq 0$, and zero otherwise.

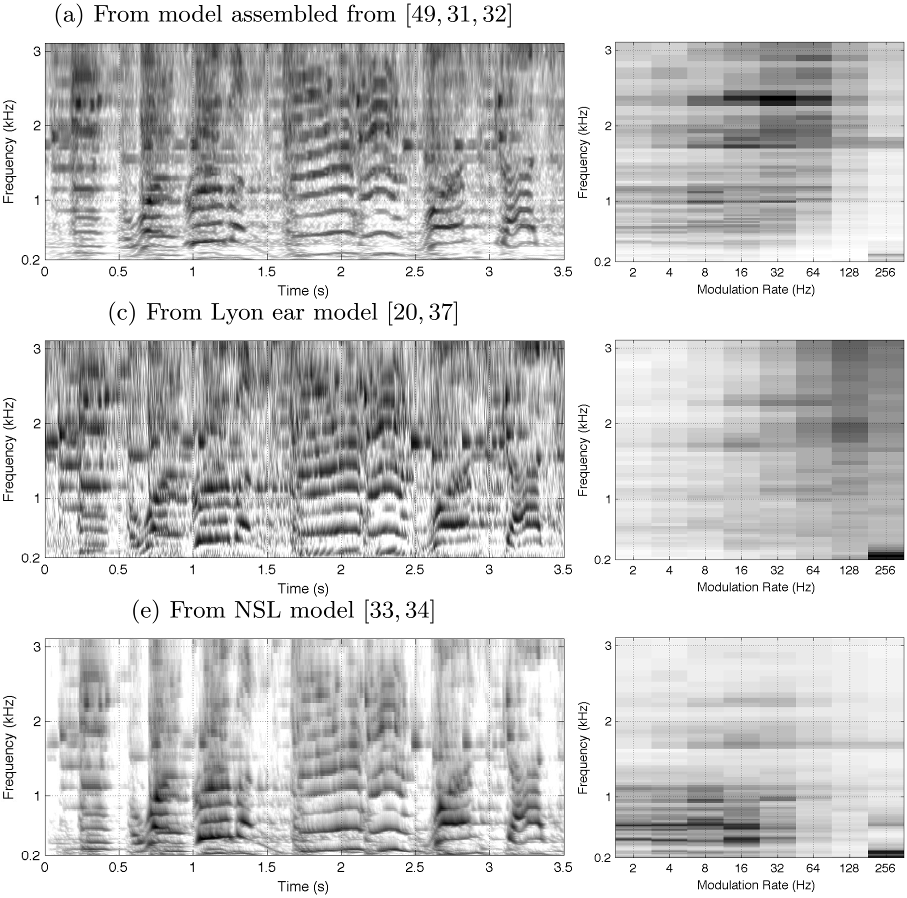
In the final step, we integrate the output with "a [possibly rectangular window with a] long time constant (10-20 ms)" [49], or a 2 − 8 ms exponential window [31, 32]. Thus, we compute the nth sample of the kth row of the AS by

$$A_k(n) := \sum_{m=0}^{\lfloor F_s \tau \rfloor} v_s(n-m) e^{-m/F_s \tau} \tag{18}$$

where we define $\tau := 8$ ms. This completes the first step of building an ATM.

Figure 3 compares the resulting AS from our model built from interpreting [49, 31, 32], that of the auditory model designed by Lyon [20, 37], and the cortical representation from the NSL model [33, 34]. The Lyon model uses 86 bands non-uniformly spread over a little more than 6.5 octaves in 80 − 7630 Hz [20, 37], whereas the NSL model covers 5.33 octaves with 24 filters per octave logarithmically spread over 180 − 7246 Hz [33, 34]. Though the frequency range of those models are larger, we only use a 4-octave range as in [31, 32].

To generate an ATM, [31, 32] describe first performing a multiresolution wavelet decomposition of each row of an AS, and then integrating the squared output across the translation axis. Based on experimental evidence [36], the authors use a set of Gabor filters sensitive to eight modulation rates $\{2, 4, 8, \ldots, 256\}$


(a) From model assembled from [49, 31, 32]

(c) From Lyon ear model [20, 37]

(e) From NSL model [33, 34]

Fig. 3. Auditory spectrograms (left) and their auditory temporal modulations (right).

Hz. We assume this Gabor filterbank can be assembled as follows. We define the sampled impulse response truncated to length N_l of our complex Gabor filter tuned to a modulation rate $f_0 2^l \geq 0$ Hz, and of scale $F_s\alpha/f_0 2^l > 0$

$$\psi(n; f_0 2^l) := \frac{f_0 2^l}{F_s \alpha} \left[e^{-(f_0 2^l/\alpha)^2((n-N_l/2)/F_s)^2} e^{j2\pi f_0 2^l n/F_s} - \mu_l \right] \qquad (19)$$

for $n = 0, \ldots, N_l - 1$, where we define μ_l such that $\psi(n; f_0 2^l)$ has zero mean. The normalization constant assures uniform attenuation at each modulation frequency, as used in joint scale-frequency analysis [40]. We set $\alpha = 256/400$ and $N_l = 4F_s\alpha/f_0 2^l$. Since a Gabor filter tuned to a low frequency has a high DC component, we make each row of the AS zero mean, thus producing $A'_k(n)$. Passing the kth row of this AS through the lth channel ($l \in \{0, 1, \ldots, 7\}$) of the Gabor filterbank produces the convolution $R_{k,l}(n) := [\psi(m; f_0 2^l) \star A'_k(m)](n)$. Finally, as in [31, 32], we sum the squared modulus of the output sampled at all wavelet translations, producing the (k,l) element of the ATM

$$[\mathbf{A}]_{kl} := \sum_{p\in\mathbb{Z}} |R_{k,l}(p\lfloor F_s\alpha/f_0 2^{l+1}\rfloor)|^2 \tag{20}$$

where p is an integer multiplying the wavelet translations, which we assume is half the wavelet scale.

To the right of each AS in Fig. 3 we see the resulting ATM. Portions of these ATMs appear similar, with major differences in scaling and feature dimensionality. Within the four octave range specified in [31, 32], the dimensionality of the vectorized features are: 768 for the ATM in [31, 32], 416 for that created from the model by Lyon [20, 37], and 800 using the NSL model [33, 34].

3.2 Classifying Genre by Auditory Temporal Modulations

Given a set $\mathcal{D}$ of vectorized ATM features, each associated with a single music genre, we can use the machinery of SRC to label an unknown vectorized ATM $\mathbf{y}$. Following [31], we first make all features of $\mathcal{D}$ have unit ℓ_2-norm, as well as the test feature $\mathbf{y}$. We next solve the BP optimization problem posed in [31]

$$\min_{\mathbf{a}\in\mathbb{R}^N} \|\mathbf{a}\|_1 \ \text{ subject to } \ \mathbf{\Phi y} = \mathbf{\Phi D a} \tag{21}$$

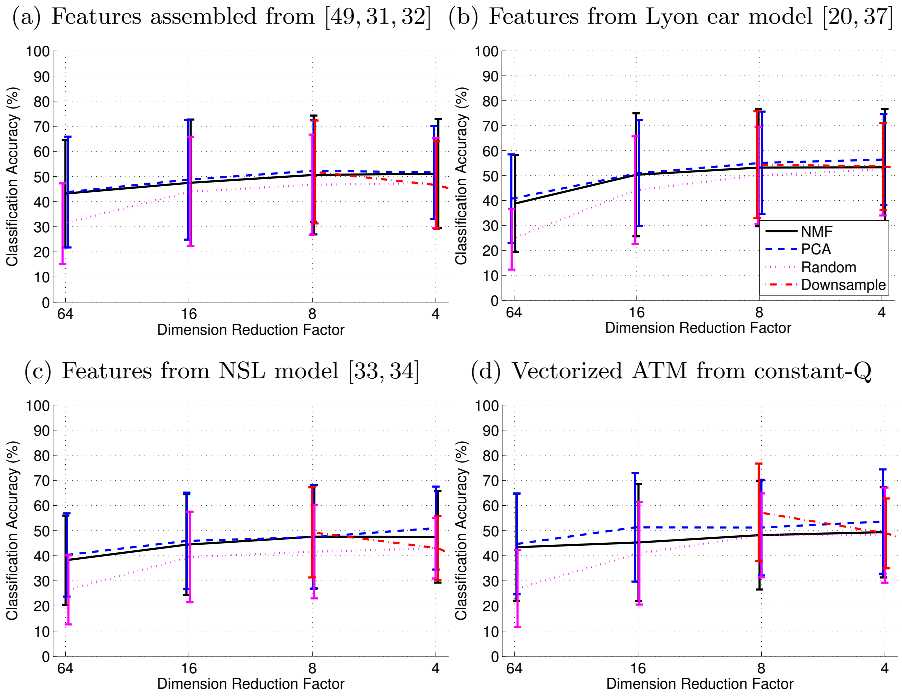
where $\mathbf{\Phi}$ reduces the features by, e.g., PCA. Finally, to classify $\mathbf{y}$, we construct the set of weights in (4), and assign a single genre label using the criterion (5).

Since we are working with real vectors, we can solve (21) as a linear program [6], for which numerous solvers have been implemented, e.g., [5, 10, 2, 14]. Because of its speed, we choose as the first step the root-finding method of the SPGL1 solver [2]. If this fails to find a solution, then we use the primal-dual method of ℓ_1-Magic [5], which takes as its starting point the minimum ℓ_2-norm solution $\mathbf{a}_2 := (\mathbf{\Phi D})^\dagger \mathbf{y}$. This initial solution satisfies the constraints of (21) as long as $\mathbf{\Phi D}$ has full rank, but probably is not the optimal solution. If the solution $\hat{\mathbf{a}}$ does not satisfy $\|\mathbf{\Phi y} - \mathbf{\Phi D}\hat{\mathbf{a}}\|_2^2 < 10^{-16}$ (numerical precision), we set $\hat{\mathbf{a}} := \mathbf{a}_2$.

4 Experimental Results

As in [31], we use the music genre dataset of [43] (GTZAN),[3] which has 1000 half-minute sound examples drawn from music in 10 broad genres: blues, classical, country, disco, hiphop, jazz, metal, pop, reggae, and rock. We define $\mathbf{\Phi}$ by PCA, NMF, or random sampling; and as in [31], we test dimension reduction by factors of $\{64, 16, 8, 4\}$, e.g., we reduce a feature vector of 768 dimensions by a factor of four to 192 dimensions. We also test downsampling the features, but we define it as vectorizing the result of lowpass filtering and decimating each column of the ATM (20). It is not clear how downsampling is done in [31]. In our case, a factor of f downsampling results in a vectorized feature of dimension $8\lceil 96/f\rceil$ when using our 96-channel features. Finally, as done in [3, 31], we use stratified 10-fold cross-validation for classifier training and testing.

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



(a) Features assembled from [49, 31, 32] (b) Features from Lyon ear model [20, 37]

(c) Features from NSL model [33, 34] (d) Vectorized ATM from constant-Q

Fig. 4. Mean classification accuracy (10 classes) of SRC based on (21) for four different feature design methods, four dimension reduction methods, and several reduction factors. Overlaid is the standard deviation. (We add a slight x-offset to each bar for readability.)

Figure 4 shows our classification results for four different features, including the vectorized modulation-analysis of the magnitude output of the constant-Q filterbank that precedes (15). Across all features and dimension reduction methods and factors, we see no mean accuracies above 57.3% — which is produced by using features that do not model the entire primary auditory cortex. Since we see all mean accuracies are within one standard deviation of each other, we cannot claim one feature, reduction method is performing significantly different from any other. This result has been observed before in the application of SRC to face recognition [47]. In the experimental results of [31], however, we see features reduced a factor of 4 by NMF give the best results: mean accuracy of around 91% with a standard deviation of 1.76%. From the plots in [31], we can surmise there to be a statistically significant (e.g., $\alpha := 0.05$) difference between the features reduction methods. This contradicts our results and those of [47], not to mention the significant difference between the best accuracies on this same dataset with the same experimental protocol.

We have verified every part of our system is working as expected. We have performed modulation analysis on synthetic signals with known modulations. We have tested and confirmed on a handwritten digits dataset [16] that the SRC classifier performs comparably to other classifiers, and that our feature reduction is working. In this context too, we find no signficant difference in performance

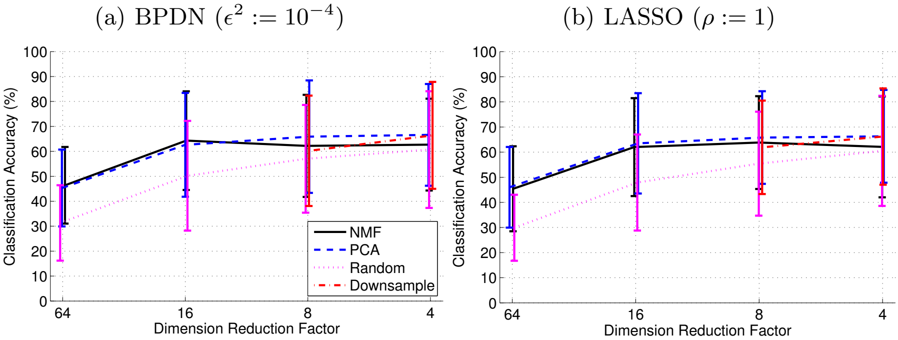


Fig. 5. Mean classification accuracy (10 classes) of SRC based on the BPDN (22) and LASSO (23), with features having dimensions mapped to $[0, 1]$, and a normalized projected dictionary, for ATM features from the Lyon model [20, 37], four dimension reduction methods, and several reduction factors. Overlaid is the standard deviation.

between feature reduction methods. From our experimentation, and conversation with the authors of [31], we believe that these differences come from several things, three of which are significant.

First, it is common in machine learning to preprocess features by accounting for dimensions with different scales. Panagakis et al. state that they make the values of each row of $\mathbf{D}$ be in $[0, 1]$ by finding and subtracting the minimum, and then dividing by the difference of the maximum and minimum.[4] When we rerun the experiments above with this modified data, we see the mean accuracy increases, but does not exceed the highest of 64% for the NSL features reduced in dimensionality a factor of 4 by PCA. Again, we see no significant difference between classifier performance with these features.

The second problem is posing the sparse representation with equality constraints in (21), which forces the sparse representation algorithm to model a feature exactly when instead we just want to find a good model of our feature. We thus pose the problem instead using BPDN [6] (7)

$$\min_{\mathbf{a} \in \mathbb{R}^N} \|\mathbf{a}\|_1 \text{ subject to } \|\mathbf{\Phi y} - \mathbf{\Phi D a}\|_2^2 \leq \epsilon^2. \tag{22}$$

or as the LASSO [41] (8)

$$\min_{\mathbf{a} \in \mathbb{R}^N} \|\mathbf{\Phi y} - \mathbf{\Phi D a}\|_2^2 \text{ subject to } \|\mathbf{a}\|_1 \leq \rho. \tag{23}$$

Solving these can produce an informative representation using few features instead of an exact fit by many.

Using features with dimensions mapped to $[0, 1]$, and a column-normalized dimension-reduced dictionary $\mathbf{\Phi D}$, Fig. 5(a) shows the results of using BPDN (22); and Fig. 5(b) shows the results when we pose the problem as the LASSO (23). (We show only the results from the Lyon model since the other features did not give significantly different results.) In both cases, we use SGPL1 [2] with

[4] Personal communication.

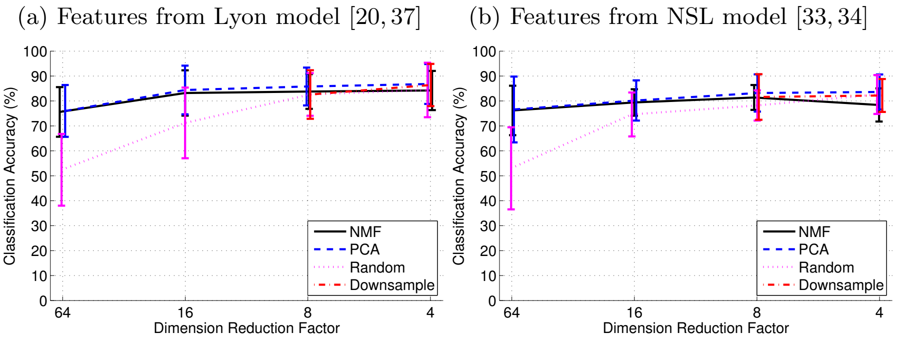

(a) Features from Lyon model [20, 37] (b) Features from NSL model [33, 34]

Fig. 6. Mean classification accuracy (5 classes) of SRC based on the LASSO (23) with $\rho := 1$, with features having dimensions mapped to $[0, 1]$ and a normalized projected dictionary, for ATM features derived from a larger frequency range, four dimension reduction methods, and several reduction factors. Overlaid is the standard deviation.

at most 100 iterations, and use the result whether it is in the feasible set or not. This is different from our approach to solving (21), where we run ℓ_1-Magic [5] if SPGL1 fails, and then use the minimum ℓ_2-norm solution if this too fails. In our experiments, we see (23) is solved nearly all the time for $\rho := 1$, and (22) is solved only about 5% of the time with $\epsilon^2 := 10^{-4}$; yet we see no significant differences between the accuracies of both cases. With these changes, we see a slight increase in mean accuracies to about 68% for the features derived from the Lyon model [20, 37], but still far from the 91% reported in [31].

The third significant problem comes from the definition of the features. We find that accuracy improves slightly if we use features from a wider frequency range than the four octaves mentioned in [31], e.g., all 86 bands of the AS from the Lyon model, covering 80 − 7630 Hz [20, 37], or all 128 bands of the AS from the NSL model [33, 34], logarithmically spread over 180 − 7246 Hz. With these changes, however, our mean accuracies do not exceed 70%.

The only way we have found to obtain something close to the 91% mean accuracy reported in [31] is to limit the classification problem to the first five genres of GTZAN: blues, classical, country, disco, and hiphop. Figure 6 shows our results using features derived from the Lyon and NSL models with a wide-frequency range, dimensions mapped to $[0, 1]$, and solving the problem posed with LASSO (23). Though we see the standard deviations are smaller, we still cannot say one feature reduction method performs signficantly different than any other, in contradiction to the findings of [31].

5 Conclusion

Were the difficult problem of music genre recognition solved, it would present a wonderful tool for exploring many interesting questions; and were it solved using solely acoustic features, it would say something significant about a process that appears influenced by much more than sound. Though the approach and results of [31] appear extremely promising in light of state of the art — it is based on a perceptually-informed acoustic feature and a classification method built upon sparse representations in exemplars, which has its own biological motivations, e.g., [19, 18] — we have not been able to reproduce their results without reducing the number of classes from 10 to 5. We have shown in as much detail possible the variety of decisions we have had to make in our work, and provide all our code for independent verification: `http://imi.aau.dk/~bst/software/`. Though our results point to a negative conclusion with regard to [31], we have confirmed the observation of [47] that the performance of SRC appears robust to the features used. We have found evidence that features modeled on the primary auditory cortex do not perform significantly different from a feature that is not perceptually based. Indeed, it does not make sense to us why perceptually-based features would be more discriminative for the recognition of genre. Finally, we have also shown that relaxing the constraints in the sparse representation component of SRC improves classification accuracy.

As a postscript, we have found in further work [39] that we can increase the mean accuracy of SRC with ATM in music genre recognition to 82% using the Lyon model and downsampling the AS by a factor of 40 (from 22, 050 Hz to 551 Hz) before performing the modulation analysis. This performance increase, however, appears irrelevant with respect to genre recognition. When we look beyond the summary statistics, we see this method confidently applies quite illogical classifications, e.g., "Why?" by Bronski Beat is supposedly Classical. We find that its results are highly sensitive to equalization of the audio, and it can be made to label the same piece of music differently if we shape the spectrum in minor ways. Furthermore, we find that the music this method claims is highly representative of a specific genre is not similarly labeled by a listener able to recognize the same genre. Thus, SRC with ATM appears to be choosing labels based on confounding factors of genre. Our future work aims at determining these factors.

Acknowledgments

B. L. Sturm is supported in part by Independent Postdoc Grant 11-105218 from Det Frie Forskningsrd. The authors would like to thank Dr. Costas Kotropoulos and Yannis Panagakis for the helpful discussions about their work.

References

1. Baumann, S., Pohle, T., Vembu, S.: Towards a socio-cultural compatibility of MIR systems. In: Proc. ISMIR. pp. 460–465. Barcelona, Spain (Oct 2004)
2. van den Berg, E., Friedlander, M.P.: Probing the pareto frontier for basis pursuit solutions. SIAM J. on Scientific Computing 31(2), 890–912 (Nov 2008)
3. Bergstra, J., Casagrande, N., Erhan, D., Eck, D., Kégl, B.: Aggregate features and adaboost for music classification. Machine Learning 65(2-3), 473–484 (June 2006)
4. Bingham, E., Mannila, H.: Random projection in dimensionality reduction: Application to image and text data. In: Proc. Int. Conf. Knowledge Discovery Data Mining. pp. 245–250. San Francisco, CA (Aug 2001)
5. Candès, E., Romberg, J.: ℓ_1-magic: Recovery of sparse signals via convex programming. Tech. rep., Caltech, Pasadena, CA, USA (2005)
6. Chen, S.S., Donoho, D.L., Saunders, M.A.: Atomic decomposition by basis pursuit. SIAM J. Sci. Comput. 20(1), 33–61 (Aug 1998)

7. Dasgupta, S.: Experiments with random projection. In: Proc. Conf. Uncertainty in Artificial Intelligence. pp. 143–151. Stanford, CA, USA (June 2000)
8. Davis, G., Mallat, S., Avellaneda, M.: Adaptive greedy approximations. J. Constr. Approx. 13(1), 57–98 (Jan 1997)
9. Fabbri, F.: A theory of musical genres: Two applications. In: Proc. First International Conference on Popular Music Studies. Amsterdam, The Netherlands (1980)
10. Figueiredo, M., Nowak, R., Wright, S.J.: Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. IEEE J. Sel. Topics Signal Process. 1(4), 586–597 (Dec 2007)
11. Gemmeke, J., ten Bosch, L., L.Boves, Cranen, B.: Using sparse representations for exemplar based continuous digit recognition. In: Proc. EUSIPCO. pp. 1755–1759. Glasgow, Scotland (Aug 2009)
12. Giacobello, D., Christensen, M., Murthi, M.N., Jensen, S.H., Moonen, M.: Enhancing sparsity in linear prediction of speech by iteratively reweighted ℓ_1-norm minimization. In: Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process. Dallas, TX (Mar 2010)
13. Gjerdingen, R.O., Perrott, D.: Scanning the dial: The rapid recognition of music genres. J. New Music Research 37(2), 93–100 (Spring 2008)
14. Grant, M., Boyd, S.: CVX: Matlab software for disciplined convex programming, version 1.21. `http://cvxr.com/cvx` (Apr 2011)
15. Greenberg, S., Kingsbury, B.E.D.: The modulation spectrogram: in pursuit of an invariant representation of speech. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. pp. 1647–1650. Munich, Germany (Apr 1997)
16. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE 86(11), 2278–2324 (Nov 1998)
17. Lena, J.C., Peterson, R.A.: Classification as culture: Types and trajectories of music genres. American Sociological Review 73, 697–718 (Oct 2008)
18. Lewicki, M.S.: Efficient coding of natural sounds. Nature Neuroscience 5(4), 356–363 (Mar 2002)
19. Lewicki, M.S., Sejnowski, T.J.: Learning overcomplete representations. Neural Computation 12, 337–365 (Feb 2000)
20. Lyon, R.F.: A computational model of filtering, detection, and compression in the cochlea. In: Proc. ICASSP. pp. 1282–1285 (1982)
21. Majumdar, A., Ward, R.K.: Robust classifiers for data reduced via random projections. IEEE Trans. Systems, Man, Cybernetics 40(5), 1359–1371 (Oct 2010)
22. Mayer, R., Neumayer, R., Rauber, A.: Rhyme and style features for musical genre classification by song lyrics. In: Proc. Int. Symp. Music Info. Retrieval (2008)
23. McKay, C., Fujinaga, I.: Music genre classification: Is it work pursuing and how can it be improved? In: Proc. Int. Symp. Music Info. Retrieval (2006)
24. Mesgarani, N., Slaney, M., Shamma, S.A.: Discrimination of speech from nonspeech based on multiscame spectro-temporal modulations. IEEE Trans. Audio, Speech, Lang. Process. 14(3), 920–930 (May 2006)
25. Mitra, S.K.: Digital Signal Processing: A Computer Based Approach. McGraw Hill, 3 edn. (2006)
26. Pachet, F., Cazaly, D.: A taxonomy of musical genres. In: Proc. Content-based Multimedia Information Access Conference. Paris, France (Apr 2000)
27. Pampalk, E., Flexer, A., Widmer, G.: Hierarchical organization and description of music collections at the artist level. In: Research and Advanced Technology for Digital Libraries. pp. 37–48 (2005)
28. Panagakis, Y., Benetos, E., Kotropoulos, C.: Music genre classification: A multilinear approach. In: Proc. ISMIR. pp. 583–588. Philadelphia, PA (Sep 2008)

29. Panagakis, Y., Kotropoulos, C.: Music genre classification via topology preserving non-negative tensor factorization and sparse representations. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. pp. 249–252. Dallas, TX (Mar 2010)
30. Panagakis, Y., Kotropoulos, C., Arce, G.R.: Music genre classification using locality preserving non-negative tensor factorization and sparse representations. In: Proc. Int. Symp. Music Info. Retrieval. pp. 249–254. Kobe, Japan (Oct 2009)
31. Panagakis, Y., Kotropoulos, C., Arce, G.R.: Music genre classification via sparse representations of auditory temporal modulations. In: Proc. European Signal Process. Conf. Glasgow, Scotland (Aug 2009)
32. Panagakis, Y., Kotropoulos, C., Arce, G.R.: Non-negative multilinear principal component analysis of auditory temporal modulations for music genre classification. IEEE Trans. Acoustics, Speech, Lang. Process. 18(3), 576–588 (Mar 2010)
33. Ru, P.: Cortical Representations and Speech Recognition. Ph.D. thesis, University of Maryland, College Park, MD, USA (Dec 1999)
34. Ru, P.: Multiscale multirate spectro-temporal auditory model. Tech. rep., Neural Systems Laboratory, University of Maryland College Park (2001), `http://www.isr.umd.edu/Labs/NSL/Software.htm`
35. Sainath, T.N., Carmi, A., Kanevsky, D., Ramabhadran, B.: Bayesian compressive sensing for phonetic classification. In: Proc. ICASSP (2010)
36. Shamma, S.A.: Encoding sound timbre in the auditory system. IETE J. Research 49(2), 145–156 (Mar-Apr 2003)
37. Slaney, M.: Auditory toolbox. Tech. rep., Interval Research Corporation (1998)
38. Sordo, M., Celma, O., Blech, M., Guaus, E.: The quest for musical genres: Do the experts and the wisdom of crowds agree? In: Proc. ISMIR (2008)
39. Sturm, B.L.: Three revealing experiments in music genre recognition. In: Proc. Int. Soc. Music Info. Retrieval. Porto, Portugal (Oct submitted 2012)
40. Sukittanon, S., Atlas, L.E., Pitton, J.W.: Modulation-scale analysis for content identification. IEEE Trans. Signal Process. 52(10), 3023–3035 (Oct 2004)
41. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. Royal Statist. Soc. B 58(1), 267–288 (Jan 1996)
42. Tropp, J.A., Wright, S.J.: Computational methods for sparse solution of linear inverse problems. Proc. IEEE 98(6), 948–958 (June 2010)
43. Tzanetakis, G., Cook, P.: Musical genre classification of audio signals. IEEE Trans. Speech Audio Process. 10(5), 293–302 (July 2002)
44. Wang, K., Shamma, S.A.: Spectral shape analysis in the central auditory system. IEEE Trans. Speech Audio Process. 3(5), 382–395 (Sep 1995)
45. Woolley, S.M.N., Fremouw, T.E., Hsu, A., Theunissen, F.E.: Tuning for spectro-temporal modulations as a mechanishm for auditory discrimination of natural sounds. Nature Neuroscience 8(10), 1371–1379 (Oct 2005)
46. Wright, J., Ma, Y., Mairal, J., Sapiro, G., Huang, T., Yan, S.: Sparse representation for computer vision and pattern recognition. Proc. IEEE 98(6), 1031–1044 (June 2009)
47. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust face recognition via sparse representation. IEEE Trans. Pattern Anal. Machine Intell. 31(2), 210–227 (Feb 2009)
48. Yang, A.Y., Ganesh, A., Zhou, Z., Sastry, S.S., Ma, Y.: A review of fast l1-minimization algorithms for robust face recognition. (preprint) (2010), `http://arxiv.org/abs/1007.3753`
49. Yang, X., Wang, K., Shamma, S.A.: Auditory representations of acoustic signals. IEEE Trans. Info. Theory 38(2), 824–839 (Mar 1992)