# MUSIC GENRE CLASSIFICATION VIA SPARSE REPRESENTATIONS OF AUDITORY TEMPORAL MODULATIONS

*Yannis Panagakis*[*], *Constantine Kotropoulos*[*†], *Gonzalo R. Arce*[†]

[*] Department of Informatics
Aristotle University of Thessaloniki
Box 451, Thessaloniki 54124, GREECE
email: {panagakis,costas}@aiia.csd.auth.gr

[†] Department of Electrical & Computer Engineering
University of Delaware
Newark, DE 19716-3130, U.S.A.
email: arce@ece.udel.edu

**ABSTRACT**

A robust music genre classification framework is proposed that combines the rich, psycho-physiologically grounded properties of slow temporal modulations of music recordings and the power of sparse representation-based classifiers. Linear subspace dimensionality reduction techniques are shown to play a crucial role within the framework under study. The proposed method yields a music genre classification accuracy of 91% and 93.56% on the GTZAN and the ISMIR2004 Genre dataset, respectively. Both accuracies outperform any reported accuracy ever obtained by state of the art music genre classification algorithms in the aforementioned datasets.

## 1. INTRODUCTION

Music genre is probably the most popular description of music content [2], although there is no a commonly agreed definition of music genre, since it depends on cultural, artistic, or market factors, and the boundaries between genres are fuzzy [27].

There is evidence that the audio signal carries information about genre [27, 29]. Most of the music genre classification algorithms resort to the so-called *bag-of-features* approach [27], which models the audio signals by their long-term statistical distribution of short-time features. Features commonly exploited for music genre classification can be roughly classified into timbral texture, rhythmic, pitch content ones, or their combinations [29]. Having extracted descriptive features, pattern recognition algorithms are employed for their classification into genres. Frequently used classifiers are the nearest-neighbor (NN), support vector machines (SVMs), or classifiers, which resort to Gaussian mixture models, linear discriminant analysis, non-negative matrix factorization (NMF), etc. Several common audio datasets have been used in experiments in order to make the reported classification accuracies comparable. Notable results on music genre classification are summarized in Table 1.

Psycho-physiological investigations indicate that the acoustic stimulus is encoded by the primary auditory cortex in terms of its spectral and temporal characteristics at various degrees of resolutions. This is accomplished by cells whose responses are selective to a range of spectral and temporal resolutions resulting into a neural representation [33]. In particular, when the acoustic stimulus is either speech or music, its perceptual properties are encoded by slow temporal modulations [28, 7, 9, 25, 24].

Recently, the interest on sparse representations of signals has revived [3]. The related research has been focused on two aspects of sparse representations: First, pursuit methods have been developed for solving the optimization problems which arise, such as the matching pursuit [18], the orthogonal matching pursuit [23], and the basis pursuit [6]. Second, overcomplete dictionaries have been derived, such as the K-SVD algorithm [1]. However, the aforementioned methods aim at representing the signals rather than classifying them. Furthermore, the dictionary atoms do not possess any particular semantic meaning as they are chosen from standard bases such as wavelet, curvelet, Gabor functions, etc. It is worth mentioning, that the sparsest representation is naturally discriminative. Indeed, among all subsets of basis vectors, the subset, which most compactly expresses the input signal, is selected and all other possible, but less compact representations, are rejected [17, 31].

Table 1: Notable classification accuracies achieved by music genre classification approaches.

| Reference | Dataset | Accuracy |
|---|---|---|
| Bergstra *et al.* [5] | GTZAN | 82.50% |
| Li *et al.* [14] | GTZAN | 78.50% |
| Panagakis *et al.* [22] | GTZAN | 78.20% |
| Lidy *et al.* [16] | GTZAN | 76.80% |
| Benetos *et al.* [4] | GTZAN | 75.00% |
| Holzapfel *et al.* [11] | GTZAN | 74.00% |
| Tzanetakis *et al.* [29] | GTZAN | 61.00% |
| Holzapfel *et al.* [11] | ISMIR2004 | 83.50% |
| Pampalk *et al.* [21] | ISMIR2004 | 82.30% |
| Panagakis *et al.* [22] | ISMIR2004 | 80.95% |
| Lidy *et al.* [15] | ISMIR2004 | 79.70% |
| Bergstra *et al.* [5] | MIREX2005 | 82.34% |
| Lidy *et al.* [16] | MIREX2007 | 75.57% |
| Mandel *et al.* [19] | MIREX2007 | 75.03% |

The appealing properties of slow temporal modulations from the human perceptual point of view and the strong theoretical foundations of sparse representations have motivated us to propose a robust framework for automatic music genre classification here. To this end, a bio-inspired auditory representation is extracted that maps a given music recording to a two-dimensional (2D) representation of its slow temporal modulations. Such a representation extends the concept of joint acoustic and modulation frequency analysis [28] by exploiting the properties of the human auditory system [9, 30] and is referred to as *auditory temporal modulation* representation. Second, these auditory temporal modulations form an overcomplete dictionary of basis signals for music genres, which is exploited for *sparse representation-based classification* (SRC) proposed in [31]. If sufficient training music recordings are available for each genre, it is possible to express any test representation of auditory temporal modulations as a compact linear combination of the dictionary atoms for the genre, where it belongs to. This representation is designed to be sparse, because it involves only a small fraction of the dictionary atoms and can be computed efficiently via $L_1$ optimization. The classification is performed by assigning each test recording to the class where the dictionary atoms, that are weighted by non-zero coefficients, belong to.

Since we are interested to build overcomplete dictionaries extracted from training representations of auditory temporal modulations, the dimensionality of such vectorized representations must be much smaller than the cardinality of the training set. Accordingly, we investigate several dimensionality reduction techniques, such as NMF [13], principal component analysis (PCA) [12], random pro-

jections [31], even downsampling, as in [31]. The features extracted by the aforementioned dimensionality techniques are next classified by SRC. Performance comparisons are made against SVMs with a linear kernel and a NN classifier, which employs the cosine similarity measure (CSM). The reported genre classification rates are juxtaposed against those achieved by the algorithms listed in Table 1 for the GTZAN and the ISMIR2004 Genre datasets. More specifically, two sets of experiments are conducted. First, stratified ten-fold cross-validation is applied to the GTZAN dataset. The proposed genre classification method yields an accuracy of 91%. Second, experiments on the ISMIR2004Genre dataset are conducted by adhering to the protocol employed during ISMIR2004 evaluation tests, which splits the dataset into two equal disjoint subsets with the first one being used for training and the second one being used for testing. The proposed genre classification method yields an accuracy of 93.56%. To the best of the authors' knowledge, the achieved classification accuracy is **the highest ever reported for both datasets**.

The remainder of the paper is as follows. In Section 2, the bio-inspired auditory representation based on a computational auditory model is briefly described. The SRC framework, that is applied to music genre classification, is detailed in Section 3. Experimental results are demonstrated in Section 4. Conclusions are drawn and future research direction are indicated in Section 5.

## 2. BIO-INSPIRED JOINT ACOUSTIC AND MODULATION FREQUENCY REPRESENTATION OF MUSIC

A key step for representing music signals in a psychophysiologically consistent manner is to focus on how the audio information is encoded in the human *primary auditory cortex*. The primary auditory cortex is the first stage of the central auditory system, where higher level mental processes take place, such as perception and cognition [20]. In this section, we briefly describe how a 2D representation of auditory temporal modulations can be obtained by modeling the path of auditory processing. The auditory representation is a joint acoustic and modulation frequency representation [28], that discards much of the spectro-temporal details and focuses on the underlying slow temporal modulations of the music signal. There is evidence that such a representation carries important time-varying information [28, 7, 9, 25, 24].

The computational model of human auditory system consists of two basic processing stages. The first stage models the early auditory system, which converts the acoustic signal into a neural representation, the so-called *auditory spectrogram*, i.e. a time-frequency distribution along a tonotopic (logarithmic frequency) axis. At the second stage, the temporal modulation content of the auditory spectrogram is estimated by wavelets applied to each row of the auditory spectrogram.

The computation of the auditory spectrogram consists of three operations, which mimic the early stages of human auditory processing. In this paper, the mathematical model of Yang *et. al* [32] is adopted. Initially, a constant-$Q$ transform is applied to the acoustic signal $s(t)$. That is, a bank of filters, such that the ratio of each filter center frequency to its resolution is kept constant. Here, the constant-$Q$ transform is implemented via a bank of 96 overlapping bandpass filters with center frequencies uniformly distributed along the tonotopic axis over 4 octaves. First, the output of cochlear filter is transduced into an auditory nerve pattern by a hair cell stage, which converts the cochlear output into inner hair cell intracellular potential. The just described process is modelled by highpass filtering corresponding to the fluid-cilia coupling, followed by an instantaneous nonlinear compression, which models the gated ionic channels, and finally lowpass filtering that models the hair cell membrane leakage. At a second step, a lateral inhibitory network (LIN) detects the discontinuities in the response along the tonotopic axis of the auditory nerve array. LIN can be approximated by a first-order derivative with respect to the logarithmic frequency followed by a half-wave rectifier. Next, the output of LIN is integrated over a short decaying exponential window with time constant 8 ms, that accounts for the further loss of phase-locking observed in the midbrain.

Higher central auditory stages, especially the primary auditory cortex, further analyze the auditory spectrogram by estimating the signal content in slow spectro-temporal modulations. In this paper, we are interested in the slow temporal modulations only present in the auditory spectrogram. In order to mimic the human perception of temporal modulation, we apply the concept of *modulation scale analysis* [28] in order to derive a compact representation that captures the underlying temporal modulations of an audio signal.

The modulation scale analysis consist of two stages. First, for discrete rate $r$, a wavelet filter is applied along each temporal row of the auditory spectrogram. This operation can be interpreted as filtering the temporal envelope of each cochlear channel output. For each audio frame, the multiresolution wavelet analysis is implemented via a bank of Gabor filters, that are selective to different temporal modulation parameters ranging from slow to fast temporal rates (in Hz). Since, the analysis yields a frequency-rate-time representation for each frame, the entire auditory spectrogram is modeled by a three-dimensional (3D) representation of frequency, rate, and time. Finally, the power of the 3D temporal modulation representation is obtained by integrating across the wavelet translation axis. Thus a joint frequency-rate representation results that has no uniform resolution in the modulation frequency indexed by the discrete rate. The resulting 2D representation is the *auditory temporal modulation*. Psychophysiological evidence [26] justifies the choice of $r \in \{2,4,8,16,32,64,128,256\}$ (Hz) to represent the temporal modulation content of sound. The cochlear model employed in the first stage, has 96 filters covering 4 octaves along the tonotopic axis (i.e. 24 filters per octave). Accordingly, the auditory temporal modulation of a music recording is naturally represented by a second-order tensor (matrix) $\tilde{\mathbf{X}} \in \mathbb{R}_+^{I_1 \times I_2}$, where $I_1 = I_f = 96$ and $I_2 = I_r = 8$. Hereafter, let $\mathbf{x} = \mathrm{vec}(\tilde{\mathbf{X}}) \in \mathbb{R}_+^{I_1 \cdot I_2} = \mathbb{R}_+^{768}$ denote the lexicographically ordered vectorial representation of the auditory temporal modulation.

The auditory temporal modulation representation is computed for each audio recording in the dataset. By vectorizing each representation, an ensemble of music recordings can be represented by the matrix $\mathbf{X} \in \mathbb{R}_+^{768 \times samples}$, where *samples* indicates the total number of the dataset recordings.

## 3. SPARSE REPRESENTATION-BASED CLASSIFICATION

The problem of determining the class label of a test auditory temporal modulation representation, given a number of labeled training temporal modulations representations from $N$ music genres is addressed based on SRC [31].

Let us denote by $\mathbf{A}_i = [\mathbf{a}_{i,1}|\mathbf{a}_{i,2}|\dots|\mathbf{a}_{i,n_i}] \in \mathbb{R}_+^{768 \times n_i}$ the dictionary that contains $n_i$ auditory modulation representations stemming from the $i$th genre as column vectors (i.e., atoms). Given a test auditory representation $\mathbf{y} \in \mathbb{R}_+^{768}$ that belongs to the $i$th class, we can assume that $\mathbf{y}$ is expressed as a linear combination of the atoms that belong to the $i$th class, i.e.

$$\mathbf{y} = \sum_{j=1}^{n_i} \mathbf{a}_{i,j} c_{i,j} = \mathbf{A}_i \mathbf{c}_i \tag{1}$$

where $c_{i,j} \in \mathbb{R}$ are coefficients, which form the coefficient vector $\mathbf{c}_i = [c_{i,1}, c_{i,2}, \dots, c_{i,n_i}]^T$.

Let us, now, define the matrix $\mathbf{A} = [\mathbf{A}_1|\mathbf{A}_2|\dots|\mathbf{A}_N] \in \mathbb{R}_+^{768 \times n}$ by concatenating the $n$ auditory modulation representations, which stem from $N$ genres. Thus the linear representation of the test auditory representation $\mathbf{y}$ in (1) can be equivalently rewritten as

$$\mathbf{y} = \mathbf{A}\,\mathbf{c} \tag{2}$$

where $\mathbf{c} = [\mathbf{0}^T|\dots|\mathbf{0}^T|\mathbf{c}_i^T|\mathbf{0}^T|\dots|\mathbf{0}^T]^T$ is the augmented coefficient vector whose elements are zero except those associated with the $i$th

genre. Thus, the entries of $\mathbf{c}$ contain information about the genre the test auditory representation $\mathbf{y}$ belongs to.

Since the genre label of any test auditory representation is unknown, we can predict it by seeking the sparsest solution to the linear system of equations $\mathbf{y} = \mathbf{A}\,\mathbf{c}$. More formally, given the matrix $\mathbf{A}$ and the test auditory representation $\mathbf{y}$, the problem of sparse representation is to find the coefficient vector $\mathbf{c}$ such that $\mathbf{y} = \mathbf{A}\,\mathbf{c}$ and $||\mathbf{c}||_0$ is to minimized, i.e.

$$\mathbf{c}^* = \arg\min_{\mathbf{c}} ||\mathbf{c}||_0 \quad \text{subject to } \mathbf{A}\mathbf{c} = \mathbf{y} \tag{3}$$

where $||.||_0$ is the $L_0$ quasi-norm returning the number of the non-zero entries of a vector. Finding the solution to optimization problem defined in (3) is NP-hard due to the nature of the underlying combinational optimization. An approximate solution to the problem (3) can be obtained by replacing the $L_0$ norm with the $L_1$ norm as follows:

$$\mathbf{c}^* = \arg\min_{\mathbf{c}} ||\mathbf{c}||_1 \quad \text{subject to } \mathbf{A}\,\mathbf{c} = \mathbf{y} \tag{4}$$

where $||.||_1$ denotes the $L_1$ norm of a vector. In [8], it has been proved that if the solution is sparse enough, then the solution of (3) is equivalent to the solution of (4), which can be solved in polynomial time by standard linear programming methods [6].

Since, we are interested to build overcomplete dictionaries derived from the auditory temporal modulation representations, the dimensionality of atoms must be much smaller than the training set cardinality. Thus, we can reformulate the optimization problem in (4) as follows:

$$\mathbf{c}^* = \arg\min_{\mathbf{c}} ||\mathbf{c}||_1 \quad \text{subject to } \mathbf{W}^T\,\mathbf{A}\,\mathbf{c} = \mathbf{W}^T\mathbf{y} \tag{5}$$

where $\mathbf{W} \in \mathbb{R}^{768\times k}$ with $k << \min(768, n)$ is a projection matrix. The projection matrix $\mathbf{W}$ can be obtained by any linear dimensionality reduction technique, such as NMF [13], PCA, a random projection matrix whose entries are independently sampled from a zero-mean normal distribution, and each column is normalized to unit length or even downsampling as proposed in [31]. The dimensionality reduction of the original auditory modulation space has two benefits: first it reduces the computational cost of linear programming solvers [6] of (4) and second it facilitates the creation of a redundant dictionary based on the training auditory temporal modulation representations.

A test auditory modulation can be classified as follows. First, $\mathbf{y}$ is projected onto the reduced dimensionality space through the projection matrix $\mathbf{W}$ as $\hat{\mathbf{y}} = \mathbf{W}^T\,\mathbf{y}$. Then the following optimization problem is solved

$$\mathbf{c}^* = \arg\min_{\mathbf{c}} ||\mathbf{c}||_1 \quad \text{subject to } \mathbf{W}^T\,\mathbf{A}\,\mathbf{c} = \hat{\mathbf{y}}. \tag{6}$$

Ideally, the coefficient vector $\mathbf{c}^*$ contains non-zero entries in positions associated with the columns of $\mathbf{W}^T\mathbf{A}$ stemming from a single genre, so that we can easily assign the test auditory representation $\mathbf{y}$ to that genre. However, due to modeling errors, there are small non-zero entries in $\mathbf{c}^*$ that are associated to multiple genres. To cope with this problem, each auditory modulation representation is classified to the genre that minimizes the $L_2$ norm residual between $\hat{\mathbf{y}}$ and $\check{\mathbf{y}} = \mathbf{W}^T\,\mathbf{A}\,\delta_i(\mathbf{c})$, where $\delta_i(\mathbf{c}) \in \mathbb{R}^n$ is a new vector whose nonzero entries are only the entries in $\mathbf{c}$ that are associated to the $i$th genre [31].

## 4. EXPERIMENTAL EVALUATION

In order to assess the discriminating power of both the auditory temporal modulations and SRC, experiments are conducted on the two publicly available datasets, which are widely used for music genre classification [5, 11, 14, 15, 21, 29]. The first dataset, abbreviated as GTZAN, was collected by G. Tzanetakis [29] and consists of 10 genre classes, namely Blues, Classical, Country, Disco, HipHop, Jazz, Metal, Pop, Reggae, Rock. Each genre class contains 100 audio recordings 30 sec long. The second dataset, abbreviated as ISMIR2004 Genre, comes from the ISMIR 2004 Genre classification contest and contains 1458 full audio recordings distributed over six genre classes as follows: Classical (640), Electronic (229), Jazz-Blues (52), MetalPunk (90), RockPop (203), World (244), where the number within parentheses refers to the number of recordings which belong to each genre class.

All the audio recordings were converted to monaural wave format at a sampling frequency of 16 kHz and quantized with 16 bits. Moreover, the audio signals have been normalized, so that they have zero mean amplitude with unit variance in order to remove any factors related to the recording conditions. Since the ISMIR2004 Genre dataset, consists of full length tracks, we extracted a segment of 30 sec just after the first 30 sec of a recording to exclude any introductory parts that may not be directly related to the music genre the recording belongs to. The auditory temporal modulations representation is computed over a segment of 30 sec duration for any recording of both datasets.

Following the experimental set-up used in [29, 16, 14, 22], stratified 10-fold cross-validation is employed for experiments conducted on the GTZAN dataset. Thus each training set consists of 900 audio files. Thus a training matrix $\mathbf{A}_{GTZAN} \in \mathbb{R}_+^{768\times 900}$ is constructed by vectorizing each auditory temporal modulations representation associated to the training set. The experiments on the ISMIR 2004 Genre dataset were conducted according to the ISMIR2004 Audio Description Contest protocol. The protocol defines training and evaluation sets, which consist of 729 audio files each. Thus the corresponding training matrix $\mathbf{A}_{ISMIR} \in \mathbb{R}_+^{768\times 729}$ is constructed is constructed by vectorizing each auditory temporal modulations representation associated to the training set. Each column of each training matrix was normalized to unit length.

The projection matrix $\mathbf{W} \in \mathbb{R}^{768\times k}$ is derived from each training matrix $\mathbf{A}_{GTZAN}$ and $\mathbf{A}_{ISMIR}$ by employing either NMF or PCA with $k \in \{12, 48, 85, 192\}$, which corresponds to downsample ratios 1/8, 1/4, 1/3, and 1/2 respectively. The same values of parameter $k$ are used in order to construct the random projection matrix. Since the low dimensional feature space obtained by the aforementioned dimensionality reduction algorithms is linear, SVMs with linear kernel and NN with CSM will be used as alternatives to SRC.

In Figure 1, the classification accuracy achieved by the three different classifiers is plotted as a function of the feature space dimension, when the various subspace analysis methods are applied to both the GTZAN and the ISMIR 2004 Genre datasets. On the GTZAN dataset the best classification accuracy (91.0 %) was obtained when NMF extracts features, that are classified by SRC. The standard deviation of the classification accuracy was estimated thanks to 10-fold cross-validation. At the best classification accuracy, its standard deviation was found to be 1.76%. The reported classification accuracy outperforms those listed in Table 1. The interval $\pm$ one standard deviation is overlaid in all plots for the various dimensions $k$. On the ISMIR 2004 Genre dataset the best classification accuracy (93.56%) was obtained, when PCA extracts features that are classified by SRC. The confidence interval for the best classification accuracy on the ISMIR 2004 Genre dataset can be estimated as $\pm z_{1-\gamma/2}\sqrt{\frac{p(1-p)}{n}}$, where $z_{1-\gamma/2}$ is the standard Gaussian percentile for confidence level 100 $(1-\gamma)$% (e.g. for $\gamma = 0.05$, $z_{1-\gamma/2} = z_{0.975}$=1.967), $p = 0.9356$ is the experimentally measured classification accuracy, and $n = 729$ is the number of test recordings. The 95% confidence interval is then 1.79%. Again, the achieved classification accuracy outperforms all previously reported rates shown in Table 1.

The experimental results reported in this paper indicate that the dimensionality reduction is crucial, when SRC is applied to music genre classification. This was not the case with face recognition in [31], where the classification accuracy achieved by SRC was independent of the dimensionality reduction algorithm used. The de-

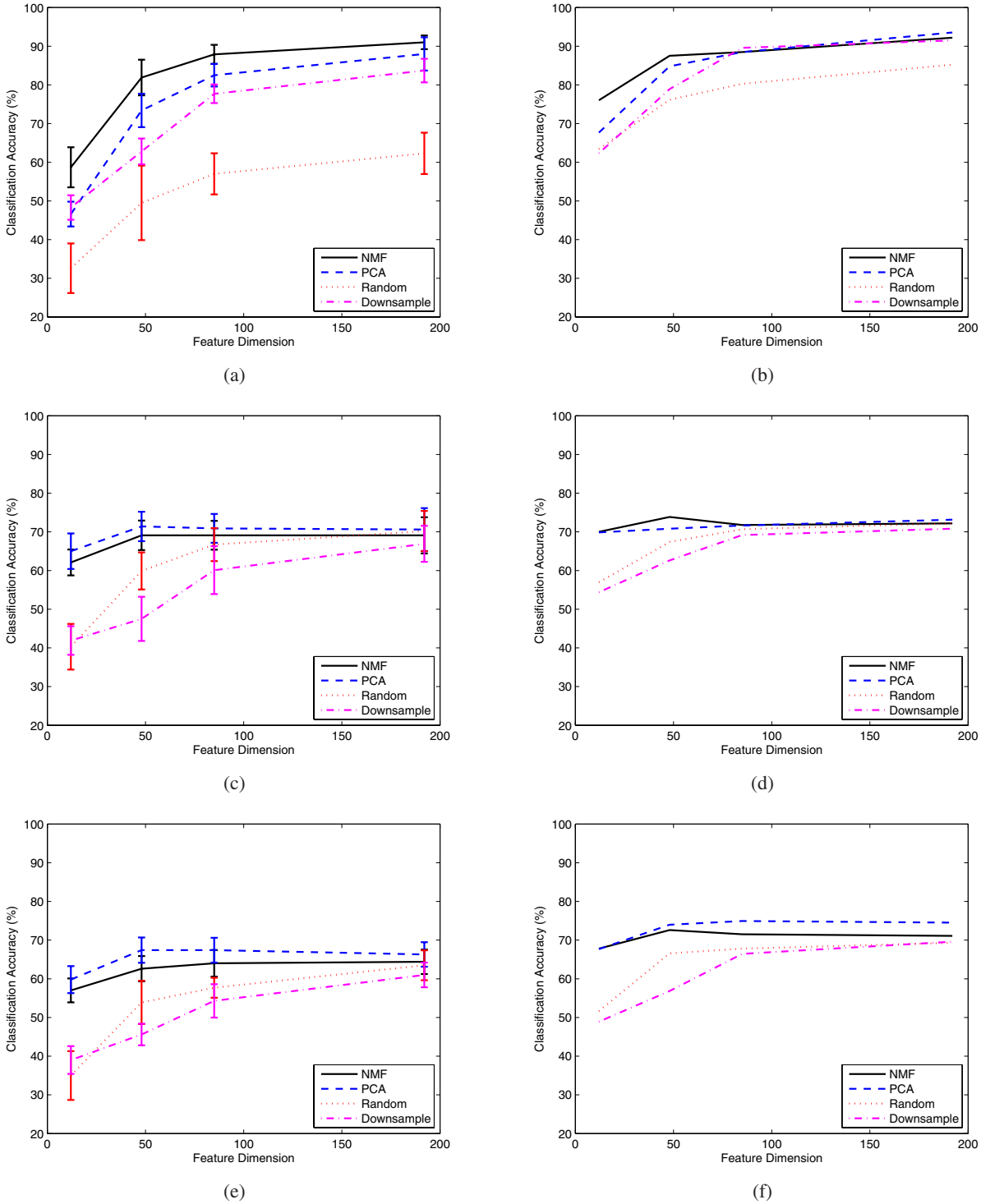


Figure 1: Classification accuracy for various features extraction methods and classifiers. (a) SRC on the GTZAN dataset; (b) SRC on the ISMIR2004 Genre dataset; (c) Linear SVM on the GTZAN dataset; (d) Linear SVM on the ISMIR2004 Genre dataset; (e) NN on the GTZAN dataset; (f) NN on the ISMIR2004 Genre dataset.

pendence of SRC on the dimensionality reduction technique used could be a point of future research.

## 5. CONCLUSIONS - FUTURE WORK

In this paper, a robust music genre classification framework has been proposed by considering the properties of the auditory human

perception. 2D auditory temporal modulations are used for music representation, while the sparse representation-based classification has been employed for genre classification. The crucial role of feature extraction and particularly dimensionality reduction for music genre classification has been demonstrated. The best classification accuracies measured in this paper outperform any rate ever reported for state of the art music genre classification algorithms applied to both the GTZAN and the ISMIR2004 Genre datasets.

In many real applications, both commercial and private, the number of available audio recordings per genre is limited. Thus, it is desirable the music genre classification algorithm to perform well in such small sample sets. Future research will address the performance of SRC framework under such conditions.

## REFERENCES


[1] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Processing*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[2] J. J. Aucouturier and F. Pachet, "Representing musical genre: A state of the art," *Journal of New Music Research*, pp. 83–93, 2003.

[3] R. G. Baraniuk, E. Candes, R. Nowak, and M. Vetterli (Guest Eds.), Special issue on Sensing, Sampling, and Compression. *IEEE Signal Processing Magazine*, vol. 25, no. 2, Mar. 2008.

[4] E. Benetos and C. Kotropoulos, "A tensor-based approach for automatic music genre classification," in *Proc. EUSIPCO 2008*, Lausanne, Switzerland, 2008.

[5] J. Bergstra, N. Casagrande, D. Erhan, D. Eck, and B. Kegl, "Aggregate features and Adaboost for music classification," *Machine Learning*, vol. 65, no. 2–3, pp. 473–484, 2006.

[6] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, no.1 pp. 33–61, 1998.

[7] T. Chi, Y. Gao, M. C. Guyton, P. Ru, and S. Shamma, "Spectro-temporal modulation transfer function and speech intelligibility," *Journal of the Acoustical Society of America*, no. 5, pp. 2719–2732, Nov. 1999.

[8] D. L. Donoho, and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Trans. Information Theory*, vol. 47, no. 7. pp. 2845–2862, 2001.

[9] S. D. Ewert and T. Dau,"Characterizing frequency selectivity for envelope fluctuations," *Journal of the Acoustical Society of America*, vol. 108, pp. 1181–1196, 2000.

[10] S. Greenberg, E. D. Brian, and Y. Kingsbury, "The modulation spectrogram: In pursuit of an invariant representation of speech," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 1647–1650, 1997.

[11] A. Holzapfel and Y. Stylianou, "Musical genre classification using nonnegative matrix factorization-based features," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 424–434, Feb. 2008.

[12] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology*, vol. 24, 417–441, 498–520.

[13] D. D. Lee and H. S. Seung, "Algoritnms for non-negative matrix factorization", in *Advances in Neural Information Processing Systems*, vol. 13, pp. 556–562, 2001.

[14] T. Li, M. Ogihara, and Q. Li, "A comparative study on content-based music genre classification," in *Proc. 26th Int. ACM SIGIR Conf. Research and Development in Informaion Retrieval*, Toronto, Canada, 2003, pp. 282–289.

[15] T. Lidy and A. Rauber, "Evaluation of feature extractors and psycho-acoustic transformations for music genre classification," in *Proc. 6th Int. Symposium Music Information Retrieval*, London, UK, 2005.

[16] T. Lidy, A. Rauber, A. Pertusa, and J. Inesta, "Combining audio and symbolic descriptors for music classification from audio," " in *Music Information Retrieval Information Exchange (MIREX)*, 2007.

[17] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Discriminative learned dictionaries for local image analysis," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, 2008.

[18] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. on Signal Processing*, vol. 41, pp. 3397–3415, 1993.

[19] M. Mandel and D. Ellis, "LABROSA's audio music similarity and classification submissions," in *Music Information Retrieval Information Exchange (MIREX)*, 2007.

[20] R. Munkong and J. Biing-Hwang, "Auditory perception and cognition," *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 98–117, May 2008.

[21] E. Pampalk, A. Flexer, and G., Widmer, "Improvements of audio-based music similarity and genre classification," in *Proc. 6th Int. Symp. Music Information Retrieval*, London, UK, 2005.

[22] I. Panagakis, E. Benetos, and C. Kotropoulos, "Music genre classification: A multilinear approach," in *Proc. 9th Int. Symp. Music Information Retrieval*, Philadelphia, USA, 2008.

[23] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition, " in *Proc. 27th Annual Asilomar Conf. on Signal, Systems, and Computers*, 1993.

[24] N. C. Singh and F. E. Theunissen, "Modulation spectra of natural sounds and ethological theories of auditory processing," *Journal of the Acoustical Society of America*, vol. 114, no. 6, pp. 3394–3411, 2003.

[25] M. Slaney and R. F. Lyon, "On the importance of time-a temporal representation of sound," in *Visual Representations of Speech Signals* (M. Cooke, S. Beet, and M. Crawford, Eds.), New York, NY: J. Wiley & Sons, 1993. pp. 95–116.

[26] S. A. Shamma, "Encoding sound timbre in the auditory system," *IETE Journal of Research*, vol. 49, no. 2–3, pp.145–156, 2003.

[27] N. Scaringella, G. Zoia, and D. Mlynek, "Automatic genre classification of music content: A survey," *IEEE Signal Processing Magazine*, vol. 23, no.2, pp. 133–141, Mar. 2006.

[28] S. Sukittanon, L. E. Atlas, and J. W Pitton, "Modulation-scale analysis for content identification," *IEEE Trans. Signal Processing*, vol. 52, no. 10, pp. 3023–3035, Oct. 2004.

[29] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech and Audio Processing*, vol. 10, no. 5, pp 293–302, Jul. 2002.

[30] S. Woolley, T. Fremouw, A. Hsu, and F. Theunissen, "Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds", *Nature Neuroscience*, vol. 8, no. 10, pp. 1371–1379, 2005.

[31] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Yi Ma "Robust Face recognition via sparse representation," *IEEE Trans. Pattern Analysis and Machine Intelligence* vol. 31, no. 2, pp. 210–227, Feb. 2009.

[32] X. Yang, K. Wang, and S. A. Shamma, "Auditory representations of acoustic signals," *IEEE Trans. Information Theory*, vol. 38, no. 2, pp. 824–839, Mar. 1992.

[33] D. N. Zotkin, T. Chi, S. A. Shamma, and R. Duraiswami, "Neuromimetic sound representation for percept detection and manipulation," *EURASIP Journal on Applied Signal Processing*, vol. 2005, no. 9, pp. 1350–1364, 2005.