Greek Folk Music Classification Using Auditory Cortical Representations

Eftychia Fotiadou, Nikoletta Bassiou, and Constantine Kotropoulos
Department of Informatics, Aristotle University of Thessaloniki
Thessaloniki 54124, GREECE
Email: e.fotiadou@imperial.ac.uk, nikoletta.basiou@sri.com, costas@aiia.csd.auth.gr

***Abstract*—In this paper, we deal with the classification of Greek folk songs into 8 classes associated with the region of origin of the songs. Motivated by the way the sound is perceived by the human auditory system, auditory cortical representations are extracted from the music recordings. Moreover, deep canonical correlation analysis (DCCA) is applied to the auditory cortical representations for dimensionality reduction. To classify the music recordings, either support vector machines (SVMs) or classifiers based on canonical correlation are employed. An average classification rate of 73.25 % is measured on a dataset of Greek folk songs from 8 regions, when the auditory cortical representations are classified by the SVMs. It is also demonstrated that the reduced features extracted by the DCCA yield an encouraging average classification rate of 66.27%. The latter features are shown to possess good discriminating properties.**

I. Introduction

In the last years, various on-line music sharing communities and several applications have emerged, which offer the users a personalized experience usually in the form of recommendations by exploiting content and context. As a result, Music Information Retrieval (MIR) has turned into a growing research field. Since the classification of music, mainly into different genres, such as classical, pop, jazz, etc., is an important aspect of such systems, a significant amount of research has been performed. A comprehensive survey of music classification methods can be found in [1].

In this paper, we study the problem of classifying Greek folk music into different genres, depending on the geographic region it stems from. Greece exhibits a long and rich music tradition [2]. Folk songs are strongly associated with customs and traditions, recounting historical events, everyday life problems and events, as well as folk tales. Greek folk music can be categorized according to the geographic region it originates from, because the music of a specific region exhibits a particular character, which is easily identified. However, in some cases, songs from neighboring regions may present strong similarities. Furthermore, folk songs can be divided into different types, depending on their content, such as laments, marriage songs, lullabies, songs of exile, and Acritic ballads, among others [3]. Apart from their lyric content, songs of the same category usually exhibit similarities in rhythm and melody.

E. Fotiadou is now with the Imperial College London and N. Basiou is with SRI International, Menlo Park, CA. This work was done when they were both with the Department of Informatics, Aristotle University of Thessaloniki, Greece.

Although there is rich literature devoted to pattern analysis of western music, the existing bibliography concerning non-western music, and especially the Greek music, is rather scarce. A web content management system is presented in [4], containing a collection of Greek folk songs and a game with a purpose is proposed, which enables the users to annotate folk music. In [5], the maximally general distinctive pattern method, originally proposed in [6], is employed in order to extract distinctive patterns from different types of Cretan folk songs. A method for retrieval of Greek, as well as African, music recordings is introduced in [7], which is based on rhythmic similarity. A rhythmic signature is extracted for each recording by means of self similarity analysis. Subsequently, the similarity between two signatures is measured using Dynamic Time Warping. Another approach for measuring the rhythmic similarity between two musical pieces can be found in [8]. The proposed method is suitable for musical pieces with similar rhythmic structure, but varying tempo, as is the case with Greek folk music. A framework for detecting similar phrases in traditional music of the Eastern Mediterranean is presented in [9], which integrates rhythmic as well as melodic aspects. In [10], a case study is presented, which deals with the classification of Greek music into different moods (e.g., “happy”, “sad”, “angry”). The authors experiment with various features and classifiers, while they utilize both audio and lyrics information.

This paper extends the preliminary work on Greek folk music reported in [11]. Here, we explore alternative representations for the music signal and assume a more complex classification scenario, which consists of eight classes instead of two. In particular, instead of mel-frequency cepstral coefficients used in [11], we employ the auditory cortical representations that are based on spectrotemporal modulations [12], whose derivation is motivated by the human auditory system. The auditory cortical representations have exhibited very good results in western music genre recognition [13]. Here, it is demonstrated that they perform equally well for Greek folk music. Deep Canonical Correlation Analysis (DCCA) [14] is applied to the extracted auditory representations in order to obtain discriminative low-dimensional feature descriptors of the music recordings in each class. The classification of the Greek folk song recordings into 8 classes is based either on Canonical Correlation Analysis (CCA) or Support Vector Machines (SVMs). An overall good performance is achieved with

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

recognition rates exceeding 66%. The top average accuracy disclosed here is 73.25%.

II. Proposed Method

The methodology applied to solve the music classification problem consists of two steps. In the first step, feature extraction is performed on the audio data in order to obtain a representation suitable for the classification task. In the second step, the audio data are classified into eight different classes.

A. Auditory cortical representations

These feature descriptors are inspired by the way sound is perceived and processed by the human auditory system [12]. The human auditory system can be modeled by a two stage process. The first stage models the cochlea, and converts the audio signal to an auditory representation (spectrogram). It has been reported that the basilar membrane across the cochlea exhibits a tonotopical organization, so that higher frequency tones stimulate peaks near the base of the cochlea, while lower frequency ones stimulate peaks near the apex of the cochlea [15]. Therefore, the basilar membrane can be modeled by a bank of bandpass filters. To this end, the constant Q transform (CQT) is employed [16]. The CQT is a technique, which transforms a signal from time to the frequency domain, such that the center frequencies of the bins are geometrically spaced and the Q factors (i.e., the ratios of the center frequencies to the bandwidths) are equal. This means that a better frequency resolution is observed for the low frequencies, while the time resolution is better for high frequencies, which resembles the frequency resolution of the auditory system.

In the second stage, the audio signal reaches the primary auditory cortex, where it is processed, perceived and interpreted. In this stage, the spectral and temporal modulation content of the auditory spectrogram is estimated. A topographical organization can be also observed in the primary auditory cortex, where the cells are organized according to their response selectivity in different spectral and temporal stimuli [15]. To model this functionality, multi-resolution two-dimensional (2D) wavelet analysis is applied on the auditory spectrogram that was extracted in the first stage. Wavelet analysis is implemented using 2D Gaussian filters, ranging from narrow to broad spectral scales and from slow to fast temporal rates. The aforementioned analysis results in a four-dimensional (4D) representation of time, frequency, rate, and scale, referred to as auditory cortical representation [12]. The auditory cortical representations extracted for each frame are averaged across time and the resulting 3D representations for each song are vectorized, as is detailed in Section III.

B. Canonical Correlation Analysis (CCA)

Canonical Correlation Analysis calculates linear transformations for two random vectors that are maximally correlated [17]. CCA uses two different views for a set of patterns (e.g., two different representations for a given dataset) and projects them into a space of lower dimension, where they are maximally correlated. Here, the CCA is treated as a pattern classification method. Furthermore, this section serves as an introduction for the next section on DCCA, which is used for discriminative dimensionality reduction.

Let us assume a dataset of n samples $\mathbf{x}_i$, $i = 1, 2, ..., n$, represented by the data matrix $\mathbf{X} = \{\mathbf{x}_1|\mathbf{x}_2|...|\mathbf{x}_n\} \in \mathbb{R}^{d\times n}$ and the label matrix $\mathbf{Y} = \{\mathbf{y}_1|\mathbf{y}_2|...|\mathbf{y}_n\} \in \mathbb{R}^{k\times n}$, where k is the number of classes. Furthermore, we assume that $\mathbf{x}_i$ and $\mathbf{y}_i$ are centered. The CCA finds the projection vectors $\mathbf{w}_x \in \mathbb{R}^{d\times 1}$ and $\mathbf{w}_y \in \mathbb{R}^{k\times 1}$, maximizing the sample correlation coefficient:

$$(\mathbf{w}_x^*, \mathbf{w}_y^*) = \underset{\mathbf{w}_x,\mathbf{w}_y}{\operatorname{argmax}} \frac{\mathbf{w}_x^T\mathbf{X}\mathbf{Y}^T\mathbf{w}_y}{\sqrt{\mathbf{w}_x^T\mathbf{X}\mathbf{X}^T\mathbf{w}_x}\sqrt{\mathbf{w}_y^T\mathbf{Y}\mathbf{Y}^T\mathbf{w}_y}}. \quad (1)$$

The above objective function is invariant to the scaling of $\mathbf{w}_x$ and $\mathbf{w}_y$. Therefore, it can be transformed into a constrained optimization problem of the form:

$$\begin{aligned} &\underset{\mathbf{w}_x,\mathbf{w}_y}{\operatorname{argmax}} \ \mathbf{w}_x^T\mathbf{X}\mathbf{Y}^T\mathbf{w}_y \quad (2)\\ &\text{subject to } \mathbf{w}_x^T\mathbf{X}\mathbf{X}^T\mathbf{w}_x = 1 \text{ and } \mathbf{w}_y^T\mathbf{Y}\mathbf{Y}^T\mathbf{w}_y = 1. \end{aligned}$$

In the case that $\mathbf{Y}\mathbf{Y}^T$ is non-singular, $\mathbf{w}_x^*$ is calculated by solving:

$$\begin{aligned} &\max_{\mathbf{w}_x} \mathbf{w}_x^T\mathbf{X}\mathbf{Y}^T(\mathbf{Y}\mathbf{Y}^T)^{-1}\mathbf{Y}\mathbf{X}^T\mathbf{w}_x \\ &\text{subject to } \mathbf{w}_x^T\mathbf{X}\mathbf{X}^T\mathbf{w}_x = 1. \end{aligned} \quad (3)$$

The aforementioned assumption for $\mathbf{Y}\mathbf{Y}^T$ can be maintained by assuming a class membership indicator matrix, where each column has an entry equal to 1 in the index corresponding to the sample's label and equal to 0, otherwise, and then apply centering. The solution of (3) is obtained as the eigenvector corresponding to the top eigenvalue η of the generalized eigenvalue problem given by:

$$\mathbf{X}\mathbf{Y}^T(\mathbf{Y}\mathbf{Y}^T)^{-1}\mathbf{Y}\mathbf{X}^T\mathbf{w}_x = \eta\mathbf{X}\mathbf{X}^T\mathbf{w}_x. \quad (4)$$

By retaining the top k eigenvectors, and under certain orthonormality constraints, multiple projection vectors can be obtained. In order to avoid the singularity of $\mathbf{X}\mathbf{X}^T$ and $\mathbf{Y}\mathbf{Y}^T$ and to prevent overfitting, two regularization terms are incorporated in (4), i.e., $\lambda_x\mathbf{I}$ and $\lambda_y\mathbf{I}$ with $\lambda_x, \lambda_y > 0$. The resulting problem, called regularized CCA [18], takes the form $\mathbf{X}\mathbf{Y}^T(\mathbf{Y}\mathbf{Y}^T + \lambda_y\mathbf{I})^{-1}\mathbf{Y}\mathbf{X}^T\mathbf{w}_x = \eta(\mathbf{X}\mathbf{X}^T + \lambda_x\mathbf{I})\mathbf{w}_x$.

Pattern classification can be considered as a least squares problem. Considering a data matrix $\tilde{\mathbf{X}} \in \mathbb{R}^{d\times n}$ and class labels $y_i \in \{1, 2, ..., k\}, i = 1, 2, ..., n$, a centered data matrix $\mathbf{X}$ and centered labels $t_i = y_i - \bar{y}$ can be obtained, where $\bar{y}$ is the average class label. The centered labels can be collected in a row vector $\mathbf{t} \in \mathbb{R}^{1\times n}$ and a projection vector $\mathbf{w} \in \mathbb{R}^{d\times 1}$ is found, which minimizes the sum of squares cost function given by [19]:

$$\min_{\mathbf{w}} \sum_{i=1}^{n} |\mathbf{w}^T\mathbf{x}_i - t_i|^2 = \|\mathbf{w}^T\mathbf{X} - \mathbf{t}\|_2^2. \quad (5)$$

After learning $\mathbf{w}^*$ that minimizes (5) for a training set (i.e., subpart of $\mathbf{X}$), the label of a test data sample $\mathbf{z}$ is calculated

by rounding:

$$\hat{y}(\mathbf{z}) = \bar{y} + (\mathbf{w}^*)^T(\mathbf{z} - \bar{\mathbf{x}}), \quad (6)$$

where $\bar{\mathbf{x}}$ denotes the average data sample for the training set. If instead of a scalar label, a vector $\mathbf{t}_i \in \mathbb{R}^{k\times 1}$ is used as label for the i-th sample, $i = 1, 2, ..., n$, a label matrix can be constructed $\mathbf{T} = \{\mathbf{t}_1|\mathbf{t}_2|...|\mathbf{t}_n\} \in \mathbb{R}^{k\times n}$. Then, the least squares cost function (5) takes the form:

$$\min_{\mathbf{W}} \sum_{i=1}^{n} \|\mathbf{W}^T\mathbf{x}_i - \mathbf{t}_i\|_2^2 = \|\mathbf{W}^T\mathbf{X} - \mathbf{T}\|_F^2, \quad (7)$$

where $\mathbf{W} \in \mathbb{R}^{d\times k}$ is the projection matrix and $\|\mathbf{A}\|_F$ is the Frobenius norm of matrix $\mathbf{A}$. The solution of (7) is obtained by [19]:

$$\mathbf{W}_{LS} = (\mathbf{X}\mathbf{X}^T)^{\dagger}\mathbf{X}\mathbf{T}^T, \quad (8)$$

where $\mathbf{A}^{\dagger}$ is the Moore-Penrose pseudo-inverse of matrix $\mathbf{A}$. After having learnt $\mathbf{W}_{LS}$ for a training set, the label of a test data sample $\mathbf{z}$ can be calculated by:

$$\underset{j=1,2,...,k}{\operatorname{argmax}} \ \bar{y}_j + \mathbf{w}_j^T(\mathbf{z} - \bar{\mathbf{x}}), \quad (9)$$

where $\bar{y}_i$ denotes the j-th element of the average class label indicator vector $\bar{\mathbf{y}}$ and $\mathbf{w}_j$ is the j-th column of the projection matrix $\mathbf{W}_{LS}$. It has been shown that for $\mathbf{T} = (\mathbf{Y}\mathbf{Y}^T)^{-1/2}\mathbf{Y}$ and under mild conditions, the solution of the least squares problem given by (8) and the matrix $\mathbf{W}_{CCA}$ constructed by the top k eigenvectors of the generalized eigenvalue problem (4) are equivalent for several classifiers, such as the k-Nearest Neighbor or the linear SVMs [20]. The aforementioned fact justifies the use of both the CCA and the linear SVM classifiers in the experiments conducted in Section III.

C. Deep Canonical Correlation Analysis (DCCA)

DCCA is an extension of CCA, whose objective is to find representations for two views of the data that are maximally correlated, by using stacked layers of non-linear transformations [14]. In more detail, the DCCA employs two deep neural networks (one for each data view), which are simultaneously trained, so that their output layers exhibit maximum correlation. In the input layer of each network, the number of nodes is equal to the dimensionality of the corresponding view, while the output layers consist of the same number of nodes for both networks. The networks may have different numbers of hidden layers, while all the hidden layers of a deep neural network have the same number of nodes.

Let us consider two deep neural networks, corresponding to the data and class label views. Assuming data samples of dimensionality d and class label vectors of dimensionality k, the two networks will have input layers of d and k nodes, respectively, while both output layers will consist of o nodes. Furthermore, we assume that the first network has L hidden layers with c_1 nodes each and the second network has M hidden layers with c_2 nodes each.

Given an input data sample $\mathbf{x}_i$ in the first network, the output of the first hidden layer is obtained by $\mathbf{h}_1 = s(\mathbf{W}_1^1\mathbf{x}_i + \mathbf{b}_1^1) \in \mathbb{R}^{c_1\times 1}$, where $\mathbf{W}_1^1 \in \mathbb{R}^{c_1\times d}$ is the weight matrix, $\mathbf{b}_1^1 \in \mathbb{R}^{c_1\times 1}$ is the vector of biases, and $s : \mathbb{R} \to \mathbb{R}$ is a non-linear activation function. The output $\mathbf{h}_1$ of the first hidden layer serves as input to the second hidden layer, which in turn has $\mathbf{h}_2$ as output, and so on. The output of each hidden layer, is thus described by:

$$\mathbf{h}_l = s(\mathbf{W}_l^1\mathbf{h}_{l-1} + \mathbf{b}_l^1), \ \ l = 2, \ldots, L \quad (10)$$

where $\mathbf{W}_l^1 \in \mathbb{R}^{c_1\times c_1}$, $l = 2, \ldots, L-1$. When $l = L$, (10) computes the final representation $f_1(\mathbf{x}_i) \in \mathbb{R}^{o\times 1}$, i.e., $\mathbf{W}_L^1 \in \mathbb{R}^{o\times c_1}$ and $\mathbf{b}_L^1 \in \mathbb{R}^{o\times 1}$. In the same way, regarding the second network, the output of the hidden layer m is given by:

$$\mathbf{h}_m = s(\mathbf{W}_m^2\mathbf{h}_{m-1} + \mathbf{b}_m^2), \ \ m = 1, 2, \ldots, M. \quad (11)$$

When $m = M$, (11) gives the output representation of $f_2(\mathbf{y}_i) \in \mathbb{R}^{o\times 1}$ for a multivariate label $\mathbf{y}_i$. The aim of the DCCA method is to jointly learn the vectors of all parameters $\mathbf{W}_l^1, \mathbf{b}_l^1, \mathbf{W}_m^2, \mathbf{b}_m^2$ for both the networks, such that the correlation between $f_1(\mathbf{X})$ and $f_2(\mathbf{Y})$ is maximized.

Let us denote with $\mathbf{H}_X \in \mathbb{R}^{o\times n}$ and $\mathbf{H}_Y \in \mathbb{R}^{o\times n}$ the matrices whose columns correspond to the output representations obtained by the two deep networks and with $\bar{\mathbf{H}}_X = \mathbf{H}_X - \frac{1}{n}\mathbf{H}_X\mathbf{1}$ and $\bar{\mathbf{H}}_Y = \mathbf{H}_Y - \frac{1}{n}\mathbf{H}_Y\mathbf{1}$ the centered matrices. The sample dispersion matrices for the output representations of the two views are calculated by $\hat{\mathbf{\Sigma}}_X = \frac{1}{n-1}\bar{\mathbf{H}}_X\bar{\mathbf{H}}_X^T + r_X\mathbf{I}$ and $\hat{\mathbf{\Sigma}}_Y = \frac{1}{n-1}\bar{\mathbf{H}}_Y\bar{\mathbf{H}}_Y^T + r_Y\mathbf{I}$, where $r_X > 0$ and $r_Y > 0$ are regularization parameters so that $\hat{\mathbf{\Sigma}}_X, \hat{\mathbf{\Sigma}}_Y$ are positive-definite. The cross-covariance matrix $\hat{\mathbf{\Sigma}}_{XY}$ is given by $\hat{\mathbf{\Sigma}}_{XY} = \frac{1}{n-1}\bar{\mathbf{H}}_X\bar{\mathbf{H}}_Y^T$. If $o = k$, the total correlation between $\bar{\mathbf{H}}_X$ and $\bar{\mathbf{H}}_Y$ is calculated as the matrix trace norm of $\mathbf{U} = \hat{\mathbf{\Sigma}}_X^{-1/2}\hat{\mathbf{\Sigma}}_{XY}\hat{\mathbf{\Sigma}}_Y^{-1/2}$:

$$\operatorname{corr}(\bar{\mathbf{H}}_X, \bar{\mathbf{H}}_Y) = \operatorname{tr}(\mathbf{U}^T\mathbf{U})^{-1/2}. \quad (12)$$

The parameters of DCCA $\mathbf{W}_l^1, \mathbf{b}_l^1, \mathbf{W}_m^2, \mathbf{b}_m^2$ are learnt using training data, so as to optimize the aforementioned quantity. To this end, a gradient-descent approach is employed, based on backpropagation, in order to compute the gradient of the total correlation (12) with respect to all parameters [14]. A quadratic penalty with weight $\lambda_b > 0$ is also added in (12) for regularization. A full-batch optimization is performed, using the limited memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) method [21], which has been successfully applied in deep learning problems.

Furthermore, in order to achieve better results, a pre-training process is applied for the initialization of the optimization parameters. In more detail, the parameters of each network layer are initialized by means of a denoising autoencoder [22]. Assuming a training data matrix $\mathbf{X}$, a distorted matrix $\tilde{\mathbf{X}}$ is constructed, by adding independent identically distributed Gaussian noise with zero mean and variance σ_a^2. The reconstructed data are formed as $\hat{\mathbf{X}} = \mathbf{W}^T s(\mathbf{W}\tilde{\mathbf{X}} + \mathbf{b}\mathbf{1}^T)$. The L-BFGS method is subsequently used to find a local minimum of the total squared reconstruction error plus a quadratic penalty $\varphi_a(\mathbf{W}, \mathbf{b}) = \|\hat{\mathbf{X}} - \mathbf{X}\|_F^2 + \lambda_a(\|\mathbf{W}\|_F^2 + \|\mathbf{b}\|_2^2)$, where λ_a is a hyperparameter.

TABLE I
AVERAGE CLASSIFICATION RATES FOR 5 FOLD CROSS-VALIDATION.

Feature type	Classifier	Classification rate (%)						Parameter C (SVM)
		Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average	
CQTM	CCA	65.06	77.11	68.67	71.08	78.31	72.05	-
CQTM	SVM	67.47	75.9	71.08	71.08	80.72	**73.25**	0.0156
CQTM + DCCA	SVM	63.86	66.27	62.65	65.06	73.49	66.27	0.0078

III. DATASET AND EXPERIMENTAL EVALUATION

In order to evaluate the performance of the proposed method, a dataset was collected, consisting of Greek folk music recordings. In more detail, we selected 415 songs from 8 different geographic regions: 70 from Asia Minor, 55 from Pontus, 45 from Dodecanese, 50 from Epirus, 45 from Peloponnese, 50 from West Macedonia, 50 from the Northeastern Aegean Islands, and 50 from Crete. From each recording, an excerpt of 30s duration was extracted after the first 30s of a recording to exclude any introductory parts. The recordings were sampled at 22.050 Hz.

Hereafter, the auditory cortical representations are denoted by CQTM. Following [13], for the CQT we employed 128 filters which cover 8 octaves between 44.9 Hz and 11 kHz. Also, the elements of the CQT matrix were raised to the power of 0.1 in order to compress the magnitude of the CQT. Regarding the wavelet analysis of the second stage, a bank of 2D Gaussian filters was employed with scales $\in \{0.25, 0.5, 1, 2, 4, 8\}$ (Cycles/Octave) and rates $\in \{\pm 2, \pm 4, \pm 8, \pm 16, \pm 32\}$ (Hz). The resulting 4D representation was averaged on time and a 3D cortical representation (frequency, rate, and scale) was obtained. Subsequently, by re-arranging the elements of the 3D representation into a single vector, each audio recording was described by a vector $\mathbf{x} \in \mathbb{R}_+^{7680\times 1}$ (i.e., 128 frequency channels $\times$ 10 rates $\times$ 6 scales). A set consisting of n audio recordings is represented by a matrix $\mathbf{X} \in \mathbb{R}_+^{7680\times n}$.

For the DCCA method[1], the number of nodes in the input layer of the first network was equal to the dimensionality of the cortical representation, i.e., 7680. The input layer of the second network had 8 nodes (i.e., equal to the dimensionality of the label vector). The number of nodes of the output layers o, was set equal to 8. For the number of hidden layers, different values in the range $[3, 10]$ were tested. Similarly, different values in $\{32, 64, 128, 256, 512, 1024\}$ were tested for the number of nodes in each hidden layer. To determine the optimal number of hidden layers and nodes, the dataset was divided into 5 folds. Each fold consists of a subset of 249 samples created by stratified sampling (i.e., retaining $\frac{3}{5}$ of the music recordings from each region) used for training, a development set formed by one fifth of the music recordings from each region used to search for the optimal values of the aforementioned parameters as well as the regularization parameter for slack variables C of the linear SVM[2], and a test set formed by one fifth of the music recordings used to measure accuracy. Optimal performance was observed when the hidden layers of the first and the second network consisted of $c_1 = 512$ and $c_2 = 32$ nodes, respectively, and both networks consisted of $L = M = 9$ hidden layers. Regarding the hyperparameter values involved in the pre-training phase of the DCCA algorithm, we used the default values, since no significant improvement was observed when altering them. In more detail, a) the regularization parameter λ_a for the input, hidden and output layers in the first network was equal to 4.711×10^{-4}, 0.052, and 2.424×10^{-4}, respectively; b) The corresponding values for the second network were 3.153×10^{-4}, 5.504×10^{-4}, and 2.125×10^{-4}; c) In the first network, the variance σ_a^2 of the Gaussian noise in the denoising autoencoder pre-training of input and hidden layers was set to values 0.1538 and 0.0264, respectively, while for the second network these values were 0.0096 and 0.1566; d) The regularization parameters λ_b, r_X and r_Y were set to values 0.045, 41.67, and 59.06, respectively; e) The convergence tolerance of the L-BFGS algorithm was set to 10^{-4} and 10^{-3} for the first and second network, respectively. The activation function for all the layers was a sigmoid function based on the cubic root.

Classification was performed using either a linear SVM classifier or the CCA method of Section II-B. In the latter case, correct classification occurs when the predicted label in (6) is the same with the true label. As can be seen in Table I, overall the auditory cortical representations achieve a good performance. This demonstrates that the auditory cortical representations are not only suitable for Western music genre classification as shown in [13], but for Greek folk music classification as well. It is worth noting that the accuracies measured for both the CCA and the linear SVM classifiers in each fold do not differ more than 2.41%. In order to check whether the accuracy difference of 2.41% is statistically significant, we apply the approximate analysis in [23]. Let us assume that the accuracies ϖ_1 and ϖ_2 are binomially distributed random variables. If $\hat{\varpi}_1, \hat{\varpi}_2$ denote the empirical accuracies, and $\overline{\varpi} = \frac{\hat{\varpi}_1 + \hat{\varpi}_2}{2}$, the hypothesis $H_0 : \varpi_1 = \varpi_2 = \overline{\varpi}$ is tested at 95% level of significance. The accuracy difference has variance $\beta = 2\frac{\overline{\varpi}(1-\overline{\varpi})}{n_t}$, where n_t is the number of test samples (i.e., 83). For $\zeta = 1.65\sqrt{\beta}$, if $\hat{\varpi}_1 - \hat{\varpi}_2 \geq \zeta$, we reject H_0 with risk 5% of being wrong. The aforementioned analysis certifies that the accuracy difference of 2.41% between the CCA classifier and the linear SVM is not statistically significant, because ζ=10.33%. This is in par with the theoretical guarantees on the equivalence of these

[1] https://homes.cs.washington.edu/~galen/files/dcca.tgz
[2] https://www.csie.ntu.edu.tw/~cjlin/libsvm/

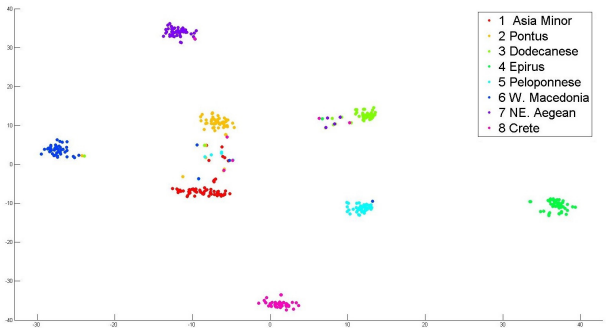

Fig. 1. Visualization of the 8-dimensional features in the output of the 1st network of DCCA using the t-SNE method.

classifiers.

A relatively good performance is observed, when the DCCA is applied to auditory cortical representations for dimensionality reduction and the obtained 8 dimensional features are classified by the SVMs. The maximum discrepancy in accuracy when the 7680 dimensional CQTM and the 8 dimensional features extracted by DCCA are classified by a linear SVM is 9.63%. Neither this accuracy difference is found to be statistically significant, because ζ=11.61%. To obtain a better insight of the DCCA performance, the t-SNE method [24] was employed in order to visualize the 8 dimensional features in the output of the first network of DCCA. The aforementioned visualization method is an embedding technique, which builds upon the Stochastic Neighbor Embedding (SNE) [25]. In Fig. 1, the features extracted from the 3rd fold are shown. As can be observed, the descriptors resulting from the DCCA are discriminated to a large extent.

IV. Conclusions and Future Work

In this paper, we have addressed the classification of Greek folk songs according to their region of origin by employing the auditory cortical representations as feature descriptors for the music signal. In addition, the application of DCCA to the auditory cortical representations has been investigated. It has been demonstrated that both the SVM and the CCA classifiers yield an equivalent performance. The auditory cortical representations have enabled us to overcome the low recognition rates measured, when mel-frequency cepstral coefficients were extracted from the music recordings [11]. The classification of Greek folk songs according to their content (e.g., laments, marriage, satiric, historical) could be a subject of future research.

References

[1] Z. Fu, G. Lu, K. M. Ting, and D. Zhang, "A survey of audio-based music classification and annotation," *IEEE Trans. Multimedia*, vol. 13, no. 2, pp. 303–319, April 2011.

[2] A. Holzapfel and G. Tzanetakis, "Why is Greek music interesting? Towards an ethics of MIR," in *Proc. Int. Conf. Music Information Retrieval*, 2014.

[3] G. Spyridakis and S. D. Peristeris, *Greek Folk Songs*. Academy of Athens, 1999, vol. 3.

[4] E. Giouvanakis, C. Kotropoulos, A. Theodoridis, and I. Pitas, "A game with a purpose for annotating greek folk music in a web content management system," in *Proc. 18th Int. Conf. Digital Signal Processing*, July 2013, pp. 1–6.

[5] D. Conklin and C. Anagnostopoulou, "Comparative pattern analysis of cretan folk songs," in *Proc. 3rd Int. Workshop Machine Learning and Music*, 2010, pp. 33–36.

[6] D. Conklin, "Discovery of distinctive patterns in music," *Intell. Data Anal.*, vol. 14, no. 5, pp. 547–554, 2010.

[7] I. Antonopoulos, A. Pikrakis, S. Theodoridis, O. Cornelis, D. Moelants, and M. Leman, "Music retrieval by rhythmic similarity applied on Greek and African traditional music," in *Proc. 8th Int. Conf. Music Information Retrieval*, Vienna, Austria, 2007, pp. 297–300.

[8] A. Holzapfel and Y. Stylianou, "Rhythmic similarity of music based on dynamic periodicity warping," in *Proc. 2008 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 2008, pp. 2217–2220.

[9] ——, "Parataxis: Morphological similarity in traditional music," in *Proc. 11th Int. Conf. Music Information Retrieval*, 2010, pp. 453–458.

[10] S. Brilis, E. Gkatzou, A. Koursoumis, K. Talvis, K. L. Kermanidis, and I. Karydis, "Mood classification using lyrics and audio: A case-study in Greek music," in *Proc. Artificial Intelligence Applications and Innovations*, ser. IFIP Advances in Information and Communication Technology. Springer Berlin Heidelberg, 2012, vol. 382, pp. 421–430.

[11] N. Bassiou, C. Kotropoulos, and A. Papazoglou-Chalikias, "Greek folk music classification into two genres using lyrics and audio via canonical correlation," in *Proc. 9th Int. Symposium Signal and Image Processing and Analysis*, 2015, pp. 240–245.

[12] N. Mesgarani, M. Slaney, and S. A. Shamma, "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 920–930, 2006.

[13] Y. Panagakis, C. L. Kotropoulos, and G. R. Arce, "Music genre classification via joint sparse low-rank representation of audio features," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1905–1917, Dec. 2014.

[14] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. 30th Int. Conf. Machine Learning*, 2013, pp. 1247–1255.

[15] R. Munkong and B.-H. Juang, "Auditory perception and cognition," *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 98–117, May 2008.

[16] C. Schörkhuber and A. Klapuri, "Constant-Q transform toolbox for music processing," in *Proc. 7th Sound and Music Computing Conf.*, 2010.

[17] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3-4, pp. 321–377, 1936.

[18] D. R. Hardoon, S. R. Szedmak, and J. R. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, 2004.

[19] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer-Verlag New York, Inc., 2006.

[20] L. Sun, S. Ji, and J. Ye, "Canonical correlation analysis for multilabel classification: A least-squares formulation, extensions, and analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 33, no. 1, pp. 194–200, Jan. 2011.

[21] J. Nocedal and S. Wright, *Numerical Optimization*. Springer New York, 2006.

[22] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. 25th Int. Conf. Machine Learning*, ser. ICML '08, 2008, pp. 1096–1103.

[23] I. Guyon, J. Makhoul, R. Schwartz, and V. Vapnik, "What size test set gives good error rate estimates?" *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 1, pp. 52–64, 1998.

[24] L. Van Der Maaten and G. E. Hinton, "Visualizing high-dimensional data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[25] G. E. Hinton and S. T. Roweis, "Stochastic neighbor embedding," in *Advances in Neural Information Processing Systems 15*. MIT Press, 2003, pp. 857–864.