MUSIC GENRE CLASSIFICATION VIA TOPOLOGY PRESERVING NON-NEGATIVE TENSOR FACTORIZATION AND SPARSE REPRESENTATIONS

Yannis Panagakis and Constantine Kotropoulos

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

ABSTRACT

Motivated by the rich, psycho-physiologically grounded properties of auditory cortical representations and the power of sparse representation-based classifiers, we propose a robust music genre classification framework. Its first pilar is a novel multilinear subspace analysis method that reduces the dimensionality of cortical representations of music signals, while preserving the topology of the cortical representations. Its second pilar is the sparse representation based classification, that models any test cortical representation as a sparse weighted sum of dictionary atoms, which stem from training cortical representations of known genre, by assuming that the representations of music recordings of the same genre are close enough in the tensor space they lie. Accordingly, the dimensionality reduction is made in a compatible manner to the working principle of the sparse-representation based classification. Music genre classification accuracy of 93.7% and 94.93% is reported on the GTZAN and the ISMIR2004 Genre datasets, respectively. Both accuracies outperform any accuracy ever reported for state of the art music genre classification algorithms applied to the aforementioned datasets.

Index Terms— Music genre classification, topology preserving, non-negative tensor factorization, sparse representations

1. INTRODUCTION

The efficient organization of large music databases is of paramount importance for the electronic music distribution. Music genre is probably the most popular description of music content [1], although it is not well-defined, since genre labels may depend on cultural, artistic, or market factors and the boundaries between genres are fuzzy [2].

The auditory representations, which are grounded on psycho-physiological investigations on human auditory system, have been proved a robust alternative to the conventional *bag-of-features* [2, 3] approach for music genre classification [4], especially when they are combined with the sparse representation-based classifier (SRC) [5]. By employing the auditory model proposed in [6], a given music recording is mapped to a three-dimensional (3D) representation of its slow spectral and temporal modulations with the same parameters as in [4]. This 3D representation is referred to as *cortical representation*. The cortical representations form a dictionary of basis signals for music genres, which is exploited next within the SRC as is proposed in [7]. That is, first each music recording is represented by its cortical representation. Second, any test cortical representation is modeled as a sparse weighted sum of dictionary atoms, which stem from cortical representations associated to training music recordings whose genre is known. The underlying assumption is that representations of the same genre are close enough in the tensor space they lie. 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 by involving only a small fraction of the dictionary atoms computed efficiently via ℓ_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. The robustness of the proposed framework is attributed to the sparsity enforced.

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. Clearly, the dimensionality reduction facilitates the treatment of missing data, noise, and outliers. Following the same reasoning as in [4], multilinear dimensionality reduction techniques, such as Non-Negative Tensor Factorization (NTF) [8], Multilinear Principal Component Analysis (MPCA) [9], General Tensor Discriminant Analysis (GTDA) [10] could be considered. However, the just mentioned methods do not take into account the topological structure of the original tensor data space. To reduce tensor dimensions in a consistent manner with the working principle of the SRC, one should guarantee that two tensorial data points, which are close in the intrinsic geometry of the original tensor space, are also close in the new tensor space after multilinear dimensionality reduction. To this end, a novel algorithm is proposed, where the topological structure of the original tensor space is incorporated into the NTF by reformulating the NTF optimization problem as minimization of the total variation norm (TVN) [11]. In particular, we extend the Topology Preserving Non-Negative Matrix Factorization (TPNMF) [12] to Topology Preserving Non-Negative Tensor Factorization (TPNTF) by minimizing the TVN [11]. The non-negativity of cortical representations is preserved maintaining their physical interpretation. A multiplicative updating algorithm for the TPNTF is derived, which extracts features from the cortical representations. For comparison purposes, NTF, MPCA, and GTDA were considered, as well.

Next, the features extracted by the aforementioned multilinear dimensionality techniques are classified by the SRC [7, 5]. 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. In particular, the proposed genre classification method, that extracts features using the TPNTF, which are then classified by the SRC (i.e. TPNTF plus SRC), yields

an accuracy of 93.7% and 94.93% on the GTZAN and ISMIR2004 Genre datasets, respectively. The aforementioned results outperform those reported in [5], where conventional linear subspace analysis techniques extract features from the auditory temporal modulations representation which are then classified by the SRC. This is mainly attributed to the topology preservation of the TPNTF, in a consistent manner to the working principle of the SRC. 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 briefly defined. The TPNTF is detailed in Section 3. The SRC framework, that is applied to music genre classification, is described 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) [13]. Throughout the paper, tensors are denoted by boldface Euler script calligraphic letters (e.g. $\mathcal{X}, \mathcal{A}$), matrices are denoted by uppercase boldface letters (e.g. $\mathbf{U}$), and vectors are denoted by lowercase boldface letters (e.g. $\mathbf{u}$). The ith row of $\mathbf{U}$ is denoted as $\mathbf{u}_{i:}$ while its jth column is denoted as $\mathbf{u}_{:j}$. A high-order real valued tensor $\mathcal{X}$ of order N is defined over the tensor space $\mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$, where $I_n \in \mathbb{Z}$ and $n = 1, 2, \ldots, N$. Mode-n unfolding of tensor $\mathcal{X}$ yields the matrix $\mathbf{X}_{(n)} \in \mathbb{R}^{I_n \times \overline{I}_n}$, where $\overline{I}_n = I_1 I_2 \ldots I_{n-1} I_{n+1} \ldots I_N$. An N-order tensor $\mathcal{X}$ 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. $\mathcal{X} = \mathbf{u}^{(1)} \circ \mathbf{u}^{(2)} \circ \ldots \circ \mathbf{u}^{(N)}$. The rank of an arbitrary N-order tensor $\mathcal{X}$ is the minimal number of rank-1 tensors that yield $\mathcal{X}$ 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 [13], for example.

3. TOPOLOGY PRESERVING NON-NEGATIVE TENSOR FACTORIZATION

Let $\{\mathcal{X}_q|_{q=1}^{Q}\}$ be a set of Q non-negative tensors $\mathcal{X}_q \in \mathbb{R}_+^{I_1 \times I_2 \times \ldots \times I_N}$ of order N. Let us also assume that these Q tensors lie in a smooth, nonlinear manifold embedded into the tensor space $\mathbb{R}_+^{I_1 \times I_2 \times \ldots \times I_N}$. Accordingly, we can represent $\{\mathcal{X}_q|_{q=1}^{Q}\}$ by a $(N+1)$-order tensor $\mathcal{A}$, which is assumed to lie in a nonlinear manifold $\mathcal{M}$ in the tensor space $\mathbb{R}_+^{I_1 \times I_2 \times \ldots \times I_{N+1}}$ with $I_{N+1} = Q$. The conventional NTF operates in the Euclidean space [8]. Therefore it is unable to find the intrinsic low-dimensionality manifold structure. To overcome the just mentioned limitation of the NTF, we propose the TPNTF by considering the constrained minimization of the TVN [11].

Let k be the desirable number of rank-1 tensors approximating $\mathcal{A}$ when linearly combined. The NTF of $\mathcal{A}$ derives $N+1$ factor matrices $\mathbf{U}^{(n)} \in \mathbb{R}_+^{I_n \times k}$, $n = 1, 2, \ldots, N+1$ by minimizing the square of the Frobenius norm $\|\mathbf{A}_{(n)} - \mathbf{U}^{(n)}[\mathbf{Z}^{(n)}]^T\|_F^2$ subject to $\mathbf{U}^{(n)} \geq \mathbf{0}$, following an alternating minimization scheme [8]. Let $\mathbf{Z}^{(n)} \triangleq \mathbf{U}^{(N+1)} \odot \ldots \odot \mathbf{U}^{(n+1)} \odot \mathbf{U}^{(n-1)} \odot \ldots \odot \mathbf{U}^{(1)}$. The mode-$n$ unfolding of $\mathcal{A}$ is decomposed as follows

$$\mathbf{A}_{(n)} = \mathbf{U}^{(n)}[\mathbf{Z}^{(n)}]^T \Leftrightarrow [\mathbf{A}_{(n)}]^T = \mathbf{Z}^{(n)}[\mathbf{U}^{(n)}]^T. \quad (1)$$

Let $[\mathbf{z}^{(n)}]_{i:} \in \mathcal{M}_{\mathbf{z}}$ and $[\mathbf{a}_{(n)}]_{:i} \in \mathcal{M}_{\mathbf{a}}$, where $\mathcal{M}_{\mathbf{z}}$ and $\mathcal{M}_{\mathbf{a}}$ are assumed to be smooth, compact, Riemmanian manifolds, non-negatively embedded in $\mathbb{R}^k$ and $\mathbb{R}^{I_n}$, respectively. (1) implies that $[\mathbf{U}^{(n)}]$ maps the rows of $\mathbf{Z}^{(n)}$ to the columns of $\mathbf{A}_{(n)}$, i.e., $f([\mathbf{z}^{(n)}]_{i:}) : \mathcal{M}_{\mathbf{z}} \mapsto \mathcal{M}_{\mathbf{a}}$. In order to preserve the local topology, two neighboring points $\mathbf{z}_{k:}^{(n)}$ and $\mathbf{z}_{l:}^{(n)} \in \mathcal{M}_{\mathbf{z}}$ should be mapped to neighboring points $\mathbf{z}_{k:}^{(n)}[\mathbf{U}^{(n)}]^T$ and $\mathbf{z}_{l:}^{(n)}[\mathbf{U}^{(n)}]^T \in \mathcal{M}_{\mathbf{a}}$, and vise versa. The mapping that best preserves the local topology can be obtained by the constrained minimization of TVN as follows [12]

$$[\mathbf{U}_*^{(n)}]^T = \underset{\substack{\mathbf{U}^{(n)} \geq \mathbf{0} \\ \mathbf{A}_{(n)} = \mathbf{U}^{(n)}[\mathbf{Z}^{(n)}]^T}}{\operatorname{argmin}} \int_{\mathcal{M}_{\mathbf{z}}} \|\Delta f(\mathbf{Z}^{(n)})\|_F^2, \quad (2)$$

where Δ denotes the discrete gradient operator. For the matrix $\mathbf{M}$, it is defined as $[\Delta \mathbf{M}]_{ij} = [\delta_x(\mathbf{M})]_{ij} + [\delta_y(\mathbf{M})]_{ij}$ with $[\delta_x(\mathbf{M})]_{ij} = m_{ij} - m_{(i-1)j}$ and $[\delta_y(\mathbf{M})]_{ij} = m_{ij} - m_{i(j-1)}$. Let $E(\mathbf{U}^{(n)}) = \frac{1}{2}\|\mathbf{A}_{(n)} - \mathbf{U}^{(n)}[\mathbf{Z}^{(n)}]^T\|_F^2 + \frac{\lambda}{2}\|\Delta \mathbf{U}^{(n)}[\mathbf{Z}^{(n)}]^T\|_F^2$. The optimization problem (2) can be expressed equivalently as [11, 12]

$$\mathbf{U}_*^{(n)} = \underset{\mathbf{U}^{(n)} \geq \mathbf{0}}{\operatorname{argmin}} E(\mathbf{U}^{(n)}), \quad (3)$$

where $\lambda > 0$ controls the trade off between the goodness of fit to the data tensor $\mathcal{A}$ and the topology preservation. The TPNTF can be obtained by solving $(N+1)$ many optimization problems (3) for $n = 1, 2, \ldots, N+1$ by following an alternating minimization scheme as in NTF [8].

The partial derivative of $E(\mathbf{U}^{(n)})$ with respect to $\mathbf{U}^{(n)}$ is given by

$$\begin{aligned} \nabla_{\mathbf{U}^{(n)}} E(\mathbf{U}^{(n)}) = \\ \mathbf{U}^{(n)}[\mathbf{Z}^{(n)}]^T\mathbf{Z}^{(n)} + \lambda\, \mathbf{U}^{(n)}[\mathbf{Z}^{(n)}]^T\mathbf{Z}^{(n)} \\ +\lambda\, \mathbf{U}^{(n)}[\delta_x(\mathbf{Z}^{(n)})]^T\delta_x(\mathbf{Z}^{(n)}) - \mathbf{A}\mathbf{Z}^{(n)}. \end{aligned} \quad (4)$$

Let t denote the iteration index, σ and ϵ be predefined small positive numbers, typically 10^{-8} [14]. Let also define $\tilde{\mathbf{U}}_{[t]}^{(n)}$ elements as $(\tilde{u}_{[t]}^{(n)})_{ij} = (u_{[t]}^{(n)})_{ij}$ if $[\nabla_{\mathbf{U}_{[t]}^{(n)}} E(\mathbf{U}_{[t]}^{(n)})]_{ij} \geq 0$ and σ otherwise. Following [14, 12] it can be proven that, given non-negative initialized factor matrices (i.e. $\mathbf{U}_1^{(n)} \geq \mathbf{0}, n = 1, 2, \ldots, N+1$) a convergent iterative update rule that solves (3) is obtained by

$$\mathbf{U}_{[t+1]}^{(n)} = \mathbf{U}_{[t]}^{(n)} - \frac{\tilde{\mathbf{U}}_{[t]}^{(n)}}{\nabla^+_{\mathbf{U}_{[t]}^{(n)}} E(\mathbf{U}_{[t]}^{(n)}) + \epsilon} * \nabla_{\mathbf{U}_{[t]}^{(n)}} E(\mathbf{U}_{[t]}^{(n)}), \quad (5)$$

where $\nabla^+_{\mathbf{U}_{[t]}^{(n)}} E(\mathbf{U}_{[t]}^{(n)})$ is defined by the first three terms in (4) and the division in (5) is elementwise.

4. SPARSE REPRESENTATION-BASED CLASSIFICATION

For each music recording, a 3D cortical representation is extracted by employing the computational auditory model of Yang *et al.*[6] with the same parameters as in [4]. 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 P music genres is addressed based on the SRC [7]. Let us denote by $\mathbf{A}_i = [\mathbf{a}_{i1}|\mathbf{a}_{i2}|\dots|\mathbf{a}_{ip_i}] \in \mathbb{R}_+^{7680\times p_i}$ the dictionary that contains p_i cortical representations stemming from the ith genre as column vectors (i.e., atoms). Given a test cortical representation $\mathbf{y} \in \mathbb{R}_+^{7680}$ that belongs to the ith class, we can assume that $\mathbf{y}$ is expressed as a linear combination of the atoms that belong to the ith class, i.e. $\mathbf{y} = \sum_{j=1}^{p_i} \mathbf{a}_{ij}\, c_{ij} = \mathbf{A}_i\, \mathbf{c}_i$, where $c_{ij} \in \mathbb{R}$ are coefficients, which form the coefficient vector $\mathbf{c}_i = [c_{i1}, c_{i2}, \dots, c_{ip_i}]^T$. Let us, now, define the matrix $\mathbf{D} = [\mathbf{A}_1|\mathbf{A}_2|\dots|\mathbf{A}_P] = \mathbf{A}_{(4)}^T \in \mathbb{R}_+^{7680\times I_{samples}}$ by concatenating $I_{samples}$ cortical representations, which are distributed across P genres. Accordingly, a test cortical representation $\mathbf{y}$ that belongs to the ith genre can be equivalently expressed as

$$\mathbf{y} = \mathbf{D}\,\mathbf{c}, \tag{6}$$

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 ith 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 (6). Formally, given the matrix $\mathbf{D}$ and the test cortical representation $\mathbf{y}$, an approximate solution to the sparsest one aims to find the coefficient vector $\mathbf{c}$ such that

$$\mathbf{c}^\star = \arg\min_{\mathbf{c}} ||\mathbf{c}||_1 \quad \text{subject to } \mathbf{D}\,\mathbf{c} = \mathbf{y}, \tag{7}$$

where $||.||_1$ denotes the ℓ_1 norm of a vector. The optimization problem (7) can be solved by standard linear programming methods in polynomial time. 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 (7) as follows:

$$\mathbf{c}^\star = \arg\min_{\mathbf{c}} ||\mathbf{c}||_1 \quad \text{subject to } \mathbf{W}\,\mathbf{D}\,\mathbf{c} = \mathbf{W}\mathbf{y}, \tag{8}$$

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 the TPNTF or any other multilinear dimensionality reduction technique, such as the NTF, the MPCA, or the GTDA. More particularly, when the TPNTF or the NTF is applied to the data tensor $\mathcal{A}$, four factor matrices $\mathbf{U}^{(n)} \in \mathbb{R}_+^{I_n\times k}$, $n = 1, 2, 3, 4$, are obtained, which are associated to scale, rate, frequency, and sample modes respectively. The projection matrix is given by $\mathbf{W} = (\mathbf{U}^{(3)} \odot \mathbf{U}^{(2)} \odot \mathbf{U}^{(1)})^T$ or $\mathbf{W} = (\mathbf{U}^{(3)} \odot \mathbf{U}^{(2)} \odot \mathbf{U}^{(1)})^\dagger$, respectively 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 the MPCA or the GTDA, three factor matrices $\mathbf{U}^{(n)} \in \mathbb{R}^{I_n\times J_n}$, with $J_n < I_n$, $n = 1, 2, 3$, are obtained, which are associated to scales, rates, and frequencies, respectively. In such a case, the reduced dimension space is obtained by applying the projection matrix $\mathbf{W} = (\mathbf{U}^{(3)} \otimes \mathbf{U}^{(2)} \otimes \mathbf{U}^{(1)})^T$ or $\mathbf{W} = (\mathbf{U}^{(3)} \otimes \mathbf{U}^{(2)} \otimes \mathbf{U}^{(1)})^\dagger$, respectively to vectorized training tensors $\text{vec}(\mathcal{X}_q)$.

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

$$\mathbf{c}^\star = \arg\min_{\mathbf{c}} ||\mathbf{c}||_1 \quad \text{subject to } \mathbf{W}\,\mathbf{D}\,\mathbf{c} = \hat{\mathbf{y}}. \tag{9}$$

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 ℓ_2 norm residual between $\hat{\mathbf{y}}$ and $\check{\mathbf{y}} = \mathbf{W}\,\mathbf{D}\,\vartheta_i(\mathbf{c})$, 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 ith genre [7].

5. EXPERIMENTAL EVALUATION

In order to assess both the discriminating power of the features derived by the TPNTF 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 [8, 15, 16, 4, 5, 3]. The first dataset, abbreviated as GTZAN, was collected by G. Tzanetakis [3] 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. 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. All the recordings were preprocessed as in [5]. 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. On the GTZAN dataset, Bergstra *et al.* [15] tested the Mel-frequency cepstral coefficients, the fast Fourier transform coefficients, the linear prediction coefficients, and the zero-crossing rate and reported classification accuracy reaching 82.5% for the Adaboost meta-classifier. Pampalk *et al.* [16] was the winner in the ISMIR2004 Genre classification contest, where a classification accuracy equal to 84.07% was obtained by combining different feature sets based on fluctuation patterns and Mel-frequency cepstral coefficients. The aforementioned classification accuracies are the best ever reported without employing sparse representations. In [5], 2D auditory temporal modulations were used for music representation, while the sparse representation-based classification has then been employed for genre classification. On the GTZAN dataset then best classification accuracy 91.0 % was obtained when Non-negative Matrix Factorization (NMF) [14] extracts features that are classified by the SRC while on the ISMIR 2004 Genre dataset then best classification accuracy 93.56% was obtained, when Principal Component Analysis (PCA) extracts features that are classified by the SRC.

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

Reference	**Dataset**	**Accuracy (%)**
Panagakis *et al.* [5]	GTZAN	91
Bergstra *et al.* [15]	GTZAN	82.5
Panagakis *et al.* [5]	ISMIR2004	93.56
Pampalk *et al.* [16]	ISMIR2004	84.07

To make our experimental results comparable with those obtained by the state-of-the-art music genre classification systems, re-

ported in Table 1, two different experimental set-ups are employed. In particular, following the experimental set-up used in [3, 15, 4, 5], 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 $\mathcal{A}_{GTZAN} \in \mathbb{R}_{+}^{6\times10\times128\times900}$ 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 $\mathcal{A}_{ISMIR} \in \mathbb{R}_{+}^{6\times10\times128\times729}$ is constructed. The projection matrix $\mathbf{W}$ is derived from each training tensor $\mathcal{A}_{GTZAN}$ and $\mathcal{A}_{ISMIR}$ by employing either the TPNTF, the NTF, the MPCA or the GTDA. Throughout the experiments the value of λ in TPNTF was empirically set to 0.5.

In Table 2, the best classification accuracies obtained by features derived from various multilinear subspace analysis techniques and classified then by the SRC for the GTZAN and the ISMIR2004 Genre Dataset are summarized. The fourth column in Table 2 indicates the number of dimensions of the vectorized cortical representation after dimensionality reduction. For the GTZAN dataset, the standard deviation of the classification accuracy was estimated thanks to the stratified 10-fold cross-validation. By inspecting Table 2 the classification accuracy obtained by TPNTF and SRC outperforms that obtained with features extracted by all other multilinear subspace analysis techniques. Clearly, the reported classification by TPNTF and SRC outperforms those listed in Table 1.

Table 2. Best classification accuracies achieved by various multilinear subspace analysis techniques plus SRC for music genre classification on standard datasets.

Method	**Dataset**	**Accuracy (%)**	**Dimension**
TPNTF + SRC	GTZAN	**93.7** (1.88)	135
NTF + SRC	GTZAN	92 (1.71)	135
MPCA + SRC	GTZAN	89.7 (2.31)	216
GTDA + SRC	GTZAN	92.1 (2.85)	216
TPNTF + SRC	ISMIR2004	**94.93**	135
NTF + SRC	ISMIR2004	94.38	135
MPCA + SRC	ISMIR2004	92.05	216
GTDA + SRC	ISMIR2004	92.05	216

6. CONCLUSIONS

In this paper, a robust music genre classification framework has been proposed. The 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. TPNTF) has been developed, which incorporates the underlying topological structure of the cortical representations with respect to the music genre into the NTF. 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 research will address the performance of SRC framework under such conditions.

ACKNOWLEDGMENT

This work has been supported by HRAKLEITOS II research project.

7. REFERENCES

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

[2] 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, March 2006.

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

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

[5] Y. Panagakis, C. Kotropoulos, and G. R. Arce, "Music genre classification via sparse representations of auditory temporal modulations," in *Proc. 17th European Signal Processing Conf.*, 2009.

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

[7] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, 2009.

[8] E. Benetos and C. Kotropoulos, "A tensor-based approach for automatic music genre classification," in *Proc. 16th European Signal Processing Conf.*, 2008.

[9] H. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "MPCA: Multilinear principal component analysis of tensor objects," *IEEE Trans. Neural Networks*, vol. 19, no. 1, pp. 18–39, 2008.

[10] D. Tao, X. Li, X. Wu, and S. J. Maybank, "General tensor discriminant analysis and Gabor features for gait recognition," *IEEE Trans. Pattern Anal. and Mach. Intell.*, vol. 29, no. 10, pp. 1700–1715, 2007.

[11] T. Chan, S. Esedoglu, F. Park, and A. Yip, "Recent developments in total variation image restoration," in *Handbook of Mathmematical Models in Computer Vision*, Springer:New York, 2005.

[12] T. Zhang, B. Fang, Y.Y. Tang, G. He, and J. Wen, "Topology preserving non-negative matrix factorization for face recognition," *IEEE Trans. Image Processing*, vol. 17, no. 4, pp. 574–584, 2009.

[13] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Review*, vol. 51, no. 3, pp. 455–500, September 2009.

[14] C. J. Lin, "On the convergence of multiplicative update for nonnegative matrix factorization," *IEEE Trans. Neural Networks*, vol. 18, no. 6, pp. 1589–1596, November 2007.

[15] 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.

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