City Research Online

City St George's, University of London

Citation: Panagakis, I., Benetos, E. & Kotropoulos, C. (2008). Music genre classification: a multilinear approach. Paper presented at the International Symposium Music Information Retrieval, 14 - 18 September 2008, Philadelphia, USA.

This is the unspecified version of the paper.

This version of the publication may differ from the final published version. To cite this item please consult the publisher's version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/2109/

Copyright and Reuse: Copyright and Moral Rights remain with the author(s) and/or copyright holders. Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge, unless otherwise indicated, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way. For full details of reuse please refer to City Research Online policy.

City Research Online: http://openaccess.city.ac.uk/ publications@citystgeorges.ac.uk

MUSIC GENRE CLASSIFICATION: A MULTILINEAR APPROACH

Ioannis Panagakis, Emmanouil Benetos, and Constantine Kotropoulos
Department of Informatics
Aristotle University of Thessaloniki
Box 451 Thessaloniki GR-54124, Greece
E-mail: {panagakis, empeneto, costas}@aiia.csd.auth.gr

ABSTRACT

In this paper, music genre classification is addressed in a multilinear perspective. Inspired by a model of auditory cortical processing, multiscale spectro-temporal modulation features are extracted. Such spectro-temporal modulation features have been successfully used in various content-based audio classification tasks recently, but not yet in music genre classification. Each recording is represented by a third-order feature tensor generated by the auditory model. Thus, the ensemble of recordings is represented by a fourth-order data tensor created by stacking the third-order feature tensors associated to the recordings. To handle large data tensors and derive compact feature vectors suitable for classification, three multilinear subspace techniques are examined, namely the Non-Negative Tensor Factorization (NTF), the High-Order Singular Value Decomposition (HOSVD), and the Multilinear Principal Component Analysis (MPCA). Classification is performed by a Support Vector Machine. Stratified cross-validation tests on the GTZAN dataset and the ISMIR 2004 Genre one demonstrate the advantages of NTF and HOSVD versus MPCA. The best accuracies obtained by the proposed multilinear approach is comparable with those achieved by state-of-the-art music genre classification algorithms.

1 INTRODUCTION

To manage large music collections, tools able to extract useful information about musical pieces directly from audio signals are needed. Such information could include genre, mood, style, and performer [13]. Aucouturier and Pachet [1] indicate that music genre is probably the most popular description of music content. Classifying music recordings into distinguishable genres is an attractive research topic in Music Information Retrieval (MIR) community.

Most of the music genre classification techniques, employ pattern recognition algorithms to classify feature vectors, extracted from short-time recording segments into genres. In general, the features employed for music genre classification are roughly classified into three classes: timbral texture features, rhythmic features, and pitch content features [20]. Commonly used classifiers are Support Vector Machines (SVMs), Nearest-Neighbor (NN) classifiers, or classifiers, which resort to Gaussian Mixture Models, Linear Discriminant Analysis (LDA), etc. Several common audio datasets have been used in experiments to make the reported classification accuracies comparable. Notable results on music genre classification are summarized in Table 1.

Reference	Dataset	Accuracy
Bergstra *et al.* [4]	GTZAN	82.50%
Li *et al.* [12]	GTZAN	78.50%
Lidy *et al.* [14]	GTZAN	76.80%
Benetos *et al.* [3]	GTZAN	75.00%
Holzapfel *et al.* [6]	GTZAN	74.00%
Tzanetakis *et al.* [20]	GTZAN	61.00%
Holzapfel *et al.* [6]	ISMIR2004	83.50%
Pampalk *et al.* [19]	ISMIR2004	82.30%
Lidy *et al.* [13]	ISMIR2004	79.70%
Bergstra *et al.* [4]	MIREX2005	82.34%
Lidy *et al.* [14]	MIREX2007	75.57%
Mandel *et al.* [16]	MIREX2007	75.03%

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

Recently, within MIR community, genre has been criticized as being a hopelessly, ambiguous, and inconsistent way to organize and explore music. Consequently users' needs would be better addressed by abandoning it in favor of more generic music similarity-based approaches [17]. From another point of view, end users are more likely to browse and search by genre than either artist similarity or music similarity by recommendation [11]. Furthermore, Aucouturier *et al*. [2] have observed that recent systems, which assess audio similarity using timbre-based features, have failed to offer significant performance gains over early systems and in addition their success rates make them unrealistic for practical use. It is clear that new approaches are needed to make automatic genre classification systems viable in practice. McKay *et al*. [17] argues on the importance of continuing research in automatic music genre classification and

encourages the MIR community to approach the problem in a inter-disciplinary manner.

The novel aspects of this paper are as follows. First, we use bio-inspired multiscale spectro-temporal features for music genre classification. Motivated by the fact that each sound is characterized by slow spectral and temporal modulations and investigations on the auditory system [22], we use the auditory model proposed in [21] in order to extract features that map a given sound to a high-dimensional representation of its spectro-temporal modulations. The auditory high-dimensional representation can be viewed as a high-order tensor that is defined on a high-dimensional space. Note that in the field of multilinear algebra, tensors are considered as the multidimensional equivalent of matrices or vectors [8]. In addition, cortical representations are highly redundant [18]. Therefore, it is reasonable to assume that the tensors are confined into a subspace of an intrinsically low dimension.

Feature extraction or dimensionality reduction thus aims to transform such a high-dimensional representation into a low-dimensional one, while retaining most of the information related to the underlying structure of spectro-temporal modulations. Subspace methods are suitable for the aforementioned goal. Indeed subspace methods, such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Non-Negative Matrix Factorization (NMF) have successfully been used in various pattern recognition problems. The just mentioned methods deal only with vectorized data. By vectorizing a typical three-dimensional cortical representation of 6 scales $\times$ 10 rates $\times$ 128 frequency bands results in a vector having dimension equal to 7680. Many pattern classifiers, can not cope with such a high-dimensionality given a small number of training samples. In addition, handling such high-dimensional samples is computationally expensive. Therefore, eigen-analysis or Singular Value Decomposition cannot be easily performed. Despite implementation issues, it is well understood that reshaping a 3D cortical representation into a vector, breaks the natural structure and correlation in the original data. Thus, in order to preserve natural structure and correlation in the original data, dimensionality reduction operating directly on tensors rather than vectors is desirable. State-of-the-art multilinear dimensionality reduction techniques are employed, e.g. Non-Negative Tensor Factorization (NTF) [3], High-Order Singular Value Decomposition (HOSVD) [9], and Multilinear Principal Component Analysis (MPCA) [15] in order to derive compact feature vectors suitable for classification. Classification is performed by an SVM.

Stratified cross-validation tests on two well-known datasets, the GTZAN dataset and the ISMIR2004Genre dataset, demonstrate that the effectiveness of the proposed approach is compared with that of state-of-the-art music genre classification algorithms on the GTZAN dataset, while its accuracy exceeds 80% on the ISMIR2004Genre one.

In Section 2, the computational auditory model and the cortical representation of sound are described. Basic multilinear algebra and multilinear subspace analysis techniques are briefly introduced in Section 3. The application of multilinear subspace analysis to cortical representations is discussed in this section as well. Datasets and experimental results are presented in Section 4. Conclusions are drawn and future research direction are indicated in Section 5.

2 COMPUTATIONAL AUDITORY MODEL AND CORTICAL REPRESENTATION OF SOUND

The computational auditory model, proposed in [21], is inspired by psychoacoustic and neurophysiological investigations in the early and central stages of the human auditory system. An acoustic signal is analyzed by the human auditory model and a four-dimensional representation of sound is obtained, the so-called *cortical representation*. The model consists of two basic stages. The first stage converts the acoustic signal into a neural representation, the so-called *auditory spectrogram*. This representation is a time-frequency distribution along a tonotopic (logarithmic frequency) axis. At the second stage, the spectral and temporal modulation content of the auditory spectrogram is estimated by multiresolution wavelet analysis. For each frame, the multiresolution wavelet analysis is implemented via a bank of two-dimensional Gabor filters, that are selective to spectrotemporal modulation parameters ranging from slow to fast temporal rates (in Hertz) and from narrow to broad spectral scales (in Cycles/Octave). Since, for each frame, the analysis yields a scale-rate-frequency representation, thus the analysis results in a four-dimensional representation of time, frequency, rate, and scale for the entire auditory spectrogram. Mathematical formulation and details about the auditory model and the cortical representation of sound can be found in [18, 21].

Psychophysiological evidence justifies the choice of $scales \in \{0.25,\ 0.5,\ 1, 2, 4, 8\}$ (Cycles / Octave) and positive and negative $rates \in \{\pm 2, \pm 4, \pm 8, \pm 16, \pm 32\}$ (Hz) to represent the spectro-temporal modulation of sound. The cochlear model, employed by the first stage, has 128 filters with 24 filters per octave, covering $5\frac{1}{3}$ octaves along the tonotopic axis. For each sound recording, the extracted four-dimensional cortical representation is averaged along time and the average rate-scale-frequency cortical representation is thus obtained, that is naturally represented by a third-order tensor. Accordingly, the feature tensor $\mathcal{D} \in \mathbb{R}_{+}^{I_1 \times I_2 \times I_3}$, where $I_1 = I_{scales} = 6$, $I_2 = I_{rates} = 10$, and $I_3 = I_{frequencies} = 128$ results. During the analysis, we have used the NSL Matlab toolbox [1] .

[1] http:// www.isr.umd.edu/CAAR/pubs.html

3 MULTILINEAR SUBSPACE ANALYSIS

Recently, extensions of linear subspace analysis methods to handle high-order tensors have been proposed. In this section, three multilinear subspace analysis methods are briefly addressed. To begin with, a short introduction in multilinear algebra is given.

3.1 Multilinear Algebra Basics

In the field of multilinear algebra, tensors are considered as the multidimensional equivalent of matrices (second-order tensors) and vectors (first-order tensors) [8]. Throughout this paper, tensors are denoted by calligraphic letters (e.g. $\mathcal{D}$), matrices by uppercase boldface letters (e.g. $\mathbf{U}$), and vectors by lowercase boldface letters (e.g. $\mathbf{u}$).

A high-order real valued tensor $\mathcal{D}$ of order N is defined over the tensor space $\mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$, $I_i \in \mathbb{Z}, i = 1, 2, \ldots, N$. Each element of tensor $\mathcal{D}$ is addressed by N indices, $\mathcal{D}_{i_1,i_2,i_3,\ldots,i_N}$. Basic operations can be defined on tensors. The mode-n vectors are column vectors of matrix $\mathbf{D}_{(n)} \in \mathbb{R}^{I_n \times (I_1 \ldots I_{n-1} I_{n+1} \ldots I_N)}$ that results by mode-n unfolding the tensor $\mathcal{D}$.

The symbol $\times_n$ stands for the mode-n product between a tensor and a matrix. The mode-n product of a tensor $\mathcal{D} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ by a matrix $\mathbf{U} \in \mathbb{R}^{J_n \times I_n}$, denoted by $\mathcal{D} \times_n \mathbf{U}$, can be computed via the matrix multiplication $\mathbf{B}_{(n)} = \mathbf{U}\mathbf{D}_{(n)}$, followed by re-tensorization to undo the mode-n unfolding.

The inner product of two tensors $\mathcal{A}$ and $\mathcal{B} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ is denoted as $< \mathcal{A}, \mathcal{B} >$. The Frobenius norm of a tensor $\mathcal{A}$ is defined as $||\mathcal{A}||_F = \sqrt{< \mathcal{A}, \mathcal{A} >}$.

An N-order tensor $\mathcal{D}$ has rank 1, when it is decomposed as the Kronecker product of N vectors $\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \ldots, \mathbf{u}^{(N)}$, i.e. $\mathcal{D} = \mathbf{u}^{(1)} \otimes \mathbf{u}^{(2)} \otimes \ldots \otimes \mathbf{u}^{(N)}$. The rank of an arbitrary N-order tensor $\mathcal{D}$, $R = rank(\mathcal{D})$, is the minimal number of rank-1 tensors that yield $\mathcal{D}$, when linearly combined.

3.2 Non-Negative Tensor Factorization

NTF using Bregman divergences is proposed in [3]. The NTF algorithm is a generalization of the NMF algorithm [10] for N-dimensional tensors. NTF is able to decompose a tensor $\mathcal{D} \in \mathbb{R}_+^{I_1 \times I_2 \times \ldots \times I_N}$ into a sum of k rank-1 tensors:

$$\mathcal{D} = \sum_{j=1}^{k} \mathbf{u}_1^j \otimes \mathbf{u}_2^j \otimes \cdots \otimes \mathbf{u}_N^j \tag{1}$$

where $\mathbf{u}_i^j \in \mathbb{R}_+^{I_i}$. In [3], the NTF is performed by minimizing various Bregman divergences, when auxiliary functions are employed. Furthermore, NTF algorithms using multiplicative update rules, for each specific Bregman divergence are proposed.

In this paper, the NTF algorithm with the Frobenius norm is used. In order to apply the NTF algorithm for an N-order tensor, N matrices $\mathbf{U}^{(i)} \in \mathbb{R}_+^{I_i \times k}$, $i = 1, 2, \ldots, N$ should be created and initialized randomly with non-negative values. Let $*$ stand for the Hadamard product and $\odot$ denote the Khatri-Rao product. The following update rule in matrix form is applied to each $\mathbf{U}^{(i)}$:

$$\mathbf{U}^{(i)} = \tilde{\mathbf{U}}^{(i)} * \frac{\mathbf{D}_{(i)}\mathbf{Z}}{\tilde{\mathbf{U}}^{(i)}\mathbf{Z}^T\mathbf{Z}} \tag{2}$$

where $\mathbf{Z} = \mathbf{U}^{(N)} \odot \ldots \odot \mathbf{U}^{(i+1)} \odot \mathbf{U}^{(i-1)} \odot \ldots \odot \mathbf{U}^{(1)}$ and $\tilde{\mathbf{U}}^{(i)}$ refers to the matrix before updating. It is worth noting, that operators such as Khatri-Rao product preserve the inner structure of data.

3.3 High Order Singular Value Decomposition

HOSVD was proposed by Lathauwer *et al*. in [9] as a generalization of Singular Value Decomposition (SVD) applied to matrix for high-order tensors. Every tensor $\mathcal{D}$ can be expressed as:

$$\mathcal{D} = \mathcal{S} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \times_3 \ldots \times_N \mathbf{U}^{(N)}. \tag{3}$$

Each $\mathbf{U}^{(n)}$ is a unitary matrix containing the left singular vectors of the mode-n unfolding of tensor $\mathcal{D}$. Tensor $\mathcal{S} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$, known as core tensor, has the properties of all orthogonality and ordering. The HOSVD of a tensor $\mathcal{D}$ according to equation (3) is computed as follows.

1. Compute matrix $\mathbf{U}^{(n)}$ by computing the SVD of the matrix $\mathbf{D}_{(n)}$ and setting $\mathbf{U}^{(n)}$ to be its left singular matrix, $n = 1, 2, \ldots, N$.

2. Solve for the core tensor:

$$\mathcal{S} = \mathcal{D} \times_1 \mathbf{U}^{(1)^T} \times_2 \mathbf{U}^{(2)^T} \times_3 \ldots \times_N \mathbf{U}^{(N)^T}. \tag{4}$$

HOSVD results in a new ordered orthogonal basis for representation of the data in subspaces spanned by each mode of the tensor. Dimensionality reduction in each subspace is obtained by projecting data on principal axes and keeping only the components that correspond to the largest singular values.

3.4 Multilinear Principal Component Analysis

Recently, Lu *et al*. [15] proposed the Multilinear Principal Component Analysis as a multilinear equivalent of PCA. Similarly to PCA, let $\{\mathcal{D}_m \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}, m = 1, 2, 3, \ldots, M\}$ be a set of M tensor samples. The total scatter of these tensors is defined as:

$$\Psi_{\mathcal{D}} = \sum_{m=1}^{M} ||\mathcal{D}_m - \overline{\mathcal{D}}||_F^2 \tag{5}$$

where $\overline{\mathcal{D}}$ is the mean tensor. MPCA aims to define a multilinear transformation that maps the original tensor space $\mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ onto a tensor subspace $\mathbb{R}^{P_1 \times P_2 \times \ldots \times P_N}$ with $P_n < I_n, n = \{1, 2, \ldots, N\}$ such that most of the variation observed in the original tensor objects is captured, assuming that the variation can be measured by the total tensor scatter. Details and an algorithm for MPCA can be found in [15].

3.5 Multilinear Dimensionality Reduction of Cortical Representations

In each step of a stratified cross-validation test, the dataset used in the experiments (see Section 4) is split into two subsets, one used for training and another used for testing. As mentioned in Section 2, each recording is represented by a third-order feature tensor $\mathcal{D} \in \mathbb{R}^{I_{scales} \times I_{rates} \times I_{frequencies}}$ defining its average cortical representation. Thus, by stacking the third order feature tensors, associated to training recordings, a fourth order data tensor $\mathcal{T} \in \mathbb{R}^{I_{samples} \times I_{scales} \times \times I_{rates} \times I_{frequencies}}$ is obtained, where $samples$ is the number of training set recordings.

3.5.1 Multilinear Dimensionality Reduction by NTF

The resulting data tensor $\mathcal{T}$ is approximated by k rank-1 tensors obtained by NTF. Without loss of generality, NTF approximation is expressed in matrix form as:

$$\begin{aligned} \mathbf{T}_{(1)} &= \mathbf{U}^{(1)}(\mathbf{U}^{(4)} \odot \mathbf{U}^{(3)} \odot \mathbf{U}^{(2)})^T \Longleftrightarrow \\ \mathbf{T}_{(1)}^T &= (\mathbf{U}^{(4)} \odot \mathbf{U}^{(3)} \odot \mathbf{U}^{(2)})\mathbf{U}^{(1)^T} \end{aligned} \quad (6)$$

where $\mathbf{T}_{(1)} \in \mathbb{R}^{samples \times (scales \cdot rate \cdot frequencies)}$ is the mode-1 unfolding of tensor $\mathcal{T}$, $\mathbf{U}^{(1)} \in \mathbb{R}^{samples \times k}$, $\mathbf{U}^{(2)} \in \mathbb{R}^{scales \times k}$, $\mathbf{U}^{(3)} \in \mathbb{R}^{rates \times k}$, and $\mathbf{U}^{(4)} \in \mathbb{R}^{frequencies \times k}$. From (6), it is clear that every column of $\mathbf{T}_{(1)}^T$, i.e. vectorized cortical representation of a sound, is a linear combination of the basis vectors, which span the columns, of the basis matrix $\mathbf{W} = \mathbf{U}^{(4)} \odot \mathbf{U}^{(3)} \odot \mathbf{U}^{(2)}$ with coefficients taken from the columns of coefficient matrix $\mathbf{U}^{(1)^T}$. Performing Gram-Schmidt orthogonalization on basis matrix $\mathbf{W}$, an orthogonal basis matrix $\mathbf{Q}$ can be obtained. The orthogonalized bases span the same space as that of learned bases. The above step was employed, because previous research [5] has shown that orthogonality increases the discriminative power of the projections. Thus, the suitable features for classification are derived from the projection $\tilde{\mathbf{x}}_i = \mathbf{Q}^T \mathbf{d}_i$, where $\mathbf{d}_i$ is the vectorized cortical representation of the ith recording of the dataset.

3.5.2 Multilinear Dimensionality Reduction by HOSVD

The resulting data tensor $\mathcal{T}$ is decomposed to its mode-n singular vectors using the algorithm described in Subsection 3.3. Then, the singular matrices $\mathbf{U}^{(scales)}$, $\mathbf{U}^{(rates)}$, and $\mathbf{U}^{(frequencies)}$ are obtained. These matrices are orthonormal ordered matrices, which contain the subspace of singular vectors. In order to produce a subspace that approximates the original data, each singular matrix is truncated by setting a threshold so as to retain only the desired principal axes for each tensor mode.

Features suitable for classification are derived as follows. Each feature tensor $\mathcal{D}_i$ corresponding to ith recording of the dataset is projected onto the truncated orthonormal axes $\hat{\mathbf{U}}^{(scales)}$, $\hat{\mathbf{U}}^{(rates)}$, and $\hat{\mathbf{U}}^{(frequencies)}$ and a new feature tensor $\hat{\mathcal{X}}_i$ is derived:

$$\hat{\mathcal{X}}_i = \mathcal{D}_i \times_1 \hat{\mathbf{U}}^{(scales)} \times_2 \hat{\mathbf{U}}^{(rates)} \times_3 \hat{\mathbf{U}}^{(frequencies)}. \quad (7)$$

The actual features for classification $\tilde{\mathbf{x}}_i$ are derived from the vectorization of $\hat{\mathcal{X}}_i$.

3.5.3 Multilinear Dimensionality Reduction by MPCA

In a similar manner to HOSVD, a multilinear transformation that maps the original tensor space $\mathbb{R}^{I_1 \times I_2 \times I_3}$ onto a tensor subspace $\mathbb{R}^{P_1 \times P_2 \times P_3}$ with $P_n < I_n$, $n = \{1, 2, 3\}$, such that the subspace captures most of the variation observed in the original tensor objects is obtained by MPCA on $\mathcal{T}$. Features for classification, $\tilde{\mathbf{x}}_i$, are derived from the vectorized form of the projected $\mathcal{D}_i$ using the multilinear transformation obtained by MPCA.

4 EXPERIMENTAL RESULTS

Experiments are performed on two different datasets widely used for music genre classification [4, 6, 12, 13, 19, 20]. The first dataset, abbreviated as GTZAN, consists of following ten genre classes: Blues, Classical, Country, Disco, HipHop, Jazz, Metal, Pop, Reggae, Rock. Each genre class contains 100 audio recordings 30 seconds long. The second dataset, abbreviated as ISMIR 2004 Genre, is from the ISMIR 2004 Genre classification contest and contains 1458 full audio recordings distributed over six genre classes as follows: Classical (640), Electronic (229), JazzBlues (52), MetalPunk (90), RockPop (203), World (244), where the number within parentheses refers to the number of recordings belong to each genre class.

Features are extracted from the cortical representation of sound using the aforementioned multilinear subspace analysis techniques. The value of parameter k in NTF algorithm was set to 150 and 140 for the GTZAN dataset and the ISMIR 2004 Genre one, respectively. The number of retained principal components for each subspace, when HOSVD is employed for feature extraction, is set to be 5 out of 6 for rate, 7 out of 10 for scale, and 12 out of 128 for frequency. The features extracted by MPCA capture 98% of the total variation in each mode. In Figures 1 and 2, the number of retained principal components in each subspace is shown as

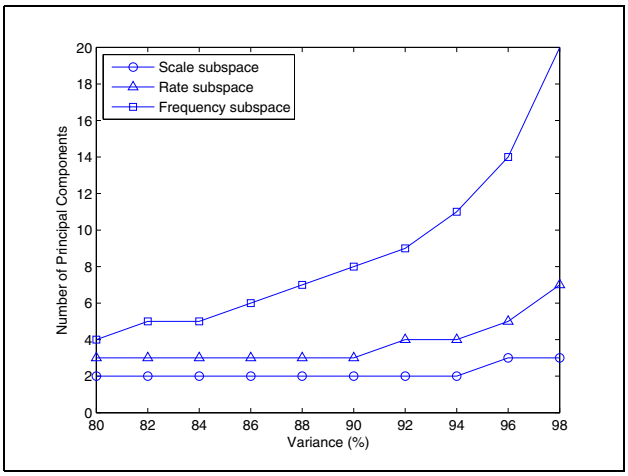

Figure 1. Total number of retained principal components in each subspace (e.g. scale, rate, and frequency) as a function of the portion of variance retained for the GTZAN dataset.

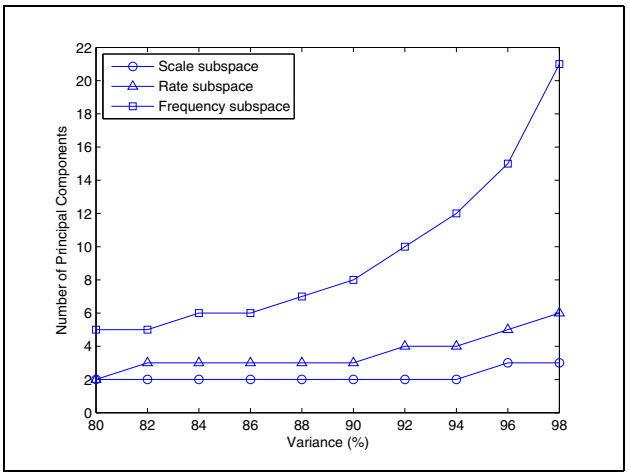

Figure 2. Total number of retained principal components in each subspace (e.g. scale, rate, and frequency) as a function of the portion of variance retained for the ISMIR2004Genre dataset.

a function of the portion of variance retained for the GTZAN and ISMIR2004 Genre dataset, respectively.

Classification was performed by SVM with an RBF kernel. In order to tune the RBF kernel parameters, a grid search algorithm similar to algorithm proposed in [7] was used. Linear and polynomials kernels were also considered, but they achieve poor performance.

The classification accuracies reported in Table 2 are the mean accuracies obtained by 10-fold stratified cross-validation on the full datasets. They are obtained by the various multilinear subspace analysis techniques followed by SVM. The row marked by NTF contains the classification results achieved by features extracted using NTF for both datasets. The row marked by HOSVD contains the classification results achieved by features extracted using HOSVD, while the row marked by MPCA contains the classification results achieved by features extracted using MPCA. The effectiveness of NTF and HOSVD as feature extraction techniques is self-evident in both datasets.

	GTZAN	**ISMIR2004Genre**
NTF	78.20%(3.82)	80.47%(2.26)
HOSVD	77.90% (4.62)	80.95% (3.26)
MPCA	75.01% (4.33)	78.53% (2.76)

Table 2. Classification accuracy on GTZAN and ISMIR2004Genre datasets. The accuracy is calculated by ten-fold stratified cross-validation. The number within parentheses is the corresponding standard deviation.

On the GTZAN dataset, the best classification accuracy outperforms the rates reported by Tzanetakis *et al.* [20] (61.0%), Lidy *et al.* [14] (76.8%), Holzapfel *et al.* [6] (74%), and it is comparable to the rate achieved by Li *et al.* [12] (78.5%). Bergstra *et al.* in [4] reported a classification accuracy equal to 82.5% on the GTZAN database, but they do not disclose details on the experimental setup (e.g. the number of folds).

On the ISMIR2004Genre dataset classification, accuracies achieved for features extracted by NTF and HOSVD are comparable and exceed 80%. It is not possible to compare directly our results with the results obtained by other researchers on this dataset, because of the quite different experimental settings [6, 13, 19].

5 CONCLUSIONS - FUTURE WORK

In this paper, the problem of automatic music genre classification is examined in a multilinear framework. Features have been extracted from the cortical representation of sound using three multilinear subspace analysis techniques. The best classification accuracies reported in this paper are comparable with the best accuracies obtained by other state-of-the-art music genre classification algorithms. The effectiveness of spectro-temporal features obtained by NTF and HOSVD has been demonstrated. It is true, that multilinear techniques applied in straightforward manner, although they provide a more accurate representation to be exploited by the subsequent classifier, do not yield a recognition accuracy much higher than state-of-the-art linear algebra techniques do. Therefore, more effort is required toward addressing the small sample case in the multilinear algebra as well. It is worth noting that the multilinear dimensionality reduction techniques employed in the paper are unsupervised. In the future, supervised multilinear subspace analysis techniques

based on NTF will be developed and tested for the automatic music genre classification.

Finally, in our experiments, we have considered that each song belongs to only one genre class. Obviously, it is realistic to use overlapping class labels for labelling music by style [4]. In general, high-order tensors are structures that are suitable for a such multi-labelling classification problem.

6 REFERENCES

[1] Aucounturier, J. J. and Pachet F. "Representing musical genre: A state of the art", *Journal of New Music Research*, pp. 83-93, 2005.

[2] Aucounturier, J. J. and Pachet F. "Improving timbre similarity: How high is the sky?", *Journal of Negative Results in Speech and Audio Sciences*,Vol. 1, No. 1, 2004.

[3] Benetos, E. and Kotropoulos C. "A tensor-based approach for automatic music genre classification", *Proceedings of the European Signal Processing Conference*, Lausanne, Switzerland, 2008.

[4] Bergstra, J., Casagrande, N., Erhan, D., Eck, D. and Kegl B. "Aggregate features and AdaBoost for music classification", *Machine Learning*, Vol. 65, No. 2-3, pp. 473-484, 2006.

[5] Duchene, J. and Leclercq, S. "An optimal transformation for discriminant and principal component analysis", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 10, No. 6, pp. 978-983, 1988.

[6] Holzapfel, A. and Stylianou Y. "Musical genre classification using nonnegative matrix factorization-based features", *IEEE Transactions on Audio, Speech, and Language Processing*,Vol. 16, No. 2, pp. 424-434, 2008.

[7] Hsu, C., Chang, C. C. and Lin, C. J. *A Practical Guide to Support Vector Classification*. Technical Report, Department of Computer Science, National Taiwan University, 2003.

[8] Lathauwer, L. *Signal Processing Based on Multilinear Algebra*. Ph.D Thesis, K.U. Leuven, E.E. Dept. - ESAT, Belgium, 1997.

[9] Lathauwer, L., Moor, B. and Vandewalle, J. "A multilinear singular value decomposition", *SIAM Journal on Matrix Analysis and Applications*, Vol. 21, No. 4, pp. 1253-1278, 2000.

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

[11] Lee, J. H. and Downie, J. S. "Survey of music information needs, uses and seeking behaviours: Preliminary findings", *Proceedings of the Fifth International Symposium on Music Information Retrieval*, Barcelona, Spain, 2004.

[12] Li, T., Ogihara, M. and Li, Q. "A comparative study on content-based music genre classification", *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pp. 282-289, Toronto, Canada, 2003

[13] Lidy, T. and Rauber, A. "Evaluation of feature extractors and psycho-acoustic transformations for music genre classification", *Proceedings of the Sixth International Symposium on Music Information Retrieval*, London, UK, 2005.

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

[15] Lu, H., Plataniotis, K. N. and Venetsanopoulos, A. N. "MPCA: Multilinear principal component analysis of tensor objects", *IEEE Transactions on Neural Networks*, Vol. 19, No. 1, pp 18-39, 2008.

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

[17] McKay, C. and Fujinaga, I. "Musical genre classification: Is it worth pursuing and how can in be improved?", *Proceedings of the Seventh International Symposium on Music Information Retrieval*, Victoria, Canada, 2006.

[18] Mesgarani, N., Slaney, M., and Shamma, S. A."Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations", *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 14, pp. 920-930, 2006.

[19] Pampalk, E., Flexer, A. and Widmer, G. "Improvements of audio-based music similarity and genre classification", *Proceedings of the Sixth International Symposium on Music Information Retrieval*, London, UK, 2005.

[20] Tzanetakis, G. and Cook, P. "Musical genre classification of audio signal", *IEEE Transactions on Speech and Audio Processing*, Vol. 10, No. 3, pp. 293-302, July 2002.

[21] Wang, K. and Shamma, S. A. "Spectral shape analysis in the central auditory system", *IEEE Transactions on Speech and Audio Processing*, Vol. 3, pp. 382-396, 1995.

[22] Woolley, S., Fremouw, T., Hsu, A. and Theunissen, F. "Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds", *Nature Neuroscience*, Vol. 8, pp. 1371-1379, 2005.