Non-negative Tensor Factorization Applied to Music Genre Classification

Emmanouil Benetos, Student Member, IEEE and Constantine Kotropoulos, Senior Member, IEEE

Abstract—Music genre classification techniques are typically applied to the data matrix whose columns are the feature vectors extracted from music recordings. In this paper, a feature vector is extracted using a texture window of one sec, which enables the representation of any 30 sec long music recording as a time sequence of feature vectors, thus yielding a feature matrix. Consequently, by stacking the feature matrices associated to any dataset recordings, a tensor is created, a fact which necessitates studying music genre classification using tensors. First, a novel algorithm for non-negative tensor factorization (NTF) is derived that extends the non-negative matrix factorization. Several variants of the NTF algorithm emerge by employing different cost functions from the class of Bregman divergences. Second, a novel supervised NTF classifier is proposed, which trains a basis for each class separately and employs basis orthogonalization. A variety of spectral, temporal, perceptual, energy, and pitch descriptors is extracted from 1000 recordings of the GTZAN dataset, which are distributed across 10 genre classes. The NTF classifier performance is compared against that of the multilayer perceptron and the support vector machines by applying a stratified 10-fold cross validation. A genre classification accuracy of 78.9% is reported for the NTF classifier demonstrating the superiority of the aforementioned multilinear classifier over several data matrix-based state-of-the-art classifiers.

Index Terms—Non-negative tensor factorization, Bregman divergences, Music genre classification.

I. INTRODUCTION

CURRENT advances in multimedia data management and retrieval have enabled the creation, distribution, and availability of vast amounts of music data including new content as well as digitized one from analog archives. Aided by the growth of the internet, these databases have become highly popular for personal as well as commercial use (e.g. online music retailers or digital libraries). Accordingly, the demand for tools to analyze and retrieve music content has emerged, leading to flourishing music information retrieval (MIR) research.

Music genres are the most popular music content descriptors, since they are employed by both the users and the music industry [1]. One may argue that the genres summarize music recordings based on some common perceptual characteristics [2]. However, music genres have not yet been precisely defined, because they are primarily determined by users’ taste and may be culturally dependent. Not to mention that more than one music genres may be associated with a certain recording. Accordingly, the creation of a universal genre taxonomy remains still infeasible. Automatic genre classification techniques classify recordings into distinguishable genres by extracting relevant features and employing pattern recognition algorithms [3]. The accuracy of such genre classification techniques often exceeds that reported for humans with moderate music training [4]. However, the research on automatic genre classification appears to have reached a local maximum recently due to the lack of carefully annotated music corpora with ground truth [5].

Most genre classification approaches represent each music recording by a feature vector and consequently employ pattern recognition algorithms in order to perform classification. In this paper, each recording is represented by a time sequence comprising feature vectors extracted every one sec, thus forming a feature matrix. Starting with a comprehensive set of features measuring spectral, temporal, perceptual, energy, and pitch characteristics of the recordings, feature selection is applied next by using a branch-and-bound search strategy in order to determine the subset of the most discriminative features with respect to the ratio of the inter-class dispersion over the intra-class dispersion [9] and keep the number of the features to be processed into a manageable size. By stacking the feature matrices associated to recordings, a tensor is created, which provides a more detailed representation of music characteristics. Tensors are considered as extensions of matrices or vectors [6]–[8]. A novel non-negative tensor factorization (NTF) algorithm is proposed whose roots are traced back in the non-negative matrix factorization (NMF). The algorithm is able to decompose a tensor in Kruskal format [8]. That is, to decompose a tensor into a sum of elementary rank-1 tensors. The algorithm can employ several cost functions, which belong to the class of Bregman divergences [10]. The Bregman divergences have previously been used to solve the non-negative matrix approximation problem [11]. Details on the derivation of the tensor element update equations are provided and the computational cost of the algorithm is estimated. In addition, a novel supervised classifier based on the NTF is proposed, which trains a basis for each class separately and employs basis orthogonalization. The proposed classifier extends a similar classifier that was based on the NMF [34]. Preliminary results for music genre classification using the NTF classifier with the Frobenius norm as cost function were reported in [33]. Here, starting from a larger feature set than that used in [33], results are reported on extended experiments performed on the GTZAN database, which contains 1000 music recordings covering...
10 music genre classes [12]. Several variants of the NTF classifier employing different feature subset sizes are tested and their music genre classification accuracy is measured using a stratified 10-fold cross-validation. For comparison purposes, support vector machines and multilayer perceptrons are also tested on the same dataset. In addition, experiments are performed using the features extracted by the MARSYAS platform [39]. An average genre classification accuracy of 78.9% with standard deviation equal to 2.6% is reported for the NTF classifier, when the Frobenius norm is utilized with a subset of 80 features. The aforementioned accuracy places the proposed NTF classifier within the most performing state-of-the-art genre classification methods. The superiority of the NTF classifier against the state-of-the-art data matrix-based classifiers is also demonstrated. Such results motivate further research using tensorial representations into audio processing applications.

The outline of the paper is as follows. Related work on automatic music genre classification is discussed in Section II. Section III details the proposed NTF method and establishes links with the NMF as well as other methods proposed for the NTF. In this section, a supervised classifier based on the proposed NTF is also described. Section IV briefly presents the dataset used, the feature set employed in the experiments, and thoroughly assesses the music genre classification accuracy of the proposed NTF classifier against that of state-of-the-art classifiers. Conclusions are drawn and future directions are indicated in Section V.

II. RELATED WORK

Several benchmark datasets have been collected making the performance of the various music genre classification approaches comparable. Such benchmark datasets are listed in Table I along with the best accuracy reported for state-of-the-art classifiers in chronological order. The GTZAN dataset was introduced by Tzanetakis and Cook [12]. It contains 1000 audio recordings split into 10 genre classes. The parameters of a Gaussian mixture model (GMM) classifier were estimated by the iterative expectation-maximization (EM) algorithm in [12]. A 61% correct classification was reported for timbre, rhythmic, and pitch features. The same dataset was used by Li et al. who employed support vector machines (SVMs) and linear discriminant analysis (LDA) for classification [13]. The Daubechies wavelet coefficient histograms were used as features and the reported classification accuracy of the SVM classifier reached 78.5%. Lidy and Rauber employed a pairwise SVM classifier applied to the GTZAN dataset [3]. The extracted features include rhythm patterns, a statistical spectrum descriptor, and rhythm histogram features. The reported best classification accuracy was 74.9%. Bergstra et al. tested the mel-frequency cepstral coefficients (MFCCs), the fast Fourier Transform coefficients, the linear prediction coefficients (LPCs), and the zero-crossing rate (ZCR) on the GTZAN dataset [14] and reported a classification accuracy reaching 82.5% for the ADABOOST meta-classifier. It should be noted, however, that the classification accuracy in [14] was measured without cross-validation. Holzapfel and Stylianou also employed the same dataset. By utilizing a spectral basis derived by the NMF, that is fed to a GMM classifier, they obtained a 74.0% classification accuracy [15].

Another collection used extensively is the MIREX 2004 dataset, released for the MIREX genre and rhythm classification contests [16]. The MIREX 2004 genre dataset contains 1458 recordings belonging into 6 genre classes. The best accuracy (i.e. 83.5%) was reported by Holzapfel and Stylianou [15]. Pampalk et al. used the nearest neighbor (NN) classifier with GMMs and obtained an accuracy of 82.3% [17]. In our experiments, we are confined to the GTZAN dataset, because it contains more genre classes than the MIREX 2004 one, thus being a more comprehensive dataset for genre classification.

Other notable music genre classification approaches include that of Burred and Lerch, who proposed a 3-level music genre taxonomy covering 13 genres [18]. In addition, 3 speech classes, and one class for background noise were also considered. A dataset was created containing 50 recordings for each genre. Timbral and rhythmic features were extracted along with MPEG-7 audio descriptors. A GMM classifier reached classification accuracy of 59.76% for all classes. In 2005, Meng et al. created two datasets for genre classification [19]. The first dataset contains 100 recordings from 5 genres and the second dataset 354 music samples from the amaz-on.com database. Short, medium, and long-time features were extracted, and two classifiers were tested. The first classifier was a single-layer neural network and the second one was a Gaussian classifier that employs full covariance matrices. The reported best accuracy was 95% on the first dataset and about 68% on the second dataset. More recently, Barbedo and Lopes proposed a 4-level hierarchical genre taxonomy covering 29 music genres [20]. Several timbre features were selected and a classification procedure was developed that uses pairwise genre comparison. Overall, a genre classification accuracy of 61% at the lowest genre level was reported. Finally, Cataltepe et al. employed 225 MIDI music pieces covering 9 genre classes [21]. Timbral, rhythmic, and pitch content features were extracted, and the recordings were classified using a 10-nearest neighbor classifier (10-NN) or a normalized compression distance classifier. Using a combination of the aforementioned classifiers, a genre classification accuracy of 62% was reported.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Dataset</th>
<th>Classifier</th>
<th>Best Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>[12]</td>
<td>GTZAN</td>
<td>GMM</td>
<td>61.0</td>
</tr>
<tr>
<td>[13]</td>
<td>GTZAN</td>
<td>SVM - LDA</td>
<td>78.5</td>
</tr>
<tr>
<td>[3]</td>
<td>GTZAN</td>
<td>SVM</td>
<td>74.9</td>
</tr>
<tr>
<td>[14]</td>
<td>GTZAN</td>
<td>ADA BOOST</td>
<td>82.5</td>
</tr>
<tr>
<td>[15]</td>
<td>GTZAN</td>
<td>GMM</td>
<td>74.0</td>
</tr>
<tr>
<td>[17]</td>
<td>MIREX 2004</td>
<td>NN - GMM</td>
<td>82.3</td>
</tr>
<tr>
<td>[3]</td>
<td>MIREX 2004</td>
<td>SVM</td>
<td>70.4</td>
</tr>
</tbody>
</table>
III. NON-NEGATIVE TENSOR FACTORIZATION

In this section, a novel non-negative tensor factorization (NTF) technique is developed. First, the non-negative matrix factorization (NMF) is briefly discussed, because the NTF could be treated as a high-order generalization of the NMF for tensorial data. Some definitions from tensor algebra follow and the motivation for using tensors is described. Previous NTF algorithms are reviewed next and the proposed one is detailed. Contrary to previous approaches, the proposed NTF algorithm is not limited to 3rd order tensors, but can be applied to nth (n > 3) order tensors. In addition, the algorithm can be formulated using a variety of objective functions. Obviously, its use is not restricted to audio processing only. Finally, a novel supervised classifier based on NTF for 3rd order tensors is proposed, which performs separate training for each class and employs basis orthogonalization.

Throughout the paper, tensors are denoted by boldface Euler script calligraphic letters (e.g. $\mathbf{A}$), matrices are denoted by uppercase boldface letters (e.g. $\mathbf{U}$), and vectors are denoted by lowercase boldface letters (e.g. $\mathbf{u}$). The elements of all the aforementioned mathematical structures are denoted by lowercase letters indexed by one or more indices. For example, the elements of matrix $\mathbf{U}$ are denoted as $u_{ij}$. Let $\mathbb{R}$ and $\mathbb{Z}$ be the sets of real and integer numbers, respectively.

A. Non-negative Matrix Factorization

Subspace analysis seeks low dimensional structures of patterns within high dimensional spaces. NMF is a subspace method able to obtain a parts-based representation of objects by imposing non-negative constraints. It was first introduced as positive matrix factorization by Paatero et al. [22] and was re-termed as NMF by Lee and Seung [23]. The problem addressed by the NMF is as follows. Given a non-negative real-valued data matrix $\mathbf{V} \in \mathbb{R}^{n \times m}$, find the non-negative matrix factors $\mathbf{W} \in \mathbb{R}^{n \times k}$ and $\mathbf{H} \in \mathbb{R}^{k \times m}$ so that

$$\mathbf{V} \approx \mathbf{WH} = \sum_{i=1}^{k} \mathbf{w}_i \mathbf{h}_i^T \triangleq \sum_{i=1}^{k} \mathbf{w}_i \circ \mathbf{h}_i$$

(1)

where $\circ$ stands for the outer product. Obviously, $\mathbf{w}_i$ are the columns of $\mathbf{W}$ and $\mathbf{h}_i^T$ are the rows of $\mathbf{H}$. $\mathbf{W}$ contains the basis vectors, while the column vectors of $\mathbf{H}$ contain the weights needed to properly approximate the corresponding column vector of $\mathbf{V}$ as a linear combination of the column vectors of $\mathbf{W}$. Usually, $k$ is chosen so that $(n + m)k < nm$, thus resulting in a compressed version of the original data matrix. To find the approximate factorization (1), a suitable objective function has to be minimized. Let $\mathbf{V} = [v_{ij}]$ and $\mathbf{WH} \triangleq \mathbf{Y} = [y_{ij}]$. In [23], the generalized Kullback-Leibler (KL) divergence between $\mathbf{V}$ and $\mathbf{WH}$ was used:

$$D(\mathbf{V} \| \mathbf{WH}) = \sum_{i=1}^{n} \sum_{j=1}^{m} [v_{ij} \log \frac{v_{ij}}{y_{ij}} - v_{ij} + y_{ij}]$$

(2)

The minimization of (2) can be solved by using iterative multiplicative rules [23]. Frequently, additional constraints are incorporated into (2). For example, the local NMF algorithm imposes spatial locality in the solution and consequently reveals local features in the data matrix $\mathbf{V}$ [24].

A more general view of the NMF is set under the so-called non-negative matrix approximation (NNMA) in [11]. In NNMA, instead of minimizing a specific objective function, the minimization of a class of objective functions, called Bregman divergences, is proposed. The same approach will be adopted for the derivation of the NTF in Subsection III-D.

B. Tensors and Multilinear Algebra Basics

Quantities addressed by more than two indices are often employed in signal processing applications. To describe such quantities, tensors need to be employed. In multilinear algebra, tensors are considered as high-order generalizations of matrices and vectors [6]–[8]. A real-valued vector $\mathbf{a} \in \mathbb{R}^I$, $I \in \mathbb{Z}$ is treated as a first-order tensor. Similarly, a real-valued matrix $\mathbf{A} \in \mathbb{R}^{I_1 \times I_2}$ with $I_1, I_2 \in \mathbb{Z}$ is defined as a second-order tensor. A real-valued tensor $\mathbf{A}$ of order $n$ is defined over the vector space $\mathbb{R}^{I_1 \times I_2 \times \cdots \times I_n}$, where $I_i \in \mathbb{Z}$, $i = 1, \ldots, n$. Each element of $\mathbf{A}$ is addressed by $n$ indices, i.e. $a_{i_1 \ldots i_n}$, where $i_1 = 1, 2, \ldots, I_1$. A 3rd order tensor is sketched in Figure 1.

Mode-$i$ unfolding of the tensor $\mathbf{A}$ yields the matrix $\mathbf{A}_{(i)} \in \mathbb{R}^{I_i \times \prod_{j \neq i} I_j}$, where $\prod_{j \neq i} I_j \equiv I_1 I_2 \cdots I_{i-1} I_{i+1} \cdots I_n$. In the following, the operations on tensors are expressed in matricized form [8]. The symbol $\times_i$ stands for the $i$-mode product between a tensor and a matrix [6]. It can be computed via the matrix multiplication $\mathbf{B}_{(i)} = \mathbf{UA}_{(i)}$ followed by re-tensorization to undo the mode-$i$ unfolding for $i = 1, 2, \ldots, n$. For example, the 2-mode product between the 3rd order tensor $\mathbf{A} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ and the matrix $\mathbf{U} \in \mathbb{R}^{I_2 \times J}$ yields the 3rd order tensor $\mathbf{B} = \mathbf{A} \times_2 \mathbf{U} \in \mathbb{R}^{I_1 \times J \times I_3}$. The inner product of two nth order tensors $\mathbf{A}$ and $\mathbf{B}$ is denoted as $\langle \mathbf{A}, \mathbf{B} \rangle$.

The norm of tensor $\mathbf{A}$ is defined as $\|\mathbf{A}\| = \sqrt{\langle \mathbf{A}, \mathbf{A} \rangle}$.

An $n$-order tensor $\mathbf{A}$ has rank 1, when it can be decomposed as the outer product of $n$ vectors $\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \ldots, \mathbf{u}^{(n)}$, i.e.

$$\mathbf{A} = \mathbf{u}^{(1)} \circ \mathbf{u}^{(2)} \circ \cdots \mathbf{u}^{(n)}.$$  

(3)

That is, each element of the tensor in (3) is given by $a_{i_1 \ldots i_n} = u_{i_1}^{(1)} u_{i_2}^{(2)} \cdots u_{i_n}^{(n)}$ for all $i_\ell = 1, 2, \ldots, I_\ell$ and $\ell = 1, 2, \ldots, n$. The rank of an arbitrary $n$-order tensor $\mathbf{A}$ is the minimal number of rank-1 tensors that yield $\mathbf{A}$ when linearly combined.

In the following, several products between matrices shall be needed, such as the Kronecker product denoted by $\otimes$ and the
Khatrier-Rao product denoted by ⊙, whose definitions can be found elsewhere, e.g. [8].

C. Motivation for Using Tensors and the Proposed Non-negative Tensor Factorization

The NMF has been used extensively in signal processing yielding promising results in the past years. A list of the numerous NMF applications can be found in [11]. However, the NMF as any other subspace method deals only with vectorized data. By vectorizing a typical 3rd order tensor stemming from 900 training recordings, which are represented by 30 feature vectors of 34 dimensions each, one obtains 900 vectors of 1020 dimensions. Many pattern classifiers cannot cope with the aforementioned dimensionality given the small number of training samples. In addition, handling such high-dimensional samples is computationally expensive. For example, eigen-analysis or singular value decomposition cannot be easily performed. Despite implementation issues, it is well understood that vectorization breaks the natural structure and correlation in the original data. Thus, in order to preserve the natural data structure and correlation, dimensionality reduction operating directly on tensors rather than vectors is desirable. The concept of low-rank decomposition of high-order signal representations is addressed in [6], [25], where several algorithms are reviewed. Thus, a high-order generalization of the NMF could be of great importance in the analysis of such high-order signal/pattern representations.

Some NMF generalizations have been proposed mostly for 3rd order tensors in face detection or recognition applications. In 2005, Shashua and Hazan proposed a generalization of the NMF for nth order tensors [26]. The problem was formulated as the decomposition of a tensor into a sum of k rank-1 tensors using the Frobenius norm as distance. Multiplicative update rules were employed and an application to sparse image coding was discussed. Hazan et al. extended the previous work by employing the KL divergence (also known as relative entropy) as distance [27].

In 2006, Boutsidis et al. introduced an algorithm for 3rd order tensor decomposition called projected alternating least squares with initialization and regularization (PALSIR) [28]. This algorithm also employed the Frobenius norm as distance and alternating least squares was used to derive the decomposition. Experiments were performed on eye image databases for biometric iris recognition applications. Heiler and Schnörr proposed a generalization of the sparse NMF algorithm for 3rd order tensors applied to face detection [29]. The Frobenius norm was used as distance in this case, too. The algorithm was termed as sparsity-constrained NTF, because a sparsity maximization algorithm was employed.

In 2007, Cichocki proposed algorithms for 3rd order NTF using alpha and beta divergences [30]. These algorithms employed alternating interior-point gradient and fixed point alternating least squares techniques incorporating sparsity constraints into the decomposition. The just described NTF method was incorporated into multilayer networks in order to improve the performance of multi-way blind source separation in EEG [31]. It should be noted that this model cannot be generalized to higher order tensors nor can degenerate to the NMF model for 2nd order tensors.

In this paper, we would like to derive a generic unified NTF algorithm, which can handle nth order tensors and can degenerate to the NMF, when \( n = 2 \).

D. Proposed NTF Algorithm

Having set our objectives, we decided to build upon the model proposed by Shashua and Hazan [26], which can be extended to nth order tensors and degenerates to the NMF, when \( n = 2 \). A preliminary version of the algorithm was first introduced in [33], which was also the basis for the Discriminant NTF algorithm in [32]. We resort to the Bregman divergences in order to offer a unified factorization framework, which includes as special cases the Frobenius norm, the KL-divergence, and the Itakura-Saito (IS) distance. The Bregman divergences, proposed by Bregman in 1967 [10], are defined as

\[
D_\phi(x, y) = \phi(x) - \phi(y) - \phi'(y)(x - y),
\]  

where \( \phi(x) \) is a strictly convex function defined on a convex set \( S \subseteq \mathbb{R} \) and \( \phi'(y) \) denotes the first derivative of \( \phi \) evaluated at \( y \). By definition, the Bregman divergences are non-negative [11] and can be extended to tensors. Let us consider the nth order tensors \( \mathbf{X}, \mathbf{Y} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_n} \).

The following identity holds:

\[
D_\phi(\mathbf{X}, \mathbf{Y}) = \sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \cdots \sum_{i_n=1}^{I_n} D_\phi(x_{i_1, i_2, \ldots, i_n}, y_{i_1, i_2, \ldots, i_n}).
\]  

For \( \phi(x) = \frac{1}{2} x^2 \), \( D_\phi(x, y) \) corresponds to the Frobenius norm.

For \( \phi(x) = x \log(x) \), the Bregman divergence coincides with the KL divergence, whereas for \( \phi(x) = -\log(x) \), the resulting \( D_\phi(x, y) \) is recognized to be the Itakura-Saito (IS) distance.

Therefore, our goal is to decompose a tensor \( \mathbf{V} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_n} \) into a sum of \( k \) rank-1 tensors:

\[
\mathbf{V} = \sum_{j=1}^{k} \mathbf{u}^{(j)} \odot \mathbf{u}^{(j)} \odot \cdots \odot \mathbf{u}^{(j)}
\]  

where \( \mathbf{u}^{(j)} \in \mathbb{R}^{I_1}, j = 1, 2, \ldots, k \). Let \( \mathbf{U}^{(i)} \triangleq [\mathbf{u}^{(i)}_{1}, \mathbf{u}^{(i)}_{2}, \ldots, \mathbf{u}^{(i)}_{n}] \), \( i = 1, 2, \ldots, n \). Obviously \( \mathbf{U}^{(i)} \in \mathbb{R}^{I_1 \times k} \).

Let us introduce the compact notation

\[
\bigoplus_i \mathbf{U} \triangleq \mathbf{U}^{(1)} \odot \cdots \odot \mathbf{U}^{(n)} \odot \mathbf{U}^{(i+1)} \odot \mathbf{U}^{(i-1)} \odot \cdots \odot \mathbf{U}^{(1)}.
\]  

In matricized form, the factorization (6) can be written as:

\[
\mathbf{V}^{(i)} = \mathbf{W}^{(i)} \bigoplus_i \mathbf{U}^{(i+1)} \odot \cdots \odot \mathbf{U}^{(n)} \odot \mathbf{U}^{(i-1)} \odot \cdots \odot \mathbf{U}^{(1)} \mathbf{H}^{(i)}
\]  

From the inspection of (8), one may readily see that the NMF results if \( \mathbf{W} = \mathbf{W}^{(1)} = \mathbf{U}^{(1)} \) and \( \mathbf{H} = \mathbf{H}^{(1)} \mathbf{H}^{(1)T} = \mathbf{U}^{(2)} \mathbf{U}^{(2)T} \).

Let

\[
\sum_{g_0=1}^{g_1} \sum_{g_2=1}^{g_2} \cdots \sum_{g_{n-1}=1}^{g_{n-1}} \sum_{g_n=1}^{g_n} \sum_{g_{n+1}=1}^{g_{n+1}} \cdots
\]
The following minimization problem with Bregman divergences is solved
\[
\min_{u^{(i)}_j} D_{\phi}(\sum_{j=1}^{k} u^{(i)}_j \circ u^{(2)}_j \circ \cdots \circ u^{(n)}_j, V)
\]
(10)
by using auxiliary functions, as is analyzed in Appendix I. In particular, for the KL divergence (i.e. \(\phi(x) = x \log(x)\)), the following multiplicative update rule is obtained for the elements of \(u^{(i)}_j\):
\[
u^{(i)}_{jl} \leftarrow \nu^{(i)}_{jl} \exp\left(\sum_{p=0}^{k} \alpha_{jpl} \circ \cdots \circ \beta_{j1l} \circ \cdots \circ \beta_{jkl} \circ \nu^{(i)}_{jl} \circ \cdots \circ \nu^{(i)}_{jn}\right) / \sum_{p=0}^{k} \alpha_{jpl} \circ \cdots \circ \beta_{j1l} \circ \cdots \circ \beta_{jkl} \circ \nu^{(i)}_{jl} \circ \cdots \circ \nu^{(i)}_{jn}
\]
(11)
where \(\nu^{(i)}_{jl}\) is the \(l\)th element of vector \(u^{(i)}_j\) before updating, \(j = 1, 2, \ldots, k\), \(i = 1, 2, \ldots, n\), \(\alpha_{jpl}\) and \(\beta_{jpl}\) are factors determined by the update rules (11), (14), or (15) applied to the column vectors \(u^{(i)}_j\) of matrix \(U^{(i)}_{j}\), where \(u^{(i)}_j\) is considered. The feature matrix is arranged to a \(k\) x \(n\) matrix, where the transpose of matrix \(V^{(i)}_{c}\) is denoted as \(V^{(i)}_{c}^T\).
(13)

Finally, for \(\phi(x) = -\log(x)\) (i.e. when the IS distance is employed), the update rule is:
\[
u^{(i)}_{jl} \leftarrow \nu^{(i)}_{jl} \sum_{p=0}^{k} \frac{\alpha_{jpl} \circ \cdots \circ \beta_{j1l} \circ \cdots \circ \beta_{jkl} \circ \nu^{(i)}_{jl} \circ \cdots \circ \nu^{(i)}_{jn}}{\sum_{p=0}^{k} \alpha_{jpl} \circ \cdots \circ \beta_{j1l} \circ \cdots \circ \beta_{jkl} \circ \nu^{(i)}_{jl} \circ \cdots \circ \nu^{(i)}_{jn}}
\]
(14)

In order to apply the aforementioned NTF algorithms to an \(n\)th order tensor \(V\), the \(n\) matrices \(U^{(i)}_{j}\), \(i = 1, 2, \ldots, n\), should be initialized by random numbers between 0 and 1. The update rules (11), (14), or (15) are applied to the column vectors \(u^{(i)}_j\) of matrix \(U^{(i)}_{j}\), \(j = 1, 2, \ldots, k\). The proof of convergence for the Frobenius NTF algorithm can be found in Appendix I. The computational cost of the various NTF algorithms is derived in Appendix II.

**E. Proposed 3rd Order NTF Classifier**

The novel NTF classifier for 3rd order tensors discussed next was inspired by the NMF classifier proposed in [34], where a basis for each class was trained separately and the test data were projected onto an orthogonalized basis. Preliminary results using the proposed classifier for 3rd order tensors in music genre classification were reported in [33]. Let \(C\) be the number of genre classes. The proposed 3rd order NTF classifier consists of a tensor \(V_c \in \mathbb{R}^{I_1 \times I_2 \times I_3}\), with \(I_1\) being the number of training recordings in class \(c\), \(c = 1, 2, \ldots, C\) (i.e. 90 for stratified 10-fold cross-validation in the GTZAN dataset), \(I_2\) being the dimensionality of feature vectors, \(I_3 = 30\) being the number of feature vectors extracted per recording (i.e. the number of 1 sec segments each recording is split to). The algorithm steps are as follows:

1. Decompose the training tensor for each genre \(V_c \in \mathbb{R}^{I_1 \times I_2 \times I_3}\), \(c = 1, 2, \ldots, C\), i.e.
\[
V_c = \sum_{j=1}^{k} u^{(1)}_{cj} \circ u^{(2)}_{cj} \circ u^{(3)}_{cj}.
\]
(16)

2. Determine the 1st mode of the tensor \(V_c\) by unfolding [6], [8]:
\[
V^{(1)}_{c} = U^{(1)}_{c} \left( U^{(3)}_{c} \circ U^{(2)}_{c} \right)^T
\]
(17)
where \(\circ\) stands for the Kronecker matrix product. Thus, \(U^{(3)}_{c} \circ U^{(2)}_{c}\) has dimensions \((I_3 I_2) \times k\), where \(V^{(1)}_{c}\) is a matrix with dimensions \((I_1 \times I_2 I_3)\). In the following, we deal with the transpose of matrix \(V^{(1)}_{c}\), i.e.
\[
[V^{(1)}_{c}]^T = \left( U^{(3)}_{c} \circ U^{(2)}_{c} \right) \left( U^{(1)}_{c} \right)^T.
\]
(18)

3. Perform QR decomposition on the basis matrix \(U^{(3)}_{c} \circ U^{(2)}_{c}\):
\[
U^{(3)}_{c} \circ U^{(2)}_{c} = Q_{c} R_{c}
\]
(19)
where \(Q_{c}\) is a \((I_3 I_2) \times k\) column-orthogonal matrix (i.e. \(Q_{c}^T Q_{c}\) is the \(k \times k\) identity matrix) and \(R_{c}\) is a \(k \times k\) upper triangular matrix. Store matrices \(Q_{c}\) and \(H_{c}\) as \(R_{c} = [U^{(1)}_{c}]^T\). It is worth noting that the Gram-Schmidt orthogonalization does not affect the non-negativity of the basis matrix. It is used to calculate correctly the \(L_2\) norms in a non-orthogonal basis.

4. For testing, the feature matrix \(V_t\) of dimensions \(I_2 \times I_3\) is considered. The feature matrix is arranged to a column vector \(V_t\) of dimensions \(I_2 I_3\) by concatenating its columns. The column vector \(V_t\) is projected onto the subspaces defined by the basis matrices of the classes:
\[
h_{ct} = Q_{c}^T V_t
\]
(20)
and has length \(k\).

5. Let \(CSM_{mn}(c)\) be the cosine similarity measure (CSM) between \(h_{ct}\) and \(h_{cm}\), \(m = 1, 2, \ldots, I_1\) (i.e. the \(m\)th column of matrix \(H_{c}\)):
\[
CSM_{mn}(c) = \frac{h_{ct}^T h_{cm}}{||h_{ct}|| \cdot ||h_{cm}||}
\]
(21)
Let \(CSM_{m}(c)\) denote the \(m\)th largest element in the set \{\(CSM_{mn}(c)\), \(m = 1, 2, \ldots, I_1\}\}. The decision taken by the classifier is based on
\[
\{ c_{\hat{c}} = \arg\max_{c} \left\{ \{ c_{\hat{c}} \} \right\} \}
(22)
where \(K \ll I_1\) (e.g. \(K = 3\)). The class label of the test pattern \(V_t\) is determined by the maximum among \(\{c_{\hat{c}}(c)\}\), i.e.:
In this Section, music genre classification experiments are discussed. In subsection IV-A, the employed dataset is described. The feature extraction is detailed in subsection IV-B, while the feature selection method is discussed in subsection IV-C. Finally, the accuracy using the various classifiers is reported in subsection IV-D.

A. Dataset

The GTZAN database was employed for genre classification experiments. The database contains 1000 audio recordings distributed across 10 music genres [12], namely: Classical, Country, Disco, HipHop, Jazz, Rock, Blues, Reggae, Pop, and Metal. 100 recordings are collected for each genre. All recordings are mono channel, are sampled at 22050 Hz rate, and have a duration of approximately 30 sec. Each recording is separated into 30 segments (i.e. texture windows) of 1 sec duration. Such a texture window has commonly been used in genre classification experiments, because it increases the classification accuracy compared to direct analysis frames [2], [12], [19]. For each 1 sec long texture window, 207 features are extracted, which are described next.

B. Feature Extraction

In music genre classification experiments, the extracted features usually belong into 3 categories, namely timbre, rhythm, and pitch-based ones [1], [2]. In this paper, a combination of descriptors measuring energy, spectral, temporal, perceptual, and pitch characteristics of the music recordings is explored [35]. The complete list of the extracted features can be found in Table II.

The 1st feature measures the energy of the audio signal. Feature 2 is computed by maximum likelihood harmonic matching. Features 3 and 4 refer to the perceptual modeling of the human auditory system [17]. The spectral shape is captured by features 5-9 and 14-15. The temporal properties of the signals are correlated with features 10-13 and 16. Feature 17 describes the amplitude of the maximum peak of the folded histogram [36]. Feature 18 was proposed in [37]. Features 1, 2, 5, 7, 11, and 12 were computed using the definitions of the MPEG-7 audio framework [38]. It should be noted that 24 Mel-frequency cepstral coefficients and 8 specific loudness sensation (SONE) coefficients are extracted for each audio frame of 10 msec duration.

Except features 10-12, the remaining features are computed on frame basis and their 1st and 2nd moments are exploited by averaging over the frames within each 1 sec long texture window. Similarly, the 1st and 2nd moments of the first-order frame-based feature differences are computed. This explains the factor 4 appearing in Table II. In total, 207 features are extracted from each texture window. All features but the MFCCs are non-negative. Accordingly, they can be employed directly into the NTF. For the MFCCs, their magnitude is retained only. The computation of the aforementioned features every 1 sec yields the tensor $V$ of dimensions $1000 \times 207 \times 30$.

For comparison purposes, a smaller feature set is also tested, which includes the features extracted by the Music Analysis, Retrieval and Synthesis for Audio Signals (MARSYAS) platform [39]. This feature set consists of the 1st order moments of the following timbral features: Spectral Centroid, Spectral Rolloff Frequency, Spectral Difference (also known as spectral flux), and 30 MFCCs per frame, which are averaged over 1 sec texture windows. Thus, the tensor of the MARSYAS features has dimensions $1000 \times 34 \times 30$.

C. Feature Selection

Careful feature selection is essential for classification. Here, the optimal feature subset maximizes the ratio of the inter-class dispersion over the intra-class dispersion: $J = \frac{\text{tr}(S_C^{-1}S_B)}{\text{tr}(S_B)}$, where $\text{tr}(\cdot)$ stands for the trace of a matrix, $S_C$ is the within-class scatter matrix, and $S_B$ is the between-class scatter matrix. Details on the computation of $S_C$ and $S_B$ can be found in any textbook on pattern recognition (e.g. [9]). Because, in our case, the number of distinct subsets having cardinality $I_2$ ($1 \leq I_2 \leq 207$) is $\binom{207}{I_2}$, the branch-and-bound search strategy is employed for complexity reduction. In this strategy, a tree structure of $207-I_2+1$ levels is created, where every node corresponds to a subset. The tree root corresponds to the full set (e.g. 207 features), while each leaf node corresponds to a subset of cardinality $I_2$. The branch-and-bound search
strategy traverses the structure using a depth-first search with backtracking [9].

In order to apply the feature selection algorithm, the data tensor should be transformed into a matrix by unfolding [6]. Thus, the unfolding \( V_2 \in \mathbb{R}^{207^2} \) is computed from tensor \( V \in \mathbb{R}^{1000 \times 207 \times 30} \). Several feature subsets were derived with respect to maximizing \( J \) comprising \( I_2 \in \{60, 70, 80, 90\} \) features out of the 207 initial features. The subset comprising 80 features listed in Table III is found to yield the highest music genre classification accuracy, when it is employed in the proposed NTF Frobenius classifier (cf. subsection IV-D). It is seen that 31 out of the 80 selected features are moments of theMFCCs or their first-order differences.

### D. Performance Assessment

Experiments were performed by employing various subsets of selected features as well as the MARSYAS feature set using a stratified 10-fold cross validation, which is widely used in genre classification experiments on the GTZAN dataset.

The rank of the 3rd order genre class-dependent tensor \( V_r \) should satisfy [8] (and references therein)

\[
k \leq \min \{ I_1 I_2, I_1 I_3, I_2 I_3 \}. \tag{24}
\]

Since by the experimental protocol \( I_1 \) and \( I_3 \) are fixed to 90 and 30, respectively, the inequality (24) implies \( k \leq 30I_2 \), if \( I_2 \leq 90 \). Various values of \( k \) were tested for the NTF algorithms within the proposed NTF classifier. The highest genre classification accuracy was obtained for the following values of \( k \): \( k = 55 \), when the feature subset comprises \( I_2 = 60 \) features; \( k = 61 \), when the number of selected features \( I_2 \) is 70 or 80; and \( k = 64 \), when \( I_2 = 90 \) features are selected. For the MARSYAS set, \( k \) was set to 22. The number of terms \( K \) taken into account in (22) was set to 3 for all NTF classifiers.

The performance of the NTF classifier was compared against that of multilayer perceptron (MLP) and SVMs. In particular, a 3-layered perceptron with the logistic activation function was used. Its training was performed by the backpropagation algorithm with learning rate equal to 0.3 and momentum equal to 0.2 for 500 training epochs. A multi-class SVM classifier with a 2nd order polynomial kernel with unit bias/offset was also tested [40]. The experiments with the aforementioned classifiers were conducted on the matrix unfolding \( V_2 \in \mathbb{R}^{207^2 \times 30000} \) using 10-fold cross validation, where \( I_2 \in \{60, 70, 80, 90\} \).

The average music genre classification accuracy achieved by the classifiers over the 10 folds is listed in Table IV, when either the subset of 80 selected features or the MARSYAS feature set was employed. In Figure 3, the average accuracy of the SVM, MLP, and NTF Frobenius classifier is plotted versus several subset cardinalities. From Figure 3, it can be seen that the highest average accuracy of 78.9% was obtained by the proposed NTF Frobenius classifier, when it is applied to the subset of 80 selected features listed in Table III. The standard deviation of the accuracy achieved by the NTF Frobenius classifier is found to be 2.60%. The aforementioned classification accuracy outperforms that reported in [12] (i.e. 61.0%), [3] (i.e. 74.9%), [15] (i.e. 74.0%), and slightly exceeds that reported in [13] (i.e. 78.5%). In [14], a greater classification accuracy than ours is reported (i.e. 82.5%) by employing boosting. However, since cross-validation was not used, the latter accuracy is not directly comparable with ours (i.e. 78.9%). The NTF classifier, when the Frobenius norm was used, attained a higher accuracy than that achieved by the SVM or the MLP for 80 selected features. This was not the case when 60 or 70 features were selected, as can be seen in Figure 3.

Concerning the classification accuracy when the full set of 207 features is used, it was measured 54.7%, 49.2%, and 40.8% for the NTF Frobenius, NTF KL, and NTF IS classifiers, respectively, with \( k \) set to 86. The corresponding accuracies for the SVM or the MLP classifiers were 51.8% and 53.5%, respectively.

The NTF Frobenius classifier outperforms the SVM for all subset feature cardinalities tested. The NTF classifier achieved a lower accuracy, when either the KL divergence or the IS distance was employed, than that when the Frobenius norm was used.

If the comparison is made across the sets of features employed, the inspection of Table IV reveals that the set of 80 features listed in Table III clearly performs better than the MARSYAS feature set within all classifiers. Using the MARSYAS features, the best classification accuracy of 69.1% was achieved by the MLP classifier. When the NTF Frobenius

### TABLE IV

<table>
<thead>
<tr>
<th>Classifier</th>
<th>80 Feature Subset</th>
<th>MARSYAS Feature Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>NTF Frobenius</td>
<td>78.9%</td>
<td>68.3%</td>
</tr>
<tr>
<td>SVM</td>
<td>77.2%</td>
<td>67.6%</td>
</tr>
<tr>
<td>MLP</td>
<td>77.0%</td>
<td>69.1%</td>
</tr>
<tr>
<td>NTF KL</td>
<td>70.4%</td>
<td>61.4%</td>
</tr>
<tr>
<td>NTF IS</td>
<td>63.6%</td>
<td>55.0%</td>
</tr>
</tbody>
</table>

![Fig. 3. Average music genre classification accuracy for the various feature subsets.](image)
classifier was used with MARSYAS features, the second best accuracy 68.3% was obtained. The superiority of the extracted features over the MARSYAS ones is partially attributed to the fact that the latter features roughly consist a subset of the former ones. For the MARSYAS features, the NTF classifier with either the KL divergence or the IS distance is less performing than the NTF classifier with the Frobenius norm.

Next, the statistical significance of the accuracy differences between the classifiers was addressed by employing the method described in [41], where the number of correctly classified patterns is assumed to be distributed according to the binomial distribution. It can easily be shown that the performance gains obtained by the NTF Frobenius classifier against the SVM and MLP classifiers is not statistically significant at 95% confidence level. On the contrary, the accuracy difference between the NTF classifier with the Frobenius norm and the same classifier, when either the KL divergence or the IS distance is used, is found to be statistically significant at 95% confidence level. It should be noted that the difference of 0.4% between the one-vs-the-rest SVMs [13] and the NTF Frobenius classifier is statistically insignificant as well. However, the performance gain obtained by the NTF Frobenius against the classifiers employed in [3], [12], [15] is statistically significant.

Insight to the performance of the NTF Frobenius, SVM, and MLP classifiers is offered by the confusion matrices averaged over the 10 splits determined by 10-fold stratified cross-validation, in Tables V, VI, and VII, respectively. The columns of the confusion matrix correspond to the predicted music genre and the rows to the actual one. For the NTF Frobenius classifier, most misclassifications occur among the Hiphop, Pop, and Rock genres. Concerning the SVM classifier, most misclassifications occur for Rock recordings, which are

<table>
<thead>
<tr>
<th>No.</th>
<th>Selected Feature</th>
<th>No.</th>
<th>Selected Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Mean of 2nd SF coefficient</td>
<td>41</td>
<td>Mean of 1st order difference of 2nd SF coefficient</td>
</tr>
<tr>
<td>2</td>
<td>Variance of 1st SONE</td>
<td>42</td>
<td>Variance of SS</td>
</tr>
<tr>
<td>3</td>
<td>Mean of SF</td>
<td>43</td>
<td>Mean of 8th SONE</td>
</tr>
<tr>
<td>4</td>
<td>Mean of 2nd SONE</td>
<td>44</td>
<td>Variance of 2nd SF coefficient</td>
</tr>
<tr>
<td>5</td>
<td>Mean of BW</td>
<td>45</td>
<td>Variance of 1st order difference of 20th MFCC</td>
</tr>
<tr>
<td>6</td>
<td>Variance of 1st order difference of 13rd MFCC</td>
<td>46</td>
<td>Variance of 9th MFCC</td>
</tr>
<tr>
<td>7</td>
<td>Mean of 12th MFCC</td>
<td>47</td>
<td>Variance of 1st order difference of 2nd MFCC</td>
</tr>
<tr>
<td>8</td>
<td>Mean of SC</td>
<td>48</td>
<td>Variance of 1st order difference of 10th MFCC</td>
</tr>
<tr>
<td>9</td>
<td>Mean of 1st SF coefficient</td>
<td>49</td>
<td>Mean of 5th MFCC</td>
</tr>
<tr>
<td>10</td>
<td>Mean of SS</td>
<td>50</td>
<td>Mean of 3rd MFCC</td>
</tr>
<tr>
<td>11</td>
<td>Variance of TL</td>
<td>51</td>
<td>Variance of 1st order difference of 1st SF coefficient</td>
</tr>
<tr>
<td>12</td>
<td>Variance of 1st order difference of 5th MFCC</td>
<td>52</td>
<td>7th AC coefficient</td>
</tr>
<tr>
<td>13</td>
<td>Mean of 3rd SF coefficient</td>
<td>53</td>
<td>Mean of SRF</td>
</tr>
<tr>
<td>14</td>
<td>Variance of 8th MFCC</td>
<td>54</td>
<td>Variance of 1st order difference of 24th MFCC</td>
</tr>
<tr>
<td>15</td>
<td>Variance of SD</td>
<td>55</td>
<td>3rd AC coefficient</td>
</tr>
<tr>
<td>16</td>
<td>Mean of RP</td>
<td>56</td>
<td>Variance of 1st order difference of 12th MFCC</td>
</tr>
<tr>
<td>17</td>
<td>Variance of PD</td>
<td>57</td>
<td>Mean of 6th SONE</td>
</tr>
<tr>
<td>18</td>
<td>Mean of 4th SF coefficient</td>
<td>58</td>
<td>4th AC coefficient</td>
</tr>
<tr>
<td>19</td>
<td>Variance of 1st order difference of SS</td>
<td>59</td>
<td>Variance of 9th MFCC</td>
</tr>
<tr>
<td>20</td>
<td>Variance of 1st order difference of 8th MFCC</td>
<td>60</td>
<td>Variance of 2nd SONE coefficient</td>
</tr>
<tr>
<td>21</td>
<td>Variance of 1st order difference of 11th MFCC</td>
<td>61</td>
<td>Variance of 1st MFCC</td>
</tr>
<tr>
<td>22</td>
<td>Variance of 1st order difference of 10th MFCC</td>
<td>62</td>
<td>Mean of 14th MFCC</td>
</tr>
<tr>
<td>23</td>
<td>Mean of 11th MFCC</td>
<td>63</td>
<td>Mean of 1st MFCC</td>
</tr>
<tr>
<td>24</td>
<td>Variance of 4th SF</td>
<td>64</td>
<td>Variance of 1st order difference of 4th SF coefficient</td>
</tr>
<tr>
<td>25</td>
<td>Mean of TL</td>
<td>65</td>
<td>Mean of 2nd MFCC</td>
</tr>
<tr>
<td>26</td>
<td>Variance of 1st order difference of 1st SONE</td>
<td>66</td>
<td>Variance of 1st order difference of 3rd SF coefficient</td>
</tr>
<tr>
<td>27</td>
<td>Mean of 1st order difference of RP</td>
<td>67</td>
<td>Variance of 1st order difference of 8th SONE coefficient</td>
</tr>
<tr>
<td>28</td>
<td>Variance of 3rd SONE</td>
<td>68</td>
<td>Variance of 1st order difference of 5th SONE coefficient</td>
</tr>
<tr>
<td>29</td>
<td>Mean of FF</td>
<td>69</td>
<td>Variance of 1st order difference of 3rd SONE coefficient</td>
</tr>
<tr>
<td>30</td>
<td>Variance of 1st order difference of 9th MFCC</td>
<td>70</td>
<td>Mean of 1st order difference of 1st SF coefficient</td>
</tr>
<tr>
<td>31</td>
<td>Mean of 1st order difference of 11th MFCC</td>
<td>71</td>
<td>Variance of STE</td>
</tr>
<tr>
<td>32</td>
<td>Variance of 1st order difference of RP</td>
<td>72</td>
<td>Variance of BW</td>
</tr>
<tr>
<td>33</td>
<td>Mean of 13th MFCC</td>
<td>73</td>
<td>Mean of 1st order difference of 12th MFCC</td>
</tr>
<tr>
<td>34</td>
<td>Variance of 1st order difference of SC</td>
<td>74</td>
<td>Mean of 16th MFCC</td>
</tr>
<tr>
<td>35</td>
<td>Mean of 1st order difference of 14th MFCC</td>
<td>75</td>
<td>Variance of 1st order difference of 7th MFCC</td>
</tr>
<tr>
<td>36</td>
<td>Variance of SC</td>
<td>76</td>
<td>Variance of SRF</td>
</tr>
<tr>
<td>37</td>
<td>Variance of 2nd MFCC</td>
<td>77</td>
<td>Variance of 1st order difference of 2nd SF coefficient</td>
</tr>
<tr>
<td>38</td>
<td>Variance of 1st SF coefficient</td>
<td>78</td>
<td>Variance of 7th MFCC</td>
</tr>
<tr>
<td>39</td>
<td>Variance of 3rd SF coefficient</td>
<td>79</td>
<td>Mean of 4th MFCC</td>
</tr>
<tr>
<td>40</td>
<td>Variance of 5th SONE</td>
<td>80</td>
<td>Mean of PD</td>
</tr>
</tbody>
</table>
TABLE V
AVERAGE CONFUSION MATRIX FOR THE NTF FROBENIUS CLASSIFIER USING THE 80 SELECTED FEATURES OVER THE 10 SPLITS DETERMINED BY 10-FOLD STRATIFIED CROSS-VALIDATION.

<table>
<thead>
<tr>
<th>Genre</th>
<th>Blues</th>
<th>Classical</th>
<th>Country</th>
<th>Disco</th>
<th>Hiphop</th>
<th>Jazz</th>
<th>Metal</th>
<th>Pop</th>
<th>Reggae</th>
<th>Rock</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blues</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Classical</td>
<td>88</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Country</td>
<td>2</td>
<td>0</td>
<td>87</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Disco</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>11</td>
<td>67</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Hiphop</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>11</td>
<td>67</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Jazz</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Metal</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Pop</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>15</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Reggae</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>11</td>
<td>67</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Rock</td>
<td>75</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

TABLE VI
AVERAGE CONFUSION MATRIX FOR THE SVM CLASSIFIER USING THE 80 SELECTED FEATURES OVER THE 10 SPLITS DETERMINED BY 10-FOLD STRATIFIED CROSS-VALIDATION.

<table>
<thead>
<tr>
<th>Genre</th>
<th>Blues</th>
<th>Classical</th>
<th>Country</th>
<th>Disco</th>
<th>Hiphop</th>
<th>Jazz</th>
<th>Metal</th>
<th>Pop</th>
<th>Reggae</th>
<th>Rock</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blues</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Classical</td>
<td>97</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Country</td>
<td>8</td>
<td>2</td>
<td>73</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Disco</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Hiphop</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Jazz</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Metal</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Pop</td>
<td>1</td>
<td>3</td>
<td>15</td>
<td>3</td>
<td>0</td>
<td>72</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Reggae</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>11</td>
<td>67</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Rock</td>
<td>72</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

TABLE VII
AVERAGE CONFUSION MATRIX FOR THE MLP CLASSIFIER USING THE 80 SELECTED FEATURES OVER THE 10 SPLITS DETERMINED BY 10-FOLD STRATIFIED CROSS-VALIDATION.

<table>
<thead>
<tr>
<th>Genre</th>
<th>Blues</th>
<th>Classical</th>
<th>Country</th>
<th>Disco</th>
<th>Hiphop</th>
<th>Jazz</th>
<th>Metal</th>
<th>Pop</th>
<th>Reggae</th>
<th>Rock</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blues</td>
<td>88</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Classical</td>
<td>93</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Country</td>
<td>1</td>
<td>1</td>
<td>74</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Disco</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Hiphop</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>80</td>
<td>0</td>
<td>2</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Jazz</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Metal</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Pop</td>
<td>1</td>
<td>0</td>
<td>6</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Reggae</td>
<td>7</td>
<td>0</td>
<td>19</td>
<td>8</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Rock</td>
<td>65</td>
<td>0</td>
<td>5</td>
<td>4</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>74</td>
</tr>
</tbody>
</table>

For example, tensors can be used for modeling any recording as a time series of potentially different genre labels. NTF offers an attractive framework for modeling data as a multilinear combination of features and can extract basis features enabling a greater interpretability of the basis contributions to the classification than SVMs and MLPs. Such factorizations can also be used for dimensionality reduction prior to the application of standard machine learning algorithms (e.g., SVMs). NTF is not limited to music genre classification, but it can be utilized in various cases associated with feature vectors computed over time, such as for multiple frequency estimation in audio recordings in order to provide a global spectral basis for a whole set instead of a basis for each recording [42].

In the future, the performance of NTF will be assessed on hierarchical music genre databases, which offer additional flexibility over the flat classification problem. In addition, the NTF algorithms could be enhanced by incorporating penalty functions into the factorization problem, which can control the outcome of the factorization. Finally, various initialization techniques similar to those proposed for the NMF algorithm [44], could be developed for the NTF algorithms aiming to reduce the number of iterations.

APPENDIX I

A. Problem Formulation

As stated in Section III-D, the following minimization problem is treated:

$$\min_{\mathbf{U}^{\phi}} D_{\phi}\left(\sum_{j=1}^{k} \mathbf{u}^{(1)}_{j} \odot \mathbf{u}^{(2)}_{j} \odot \cdots \odot \mathbf{u}^{(n)}_{j}, \mathbf{V}\right).$$

Let \( l = 1, 2, \ldots, I_l \) and \( i = 1, 2, \ldots, n \). The goal is to find a multiplicative updating rule for the elements of vectors \( \mathbf{u}^{(i)}_{j} \) denoted as \( u^{(i)}_{j,l} \), \( j = 1, 2, \ldots, k \). From (5), it can be seen that:

$$D_{\phi}\left(\sum_{j=1}^{k} \mathbf{u}^{(1)}_{j} \odot \mathbf{u}^{(2)}_{j} \odot \cdots \odot \mathbf{u}^{(n)}_{j}, \mathbf{V}\right) = \sum_{l=1}^{I_l} D_{\phi}\left(\sum_{j=1}^{k} \mathbf{u}^{(1)}_{j,l} \odot \mathbf{u}^{(2)}_{j,l} \odot \cdots \odot \mathbf{u}^{(n)}_{j,l}, \mathbf{V}_{\ell_l=l}\right)$$

where \( \mathbf{V}_{\ell_l=l} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_{l-1} \times I_{l+1} \times \cdots \times I_{\ell} \times \cdots \times I_N} \) is a sub-tensor with the \( \ell \)th index fixed to \( l \) whose elements are denoted by \( v_{\ell_l=l} \), \( \ell = 1, 2, \ldots, I_l \) and \( \ell = 1, 2, \ldots, i-1, i+1, \ldots, n \).

B. Auxiliary function

The minimization problem can be solved using auxiliary functions [11], [23]. Let \( F(\mathbf{u}^{(i)}_{j,l}) \) denote the divergence term in (25), i.e.

$$F(\mathbf{u}^{(i)}_{j,l}) = D_{\phi}\left(\sum_{j=1}^{k} \mathbf{u}^{(1)}_{j,l} \odot \mathbf{u}^{(2)}_{j,l} \odot \cdots \odot \mathbf{u}^{(i-1)}_{j,l} \odot \mathbf{u}^{(i)}_{j,l} \odot \cdots \odot \mathbf{u}^{(n)}_{j,l}, \mathbf{V}_{\ell_l=l}\right).$$

$$(26)$$
The application of (4) and (5) to (26) yields:

\[
F(\mathbf{u}_j^{(i)}) = \sum_{t_0} \phi \left( \sum_{j=1} u_j^{(i)} - \psi \left( u_j^{(i)} - i^{(i)} \right) \right) - \psi ' \left( u_j^{(i)} - i^{(i)} \right)
\]

where \(\psi(x) = \phi(x)\). The following auxiliary function for \(F(\mathbf{u}_j^{(i)})\) is proposed:

\[
G(\mathbf{u}_j^{(i)}, \mathbf{\tilde{u}}_j^{(i)}) = \sum_{t_0} \left( \sum_{j=1} \lambda_{j_1 \ldots j_n} \cdot \phi \left( \sum_{j=1} u_j^{(i)} - \psi \left( u_j^{(i)} - i^{(i)} \right) \right) \right)
\]

where

\[
\lambda_{j_1 \ldots j_n} = \sum_{t_0} \left( \sum_{j=1} u_j^{(i)} - \psi \left( u_j^{(i)} - i^{(i)} \right) \right)
\]

It can easily be shown that \(G(\mathbf{u}_j^{(i)}, \mathbf{\tilde{u}}_j^{(i)}) = F(\mathbf{u}_j^{(i)})\). In addition, using Jensen’s inequality for convex functions, it can be verified that \(G(\mathbf{u}_j^{(i)}, \mathbf{\tilde{u}}_j^{(i)}) \geq F(\mathbf{u}_j^{(i)})\). Accordingly, indeed \(G(\mathbf{u}_j^{(i)}, \mathbf{\tilde{u}}_j^{(i)})\) is an auxiliary function for \(F(\mathbf{u}_j^{(i)})\).

C. Minimization of the auxiliary function

In order to derive a multiplicative update rule for \(u_j^{(i)}\), the auxiliary function \(G(\mathbf{u}_j^{(i)}, \mathbf{\tilde{u}}_j^{(i)})\) should be minimized with respect to \(u_j^{(i)}\). The partial derivative of \(G(\mathbf{u}_j^{(i)}, \mathbf{\tilde{u}}_j^{(i)})\) with respect to \(u_j^{(i)}\) is set to zero:

\[
\frac{\partial G(\mathbf{u}_j^{(i)}, \mathbf{\tilde{u}}_j^{(i)})}{\partial u_j^{(i)}} = 0
\]

By replacing (29) into (30) and after performing some algebraic manipulations, we obtain:

\[
\frac{\partial G(\mathbf{u}_j^{(i)}, \mathbf{\tilde{u}}_j^{(i)})}{\partial u_j^{(i)}} = \sum_{t_0} \left( \psi \left( u_j^{(i)} - i^{(i)} \right) \right) - \psi \left( u_j^{(i)} - i^{(i)} \right) \cdot \psi \left( u_j^{(i)} - i^{(i)} \right) \cdot \left( u_j^{(i)} - i^{(i)} \right)
\]

The equation

\[
\frac{\partial G(\mathbf{u}_j^{(i)}, \mathbf{\tilde{u}}_j^{(i)})}{\partial u_j^{(i)}} = 0
\]

cannot be analytically solved for any \(\psi()\). If \(\psi()\) is assumed to be multiplicative as in [11] (i.e. \(\psi(x) = \phi(x)\)), the substitution of (31) in (32) yields the following update rule:

\[
u_j^{(i)} = 0
\]

(33)

D. Proof of Convergence

To prove the convergence of the multiplicative update rule for the Frobenius NTB algorithm (14), one needs to show that \(F(\mathbf{u}_j^{(i)}) \leq F(\mathbf{\tilde{u}}_j^{(i)})\). Using the definition of \(F(\mathbf{u}_j^{(i)})\), \(\alpha_{j_1 \ldots j_n}^{(i)}\) and \(\beta_{j_1 \ldots j_n}^{(i)}\) defined in (27), (12), and (13), respectively, it can be shown that:

\[
F(\mathbf{u}_j^{(i)}) - F(\mathbf{\tilde{u}}_j^{(i)}) = \sum_{j=1}^{k} \left( \sum_{j_1 \ldots j_n} \alpha_{j_1 \ldots j_n}^{(i)} \cdot u_{j_1 \ldots j_n}^{(i)} - \psi \left( u_{j_1 \ldots j_n}^{(i)} - i^{(i)} \right) \right) \cdot \sum_{j_1 \ldots j_n} \beta_{j_1 \ldots j_n}^{(i)} \cdot u_{j_1 \ldots j_n}^{(i)}
\]

Since the 2nd term is by definition non-negative, it suffices to prove that the first term in (35) is non-negative. For the Frobenius norm, \(\psi(x) = x\), a fact that facilitates further the derivations. Moving the denominator in (14) to the left hand side part and summing over \(j\), we get

\[
\sum_{j} \beta_{j_1 \ldots j_n}^{(i)} \cdot u_{j_1 \ldots j_n}^{(i)} \cdot \psi \left( u_{j_1 \ldots j_n}^{(i)} - i^{(i)} \right)
\]

\[
= \sum_{j} \left( \sum_{j_1 \ldots j_n} \alpha_{j_1 \ldots j_n}^{(i)} \cdot u_{j_1 \ldots j_n}^{(i)} \right) \cdot \beta_{j_1 \ldots j_n}^{(i)} \cdot u_{j_1 \ldots j_n}^{(i)}
\]
Using (36) into the algebraic manipulations of the first term in (35), we conclude that it is sufficient to prove
\[ \sum_{\phi_1} v_{\phi_1 \ldots \phi_{k-1} \phi_k} \left( \sum_{j=1}^{k} \alpha_{\phi_1 \ldots \phi_{k-1} \phi_k} x_{\phi_1 \ldots \phi_{k-1} \phi_k} \cdot u_{j}^{(i)} \right) \geq \sum_{\phi_1} \left( \sum_{j=1}^{k} \alpha_{\phi_1 \ldots \phi_{k-1} \phi_k} x_{\phi_1 \ldots \phi_{k-1} \phi_k} \cdot u_{j}^{(i)} \right)^{\frac{1}{2}}. \]

(37)

(14) implies that
\[ \sum_{\phi_1} \alpha_{\phi_1 \ldots \phi_{k-1} \phi_k} x_{\phi_1 \ldots \phi_{k-1} \phi_k} \cdot v_{\phi_1 \ldots \phi_{k-1} \phi_k} \cdot u_{j}^{(i)} = \sum_{\phi_1} \alpha_{\phi_1 \ldots \phi_{k-1} \phi_k} x_{\phi_1 \ldots \phi_{k-1} \phi_k} \cdot \beta_{\phi_1 \ldots \phi_{k-1} \phi_k} \cdot u_{j}^{(i)} \]

(38)

Using (38), the inequality (37) is rewritten as
\[ \sum_{\phi_1} \left( \sum_{j=1}^{k} \alpha_{\phi_1 \ldots \phi_{k-1} \phi_k} x_{\phi_1 \ldots \phi_{k-1} \phi_k} \cdot u_{j}^{(i)} \right)^{\frac{1}{2}} \cdot \beta_{\phi_1 \ldots \phi_{k-1} \phi_k} \cdot u_{j}^{(i)} \geq \sum_{\phi_1} \left( \sum_{j=1}^{k} \alpha_{\phi_1 \ldots \phi_{k-1} \phi_k} x_{\phi_1 \ldots \phi_{k-1} \phi_k} \cdot u_{j}^{(i)} \right). \]

Inequality (39) holds thanks to Lemma 4 [11], which concludes the proof.

**Appendix II**

The computational cost of the NTF is derived for \( n = 3 \) (3rd order tensors), when the Frobenius norm is used. We assume that one flop corresponds to a single floating point operation, i.e. a floating point addition or a floating point multiplication [43]. The computational cost is detailed in Table VIII. We explicitly calculate the cost for the computation of the matrix \( U^{(1)} \in \mathbb{R}^{1 \times k} \) having columns \( u_{j}^{(i)} \). The first entry refers to terms \( \alpha_{j \phi_1 \phi_2} \), defined by (12), which are \( kI_2I_3 \) in total and each of them requires 1 multiplication. The second entry refers to terms \( \beta_{j \phi_1 \phi_2} \), defined by (13), which are \( I_1I_2I_3 \) in total and each of them requires \( 2k \) multiplications and \( k - 1 \) additions. The cost for one update of \( u_{j}^{(i)} \) given by (14) is \( 2(I_2I_3 + 1) \) multiplications and \( 2(I_2I_3 - 1) \) additions, therefore in total \( 4I_2I_3 \). There are \( I_1k \) elements that should be computed. In total, one needs \( (7k - 1)I_1I_2I_3 + kI_2I_3 \) flops per iteration in order to compute the full matrix \( U^{(1)} \). By repeating the same computation for \( U^{(2)} \) and \( U^{(3)} \), and multiplying by the number of iterations \( r \) needed for convergence, we obtain:
\[ 3r(7k - 1)I_1I_2I_3 + r k(I_2I_3 + I_1I_3 + I_1I_2). \]

(40)

It can be said that the Itakura-Saito NTF and the Kullback-Leibler NTF algorithms have a computational cost of the same order to that given by (40), if a constant cost is assumed for the computation of logarithms and exponentials.

The inspection of (40) reveals that the cost of the non-negative training tensor \( V_r \) factorization depends linearly on the number of the training recordings \( I_{c1} \) for each genre class. Besides the NTF, the computational cost of the proposed NTF classifier training in Section III-E involves tensor unfolding (of no cost), the Khatri-Rao matrix product \( U^{(3)} \odot U^{(2)} \) its QR decomposition, and the matrix product \( H_c = R_c [U^{(1)}] T \) for each genre. The just mentioned Khatri-Rao matrix product is computed at a cost of \( I_3I_2k \) flops. The QR decomposition can be performed at a cost of \( 2I_2I_3k^2 \) flops, if the modified Gram-Schmidt method is used [43]. \( H_c \) can be computed at a cost of \( I_2(I_2 - 1)k \) flops. The test phase involves (20)-(22), which implies a computational cost of \( O(I_1I_2^2) \) for each genre.

**Table VIII**

<table>
<thead>
<tr>
<th>Term</th>
<th>Frobenius NTF Computational Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha_{j \phi_1 \phi_2} ) ( \beta_{j \phi_1 \phi_2} )</td>
<td>( 5k - I_1I_2I_3 )</td>
</tr>
<tr>
<td>( u_{j}^{(i)} )</td>
<td>( 2k(4I_2I_3 + 1) )</td>
</tr>
<tr>
<td>( I_1k ) per iteration</td>
<td>( 3I_1k(7k - 1)I_2I_3 + I_2I_3(7k - 1)I_1 + I_1I_2 )</td>
</tr>
</tbody>
</table>

**References**


Emmanouil Benetos received the B.Sc. degree in Informatics in 2005 and the M.Sc. degree in Digital Media in 2007, both from the Aristotle University of Thessaloniki. In 2008 he was with the Multimedia Informatics Lab at the Department of Computer Science, University of Crete. Currently, he is pursuing his Ph.D. at the Department of Electronic and Computer Science, Queen Mary University of London, in the field of automatic music transcription. His research interests include music and speech processing, music information retrieval, and computational intelligence. He is a member of the Alexander S. Onassis Scholars’ Association and a student member of IEEE.

Constantine Kotropoulos was born in Kavala, Greece in 1965. He received the Diploma degree with honors in Electrical Engineering in 1988 and the PhD degree in Electrical & Computer Engineering in 1993, both from the Aristotle University of Thessaloniki.

He is currently an Associate Professor in the Department of Informatics at the Aristotle University of Thessaloniki. From 1989 to 1993 he was a research and teaching assistant in the Department of Electrical & Computer Engineering at the same university. In 1995, he joined the Department of Informatics at the Aristotle University of Thessaloniki as a senior researcher and served then as a Lecturer from 1997 to 2001 and as an Assistant Professor from 2002 to 2007. He was a visiting research scholar in the Department of Electrical and Computer Engineering at the University of Delaware, USA during the academic year 2008-2009 and he conducted research in the Signal Processing Laboratory at Tampere University of Technology, Finland during the summer of 1993. He has co-authored 41 journal papers, 146 conference papers, and contributed 6 chapters to edited books in his areas of expertise. He is co-editor of the book “Nonlinear Model-Based Image/Video Processing and Analysis” (J. Wiley and Sons, 2001). His current research interests include audio, speech, and language processing; signal processing; pattern recognition; multimedia information retrieval; biometric authentication techniques, and human-centered multimodal computer interaction. Prof. Kotropoulos was a scholar of the State Scholarship Foundation of Greece and the Bodossaki Foundation. He is a senior member of the IEEE and a member of EURASIP, IAPR, and the Technical Chamber of Greece.

He is a member of the Editorial Board of Advances in Multimedia journal and serves as a EURASIP local liaison officer for Greece.