# MUSIC GENRE CLASSIFICATION USING LOCALITY PRESERVING NON-NEGATIVE TENSOR FACTORIZATION AND SPARSE REPRESENTATIONS

**Yannis Panagakis**[*] **Constantine Kotropoulos**[*,†]
[*]Dept. of Informatics
Aristotle University of Thessaloniki
Box 451 Thessaloniki, GR-54124, Greece
{panagakis,costas}@aiia.csd.auth.gr

**Gonzalo R. Arce**[†]
[†]Dept. of Electrical & Computer Engineering
University of Delaware
Newark, DE 19716-3130, U.S.A.
arce@ece.udel.edu

## ABSTRACT

A robust music genre classification framework is proposed that combines the rich, psycho-physiologically grounded properties of auditory cortical representations of music recordings and the power of sparse representation-based classifiers. A novel multilinear subspace analysis method that incorporates the underlying geometrical structure of the cortical representations space into non-negative tensor factorization is proposed for dimensionality reduction compatible to the working principle of sparse representation-based classification. The proposed method is referred to as *Locality Preserving Non-Negative Tensor Factorization* (LPNTF). Dimensionality reduction is shown to play a crucial role within the classification framework under study. Music genre classification accuracy of 92.4% and 94.38% on the GTZAN and the ISMIR2004 Genre datasets is reported, respectively. Both accuracies outperform any accuracy ever reported for state of the art music genre classification algorithms applied to the aforementioned datasets.

## 1. INTRODUCTION

Despite the lack of a commonly agreed definition of music genre due to genre dependence on cultural, artistic, or market factors and the rather fuzzy boundaries between different genres, music genre is probably the most popular description of music content [1].

Psycho-physiology indicates that the acoustic stimulus is encoded in the primary auditory cortex by its spectral and temporal characteristics. This is accomplished by cells whose responses are selective to a range of spectral and temporal resolutions resulting into a neural representation. In particular, when the acoustic stimulus is either speech or music, its perceptual properties are encoded by slow spectral and temporal modulations [13, 18].

The appealing properties of slow spectro-temporal modulations from the human perceptual point of view and the strong theoretical foundations of sparse representations [4, 6] have motivated us to propose a robust framework for automatic music genre classification here. To this end, the auditory model [17] is used in order to map a given music recording to a three-dimensional (3D) representation of its slow spectral and temporal modulations with the same parameters as in [15]. This 3D representation is referred to as *cortical representation* and exploits the properties of the human auditory system [18]. The cortical representations form an overcomplete dictionary of basis signals for music genres, which is exploited for *sparse representation-based classification* (SRC) as proposed in [19]. That is, first each music recording is represented by its cortical representation. Second, each cortical representation is modeled as a sparse weighted sum of the basis elements (atoms) of an overcomplete dictionary, which stems from the cortical representations associated to training music recordings whose genre is known. If sufficient training music recordings are available for each genre, it is possible to express any test cortical representation as a compact linear combination of the dictionary atoms of the genre, where it actually 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 $\ell_1$ optimization. The classification is performed by assigning each test recording to the class associated with the dictionary atoms, that are weighted by non-zero coefficients.

Since we would like to build an overcomplete dictionary extracted from training cortical representations, the dimensionality of dictionary atoms must be much smaller than the cardinality of the training set. Such a dimensionality reduction facilitates the treatment of missing data, noise, and outliers. Conventional linear subspace analysis methods, such as Principal Component Analysis, Linear Discriminant Analysis, and Non-Negative Matrix Factorization (NMF) deal only with vectorial data. By vectorizing a typical 3D cortical representation of 6 scales, 10 rates, and 128 frequency bands, a vector of dimensions $7680 \times 1$ results. Handling such high-dimensional patterns is computationally expensive not to mention that eigenanalysis cannot be easily performed. Despite the implementation issues, by reshaping a 3D cortical representation into a vector the natural structure of the original data

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page.
© 2009 International Society for Music Information Retrieval.

is destroyed. Thus, dimensionality reduction applied directly to tensors rather than their vectorized versions is desirable. Unsupervised multilinear dimensionality reduction techniques, such as Non-Negative Tensor Factorization (NTF) [2] or Multilinear Principal Component Analysis (MPCA) [12] as well as supervised ones including General Tensor Discriminant Analysis (GTDA) [20] or Discriminant Non-Negative Tensor Factorization (DNTF) [21] could be considered. However, the just mentioned methods do not take into account the geometrical structure of the original data space. To reduce tensor dimensions in a consistent manner with the working principle of SRC, we should guarantee that two data points, which are close in the intrinsic geometry of the original data space are also close in the new data space after multilinear dimensionality reduction. To this end, we propose a novel algorithm, where the geometrical information of the original data space is incorporated into the objective function optimized by NTF. In particular, we encode the geometrical information by constructing a nearest neighbor graph. Furthermore, the non-negativity of cortical representations is preserved to maintain their physical interpretation. The proposed method is referred to as Locality Preserving Non-Negative Tensor Factorization (LPNTF). We derive a multiplicative updating algorithm for LPNTF, which extracts features from the cortical representations. For comparison purposes, NTF, MPCA, GTDA, DNTF, and random projections are also tested.

Next, the features extracted by the aforementioned multilinear dimensionality techniques are classified by SRC. Performance comparisons are made against the SVMs employing a linear kernel. The reported genre classification accuracies are juxtaposed against the best ones achieved by the state of the art algorithms applied to the GTZAN and 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, that extracts features using the LPNTF, which are then classified by SRC (i.e. LPNTF plus SRC), yields an accuracy of 92.4%. Second, experiments on the ISMIR2004Genre dataset are conducted by adhering to the protocol employed during ISMIR2004 evaluation tests. This protocol splits the dataset into two equal disjoint subsets with the first one being used for training and the second one being used for testing. Features extracted by NTF, which are then classified by SRC, yield an accuracy of 94.38%. An accuracy of 94.25% was achieved when the LPNTF plus SRC framework is employed. To the best of the authors' knowledge, the just quoted genre classification accuracies are **the highest ever reported for both datasets**.

The paper is organized as follows. In Section 2, basic multilinear algebra concepts and notations are defined. The LPNTF is introduced in Section 3. The SRC framework, that is applied to music genre classification, is detailed in Section 4. Experimental results are demonstrated in Section 5 and conclusions are drawn in Section 6.

## 2. NOTATION AND MULTILINEAR ALGEBRA BASICS

Tensors are considered as the multidimensional equivalent of matrices (i.e., second-order tensors) and vectors (i.e., first-order tensors) [9]. Throughout this paper, tensors are denoted by boldface Euler script calligraphic letters (e.g. $\mathbf{\mathcal{X}}$, $\mathbf{\mathcal{A}}$), matrices are denoted by uppercase boldface letters (e.g. $\mathbf{U}$), and vectors are denoted by lowercase boldface letters (e.g. $\mathbf{u}$).

A high-order real valued tensor $\mathbf{\mathcal{X}}$ of order $N$ is defined over the tensor space $\mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$, where $I_i \in \mathbb{Z}$ and $i = 1, 2, \ldots, N$. Each element of tensor $\mathbf{\mathcal{X}}$ is addressed by $N$ indices, i.e. $x_{i_1 i_2\ i_3 \ldots i_N}$. Mode-$n$ unfolding of tensor $\mathbf{\mathcal{X}}$ yields the matrix $\mathbf{X}_{(n)} \in \mathbb{R}^{I_n \times (I_1\ \ldots I_{n-1} I_{n+1} \ldots I_N)}$. In the following, the operations on tensors are expressed in matricized form [9].

An $N$-order tensor $\mathbf{\mathcal{X}}$ has rank 1, when it is decomposed as the outer product of $N$ vectors $\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \ldots, \mathbf{u}^{(N)}$, i.e. $\mathbf{\mathcal{X}} = \mathbf{u}^{(1)} \circ \mathbf{u}^{(2)} \circ \ldots \circ \mathbf{u}^{(N)}$. That is, each element of the tensor is the product of the corresponding vector elements, $x_{i_1 i_2 \ldots i_N} = u^{(1)}_{i_1}\ u^{(2)}_{i_2}\ \ldots\ u^{(N)}_{i_N}$ for all $i_n = 1, 2, \ldots, I_n$. The rank of an arbitrary $N$-order tensor $\mathbf{\mathcal{X}}$ is the minimal number of rank-1 tensors that yield $\mathbf{\mathcal{X}}$ when linearly combined. Next, several products between matrices will be used, such as the Kronecker product denoted by $\otimes$, the Khatri-Rao product denoted by $\odot$, and the Hadamard product denoted by $*$, whose definitions can be found in [9] for example.

## 3. LOCALITY PRESERVING NON NEGATIVE TENSOR FACTORIZATION

Let $\{\mathbf{\mathcal{X}}_q|_{q=1}^{Q}\}$ be a set of $Q$ non-negative tensors $\mathbf{\mathcal{X}}_q \in \mathbb{R}_+^{I_1 \times I_2 \times \ldots \times I_N}$ of order $N$. Let us also assume that these $Q$ tensors lie in a nonlinear manifold $\mathcal{A}$ embedded into the tensor space $\mathbb{R}_+^{I_1 \times I_2 \times \ldots \times I_N}$. Accordingly, we can represent such a set by a $(N+1)$-order tensor $\mathbf{\mathcal{A}} \in \mathbb{R}_+^{I_1 \times I_2 \times \ldots \times I_N \times I_{N+1}}$ with $I_{N+1} = Q$. Conventional NTF operates in the Euclidean space and does not consider the intrinsic geometrical structure of the data manifold [2]. To overcome the just mentioned limitation of NTF, we propose LPNTF by incorporating a geometrically-based regularizer stemming from locality preserving projections [7] into the optimization problem to be solved.

Given $\{\mathbf{\mathcal{X}}_q|_{q=1}^{Q}\}$, one can model the local structure of $\mathcal{A}$ by constructing the nearest neighbor graph $\mathcal{G}$. By exploiting the heat kernel function [7], one can define the elements of the weight matrix $\mathbf{S}$ of $\mathcal{G}$ as $s_{qp} = e^{-\frac{||\mathbf{\mathcal{X}}_q - \mathbf{\mathcal{X}}_p||^2}{\tau}}$ if $\mathbf{\mathcal{X}}_q$ and $\mathbf{\mathcal{X}}_p$ belong to the same class and 0 otherwise, where $||\ ||^2$ denotes the tensor norm [9]. Accordingly, the Laplacian matrix is defined as $\mathbf{L} = \mathbf{\Gamma} - \mathbf{S}$, where $\mathbf{\Gamma}$ is a diagonal matrix with elements $\gamma_{qq} = \sum_p s_{qp}$, i.e. the column sums of $\mathbf{S}$. Let $\mathbf{Z}^{(i)} = \mathbf{U}^{(N+1)} \odot \ldots \odot \mathbf{U}^{(i+1)} \odot \mathbf{U}^{(i-1)} \odot \ldots \odot \mathbf{U}^{(1)}$. Since the Laplacian matrix is analogous to Laplace-Beltrami operator on compact Riemannian manifolds [7], one can incorporate the local geometry of $\mathcal{A}$ into NTF by constructing the following objective function for LPNTF

in matrix form:

$$f_{LPNTF}\big(\mathbf{U}^{(i)}|_{i=1}^{N+1}\big) = ||\mathbf{A}_{(i)} - \mathbf{U}^{(i)}\big[\mathbf{Z}^{(i)}\big]^T||^2 + \lambda \,\mathrm{tr}\Big\{\big[\mathbf{U}^{(N+1)}\big]^T \mathbf{L}\,\mathbf{U}^{(N+1)}\Big\}, \quad (1)$$

where $\lambda > 0$ is a parameter, which controls the trade off between goodness of fit to the data tensor $\mathcal{A}$ and locality preservation. Consequently, we propose to minimize (1) subject to the non-negativity constraint on factor matrices $\mathbf{U}^{(i)} \in \mathbb{R}_+^{I_i \times k}$, $i = 1, 2, \ldots N+1$, where $k$ is the desirable number of rank-1 tensors approximating $\mathcal{A}$ when linearly combined.

Let $\nabla_{\mathbf{U}^{(i)}} f_{LPNTF} = \frac{\partial f_{LPNTF}}{\partial \mathbf{U}^{(i)}}$ be the partial derivative of the objective function $f_{LPNTF}(\mathbf{U}^{(i)}|_{i=1}^{N+1})$ with respect to $\mathbf{U}^{(i)}$. Since $\mathbf{U}^{(i)}$, $i = 1, 2, \ldots, N+1$, $\mathbf{\Gamma}$, and $\mathbf{S}$ are non-negative, the partial derivatives of the objective function can be decomposed as differences of two non-negative components denoted by $\nabla^+_{\mathbf{U}^{(i)}} f_{LPNTF}$ and $\nabla^-_{\mathbf{U}^{(i)}} f_{LPNTF}$, respectively. It can be shown that for $i = 1, 2, \ldots, N$ we have

$$\nabla_{\mathbf{U}^{(i)}} f_{LPNTF} = \underbrace{\mathbf{U}^{(i)}\big[\mathbf{Z}^{(i)}\big]^T\mathbf{Z}^{(i)}}_{\nabla^+_{\mathbf{U}^{(i)}} f_{LPNTF}} - \underbrace{\mathbf{A}_{(i)}\mathbf{Z}^{(i)}}_{\nabla^-_{\mathbf{U}^{(i)}} f_{LPNTF}}, \quad (2)$$

while for $i = N+1$ by invoking the definition of the Laplacian we obtain

$$\nabla_{\mathbf{U}^{(N+1)}} f_{LPNTF} = \underbrace{\mathbf{U}^{(N+1)}\big[\mathbf{Z}^{(N+1)}\big]^T\mathbf{Z}^{(N+1)} + \lambda\,\mathbf{\Gamma}\mathbf{U}^{(N+1)}}_{\nabla^+_{\mathbf{U}^{(N+1)}} f_{LPNTF}} - \underbrace{\big(\mathbf{A}_{(N+1)}\mathbf{Z}^{(N+1)} + \lambda\,\mathbf{S}\,\mathbf{U}^{(N+1)}\big)}_{\nabla^-_{\mathbf{U}^{(N+1)}} f_{LPNTF}}. \quad (3)$$

Following the strategy employed in the derivation of NMF [10], we can obtain an iterative alternating algorithm for LPNTF as follows. Given $N+1$ randomly initialized non-negative matrices $\mathbf{U}^{(i)}|_{i=1}^{N+1} \in \mathbb{R}_+^{I_i \times k}$, a local minimum of the optimization problem (1) subject to non-negativity constraints can be found by the multiplicative update rule:

$$\mathbf{U}^{(i)}_{[t+1]} = \mathbf{U}^{(i)}_{[t]} * \frac{\nabla^-_{\mathbf{U}^{(i)}_{[t]}} f_{LPNTF}}{\nabla^+_{\mathbf{U}^{(i)}_{[t]}} f_{LPNTF}}, \quad (4)$$

where the division in (4) is elementwise and $t$ denotes the iteration index. The multiplicative update rule (4) suffers from two drawbacks: (1) The denominator may be zero; (2) $\mathbf{U}^{(i)}_{[t+1]}$ does not change when $\mathbf{U}^{(i)}_{[t]} = \mathbf{0}$ and $\nabla_{\mathbf{U}^{(i)}[t]} f_{LPNTF} < \mathbf{0}$. In order to overcome these drawbacks, we can modify (4) as in [11]. A robust multiplicative update rule for LPNTF is then

$$\mathbf{U}^{(i)}_{[t+1]} = \mathbf{U}^{(i)}_{[t]} - \frac{\bar{\mathbf{U}}^{(i)}_{[t]}}{\nabla^+_{\mathbf{U}^{(i)}_{[t]}} f_{LPNTF} + \delta} * \nabla_{\mathbf{U}^{(i)}_{[t]}} f_{LPNTF}, \quad (5)$$

where $\bar{\mathbf{U}}^{(i)}_{[t]} = \mathbf{U}^{(i)}_{[t]}$ if $\nabla_{\mathbf{U}^{(i)}_{[t]}} f_{LPNTF} \geq \mathbf{0}$ and $\sigma$ otherwise. The paremeters $\sigma$, $\delta$ are predefined small positive numbers, typically $10^{-8}$ [11].

## 4. SPARSE REPRESENTATION-BASED CLASSIFICATION

For each music recording a 3D cortical representation is extracted by employing the computational auditory model of Wang *et al.* [17] with the same parameters as in [15]. Thus, each ensemble of recordings is represented by a 4th-order data tensor, which is created by stacking the 3rd-order feature tensors associated to the recordings. Consequently, the data tensor $\mathcal{A} \in \mathbb{R}_+^{I_1 \times I_2 \times I_3 \times I_4}$, where $I_1 = I_{scales} = 6$, $I_2 = I_{rates} = 10$, $I_3 = I_{frequencies} = 128$, and $I_4 = I_{samples}$ is obtained.

Determining the class label of a test cortical representation, given a number of labeled training cortical representations from $N$ music genres is addressed based on SRC [19]. Let us denote by $\mathbf{A}_i = [\mathbf{a}_{i1}|\mathbf{a}_{i2}|\ldots|\mathbf{a}_{in_i}] \in \mathbb{R}_+^{7680 \times n_i}$ the dictionary that contains $n_i$ cortical representations stemming from the $i$th genre as column vectors (i.e., atoms). Given a test cortical representation $\mathbf{y} \in \mathbb{R}_+^{7680}$ 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.

$$\mathbf{y} = \sum_{j=1}^{n_i} \mathbf{a}_{ij}\, c_{ij} = \mathbf{A}_i\, \mathbf{c}_i, \quad (6)$$

where $c_{ij} \in \mathbb{R}$ are coefficients, which form the coefficient vector $\mathbf{c}_i = [c_{i1}, c_{i2}, \ldots, c_{in_i}]^T$. Let us, now, define the matrix $\mathbf{D} = [\mathbf{A}_1|\mathbf{A}_2|\ldots|\mathbf{A}_N] = \mathbf{A}_{(4)}^T \in \mathbb{R}_+^{7680 \times I_{samples}}$ by concatenating $I_{samples}$ cortical representations, which are distributed across $N$ genres. Accordingly, a test cortical representation $\mathbf{y}$ that belongs to the $i$th genre can be equivalently expressed as

$$\mathbf{y} = \mathbf{D}\,\mathbf{c}, \quad (7)$$

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

Since the genre label of any test cortical representation is unknown, we can predict it by seeking the sparsest solution to the linear system of equations (7). Let $||.||_0$ be the $\ell_0$ quasi-norm of a vector, which returns the number of its non-zero elements. Formally, given the matrix $\mathbf{D}$ and the test cortical representation $\mathbf{y}$, sparse representation aims to find the coefficient vector $\mathbf{c}$ such that (7) holds and $||\mathbf{c}||_0$ is minimum, i.e.

$$\mathbf{c}^\star = \arg\min_{\mathbf{c}} ||\mathbf{c}||_0 \quad \text{subject to } \mathbf{D}\mathbf{c} = \mathbf{y}. \quad (8)$$

(8) is NP-hard due to the underlying combinational optimization. An approximate solution to (8) can be obtained by replacing the $\ell_0$ norm with the $\ell_1$ norm, i.e.

$$\mathbf{c}^\star = \arg\min_{\mathbf{c}} ||\mathbf{c}||_1 \quad \text{subject to } \mathbf{D}\,\mathbf{c} = \mathbf{y}, \quad (9)$$

where $||.||_1$ denotes the $\ell_1$ norm of a vector. The optimization problem (9) can be solved by standard linear programming methods in polynomial time [5].

Since we are interested in creating overcomplete dictionaries derived from the cortical representations, the dimensionality of atoms must be much smaller than the training

set cardinality. Thus, we can reformulate the optimization problem in (9) as follows:

$$\mathbf{c}^{\star} = \arg\min_{\mathbf{c}} ||\mathbf{c}||_1 \quad \text{subject to } \mathbf{W}\,\mathbf{D}\,\mathbf{c} = \mathbf{W}\mathbf{y}, \quad (10)$$

where $\mathbf{W} \in \mathbb{R}^{k \times 7680}$ with $k \ll \min(7680, I_{samples})$ is a projection matrix. The projection matrix $\mathbf{W}$ can be obtained by LPNTF or any other multilinear dimensionality reduction technique, such as NTF [2], MPCA [12], GTDA [20], or DNTF [21]. Alternatively, one can even employ a random projection matrix whose elements are independently sampled from a zero-mean normal distribution, and each column is normalized to unit length as proposed in [19]. More particularly, when LPNTF, NTF, or DNTF is applied to the data tensor $\mathcal{A}$, four factor matrices $\mathbf{U}^{(i)} \in \mathbb{R}_{+}^{I_i \times k}$, $i = 1, 2, 3, 4$, are obtained, which are associated to scale, rate, frequency, and sample modes respectively. The projection matrix $\mathbf{W}$ is given by either $\mathbf{W} = (\mathbf{U}^{(3)} \odot \mathbf{U}^{(2)} \odot \mathbf{U}^{(1)})^T$ or $\mathbf{W} = (\mathbf{U}^{(3)} \odot \mathbf{U}^{(2)} \odot \mathbf{U}^{(1)})^{\dagger}$, where $(.)^{\dagger}$ denotes the Moore-Penrose pseudoinverse. Accordingly, every column of $\mathbf{D}$ (i.e. vectorized cortical representation of a music recording) is a linear combination of the basis vectors, which span the columns of the basis matrix $\mathbf{W}^T$ with coefficients taken from the columns of matrix $[\mathbf{U}^{(4)}]^T$. That is, $\mathbf{D} = \mathbf{A}_{(4)}^T = \mathbf{W}^T[\mathbf{U}^{(4)}]^T$. For MPCA or GTDA, three factor matrices $\mathbf{U}^{(i)} \in \mathbb{R}^{I_i \times J_i}$, with $J_i < I_i$, $i = 1, 2, 3$, are obtained, which are associated to scales, rates, and frequencies, respectively. The columns of $\mathbf{D}$ are obtained by applying the projection matrix $\mathbf{W} = (\mathbf{U}^{(3)} \otimes \mathbf{U}^{(2)} \otimes \mathbf{U}^{(1)})^T$ or $\mathbf{W} = (\mathbf{U}^{(3)} \otimes \mathbf{U}^{(2)} \otimes \mathbf{U}^{(1)})^{\dagger}$ to vectorized training tensors $\mathrm{vec}(\mathcal{X}_q)$. The dimensionality reduction of the original cortical representations data space has two benefits: (1) It reduces the computational cost of linear programming solvers for (9) [5]; (2) It facilitates the creation of a redundant dictionary out of training cortical representations.

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

$$\mathbf{c}^{\star} = \arg\min_{\mathbf{c}} ||\mathbf{c}||_1 \quad \text{subject to } \mathbf{W}\,\mathbf{D}\,\mathbf{c} = \hat{\mathbf{y}}. \quad (11)$$

Ideally, the coefficient vector $\mathbf{c}^{\star}$ contains non-zero entries in positions associated with the columns of $\mathbf{WD}$ associated with 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 elements in $\mathbf{c}^{\star}$ that are associated to multiple genres. To cope with this problem, each auditory modulation representation is classified to the genre that minimizes the $\ell_2$ norm residual between $\hat{\mathbf{y}}$ and $\breve{\mathbf{y}} = \mathbf{W}\,\mathbf{D}\,\vartheta_i(\mathbf{c})$, where $\vartheta_i(\mathbf{c}) \in \mathbb{R}^n$ is a new vector whose non-zero entries are only the elements in $\mathbf{c}$ that are associated to the $i$th genre [19].

## 5. EXPERIMENTAL EVALUATION

In order to assess both the discriminating power of the features derived by LPNTF applied to cortical representations for dimensionality reduction and the accuracy of sparse representation-based classification, experiments are conducted on two widely used datasets for music genre classification [3, 8, 14, 16]. The first dataset, abbreviated as GTZAN, was collected by G. Tzanetakis [16] and consists of 10 genre classes. 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 across 6 genre classes. All the recordings were converted to monaural wave format at a sampling frequency of 16 kHz and quantized with 16 bits. Moreover, the music 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 in order to exclude any introductory parts that may not be directly related to the music genre the recording belongs to. The cortical representation is extracted for the aforementioned segment of 30 sec duration for any recording from both datasets. The best reported music genre classification accuracies obtained for the aforementioned datasets are summarized in Table 1.

| **Reference** | **Dataset** | **Accuracy** |
|---|---|---|
| Bergstra *et al.* [3] | GTZAN | 82.5% |
| Holzapfel *et al.* [8] | ISMIR2004 | 83.5% |
| Pampalk *et al.* [14] | ISMIR2004 | 82.3% |

**Table 1**. Best classification accuracies achieved by music genre classification approaches on standard datasets.

Following the experimental set-up used in [3, 15, 16], stratified 10-fold cross-validation is employed for experiments conducted on the GTZAN dataset. Thus each training set consists of 900 audio files. Accordingly, the training tensor $\mathcal{A}_{GTZAN} \in \mathbb{R}_{+}^{6 \times 10 \times 128 \times 900}$ is constructed by stacking the cortical representations. The experiments on 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 tensor $\mathcal{A}_{ISMIR} \in \mathbb{R}_{+}^{6 \times 10 \times 128 \times 729}$ is constructed.

The projection matrix $\mathbf{W}$ is derived from each training tensor $\mathcal{A}_{GTZAN}$ and $\mathcal{A}_{ISMIR}$ by employing either LPNTF, NTF, DNTF, MPCA or GTDA. Throughout the experiments the value of $\lambda$ in LPNTF was empirically set to 0.5, while the parameter $\tau$ of the heat kernel was set equal to 1. In order to determine automatically the parameters $\lambda$ and $\tau$ one can apply cross-validation to the training set. However, the systematic setting of these parameters could be a subject of future research. In order to determine the dimensionality of factor matrices, the ratio of the sum of eigenvalues retained over the sum of all eigenvalues for each mode-$n$ tensor unfolding is employed as in [12]. By using this ratio as a specification parameter, the number of retained principal components for each mode (e.g. scale, rate, and frequency) was determined, as is demon-

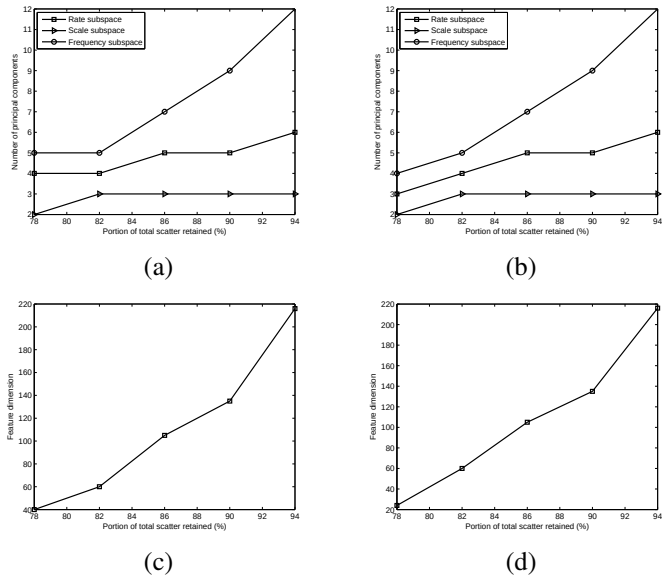


**Figure 1**. Total number of retained principal components in each mode (e.g. rate, scale, and frequency) as a function of the portion of total scatter retained for the: a) GTZAN dataset and b) ISMIR 2004 Genre dataset. Feature dimension as a function of the portion of the total scatter retained for the: c) GTZAN dataset and d) ISMIR 2004 Genre dataset.

strated in Figure 1 for the GTZAN and the ISMIR Genre 2004 datasets. The different subspace analysis methods are compared for equal dimensionality reduction. That is, the same $J_1 = J_{scales}$, $J_2 = J_{rates}$ and $J_3 = J_{frequencies}$ were used in MPCA and GTDA, while $k = J_1 J_2 J_3$ for LPNTF, NTF, and DNTF. The same value of parameter $k$ is used in order to construct the random projection matrix. Since the low dimensional features obtained by the aforementioned multilinear dimensionality reduction algorithms are linearly combined for classification, SVMs with linear kernel are tested as alternatives to SRC.

In Figure 2, the classification accuracy achieved by the three different classifiers is plotted as a function of the portion of the total scatter retained, when various subspace analysis methods are applied to both GTZAN and ISMIR 2004 Genre datasets. On the GTZAN dataset the best classification accuracy (92.4%) was obtained when LPNTF extracts features, that are classified by SRC. In this case, $k = 135$, as shown in Figure 1(c). 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 2%. The reported classification accuracy outperforms those listed in Table 1. The interval $\pm$ one standard deviation is overlaid in all plots for the various values of the portion of the total scatter retained.

On the ISMIR 2004 Genre dataset the best classification accuracy (94.38%) was obtained, when the NTF with $k = 135$ extracts the low dimensional features that are classified by SRC next. When the LPNTF with $k = 105$ extracts features that are classified by SRC next, the classification accuracy is found equal to 94.25%, that is very close to the best accuracy. Both accuracies outperform the previously reported ones, which are listed in Table 1.

It is seen that the classification accuracy obtained by LPNTF and SRC outperforms the accuracy obtained with features extracted by all other multilinear subspace analysis techniques, which are next classified by either SRC or linear SVMs, for all the values of the portion of the total scatter retained but one. Moreover, the classification accuracy obtained with features extracted by LPNTF, NTF, MPCA or GTDA that are subsequently classified by SRC exceeds 80% for both datasets despite the reduced dimensions of the feature space extracted that are plotted in Figure 1(c) and (d). 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 for face recognition [19].

## 6. CONCLUSIONS

In this paper, a robust music genre classification framework has been proposed. This framework resorts to cortical representations for music representation, while sparse representation-based classification has been employed for genre classification. A multilinear subspace analysis technique (i.e. LPNTF) has been developed, which incorporates the underlying geometrical structure of the cortical representations with respect to the music genre into the NTF. The crucial role of feature extraction and dimensionality reduction for music genre classification has been demonstrated. The best classification accuracies reported in this paper outperform any accuracy ever obtained by state of the art music genre classification algorithms applied to both GTZAN and 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 that the music genre classification algorithm performs well for such small sets. Future

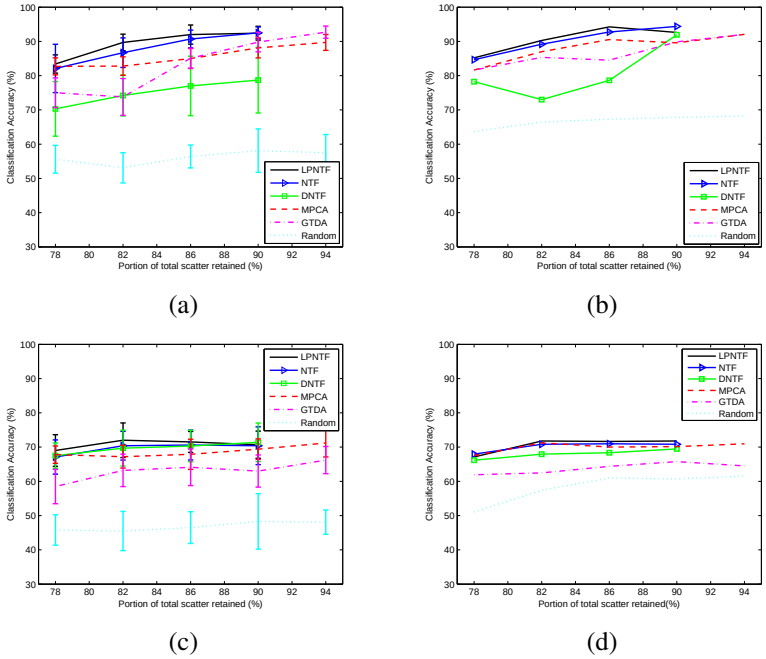


**Figure 2**. Classification accuracy for various feature extraction methods and classifiers. (a) SRC on GTZAN dataset; (b) SRC on ISMIR2004 Genre dataset; (c) Linear SVM on GTZAN dataset; (d) Linear SVM on ISMIR2004 Genre dataset.

research will address the performance of SRC framework under such conditions.

## 7. REFERENCES


[1] J. J. Aucouturier and F. Pachet: "Representing Musical Genre: A State of the Art," *Journal of New Music Research*, pp. 83–93, 2003.

[2] E. Benetos and C. Kotropoulos: "A Tensor-Based Approach for Automatic Music Genre Classification," *Proceedings of the European Signal Processing Conference*, Lausanne, Switzerland, 2008.

[3] 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.

[4] E. J. Candès, J. Romberg, and T. Tao: "Robust Uncertainty Principles: Exact Signal Reconstruction From Highly Incomplete Frequency Information," *IEEE Transactions on Information Theory*, Vol. 52, No. 2, pp. 489–509, 2006.

[5] S. S. Chen, D. L. Donoho, and M. A. Saunders: "Atomic Decomposition by Basis Pursuit," *SIAM Journal Scientific Computing*, Vol. 20, No.1 pp. 33–61, 1998.

[6] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, Vol. 52, No. 4, pp. 1289–1306, 2006.

[7] X. He and P. Niyogi: "Locality Preserving Projections," *Advances in Neural Information Processing Systems*, Vol. 16, MIT Press, 2004.

[8] A. Holzapfel and Y. Stylianou: "Musical Genre Classification Using Nonnegative Matrix Factorization-Based Features," *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 16, No. 2, pp. 424–434, 2008.

[9] T. Kolda and B. W. Bader: "Tensor Decompositions and Applications," *SIAM Review*, Vol. 51, No. 3, to appear.

[10] D. D. Lee and H. S. Seung: "Algorithms for Non-negative Matrix Factorization," *Advances in Neural Information Processing Systems*, Vol. 13, pp. 556–562, 2001.

[11] C. -J. Lin: "On the Convergence of Multiplicative Update Algorithms for Nonnegative Matrix Factorization," *IEEE Transactions on Neural Networks*, Vol. 18, No. 6, pp. 1589–1596, 2007.

[12] H. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos: "MPCA: Multilinear Principal Component Analysis of Tensor Objects," *IEEE Transactions on Neural Networks*, Vol. 19, No. 1, pp 18–39, 2008.

[13] N. Mesgarani, M. Slaney, and S. A. Shamma: "Discrimination of Speech from Nonspeech Based on Multiscale Spectro-temporal Modulations," *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 14, No. 3, pp. 920–930, 2006.

[14] E. Pampalk, A. Flexer, and G. Widmer: "Improvements of Audio-based Music Similarity and Genre Classification," *Proceedings of the Sixth International Symposium on Music Information Retrieval*, London, UK, 2005.

[15] I. Panagakis, E. Benetos, and C. Kotropoulos: "Music Genre Classification: A Multilinear Approach," *Proceedings of the Seventh International Symposium on Music Information Retrieval*, Philadelphia, USA, 2008.

[16] G. Tzanetakis and P. Cook: "Musical Genre Classification of Audio Signal," *IEEE Transactions on Speech and Audio Processing*, Vol. 10, No. 3, pp. 293–302, 2002.

[17] K. Wang and S. A. Shamma: "Spectral Shape Analysis in the Central Auditory System," *IEEE Transactions on Speech and Audio Processing*, Vol. 3, No. 5, pp. 382–396, 1995.

[18] 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, pp. 1371–1379, 2005.

[19] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma "Robust Face Recognition via Sparse Representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 31, No. 2, pp. 210–227, 2009.

[20] D. Tao, X. Li, X. Wu, and S. J. Maybank: "General Tensor Discriminant Analysis and Gabor Features for Gait Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 29, No. 10, pp. 1700–1715, 2007.

[21] S. Zafeiriou: "Discriminant Nonnegative Tensor Factorization Algorithms," *IEEE Transactions on Neural Networks*, Vol. 20, No. 2, pp. 217–235, 2009.