Musical Genre Classification Using Nonnegative Matrix Factorization-Based Features

André Holzapfel and Yannis Stylianou, *Member, IEEE*

***Abstract*—Nonnegative matrix factorization (NMF) is used to derive a novel description for the timbre of musical sounds. Using NMF, a spectrogram is factorized providing a characteristic spectral basis. Assuming a set of spectrograms given a musical genre, the space spanned by the vectors of the obtained spectral bases is modeled statistically using mixtures of Gaussians, resulting in a description of the spectral base for this musical genre. This description is shown to improve classification results by up to 23.3% compared to MFCC-based models, while the compression performed by the factorization decreases training time significantly. Using a distance-based stability measure this compression is shown to reduce the noise present in the data set resulting in more stable classification models. In addition, we compare the mean squared errors of the approximation to a spectrogram using independent component analysis and nonnegative matrix factorization, showing the superiority of the latter approach.**

***Index Terms*—Audio classification, audio feature extraction, music information retrieval, nonnegative matrix factorization.**

I. Introduction

In THE 1960s, in one of his last interviews, the brilliant saxophone player Eric Dolphy uttered the phrase: "When you hear music, after it's over, it's gone in the air; you can never recapture it again." Luckily he was wrong. Nowadays almost all music recordings are available in digital format, we can listen to them on our computers, we can buy them from the internet. This way, each kind of music went out of its traditional place of performance. We enjoy Mozart in the shopping mall and listen to the latest performance of the Rolling Stones at our computer at work. Every kind of music has gone to all the places. Musical genres interact and new styles are created.

With the growing availability of music on the Internet, this interaction grows even further. At the same time, there is an amazing opportunity in this widespread distribution and diversity of media. With the old distribution system of physical media on disks, the main focus was always restricted to some artists that were massively promoted, while much music was either only published in a limited edition or even never produced by any company. Thus, availability of music was strongly limited. However, throughout the recent years, many internet based distributors made recordings available for download.[1] Nowadays, every musician doing a recording is able to publish his/her work on the Internet. Obviously, in order for the listener to have a chance to find the music he likes, an automatic tool to retrieve information about the content of music pieces is necessary. A way to describe music by generating meta information in text format would fail for a decentralized system, as noted by Huron in [1], because of the strongly different ways members in a decentralized system describe their data. Again, according to [1], among the most suitable characteristics to get a description for musical data are style, instrumentation, tempo, and mood. The research in the automatic detection of the mood of a piece of music has first been approached systematically quite recently by Li and Ogihara [2]. However, the way humans react emotionally to specific pieces of music is still to be examined in a large-scale study, while there are no available data that could give a ground truth for evaluation.

The other mentioned characteristics are directly connected with the structure of the musical piece. This structure can generally be assumed to have a horizontal and a vertical direction. The horizontal direction contains information about onsets of the different instruments, and thus tells us about tempo and rhythm. Also, melody is partly described in the horizontal structure of music as it develops over time as well. Ways to automatically describe tempo and rhythm of musical pieces have been shown in [3], and recently a system for the classification of dance music based on the recognition of its rhythmic characteristics has been presented [4].

The vertical structure of music contains information about harmonic relations of the notes. The notes are reproduced by organs with characteristic frequency structures, which is referred to as the formant structure of an instrument [5]. These sounds have all been processed individually and/or together in a studio environment, thus changing their spectral characteristics. As such, we find information about instrumentation and production in the vertical structure; in music information retrieval (MIR), this is often referred to as the timbre of music. Moreover, experimental results lead to the conclusion that musical style is a characteristic found in the vertical structure as well. For example in [6], listeners were able to assign a piece of music to a style given an excerpt of duration less than one second. Recently, Li and Ogihara [2] received improved results in a genre classification task by using only spectral descriptors and neglecting temporal information. This can be interpreted as a supporting result for [6], since a musical genre is defined as a category of pieces that share a certain style [33]. Therefore, a system to automatically retrieve information about the vertical structure of music

Manuscript received December 15, 2006; revised August 2, 2007. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Dan Ellis.

The authors are with the Institute of Computer Science, Foundation for Research and Technology-Hellas (FORTH), GR-700 13 Heraklion, Crete, Greece, and also with the Computer Science Department, Multimedia Informatics Lab, University of Crete, 71409 Heraklion, Greece (e-mail: hannover@csd.uoc.gr; yannis@csd.uoc.gr).

Digital Object Identifier 10.1109/TASL.2007.909434

[1][Online]. Available: www.magnatune.com; www.freemusicdownload.com

will be capable of describing style, genre, timbre, and harmonic concept of the composition.

In many publications, the vertical dimension of music has been described by using a feature set consisting of Mel frequency cepstral coefficients (MFCCs). These features have been successfully applied to the task of speech recognition [34]. They have also found wide application in the classification of music into genres or in developing measures for the similarity of musical pieces as reviewed in [8]. In [8], it has been shown that systems following the general model of using MFCC-based features are upper bounded in their recognition performance.

An aspect that has not been considered in the development of the previously reported representation approaches is the fact that the characteristic timbre of the recordings is usually created by mixing several instruments into a single signal. Thus, an approach to derive descriptions of these components from the mixture signal could provide a more versatile feature set for the genre classification task. In [9], a method for the classification of sounds has been presented, where the spectral space of a signal is described using techniques based on independent component analysis (ICA) [16] applied to the spectrogram of the signal. Considering musical signals, methods based on a nonnegative matrix factorization (NMF) [10] have recently shown success in separating instruments from a mixture [12], [11]. NMF has been used as well for the classification of sounds in [13]–[15]. The classification approaches based on these techniques follow a deterministic path by first defining a set of spectral bases for the sounds and then projecting new sounds into these spaces.

In this paper, we first evaluate the factorization of spectrograms by using ICA- and NMF-related techniques. As NMF is shown to yield a compact representation and, compared to ICA, superior results in a mean squared error sense, we describe a signal spectrogram with the spectral space spanned by the vectors computed by this factorization approach. For a given musical genre, a Gaussian mixture model (GMM) is built on all the spectral base vectors that have been computed for the spectrograms of the training data for a particular class. In this way, we get a description for the spectral base of the particular genre. The classification is based on the maximum-likelihood (ML) considering all the spectral base vectors from a test signal. Extended classification tests were conducted on two widely used datasets for music classification (Tzanetakis *et al.* [21] and from the ISMIR 2004 contest[2]) comparing the performance of the proposed NMF-based features and that of MFCCs. The proposed NMF-based features constantly outperformed the MFCCs in terms of classification score. The proposed classification system was also compared to reference systems [21], [23], [25] for the task of music genres classification. The proposed classification system achieved higher classification score compared to these systems, in most of the conducted experiments, although [21] employs features that model both the vertical and horizontal structure of music.

The paper is structured as follows. Section II reviews and compares the approaches of ICA and NMF for the factorization of a music spectrogram providing evidence for choices like the number of components used in these types of factorization and the length of the input spectrogram. Section III presents the computation of the proposed NMF-based features along with the classification system based on these features. In Section IV, a baseline system using MFCC is presented, and a stability measure for GMM-based classifiers is developed. The conducted experiments are described in detail in Section V, while conclusions and discussions for future work are provided in Section VI.

[2][Online]. Available: http://ismir2004.ismir.net/ISMIR_Contest.html

II. Matrix Factorization

Our goal is to describe the vertical dimension of music in a compact and salient way. Optimally, this description should give us information about the components that have been mixed together in the musical sound. We suggest to obtain these descriptors from a temporal/spectral description of the sound using a matrix factorization. For this, the optimum approach to use has to be determined.

Let us assume a real signal to be stationary within a temporal window of length t_{win} (sec). After sampling the windowed signal at a frequency f_s, its *discrete Fourier transform* (DFT) will provide $N_{\mathrm{fft}} = t_{\mathrm{win}} f_s$ coefficients if no zero padding is used. Let $\mathbf{x}$ be an N_c dimensional column vector containing the magnitudes of the Fourier transform of the signal for frequencies up to the Nyquist frequency, where $N_c = N_{\mathrm{fft}}/2 + 1$. Assuming that $\mathbf{x}$ has been produced by linearly combined components as

$$\mathbf{x} = \mathbf{W}\mathbf{h} = \sum_{i=1}^{d} \mathbf{w}_i h_i \tag{1}$$

with $\mathbf{W}$ being an $N_c \times d$ matrix containing the description of the spectral content of the d mixture components in its columns $\mathbf{w}_i$, and $\mathbf{h}$ being a d dimensional weighting vector. Then, the problem of finding these components can be described in a blind source separation [30] context. We consider the value of d in the present problem to be smaller than the number of the frequency bins N_c as we want to get a compact representation of the signal. Taking k observation vectors $(\mathbf{x}_1, \ldots, \mathbf{x}_k)$ a matrix $\mathbf{X} \in \mathbb{R}^{N_c \times k}$, containing the observations in its columns, may be constructed. This matrix is usually referred to as spectrogram, and it describes the spectral content of the signal in a temporal range denoted by t_{Timbre} in this paper.[3] Setting the number of mixture components to a value $d \ll N_c$ we will usually not achieve equality as in (1) because of the time-varying spectral content of the initial components throughout the spectrogram. From a mathematical point of view, every column of $\mathbf{X}$ would have to be representable as a linear combination of the columns of $\mathbf{W}$, which is unlikely to happen for a nonartificial signal and $d \ll N_c$. Thus, (1) in matrix notation becomes

$$\mathbf{X} \approx \mathbf{W}\mathbf{H} \tag{2}$$

with the matrix $\mathbf{H} \in \mathbb{R}^{d \times k}$ containing the weighting vectors for time instances $1, \ldots, k$ in its columns. We can pursue this approximation task with a number of error functions and assumptions on the variables.

One approach is to choose a statistical framework. In this framework, $\mathbf{H}$ contains random variables (in $\mathbb{R}^d$) in its columns that are statistically independent. Then, given $\mathbf{X}$, we have to

[3]The term *timbre* is used here since within this window the description of the spectral space of the signal will be derived

search for a matrix $\mathbf{W}^{-1}$ that minimizes the mutual information between these independent components. This approach is based on independent component analysis and has been presented as independent subspace analysis (ISA) in [9]. A necessary condition in this framework is that the distributions of the d sources that are to be estimated remain stationary throughout the length t_{Timbre} of the spectrogram under consideration. It is worth to note that the values for t_{Timbre} range from 0.25 s up to 10 s, according to [9]. However, experiments to determine the influence of t_{Timbre} and d, when constrained to $d \ll N_c$, on the mean squared error (mse)

$$\mathrm{mse}(\mathbf{X}||\mathbf{WH}) = \sum_{i}^{N_c} \sum_{j}^{k} \frac{(\mathbf{X}_{i,j} - [\mathbf{WH}]_{i,j})^2}{(N_c k)} \tag{3}$$

of the approximation in (2) have not been conducted yet.

Without considering a statistical framework the NMF minimizes an error function like

$$D(\mathbf{X}||\mathbf{WH}) = \sum_{i,j} \left(\mathbf{X}_{i,j} \log \frac{\mathbf{X}_{i,j}}{[\mathbf{WH}]_{i,j}} - \mathbf{X}_{i,j} + [\mathbf{WH}]_{i,j} \right) \tag{4}$$

and constrains all the values in $\mathbf{W}$, $\mathbf{H}$, and $\mathbf{X}$ to be nonnegative [10]. Also for the NMF approach, experiments considering the influence of the length of the input spectrogram on the mse are not known to the authors. Nevertheless, it can be assumed that constraining the number of observations k is likely to cause $\mathbf{X}$ to span a vector subspace of $\mathbb{R}^{N_c}$ that can be spanned by a small number of d columns of $\mathbf{W}$. In terms of musical content, due to a shorter duration t_{timbre} less different instrumental sounds will be present in the spectrogram, which causes its columns to span a more compact subspace.

We evaluated both ISA and NMF on a set of music samples taken from a database used in [21]. The set consisted of 20 musical pieces of 30-s length each, two pieces randomly chosen from each of the ten classes contained in the data set. The software for evaluation was taken from the MPEG-7 reference software [17] as implemented by Casey. This includes the *fastICA* algorithm [26] for the calculation of ICA. The reference software was expanded by including an implementation of NMF without sparseness constraint as implemented in [18], that minimizes the cost function shown in (4). The choice of this cost function has been motivated by [19], where it was found to be subjectively superior to a squared error function in measuring spectral distances. This is assigned to the property of (4) to emphasize differences in regions with high energy, representing therefore a weighted contrast function. The block diagram of the evaluation algorithm is shown in Fig. 1. The power spectrum is estimated through the DFT of the signal, computed on a 40-ms Hamming window with 50% overlap. The next step is a conversion from the linear frequency abscissa to a logarithmic axis. Using eight bands per octave ranging from 65.5 Hz to 8 kHz results in $N_{\mathrm{bands}} = 56$ coefficients for each DFT window. This conversion is following the AudioSpectrumEnvelope descriptor (ASE) of the MPEG-7 standard. It enables a more compact description of the signal, i.e., it reduces dimensionality from the number of coefficients N_c on linear scale to N_{bands} . The choice of eight bands per octave has been motivated by the equal tempered musical system of western music, in which the most common tonal scales contain seven steps from the fundamental tone until its octave. Having computed the ASE vectors for a whole sample, a spectrogram representation is then obtained. This is segmented into smaller nonoverlapping subspectrograms that represent k ASE descriptors, a step denoted as *timbre windowing* in Fig. 1. Note that the number of observation vectors k defines the length of the timbre window (t_{timbre}). Varying the length of the timbre window t_{timbre} as well as the number of components d, while fixing the number of bands, $N_{\mathrm{bands}} = 56$, we may determine the mse of the factorizations produced by ISA and NMF. The samples of 30-s length were split into $N_B = [1, 2, 4, 8, 12, 16, 20, 30]$ segments of equal size. Spectrograms computed from these partitions were factorized with $d = [3, \ldots, 30]$ components. For example, for $N_B = 4$ segments, each segment is 7.5 s long (segments were obtained without overlap), resulting in $k = 7500$ ms$/20$ ms $= 375$, where a frame rate of 20 ms is assumed. For a given choice of splitting (i.e., $N_B = 4$) the corresponding mse was computed as the sum of mse from all segments. The number of components as well as the length of the input spectrogram influences the quality of the approximation provided by the two considered factorization methods (NMF and ISA). Increasing the number of components improved the approximation in both methods. This is because, with d increasing, the columns of $\mathbf{W}$ are more likely to construct a basis for the subspace of $\mathbb{R}^{N_{\mathrm{bands}}}$ spanned by the columns of $\mathbf{X}$. Two example error functions averaged over the parameter d are depicted in Fig. 2, showing that NMF is superior to ISA in the mean squared error sense for all numbers of partitions. This was consistently the case for all the songs in the set of music samples. Additionally, it can be seen that for shorter spectrograms (i.e., more partitions), the error gets smaller for NMF while it increases for ISA. Indeed, for shorter timbre windows, the value of k gets closer to d and in the extreme case of $k = d$, NMF will reach a perfect result by setting $\mathbf{W} = \mathbf{X}$ while $\mathbf{H}$ being the $k \times k$ identity matrix. On the other hand, the updates in *fastICA* use sample means in order to estimate expectation values, and because of this a short timbre window leads to worse approximations (see [26] for a description of the algorithm).

We conclude that computing NMF on short spectrograms leads to more adequate spectral representations for the signals under consideration. The optimal length and number of components in the classification task will be determined in Section V-B.

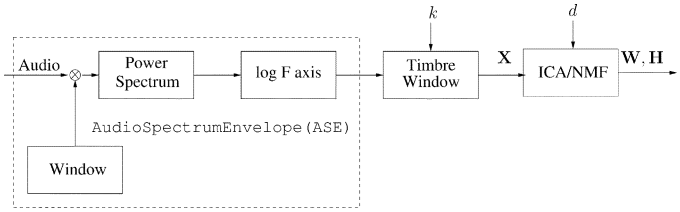

Fig. 1. Computation of spectral bases in the MPEG-7 reference.

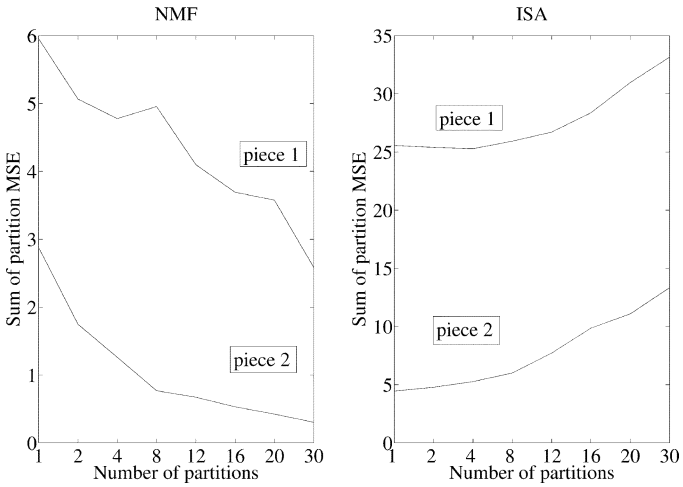

Fig. 2. Example of error curves of NMF and ISA for two pieces of music. Approximation by NMF has generally a smaller error than approximation by ISA.

Fig. 3. Calculation of the features used for the statistical model of musical genres.

III. System Description

A. Feature Calculation

The features describing the spectral space are calculated as shown in Fig. 3. The preprocessing steps avoid the influence of recording conditions which are not significant for classification. They include removal of mean values and normalization to an average sound pressure level of $L = 96$ dB. The next step is the computation of the ASE descriptors, as described in Section II. Then, the timbre window is applied to segment the spectrogram of the audio signal into nonoverlapping subspectrograms of size $N_{\text{bands}} \times k$, with $N_{\text{bands}} = 56$ and k represents the number of descriptors per subspectrogram. Each subspectrogram is then factorized using NMF providing a spectral base consisting of d vectors in the columns of matrix $\mathbf{W}$ in (2), with $d \ll k$. The next step transforms the energy values of the spectral bases into decibel scale, which has been shown to be crucial for an audio description task in [35]. The final step of the feature calculation is a *Discrete Cosine Transform* (DCT) on the dB-scale spectral base vectors; the size of the used DCT matrix is 20×56, containing the first 20 cosine bases $\sqrt{2/56}\cos[(2j+1)i\pi/(2\cdot 56)]$, $j = 0, \ldots, 55$, $i = 1, \ldots, 20$, in its rows. This helps to reduce the dimensionality of the space from 56 to 20. The resulting 20–dimensional vectors $\mathbf{v}_1, \ldots, \mathbf{v}_d$ represent the features of the presented system, and describe the spectral base of a subspectrogram in a compact way. The spectral space of the audio signal is described by the feature vectors computed from all its subspectrograms. Since the length of the timbre window is fixed, the number of subspectrograms computed from every song depends on its duration.

1) Psychoacoustic Model: Instead of using a logarithmic frequency axis in the $\log F$ *axis* box of Fig. 1, the introduction of a psychoacoustic model was evaluated as well. It consists of three elements:

Outer Ear Model: At each time instance a weighting is applied to the spectrum that adapts the calculated coefficients to the actually perceived loudness of the signal. The function presented by Terhardt [27] has been used

$$L_{TH} = \{3.64 f^{-0.8} - 6.5 \exp\left[-0.6(f-3.3)^2\right] + 10^{-3} f^4\} \text{ dB} \quad (5)$$

where L_{TH} represents the sound pressure level at hearing threshold and f denotes frequencies in kilohertz. It has the effect of emphasizing frequencies around 3 kHz and damping low frequencies.

Bark Scale: The linear frequency scale is converted to the *Bark* scale or critical band rate scale. This scale describes best the critical bandwidths of the human ear that lead to spectral masking when two frequencies are close enough to stimulate the same region of the basilar membrane. For an exact definition of this terminology, see [28]. The critical bandwidths remain constant for frequencies below 500 Hz and grow then in a nonlinear fashion, thus being different from the logarithmic frequency axis used in the experimental setups above. This leads to a conversion from frequencies in kilohertz to Bark which can be calculated as

$$\frac{z}{\text{Bark}} = 13 \arctan(0.76 f) + 3.5 \arctan\left(\frac{f}{7.5}\right)^2. \quad (6)$$

Using (6), the lower and upper frequency limits of critical bands smaller half the sampling frequency have been calculated. Because the sampling frequency of all used data is 16 kHz, the number of critical bands to be considered is 22. The values of the power spectrum within the frequency limits of the ith critical band, z_i, have been summed up for all bands to get the representation on the Bark scale.

Inner Ear Model: The model estimates the spread of masking between the critical bands caused by the structure of the ear's basilar membrane. The basilar membrane spreading function used to model the influence of the jth critical band on the ith band was derived by Schroeder in [29]

$$10 \log_{10} B(z_i, z_j) = 15.81 + 7.5((z_i - z_j) + 0.474) - 17.5(1 + ((z_i - z_j) + 0.474)^2)^{1/2} \text{ dB}. \quad (7)$$

A function for a specific Bark band is steeper to the side of low frequencies which indicates that spectral masking is more present towards higher frequencies. For each of the 56 bands, a function was computed using (7), resulting in a 22×22 matrix that was multiplied with the power spectrum on Bark scale. For all steps of the psychoacoustic model, the implementation of [25] has been used.

If the features used in this paper have some connection to the characteristics that are used by humans to categorize sounds, a further improvement by this alternative preprocessing procedure may be expected.

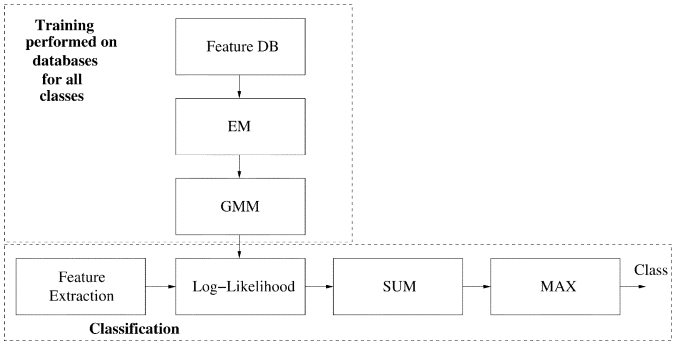

Fig. 4. Model estimation and classification of data.

B. Statistical Model and Classification

In order to construct the models for the musical genres we calculate the features for all audio signals of the database, i.e., the features $\mathbf{v_1}, \ldots, \mathbf{v_d}$ are computed for each subspectrogram, and then the features are stored for each class separately regardless of their temporal order in the samples. This is referred to as a *bag of frames* model in [7]. Then, a GMM θ^i for each genre is built (i.e., with $i = 1, \ldots, C$, where C denotes the number of genres), using a standard *expectation-maximization* (EM) algorithm [32]. The EM algorithm is initialized by a deterministic procedure based on the Gaussian means algorithm presented in [20]. A new song is classified into a genre by applying a maximum-likelihood criterion: For this, for all S feature vectors $\mathbf{v}_1, \ldots, \mathbf{v}_S$ collected from the subspectrograms of a test song, the likelihoods $p(\mathbf{v}_j|\theta^i)$, with $i = 1, \ldots, C$ and $j = 1, \ldots, S$, are computed. Summing up the log-likelihood values for each class, the song is assigned to the genre K that has the maximum score

$$K = \underset{i}{\operatorname{argmax}} \sum_{j=1}^{S} \log p(\mathbf{v}_j|\theta^i). \tag{8}$$

The principle of the model training and classification is depicted in Fig. 4. Our classification method differs from [7] as we do not build a statistical model for the song to classify. In this way, detailed information contained in the features is preserved. Design parameters of the GMM are provided in Section V.

IV. Performance Evaluation

In this paper, the performance of the presented system is evaluated in two different ways. At first, we compare its classification accuracy with the accuracy achieved by two alternative features sets, one using MFCC, and the other using randomly chosen spectral bases. Furthermore, a stability measure is suggested based on the distances between the statistical models built on the datasets used for the evaluation.

A. Two Alternative Feature Sets

In order to evaluate the performance of our classification approach based on NMF, it is necessary to compare with some kind of standard procedures used in many recent publications. For this purpose, a *baseline* system was implemented that is as close as possible to our classification system except for the feature calculation approach. The form of the baseline system was motivated by [8] which presents a frequently applied system for capturing the vertical structure of music. The model estimation and classification follow exactly the procedure depicted in Fig. 4. However, in the baseline system, 20 MFCCs are used instead of the NMF-based features. Note that in contrast to [7] and [8], no model is constructed for a song to be classified. Every feature vector is considered in the same ML-classification approach as described for NMF in Section III-B.

The second system to compare with differs from the NMF system only in the choice of spectral bases. These are simply d randomly chosen columns from each subspectrogram, which contains k columns as described in Section III-A. Comparing accuracies between this system, that will be referred to as a *random base* system, and a NMF-based system should clarify the impact of the matrix factorization in the whole classification concept.

B. Measure of Stability

In addition to comparing the performance of the proposed classification system with those of baseline and random base system, we suggest a method to quantify the quality of the classifiers based on a measure that estimates their sensitivity (or stability).

In order to judge the stability of the trained GMM, a method based on Kullback–Leibler divergence (KLD) was implemented. The Kullback–Leibler divergence between two distributions f and g is given by

$$KL(f\|g) = \int f(x) \log \frac{f(x)}{g(x)} dx. \tag{9}$$

Since there is no closed-form expression for KLD in a GMM context, a possible way to get a distance measure in this case is by generating M samples $x_1, \ldots, x_M$ from $f(x)$ and then approximate KLD, by [7]

$$KL(f\|g) \approx \frac{1}{M} \sum_{t=1}^{M} \log \frac{f(x_t)}{g(x_t)}. \tag{10}$$

Based on (10), a symmetric distance measure is constructed as

$$D_{KL}(f, g) = KL(f\|g) + KL(g\|f). \tag{11}$$

Let us assume that our dataset consists of C classes. Performing an n-fold cross validation, we will get a set of $n \times C$ GMMs described by their parameters θ_i^j, $1 \leq i \leq n$, and $1 \leq j \leq C$. For convenience, this set is shown as an $n \times C$ matrix in Fig. 5. We can now determine the distances between the GMMs of different classes using (11) for each of the n cross validation runs separately. For example, for the first run we would consider the mixture models marked by the horizontal ellipse. The minimum of these values throughout the cross validation runs gives us the least distance D_{inter} between two different classes. Then, the distances within the classes throughout the different cross validation runs are computed, for example for the first class the mixture models marked by the vertical ellipse would be considered. The biggest value along all classes D_{intra} gives us a measure of how much the model differs throughout the cross validation

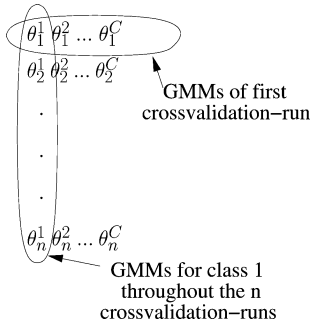

Fig. 5. Resulting GMMs from an n-fold cross validation.

due to diversity of the data set. We can now define a condition measure for a specific feature set, computed by

$$Cond_\theta = \frac{D_{\text{inter}}}{D_{\text{intra}}}. \tag{12}$$

Obviously, values for $Cond_\theta$ smaller than 1 for a specific feature set imply that a classification with this feature set might be unreliable. This is because there is a high variability between models built from a different set of data for a specific class, while at the same time there is a relatively small distance between the models for different classes. Note that using minimum and maximum values for D_{inter} and D_{intra} is a rather pessimistic approach. It penalizes a single outlier in the distances. For the intra class distance, this outlier could be the result of a single song that differed from the others in the training set and caused the model to vary strongly once it was moved from the training to the test set.

V. Experiments

A. Databases

For the experiments, two different data sets have been used. All the audio files of the databases have been converted to monaural wave files at a sampling frequency of 16 000 Hz quantized with 16 bits. The first database (D1) consists of ten classes,[4] each containing 100 subsections of musical pieces of 30-s length. The database was collected by Tzanetakis [21] and has been used for performance evaluation also by other researchers [23]. The second database (D2) was downloaded from the website of the ISMIR contest in 2004,[5] where it served as training set for the genre classification contest. The songs had been selected from the *magnatune*[6] collection. D2 consists of six classes.[7] It contains 729 songs that are not equally distributed among the classes as they are in D1. Also, the pieces are full musical pieces and not snapshots as in D1; therefore, the lengths of the pieces in D2 differ. As proposed for the MIREX 2005 evaluation,[8] a fivefold cross validation has been used. The whole data set has been used; stratified cross validation has been applied. All the classification accuracies shown in this paper are results of cross validations.

[4]Blues, Classical, Country, Disco, Hip Hop, Jazz, Metal, Pop, Reggae, Rock.

[5][Online]. Available: http://ismir2004.ismir.net/genre_contest/index.htm.

[6][Online]. Available: www.magnatune.com.

[7]Classical, Electronic, Jazz, Metal/Punk, Rock/Pop, World.

[8][Online]. Available: http://www.music-ir.org/mirex2005/index.php/Audio_Genre_Classification.

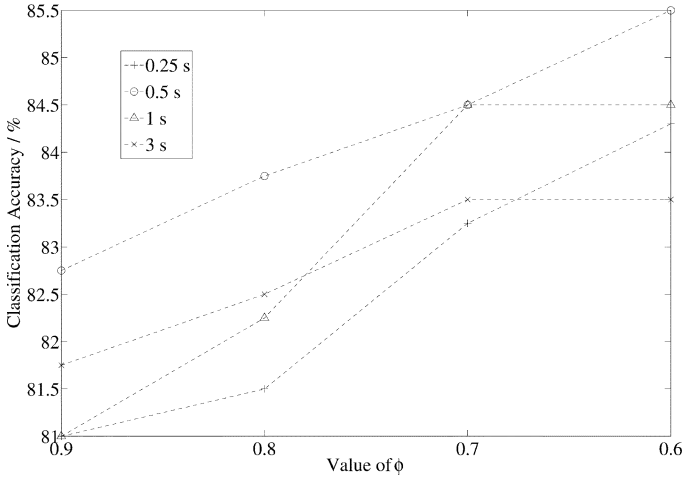

Fig. 6. Classification accuracies for varying timbre window length and value of ϕ.

B. System Parameters

For classification purposes, the optimum values for the temporal length t_{Timbre} of the timbre window and the number d of spectral base vectors to compute, should be defined. Values for t_{Timbre} from 0.25 to 3 s have been tested. A value for d is computed by varying the values of ratio ϕ defined as

$$\phi \leq \frac{\sum_{j=1}^{d} \sigma_j}{\sum_{i=1}^{N_{\text{bands}}} \sigma_i} \tag{13}$$

from 0.9 to 0.6, where σ_i denotes the ith singular value of the *singular value decomposition* (SVD) of the spectrogram to be factorized. Therefore, d provides an estimation of the minimum number of components necessary for preserving the amount of variance in the spectral basis as defined by ϕ.

These two system parameters have been defined using a subset of four classes (classical, disco, metal, rock) from the first database. A subset was chosen for computational efficiency and in order to avoid overfitting the system parameters to the whole data set. The subset contains two classes that revealed to be easily classified in preliminary experiments (classic and metal), as well as two problematic classes (rock and disco). A mixture of Gaussians with five components using full covariance matrices has been built for each genre (see Section III-B for details). Fig. 6 depicts the accuracies depending on ϕ and d. The optimum length of the timbre window is half a second while the rising accuracy for reduced values of ϕ implies that further decrease may provide even better results. However, this often leads to a value for d equal to one, especially when t_{Timbre} takes a small values. Indeed, in this case one eigenvector of the sample covariance matrix $\mathbf{X}^T\mathbf{X}$ describes a sufficient amount of the data variance [according to (13)]. Setting d to 1 leads to numerical problems in the EM algorithm because some covariance matrices are close to be singular. From this we conclude that we have to assure that $d > 1$, taking therefore into account

TABLE I
MEAN VALUES FOR THE NUMBER OF SPECTRAL BASE VECTORS

		ϕ			
		0.9	0.8	0.7	0.6
$t_{Timbre}(s)$	0.25	5.38	3.60	2.62	1.96
	0.5	8.01	5.06	3.55	2.60
	1	11.30	6.95	4.76	3.41
	3	15.92	9.90	6.71	3.77

TABLE II
CLASSIFICATION ACCURACIES (%) AFTER FIVEFOLD CROSS VALIDATION

	Database 1	Database 2
NMF(5)	71.7	75.7
NMF(10)	74.0	83.5
NMF(15)	73.9	77.7
NMF(20)	73.2	78.6
NMF(30)	-	78.5
NMF(40)	-	78.4
MFCC(10)	70.3	60.0
MFCC(20)	71.6	61.1
MFCC(30)	73.0	67.7
MFCC(40)	72.3	67.3

also directions of additional eigenvectors. We did experiments on the same dataset fixing t_{Timbre} to 0.5 s and set $d = [2, 3, 4]$. We found that the classification accuracies were best for $d = 3$. This result is supported by considering the values listed in Table I, which are the mean values of d determined using (13) to achieve the results displayed in Fig. 6. In Table I, the value of d corresponding to the best classification accuracy score ($\phi = 0.6$, $t_{\mathrm{Timbre}} = 0.5$ s) in Fig. 6 is close to 3. Therefore, in the following t_{Timbre} was set to 0.5 s, and d was set to 3. In this way, a meaningful representation of the signal space is achieved while the stability of the EM algorithm is assured.

C. Classification Results

Table II shows the classification accuracies on the two databases in per cent. The rows marked with NMF contain results achieved with the system presented in Sections III-A and III-B, while rows marked with MFCC contain results achieved with the baseline system as outlined in Section IV-A. The values in parentheses denote the number of Gaussians used. Full covariance matrices have been used for all experiments. We observed covariance matrices to have strong diagonals but we estimate full matrices in order to model possible covariances between the variables. For both feature sets (MFCC and NMF), the number of Gaussians had been varied in steps of five from 5 to 40. In the following Tables results that do not provide additional information have been left out to improve comprehensibility of the representation (i.e., for instance MFCC with 15 Gaussian components). For the fields with missing values for D1, training was not possible, because of the high compression performed by NMF on the training dataset. Using the bigger database D2, we were able to increase the number of components without serious estimation problems. In this case, the influence of the number of Gaussians on the classification accuracy may be observed.

The results show that our system outperforms the baseline system on both databases. On D1, the NMF-based system outperforms the baseline system slightly, but only ten Gaussian components are necessary to reach optimum performance for the presented system, while the baseline performs best using 30 mixture components. For D2, the performance superiority of the NMF system is more noticeable. Also here, the proposed system achieves best results using ten components, while for the baseline system (MFCC) this is achieved using 30 components. The decline of the classification accuracy with the increased number of Gaussians may be attributed to overfitting. The dependency of the classification accuracy on the number of Gaussians for MFCC agrees with the findings in [8]. There, for 20 MFCC the best performance of the system was reached with 50 components, with slightly decreasing results when exceeding this value. Probably the lower number of components used in the baseline system for achieving the highest score can be assigned to the usage of full covariance matrices that capture correlations not extincted by the orthogonal basis of the DCT matrix used in the MFCC calculation. For the NMF features, the optimum number of Gaussians is 10. This shows that more complex models do not capture significant structure in the data anymore. Thus, the usage of NMF simplified the densities of the data while keeping the significant differences between the classes.

The accuracies of the random base system have been extremely low for all used number of Gaussians. When comparing to the best performing systems, i.e., NMF(10), the random base system with ten Gaussian components achieved accuracies of 20.2% (compared to 74.0%) and 22.8% (compared to 83.5%) on D1 and D2, respectively. This proves the importance of using of NMF in the computation of the spectral bases.

It is worth to note that the NMF system is trained very fast. The data reduction performed by the matrix factorization reduces a spectrogram of half a second length (25 DFT-coefficient vectors using a frame rate of 20 ms) to three spectral base vectors. This yields a data compression of 88%. This is advantageous regarding training times: training a 20-component model on the first database took about 20 times longer using the baseline system (MFCC) instead of the NMF-based system. The computation of the features for NMF took longer than computing MFCC because of the rescaled gradient descent algorithm used in NMF (about 2.3 times longer). However, summing up times for feature calculation and training, the NMF-based system is still about six times faster than the MFCC-based system. This difference in time grows nonlinearly with the number of Gaussians.

Even though the system suggested in this paper captures only information about the vertical characteristics of music, it also performs well in comparison with approaches incorporating more versatile feature sets that partly include *both* vertical and horizontal directions. On D1, Li and Tzanetakis [21] report an accuracy of 71% using a feature set containing MFCC and FFT-derived characteristics as well as information about beat and pitch, and linear discriminant analysis as classifier. The first author of [21] presents a score of 79.5% using DWCH[9] as best performing feature and SVM as a classifier, while using GMM with three Gaussian components, an accuracy of 63.5% is reported [22]. Lidy and colleagues [35] report an

[9]Daubechies wavelet coefficient histogram.

TABLE III
CONFUSION MATRIX FOR DATABASE 1, USING NMF-BASED FEATURES [NMF(10)]

	Bl	Cl	Co	Di	Hi	Ja	Me	Po	Re	Ro
Bl	68	1	3	0	1	4	0	1	8	3
Cl	0	94	0	0	0	4	0	0	0	0
Co	12	1	73	6	0	2	1	7	5	16
Di	3	0	10	69	8	5	4	6	2	11
Hi	0	0	0	6	69	2	1	2	12	2
Ja	1	2	0	0	1	79	0	1	1	0
Me	2	0	2	1	2	2	83	0	0	5
Po	1	0	4	10	3	1	0	79	2	2
Re	3	0	0	2	13	1	0	2	69	4
Ro	10	2	8	6	3	0	11	2	1	57

TABLE IV
CONFUSION MATRIX FOR DATABASE 2, USING NMF-BASED FEATURES [NMF(10)]

	cl	el	ja	mp	rp	wo
cl	300	1	0	0	0	10
el	0	103	0	1	8	24
ja	0	0	25	0	0	0
mp	1	0	0	32	16	2
rp	6	7	0	10	69	8
wo	13	4	0	2	7	76

accuracy of 74.9% on D1, using an SVM classifier on features describing spectral and temporal structure of a song. Pampalk and colleagues presented an accuracy on D2 of 81% using a combination of spectral descriptors and a descriptor for modulations present in the signal, which are referred to as *fluctuation patterns* [24]. Using the training and development set of the ISMIR 2004 Audio description contest as a data set, the system presented in [35] was reported to achieve an accuracy of 80.3%.

For sound classification approaches that are based on spectral projections and HMM, as for example [14] and [15], no results on the presented databases are known to the authors. Nevertheless, the approach presented in [14] has been implemented by the authors and tested on D1, resulting in an accuracy of 50% in a fivefold cross validation. This indicates the superiority of the approach presented in this paper to the mentioned projection-based approaches, at least in the context of musical genre classification.

Another important conclusion can be drawn by comparing the results of the baseline system on D2 with the results of [24], where MFCC have been used as an alternative feature set as well. The baseline system presented in this paper does not build a statistical model of a song, but considers each MFCC vector separately by calculating its likelihood given the class models. In [24], songs have been modeled by Gaussians. This leads to an improvement in the classification accuracy of about 17% compared to our baseline system. Thus, it seems that by modeling the feature distribution for a song using GMM, results are improved, a finding confirmed in [7] in an artist identification task. Based on the above observations it would be interesting to check if such a modeling approach will be also beneficial for the NMF-based system, although such an approach is computationally quite expensive.

Confusion matrices using NMF-based features are provided in Tables III and IV for D1 and D2, respectively, using ten Gaussians [NMF(10)]. The columns contain the actual genres of the test data and rows contain the predicted classification. Apart from illustrating the above referred results and observations, Table IV can be contrasted with the matrices shown in the ISMIR 2004 genre classification contest.[10] In most cases, misclassifications have musical sense. For example, the genre Rock in D1 was confused most of the time with Country, while a Disco track is quite possible to be classified as a Pop music piece. In D2, the Rock/Pop genre was mostly misclassified as Metal/Punk. Genres which are assumed to be very different, like Metal and Classic, were never confused. The worst classification performance for the proposed system was: Rock in D1 [57%, NMF(10)] and World in D2 [63.3%, NMF(10)]. It is worth to note that this behavior in performance is similar to other systems as well (see ISMIR genre contest results). The low performance for these genres may be assigned to their large intravariance of music style (at least for the analyzed data).

TABLE V
PERFORMANCE WITH AND WITHOUT A PSYCHOACOUSTIC MODEL (%), NMF(10)

	Database 1	Database 2
Psychoacoustic Model	68.1	72.1
Best Psychoacoustic	72.8	77.1
Log frequency scale	74.0	83.5

1) Psychoacoustic Model: The psychoacoustic processing described in Section III-A was included into the feature calculation as depicted in Fig. 1 in the place of the simple log frequency conversion rule. All the other components of the system have been left as before, and the results of the classification have been compared with the best performing systems on D1 and D2, i.e., NMF(10) in both cases. Classification results are shown in the first row of Table V. For convenience, the best scores from Table II for log frequency rule are repeated in the third row. The introduction of the psychoacoustic preprocessing deteriorated the performance of the system noticeably. Experiments have been conducted in order to evaluate the influence of the individual steps of the preprocessing, i.e., the outer ear model, the Bark scale, and the inner ear model. On D1, using only Bark scale without inner/outer ear models performed best. On D2, Bark scale used together with the outer ear model slightly outperformed the complete psychoacoustic model. The accuracies of these two settings are denoted in the second row of Table V. It can be resumed that neither a partial usage of the psychoacoustic preprocessing lead to improved performance. If the psychoacoustic model efficiently describes the perception system, we would expect the classification results to be better than in the case of using the simple log frequency conversion rule. Therefore, either the model does not describe the perception process efficiently, or the features as input to the system have nothing to do with the cues used by humans for classifying a musical piece. Note that in [35], the influence of the particular parts of psychacoustic preprocessing on the accuracy in a genre classification task has been analyzed. The result is the outer ear model being a crucial part of the preprocessing, which is contradictory to our results. As the psychoacoustic model used in [35] is similar with the one used in this paper, a reason for the bad performance of

[10][Online]. Available: http://ismir2004.ismir.net/genre_contest/results.htm.

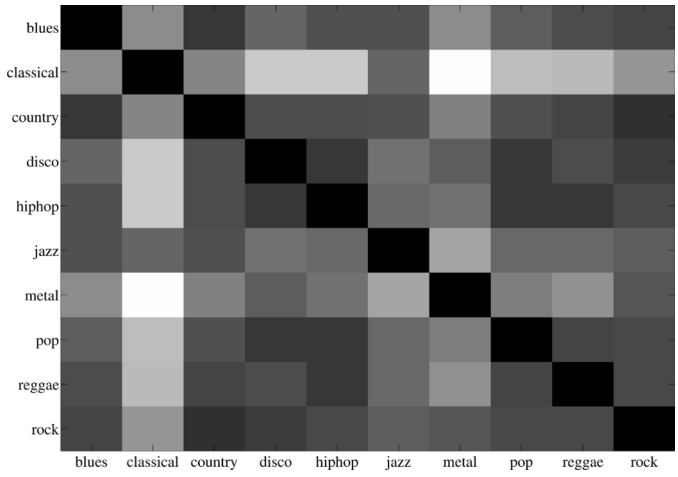

Fig. 7. Inter class distance matrix for NMF(10) on D1.

the psychoacoustic model could be the combination of this specific preprocessing with NMF.

D. Stability Measures

As introduced in Section IV-B, the stability of a given GMM-based classifier is estimated based on distances between the models for the particular classes according to (12). Table VI shows these condition numbers for all different configurations that had been depicted in Table II. The condition numbers are always bigger for the proposed NMF-based model than for the MFCC-based model. Only for five components the NMF-based features have a condition number less than 1. This can be attributed to the existence of components with large variance. Moreover, with more than ten components, the condition numbers for the NMF features are consistently bigger than one, while for the baseline system all the condition numbers are smaller than one. This indicates that for the NMF-based features, the smallest inter class distance is always bigger than the biggest intra class distance; this is not the case for MFCC. This provides a further proof of the superiority of the proposed feature set compared to MFCC. As an example, we show a graphical representation of the inter class distances for NMF(10) model on D1 in Fig. 7. The mean values of the inter class distances from the fivefold cross validations have been calculated; dark areas indicate a low distances and light areas indicate higher distances. It is evident that there is a high correlation between the confusion matrix in Table III and the distances depicted in Fig. 7 [computed using (11)]. Note that for the NMF-based features, there is also a high correlation between the condition numbers in Table VI and the classification accuracies in Table II: The condition numbers of the NMF-based system rise until a certain number of Gaussians that is bigger than the optimal in the classification accuracy sense (15 instead of 10 for D1, 20 instead of 10 for D2, compare with Table II). Beyond this maximum, the condition numbers decrease. A similar pattern may be observed for the classification score in Table II. However, this structure is not clear for the MFCC based system.

Taking a detailed look at all the measured inter and intra class distances reveals a more informative insight into the different characteristics of the feature space modeling. Sorting all the

TABLE VI
CONDITION NUMBERS

	Database 1	Database 2
NMF(5)	0.85	0.69
NMF(10)	1.33	1.27
NMF(15)	1.62	1.29
NMF(20)	1.53	1.37
NMF(30)	-	1.20
NMF(40)	-	1.15
MFCC(10)	0.88	0.56
MFCC(20)	0.86	0.55
MFCC(30)	0.89	0.64
MFCC(40)	0.92	0.52

intra class distances in increasing order gives the plots shown in Fig. 8 for D1 and in Fig. 9 for D2. The total number of computed distances in Figs. 8 and 9 is given by $C(n(n-1)/2)$, where $n = 5$ is the number of cross validations, and C is the number of classes ($C = 10$ for D1 and $C = 6$ for D2). As a common difference between the two feature sets, we can recognize that the intra class distances between the NMF-based models are more evenly distributed. This is indicated by a less steep gradient of the corresponding curves in Figs. 8 and 9. In these figures, we show the intra class distances for the number of components that provided the best classification score for both features; 30 for MFCC and 10 for NMF-based features. A similar behavior for both features has been observed for other numbers of components. However, for five components in the case of NMF-based features, the steepness of the corresponding curve was high, which caused the condition number to be smaller than one. The more evenly distribution of the intra class distances can be also observed from their detailed illustration in Fig. 10. Increasing the number of Gaussians results in more uniform distributed intra class distances (Fig. 10). This is not the case for MFCC features (Fig. 11). Similar observations can be also made for the inter class distances. The sorted inter class distances for both features are depicted in Figs. 12 and 13 for D1 and D2, respectively. The total number of computed distances in Figs. 12 and 13 is given by $n(C(C-1)/2)$, where $n = 5$ is again the number of cross validations, and C is the number of classes ($C = 10$ for D1 and $C = 6$ for D2).

VI. CONCLUSION

We suggest a new feature set based on NMF of the spectrogram of a music signal for the description of the vertical structure of music for the task of automatic musical genre classification. Extended experiments on two widely used databases showed the superiority of the proposed features compared to the standard feature set of MFCC. By using Kullback–Leibler-based distance measures, we were able to connect the superiority of the NMF-based features in the classification task with more uniform, compared to the MFCC case, intra class distances. In addition, the proposed feature extraction algorithm has the advantage of low training times of the mixture models due to the data compression and the lower number of Gaussians necessary to reach the optimum classification accuracy. Tests with a psychoacoustic preprocessing did not improve the classification accuracy. As mentioned in the previous sections, the feature set developed here is capable of describing the vertical

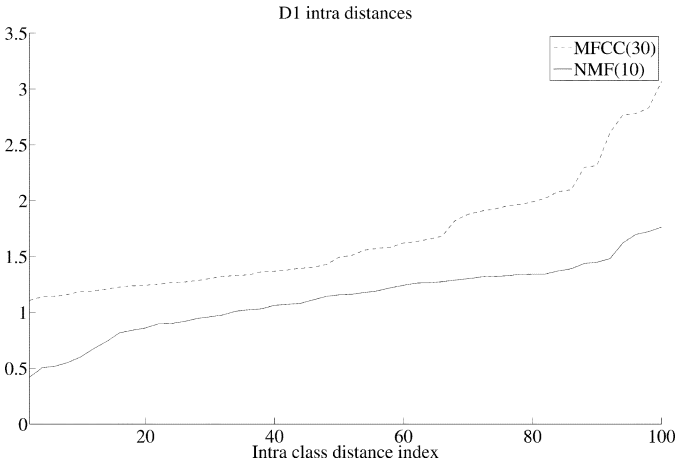

Fig. 8. Sorted intra class distances for D1, NMF: solid line, MFCC: dotted line.

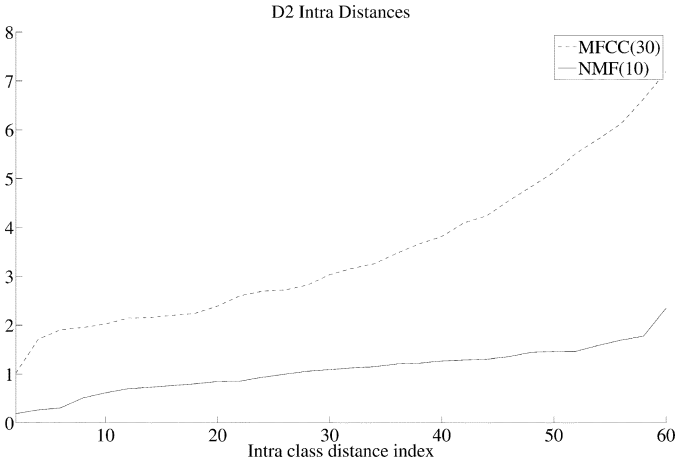

Fig. 9. Sorted intra class distances for D2, NMF: solid line, MFCC: dotted line.

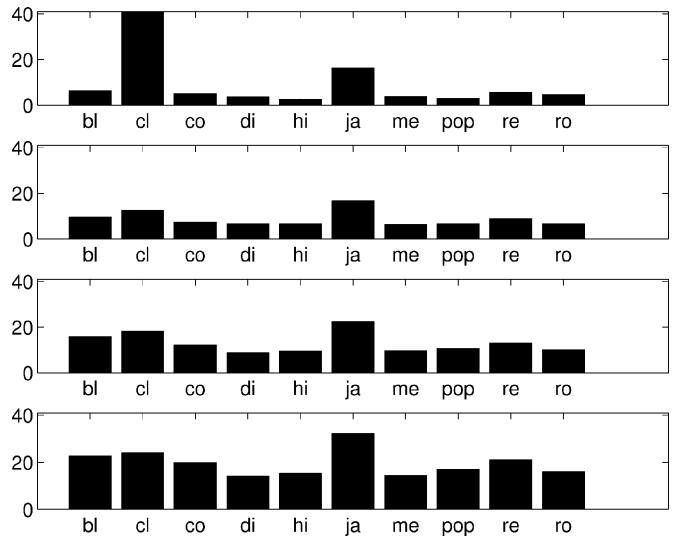

Fig. 10. Distribution of the intra class distances for NMF on D1 using 5, 10, 15, and 20 Gaussians (from top to bottom).

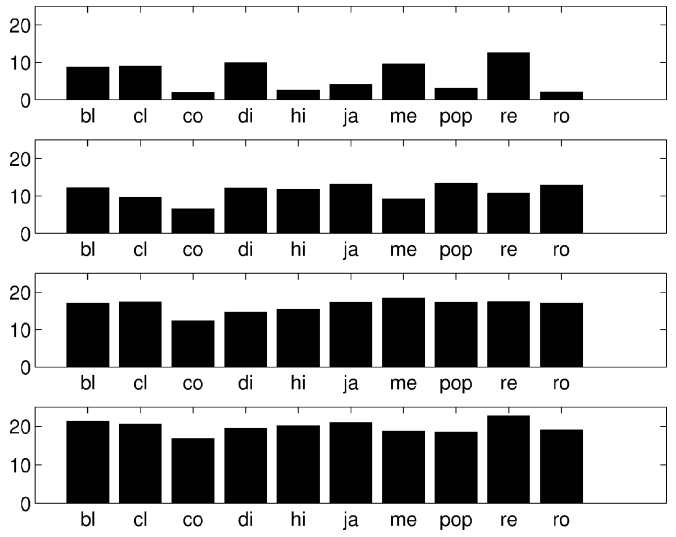

Fig. 11. Distribution of the intra class distances for MFCC on D1 using 5, 10, 20, and 40 Gaussians (from top to bottom).

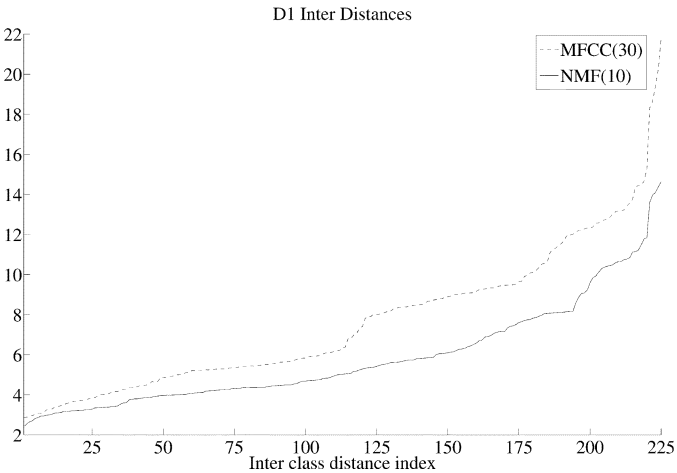

Fig. 12. Sorted inter class distances on D1.

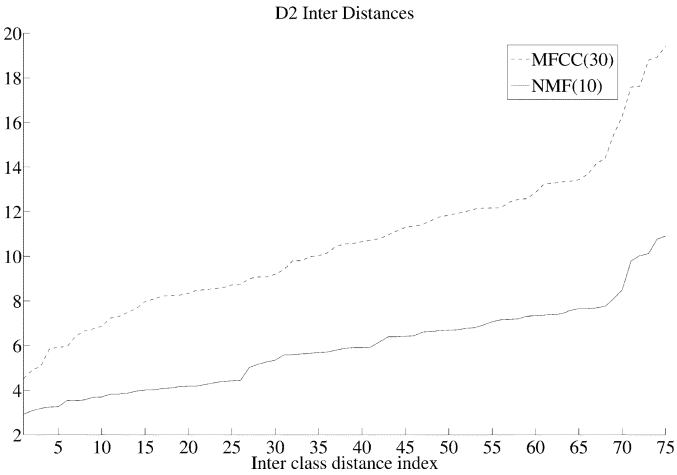

Fig. 13. Sorted inter class distances on D2.

structure of music. The next step will be to derive descriptors for the horizontal dimension. Therefore, future work includes the modeling of rhythm and modulation characteristics for a piece of music based on the NMF approach. A possible starting point for this work is the use of the rows of matrix $\mathbf{H}$ in (2).

REFERENCES

[1] D. Huron and B. Aarden, "Cognitive Issues and approaches in music information retrieval," 2002 [Online]. Available: http://www.musicog.ohio-state.edu/Huron/publications.html, unpublished.

[2] T. Li and M. Ogihara, "Toward intelligent music information retrieval," *IEEE Trans. Multimedia*, vol. 8, no. 3, pp. 564–574, Jun. 2006.

[3] E. D. Scheirer, "Music listening systems," Ph.D. dissertation, Mass. Inst. Technol., Cambridge , MA, 2000.

[4] G. Peeters, "Rhythm classification using spectral rhythm patterns," in *Proc. 6th Int. ISMIR Conf.*, London, U.K., 2005, pp. 644–647.

[5] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*. New York: Prentice-Hall, 2001.

[6] D. Perrott and R. Gjerdingen, "Scanning the dial: An exploration of factors in identification of musical styles," in *Presentation 1999 Soc. Music Perception Cognition Conf.*, Evanston, IL, p. 88.

[7] M. I. Mandel and D. P. W. Ellis, "Song-level features and support vector machines for music classification," in *Proc. 6th Int. ISMIR Conf.*, London, U.K., 2005, pp. 594–599.

[8] F. Pachet and J.-J. Aucouturier, "Improving timbre similarity: How high is the sky?," *J. Negative Results Speech Audio Sci.*, vol. 1.1, 2004.

[9] M. Casey, "General sound classification and similarity in MPEG-7," *Organized Sound*, vol. 6, no. 2, pp. 153–164, 2001.

[10] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[11] B. Wang and M. D. Plumbley, "Musical audio stream separation by non-negative matrix factorization," in *Proc. Digital Music Res. Netw. Summer Conf. (DMRN)*, Glasgow, U.K., 2005.

[12] P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," in *Proc. 5th Int. Conf. Independent Compon. Anal. Blind Signal Separation*, Granada, Spain, 2004, pp. 494–499.

[13] E. Benetos, M. Kotti, and C. Kotropoulos, "Musical instrument classification using non-negative matrix factorization algorithms and subset feature selection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Toulouse, France, 2006, pp. V-221–V-224.

[14] H. G. Kim, J. J. Burred, and T. Sikora, "How efficient is MPEG-7 for general sound recognition?," in *Proc. 25th Int. AES Conf.*, London, U.K., 2004.

[15] Y. C. Cho, S. Choi, and S. Y. Bong, "Non-negative component parts of sound for classification," in *Proc. IEEE Int. Symp. Signal Process. Inf. Technol. (ISSPIT)*, Darmstadt, Germany, 2003, pp. 633–636.

[16] P. Comon, "Independent component analysis, A new concept?," *Signal Process.*, vol. 36, pp. 287–314, 1994.

[17] B. S. Manjunath, P. Salembier, and T. Sikora, *Introduction to MPEG-7*. New York: Wiley, 2002.

[18] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *J. Mach. Learn. Res.*, vol. 5, pp. 1457–1469, 2004.

[19] E. Klabbers and R. Veldhuis, "Reducing audible spectral discontinuities," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 1, pp. 39–51, Jan. 2001.

[20] G. Hamerly and C. Elkan, "Learning the k in kmeans," *Adv. Neural Inf. Process. Syst.*, vol. 16, 2003.

[21] T. Li and G. Tzanetakis, "Factors in automatic musical genre classification of audio signals," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust.*, New Paltz, NY, 2003, pp. 143–146.

[22] T. Li, M. Ogihara, and Q. Li, "A comparative study on content-based music genre classification," in *Proc. 26th ACM SIGIR Conf.*, Toronto, ON, Canada, 2003, pp. 282–289.

[23] J. Bergstra, N. Casagrande, D. Erhan, D. Eck, and B. Kégl, *Aggregate Features and ADABOOST for Music Classification*. Norwell, MA: Kluwer, 2006.

[24] E. Pampalk, A. Flexer, and G. Widmer, "Improvements of audio-based music similarity and genre classification," in *Proc. 6th Int. ISMIR Conf.*, London, U.K., 2005.

[25] E. Pampalk, "A matlab toolbox to compute music similarity from audio," in *Proc. 5th Int. ISMIR Conf.*, Barcelona, Spain, 2004.

[26] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Netw.*, vol. 35, no. 4–5, pp. 411–430, 2000.

[27] E. Terhardt, "Calculating virtual pitch," *Hear. Res.*, vol. 1, pp. 155–182, 1979.

[28] E. Zwicker and H. Fastl, *Psychoacoustics—Facts and Models*. Springer: New York, 1990.

[29] M. R. Schroeder, B. S. Atal, and J. L. Hall, "Optimizing digital speech coders by exploiting masking properties of the human ear," *J. Acoust. Soc. Amer.*, vol. 66, no. 6, pp. 1647–1652, 1979.

[30] S. Haykin, *Adaptive Filter Theory*, 4th ed. Upper Saddle River, NJ: Prentice-Hall, 2002.

[31] K. Kreutz-Delgado, J. F. Murray, B. D. Rao, K. Engan, T. W. Lee, and T. J. Sejnowski, "Dictionary learning algorithms for sparse representations," *Neural Comput.*, vol. 15, pp. 349–396, 2003.

[32] L. Baum and J. Eagon, "An inequality with applications to statistical estimation for probalistic functions of Markov processes and to a model for ecology," *Amer. Math. Soc. Bull.*, vol. 73, pp. 360–363, 1967.

[33] P. van der Merwe, *Origins of the Popular Style: The Antecedents of Twentieth-Century Popular Music*. Oxford, U.K.: Clarendon, 1989.

[34] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[35] T. Lidy and A. Rauber, "Evaluation of feature extractors and psychoacoustic transformations for music genre classification," in *Proc. 6th Int. ISMIR Conf.*, London, U.K., 2005, pp. 34–41.

Andre Holzapfel received the graduate engineer degree in media technology from the University of Applied Sciences, Duesseldorf, Germany, and the M.Sc. degree in computer science from University of Crete, Heraklion, Crete, Greece, where is currently pursuing the Ph.D. degree.

His research interests are in the field of speech processing, music information retrieval, and ethnomusicology.

Yannis Stylianou (M'03) received the diploma of electrical engineering degree from the National Technical University of Athens (NTUA), Athens, Greece, in 1991 and the M.Sc. and Ph.D. degrees in signal processing from the Ecole National Superieure des Telecommunications (ENST), Paris, France, in 1992 and 1996, respectively.

He is an Associate Professor in the Department of Computer Science, University of Crete, Heraklion, Crete, Greece. From 1996 to 2001, he was with AT&T Labs Research, Murray Hill and Florham Park, NJ, as a Senior Technical Staff Member. In 2001, he joined Bell-Labs Lucent Technologies, Murray Hill, NJ. Since 2002, he has been with the Computer Science Department, University of Crete. He holds nine patents and has many publications in edited books, journals, and conference proceedings. Currently, he is Associate Editor of the *EURASIP Journal on Speech, Audio, and Music Processing* and of the *EURASIP Research Letters in Signal Processing*

He is member of the IEEE Speech and Language Technical Committee and Vice Chairman of the Cost Action 2103: "Advanced Voice Function Assessment." He is a member of the Technical Chamber of Greece.