MUSICAL GENRE CLASSIFICATION VIA GENERALIZED GAUSSIAN AND ALPHA-STABLE MODELING

C. Tzagkarakis, A. Mouchtaris, and P. Tsakalides

Department of Computer Science, University of Crete and
Institute of Computer Science (FORTH-ICS)
711 10 Heraklion, Crete, Greece

ABSTRACT

This paper describes a novel methodology for automatic musical genre classification based on a feature extraction /statistical similarity measurement approach. First, we perform a 1-D wavelet decomposition of the music signal and we model the resulting subband coefficients using the Generalized Gaussian Density (GGD) and the Alpha-Stable distribution. Subsequently, the GGD and Alpha-Stable distribution parameters are estimated during the feature extraction step, while the similarity between two music signals is measured by employing the Kullback-Leibler Divergence (KLD) between their corresponding estimated wavelet distributions. We evaluate the performance of the proposed methodology by using a dataset consisting of six different musical genre sets.

1. INTRODUCTION

In recent years, there has been a rapid proliferation of multimedia databases, which caused an urgent need to create effective methods for classification and retrieval of multimedia data. In the music domain, we would like to characterize the genre of a music track from the information that "hides" inside its content. This means that we must exploit the internal attributes of a music track by processing it either in the time or in the frequency domain. A successful way to compare and classify music tracks would allow for constructing better browsing systems and would lead to better performance of Music Information Retrieval (MIR) systems.

Automatic musical genre classification is a fundamental component of MIR systems. Typically in a genre categorization process there are two major tasks: Feature Extraction (FE) and Similarity Measurement (SM). In the FE task, a set of features is generated to accurately represent the content of a given music signal. The dimensionality of this set has to be smaller than the original signal while capturing as much as possible of the signal information. In the SM task, a distance function is employed to measure how close a query music signal is to each of the music categories in the dataset, by comparing their features. The query signal is then classified to the genre that is associated with the minimal of the measured distances.

There has been a lot of research in feature extraction for classification of speech signals (*e.g.* speaker identification), but work on music signal classification has only recently gained momentum. In [1], features that describe characteristics of the music such as rhythm, timbral texture, and pitch content are derived for musical genre classification. In [2], the problem of automatic music artist classification (artist identification) is considered in a statistical framework, where a set of features is modeled using a Gaussian Mixture Model (GMM). The SM task is performed using an approximation of the KLD between GMMs (Asymptotic Likelihood Approximation (ALA) [3]), since no closed form expression exists for the KLD between GMMs. In [4], Mel-Frequency Cepstral Coefficients (MFCCs) and beat histograms are used as features, resulting in classification accuracy of about 57% for the MAMI dataset. In [5], support vector machines are used for musical genre classification acting on a variety of temporal and frequency-domain features, achieving 73% average classification over two different datasets.

In this paper, we approach the automatic musical genre classification problem in a statistical framework, motivated by two recently introduced texture image retrieval methods. The first is based on the modeling of the marginal distribution of wavelet coefficients using a GGD and a closed form KLD between GGDs [6]; the second is based on modeling the marginal distribution of wavelet coefficients using an Alpha-Stable distribution and a closed form KLD between the characteristic functions [7]. The development of retrieval/classification models in a transform-domain is based on the observation that often the transform restructures the signal, resulting in a set of coefficients that are simpler to model. The wavelet transform in particular has been found to be very useful for extracting patterns in audio signals [8]. In this paper, we apply the wavelet transform to the music signals and we model the probability density function (PDF) of the wavelet subband coefficients in two ways: (i) using a GGD and (ii) using an Alpha-Stable distribution. The model parameters are estimated using a Maximum Likelihood (ML) estimator, as opposed to the parameters of a GMM which are estimated using the iterative Expectation-Maximization algorithm. Our objective is to introduce a new feature set that compactly models the raw signal space and thus includes properties of the signal that might be discarded by other feature extraction methods (*e.g.* mfcc features). In this sense, it is not within our immediate goals to produce classification performance superior of all previously proposed methods on the subject. Our goal is to show that the proposed statistically-derived features indeed characterize the properties of the raw signal space in a compact manner, which can be effectively demonstrated by their resulting classification performance.

This work has been funded by the Greek General Secretariat for Research and Technology, Program ΕΠΑΝ Code 01EP111, and by a Marie Curie International Reintegration Grant within the 6$^{\text{th}}$ European Community Framework Program.

Additionally, an important fact to note is that the KLD between GGDs and between characteristic functions of Alpha-Stable distributions has closed form and is not approximate (contrary to ALA for GMMs), meaning that our method is more accurate from a statistical point of view.

2. STATISTICAL MODELING

2.1. Signal Transform

As a first step in the FE task, we employ the 1-D orthogonal Discrete Wavelet Transform (DWT) which expands a signal using a certain basis, with elements that are are scaled and translated versions of a single prototype filter ("mother wavelet"). In our experiments, we used the biorthogonal wavelet; the best choice of wavelet for the task examined is outside the scope of this paper.

2.2. GGD Modeling of Wavelet Coefficients

Our method is based on fitting a Generalized Gaussian Density (GGD) on the PDF of the wavelet coefficients of a particular subband. This task can be achieved by estimating the two parameters of the GGD (α, β), which is defined as

$$p(x;\, \alpha, \beta) = \frac{\beta}{2\alpha\Gamma(1/\beta)} e^{-(|x|/\alpha)^{\beta}}, \qquad (1)$$

where $\Gamma(\cdot)$ is the Gamma function defined as

$$\Gamma(x) = \int_0^{\infty} t^{x-1}\, e^{-t} dt.$$

In (1), α models the width of the PDF peak (standard deviation), while β is inversely proportional to the decreasing rate of the peak. Usually, α is referred to as the scale parameter, while β as the shape parameter. The GGD model contains the Gaussian and Laplacian PDFs as special cases, corresponding to $\beta = 2$ and $\beta = 1$, respectively. During the FE step, we estimate the GGD model parameters (α, β) using the ML method described in [6].

Following the 1-D wavelet transform, the marginal statistics of the coefficients at each decomposition level are modeled via a GGD. Then, we simply estimate the (α, β) pairs at each subband. Thus, for a given music signal, decomposed in K levels, we estimate the set of the $K+1$ pairs

$$\{(\alpha_1, \beta_1), (\alpha_2, \beta_2), ..., (\alpha_{K+1}, \beta_{K+1})\}, \qquad (2)$$

where (α_i, β_i) are the estimated model parameters of the i^{th} subband. Note that we follow the convention that $i = 1$ corresponds to the detail subband at the first decomposition level and so forth, while $i = K + 1$ corresponds to the approximation subband at the K^{th} level. The total size of the above vector equals $2(K+1)$ which means that the content of a music signal can be represented by only a few parameters, in contrast with the large number of the transform coefficients.

2.3. Statistical Modeling of Wavelet Subband Coefficients via Alpha-Stable Modeling

As an alternative model to the GGD, the wavelet subband coefficients in various scales are modeled as Symmetric Alpha-Stable ($S\alpha S$) random variables [9]. The $S\alpha S$ distribution, which does not have a closed form expression, except for the Cauchy and Gaussian cases, is best defined by its characteristic function as follows [10]

$$\phi(t) = \exp(\imath\delta t - \gamma|t|^{\alpha}) \,, \qquad (3)$$

where α is the characteristic exponent, taking values $0 < \alpha \leq 2$, δ $(-\infty < \delta < \infty)$ is the location parameter, and γ $(\gamma > 0)$ is the dispersion of the distribution. The characteristic exponent is a shape parameter which controls the "thickness" of the tails of the density function. The smaller the α is, the heavier the tails of the $S\alpha S$ density function. The dispersion parameter determines the spread of the distribution around its location parameter, similar to the variance of the Gaussian.

3. CLASSIFICATION

The automatic musical genre classification problem can be formulated as a multiple hypothesis problem. Let us assume that there are M number of genres and that we have represented the query music signal S_q by its query data set $\mathbf{x} = (x_1, x_2, ..., x_L)$, which is typically obtained after a pre-processing stage (FE stage).

Each genre G_i, $i = 1, ..., M$, is assigned with a hypothesis H_i. The goal is to select one hypothesis out of M, which best describes the data from S_q. Under the common assumption of equal prior probabilities of the hypotheses, the optimum rule resulting in the minimum probability of classification error, is to select the hypothesis with the highest likelihood among the M. Thus, the query music signal is assigned to the genre corresponding to the hypothesis H_k if

$$p(\mathbf{x}|H_k) \geq p(\mathbf{x}|H_i),\ i \neq k\, (\forall\ i = 1, ..., M). \qquad (4)$$

The problem with (4) is that, in most cases, it is computationally expensive to compute. This turns out to be impractical in most applications since this operation in many cases has to be done on-line in an interactive mode. Therefore, we need to find an approximation with much less computational cost. A computationally efficient implementation of this setting is to adopt a parametric approach. Then, each conditional probability density $p(\mathbf{x}|H_i)$ is modeled by a member of a family of PDFs, denoted by $p(\mathbf{x}; \boldsymbol{\theta}_i)$ where $\boldsymbol{\theta}_i$ is a set of model parameters. Under this assumption, the extracted features for the songs in the musical genre S_i are represented by the estimated model parameter $\hat{\boldsymbol{\theta}}_i$, computed in the FE stage.

For classifying the query signal S_q to the closest genre:

1. We compute the KLDs between the query density $p(\mathbf{x}; \boldsymbol{\theta}_q)$ and the density $p(\mathbf{x}; \boldsymbol{\theta}_i)$ associated with genre G_i in the dataset, $\forall\, i = 1, \ldots, M$:

$$D(p(\mathbf{x}; \boldsymbol{\theta}_q) \| p(\mathbf{x}; \boldsymbol{\theta}_i)) = \int p(x; \boldsymbol{\theta}_q) \log \frac{p(x; \boldsymbol{\theta}_q)}{p(x; \boldsymbol{\theta}_i)}\, dx. \qquad (5)$$

2. We classify S_q in the genre corresponding to the smallest value of the KLD.

The KLD in (5) can be computed using consistent estimators $\hat{\boldsymbol{\theta}}_q$ and $\hat{\boldsymbol{\theta}}_i$, for the model parameters. The ML estimator is consistent and for the query signal it gives

$$\hat{\boldsymbol{\theta}}_q = arg \max_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}). \qquad (6)$$

We can also apply a chain rule, in order to combine the KLDs from multiple data sets. This rule states that the KLD between two joint PDFs, $p(\mathbf{X}, \mathbf{Y})$ and $q(\mathbf{X}, \mathbf{Y})$, where $\mathbf{X}, \mathbf{Y}$ are assumed to be independent data sets, is given by

$$D(p(\mathbf{X}, \mathbf{Y}) \| q(\mathbf{X}, \mathbf{Y})) = D(p(\mathbf{X}) \| q(\mathbf{Y})) + D(p(\mathbf{X}) \| q(\mathbf{Y})). \qquad (7)$$

Given the GGD model, the PDF of the wavelet coefficients in each subband can be completely defined by the

parameters (α, β). Substitution of (1) into (5) gives the following closed form for the KLD between two GGDs [6]

$$D(p_{\alpha_1,\beta_1}||p_{\alpha_2,\beta_2}) = \log\Big(\frac{\beta_1\alpha_2\Gamma(1/\beta_2)}{\beta_2\alpha_1\Gamma(1/\beta_1)}\Big) + \Big(\frac{\alpha_1}{\alpha_2}\Big)^{\beta_2}\frac{\Gamma(\frac{\beta_2+1}{\beta_1})}{\Gamma(\frac{1}{\beta_1})} - \frac{1}{\beta_1}. \quad (8)$$

For the case of the $S\alpha S$ model, the parameters (α, γ) can define the PDF of the wavelet subband coefficients. We expect that the KLD between normalized characteristic functions will be a good similarity measure between $S\alpha S$ distributions, because there is an one-to-one correspondence between a $S\alpha S$ density and its associated characteristic function. By employing the KLD between a pair of normalized characteristic functions for the parameterization (3), we obtain the following closed form expression [7]

$$D(\hat{\phi}_{\alpha_1,\gamma_1}\|\hat{\phi}_{\alpha_2,\gamma_2}) = \ln\Big(\frac{c_2}{c_1}\Big) - \frac{1}{\alpha_1} + \frac{2\gamma_2\Gamma\Big(\frac{\alpha_2+1}{\alpha_1}\Big)}{c_1\alpha_1\gamma_1^{\frac{\alpha_2+1}{\alpha_1}}}, \quad (9)$$

where (α_i, γ_i) are the parameters of the characteristic function $\phi_i(\omega)$ and c_i is its normalizing factor. The normalized characteristic function is defined as

$$\hat{\phi}(\omega) = \frac{\phi(\omega)}{c}, \quad \text{with} \quad c = \frac{2\Gamma\Big(\frac{1}{\alpha}\Big)}{\alpha\,\gamma^{1/\alpha}}.$$

The implementation of a K-level DWT on each music signal, results in its representation by $K+1$ subbands, $(D_1, D_2, \ldots, D_K, A_K)$, where D_i, A_i denote the i^{th} level detail and approximation subband coefficients, respectively. Assuming that the wavelet coefficients belonging to different subbands are independent, (7) yields the following expression for the overall normalized distance between two music signals S_1, S_2

$$D(S_1\|S_2) = \frac{1}{K+1}\sum_{m=1}^{K+1} D(p_{S_1,m}\|p_{S_2,m}). \quad (10)$$

4. EXPERIMENTAL RESULTS

In this section, we present results for three different experiments. For all experiments, we used a part of the ISMIR 2004 dataset for the contest of genre classification [1]. Our training dataset contained the following numbers of songs per category: 119 classical, 53 electronic, 26 jazz-blues, 30 metal-punk, 63 rock-pop, and 71 world songs. For obtaining the classification results, we used the database query set, which is a different set of songs containing 101 classical, 58 electronic, 26 jazz-blues, 41 metal-punk, 45 rock-pop, and 53 world songs. We converted the .mp3 format dataset songs in mono audio .wav format, using a sampling frequency of 44.1 kHz. Even though we used only a few seconds of each song, the size of the training and testing datasets was quite large and it guaranteed that enough coefficients were available at each subband for accurate estimation of the GGD and $S\alpha S$ parameters needed.

4.1. Model Accuracy

We conducted some initial experiments for testing the statistical modeling, by employing the amplitude probability density (APD) curves ($P(|X| > x)$). APD curves give a good indication of whether the proposed model (GGD or $S\alpha S$) matches the actual density of the data. An example for a part of our dataset is given in Fig. 1, where we compare the empirical APD (dashed line) against the APD curves obtained for the GGD (dotted line), $S\alpha S$ (solid line), and the Gaussian (dashed-dotted line) models. The results in the figure correspond to the approximation subband of the 3-level wavelet decomposition of the classical music training dataset we used (extracted from the database as explained in Section 4.2). Clearly, the GGD follows more closely the empirical APD than the decaying Gaussian density. We also observe that the $S\alpha S$ curve is closer to the empirical APD than the decaying Gaussian density, but it assumes a heavier-tailed behavior than the one followed by the actual data. This trend was observed in the majority of the music data we used in our experiments. Thus, we can expect that the GGD model will give better results than the $S\alpha S$ when applied directly to the wavelet coefficients.

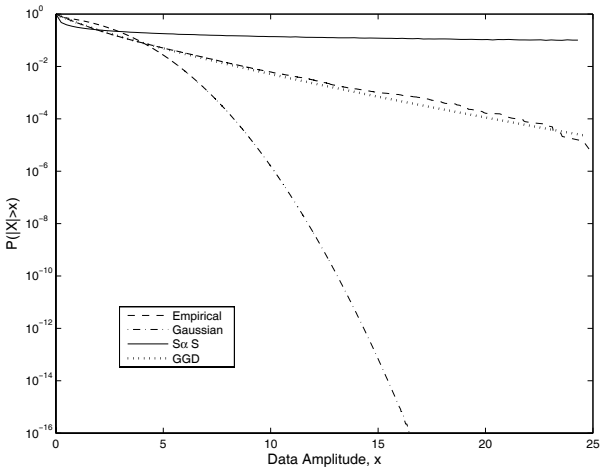

Fig. 1. Example Amplitude Probability Density curves of the approximation subband at the third level of decomposition of a classical music signal. The GGD model can be seen to best follow the empirical APD for this data.

4.2. Experiment 1 - Wavelets and GGD modeling

For each song in the database, we (manually) derived a small excerpt (few seconds) corresponding to a characteristic part of the song (*e.g.* a part of the song that is repeated most often). Then, we concatenated all these small segments in order to create a "hyper song" (one "hyper song" per genre). This procedure was necessary for reducing the dataset size. We resampled each "hyper song" at 1/4 times the original sampling rate. Subsequently, we performed a 3-level wavelet decomposition, in which a biorthogonal (bior4.4) wavelet was used as the "mother wavelet". The (α, β) GGD parameters of the wavelet subband coefficients for each "hyper song" were estimated. The result of this training procedure was a 8×1 vector of coefficients per genre.

In the case of a query song, we followed the same procedure as above. A 8×1 vector containing the estimated GGD parameters of the subbands was derived for each query song. In order to classify the query song to one of the musical genres, we computed the KLD for GGDs (8) using the parameters obtained during the training and testing procedures. We computed the KLD between the two estimated GGDs

[1] http://ismir2004.ismir.net/genre_contest/index.htm

of the 1^{st} subband, of the 2^{nd} subband, and so forth, for each genre. This procedure resulted in four distances, and we computed the mean of these numbers, as justified by (10). Finally, we calculated the six distances for the six different musical genres in our database, for each query song. The query song was then classified to the musical genre corresponding to the minimal distance.

4.3. Experiment 2 - Wavelets and $S\alpha S$ modeling

We followed the same procedure as in Experiment 1. Specifically, we used the same "hyper song" for each genre as in Experiment 1, resampled at 1/4 times the original sampling rate. Then, we estimated the $S\alpha S$ parameters of the wavelet subbands. The classification was performed using the derived KLD for $S\alpha S$ densities, shown in (9).

4.4. Experiment 3 - MFCCs and GGD modeling

In this experiment we resampled each "hyper song" at 1/2 times the original sampling rate (since MFCCs result in a dataset of smaller size). We calculated the MFCCs of each "hyper song" by using a 15-msec window with 50% overlapping. The order of the MFCCs was 19 (we discarded the first coefficient). Thus, for each row of the matrices, we estimated the GGD parameters (α, β). In this manner, we obtained a 19×2 matrix of parameters for each genre. In order to classify a query song to a genre, we computed the KLD for the GGD models for each row, *i.e.*, for each MFCC coefficient separately, assuming independence. This procedure resulted in 19 distance results, which were averaged as in (10) to provide the distance between the query song and each of the musical genres.

4.5. Performance Evaluation

We used the following criterion in order to evaluate the musical genre classification rate

$$Criterion = \frac{\#\text{ correctly classified query songs}}{\#\text{ query songs from a genre}}. \qquad (11)$$

We evaluated this expression for all the genres of the testing dataset. The results in Fig. 2 are the means of this expression for each genre. The total percentage of correct classification for Experiment 1 is 45%, for Experiment 2 is 22%, and for Experiment 3 is 24%. It is apparent that overall, the KLD classifier based on the wavelet decomposition and the GGD statistical model gave the best results. On the other hand, the GGD-modeled MFCCs gave low accuracy in many categories. We can conclude that the results obtained with the GGD model of the wavelet coefficients are encouraging. It is important to note that the GGD features proposed here can be used jointly with other features, such as the ones proposed in [1]. Our objective here was to introduce a new feature set that compactly models the properties of the raw signal space.

5. CONCLUSIONS

In this paper, we presented a statistical methodology for automatic musical genre classification. Our approach was based on modeling the wavelet coefficients of the music signals by means of heavy-tailed non-Gaussian densities, be it GGD or $S\alpha S$. The GGD-based method was shown to perform well in most cases. Our objective was to introduce a compact set of parameters, which capture signal properties

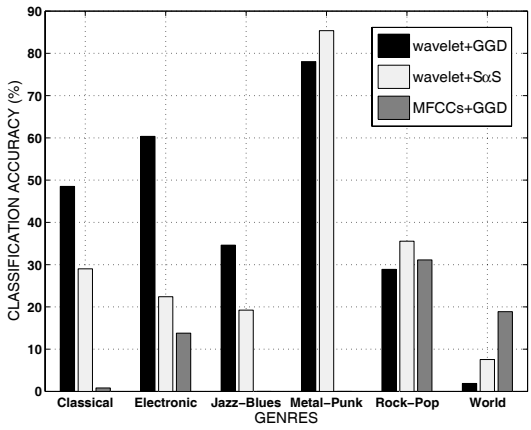

Fig. 2. Classification rates for each individual genre.

of the raw audio signals and can be used jointly with other features for improved classification results. Our future directions include identifying those additional features that can result in improved classification accuracy.

6. REFERENCES

[1] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. on Speech and Audio Processing*, vol. 10, pp. 293–302, July 2002.

[2] A. Berenzweig, D. P. W. Ellis, and S. Lawrence, "Anchor space for classification and similarity measurement of music," *Proc. IEEE Int. Conf. Multimedia and Expo (ICME), (Baltimore, MD)*, pp. 29–32, July 2003.

[3] N. Vasconcelos, "On the efficient evaluation of probabilistic similarity functions for image retrieval," *IEEE Trans. Information Theory*, vol. 50(7), pp. 1482–1496, July 2004.

[4] S. Lippens, J. P. Martens, T. DeMulder, and G. Tzanetakis, "A comparison of human and automatic genre classification," *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pp. 233–236, May 2004.

[5] N. Scaringella and D. Mlynek, "A mixture of support vector machines for audio classification," *1^{st} Music Information Retrieval Evaluation Exchange (MIREX)*, September 2005.

[6] M. N. Do and M. Vetterli, "Wavelet-based texture retrieval using generalized Gaussian density and Kullback-Leibler distance," *IEEE Trans. Image Processing*, vol. 11, pp. 146–158, February 2002.

[7] G. Tzagkarakis and P. Tsakalides, "A statistical approach to texture image retrieval via alpha-stable modeling of wavelet decompositions," *5th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, April 21-23 2004.

[8] R. Kronland-Martinet, J. Morlet, and A. Grossman, "Analysis of sound patterns through wavelet transform," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 1(2), pp. 237–301, 1988.

[9] C. L. Nikias and M. Shao, *Signal Processing with Alpha-Stable Distributions and Applications*. New York: John Wiley and Sons, 1995.

[10] J. P. Nolan, "Parameterizations and modes of stable distributions," *Statistics & Probability Letters*, no. 38, pp. 187–195, 1998.