A Music Similarity Function Based on the Fisher Kernels

Jin S. Seo[1], Nocheol Park[1], and Seungjae Lee[2] *

[1] Dept. of EE, Gangneung-Wonju National University, Korea
[2] Creative Content Research Laboratory, Electronics and Telecommunications Research Institute, Daejeon, Korea
jsseo@gwnu.ac.kr, seungjlee@etri.re.kr

Abstract. Music-similarity computation is an essential building block for browsing, retrieval, and indexing of digital music archives. This paper presents a music similarity function based on the Fisher-vector representation of the spectral features extracted from a song. The distance between the Fisher vectors of two songs is used as the similarity of the two songs. The Fisher vector has a closed-form representation and can be readily incorporated with simple vector distance measures. Experimental results show that the Fisher-vector representation of the auditory features is promising for the music-similarity computation.

Keywords: music similarity, music retrieval, music browsing, Fisher kernel

1 Introduction

Computing similarity between two songs is essential for browsing, retrieval, and indexing of digital music archives. Music similarity can be inferred in two different ways; collaborative filtering and content-based approach. In collaborative filtering, based on the musical tastes of many people, the musical preference of one person is predicted by those of other people [1]. In content-based approach, based on the perceptual auditory features, music similarity is directly computed from the distance between features from two songs. Both approaches have pros and cons. For example, the collaborative filtering cannot be adopted for new songs, and the content-based approach requires perceptually-meaningful feature extraction and computationally-efficient distance measure. This paper deals with the content-based approach.

The difficulty in computing the music similarity lies in the fact that the criteria used to determine the level of the similarity between two songs are subjective and hard to be described quantitatively. For the content-based music similarity, auditory features representing the music timbre, such as mel-frequency cepstral

* The authors would like to thank Seokjeong Lee for insightful discussions and help in collecting a dataset. This research project was supported by Government Fund from Korea Copyright Commission.

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

coefficients (MFCC) or other spectrum descriptors, has been adopted. In [1][2], the low-level spectral features extracted from a song are modeled by the k-means cluster or Gaussian Mixture Model (GMM). The distance between the song-level representations is estimated by either KL divergence [2][3], or earth-mover distance (EMD) [1], which is used as a metric for music similarity. Despite their excellent performance, the above mentioned methods are characterized by several short-comings. First of all, the construction of the song-level representations is based on an iterative process, which may not converge in some cases. Second, the pairwise distance using the KL or the EMD is computationally expensive and does not have a closed-form solution in most of the cases. To mitigate these problems, we employ the Fisher kernel to represent the feature distribution of a song instead of the iterative modeling process. The Fisher kernel was first introduced by Jaakkola and Haussler [4] and further studied by Perronnin and Dance [5] for image classification [6] and retrieval [7]. To combine the benefits of generative and discriminative approaches, the key idea of the Fisher kernel is to characterize a signal with a gradient vector derived from a probability density function which models the generation process of the signal [5][6][7]. In this paper, the closed-form vector representation of the Fisher kernel (the Fisher vector) derived in [5] is applied to represent the auditory features of the songs and compared with each other using simple distance measures, such as the Euclidean or the Cosine distance. In the experiments, the music similarity function based on the Fisher-vector representation showed retrieval performance comparable to the previous one [1].

This paper is organized as follows. Section 2 describes the music-similarity computation based on the Fisher-vector representation. Section 3 presents the experimental results of music retrieval tests. Finally, section 4 summarizes the paper.

2 Music Similarity Based on the Fisher-Vector Representation

The overview of the content-based music similarity computation is shown in Fig. 1. In the previous methods [1][2], the underlying distribution of the spectral features from a music clip is used as a signature for the music clip. Usually k-means clustering or GMM is used to fit the underlying distribution of the features. The music similarity of two songs is calculated as the statistical distance between the feature distributions of the two songs. As noted in Section 1, the previous methods mentioned above have several shortcomings associated to the iterative fitting of a mixture model and the computation of the pairwise distance. In contrast, the closed-form vector representation of the Fisher kernel in [5] can be readily extended to represent the auditory features and easily incorporated with simple distance measures, such as Euclidean or Cosine distance. We introduce the Fisher kernel in Section 2.1 and apply it to the music similarity in Section 2.2.

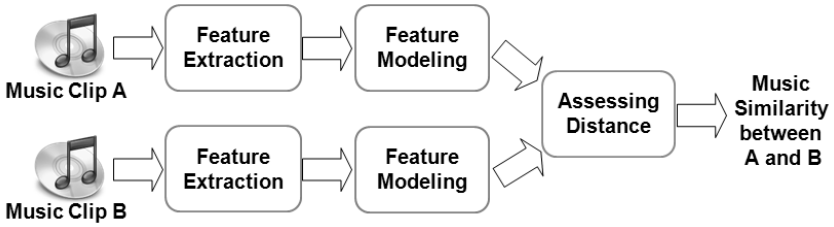

Fig. 1. Overview of the content-based music similarity computation.

2.1 Fisher Kernel

The followings are the introduction to the Fisher kernel as was proposed in [4][5]. Let X be a sample whose generation process can be modeled by a probability density function p with parameters λ [6]. In this paper, X corresponds to feature vectors from a music clip. With respect to the parameters λ, the gradient vector of X is denoted by

$$G_\lambda^X = \nabla_\lambda \log p(X|\lambda) \ . \tag{1}$$

The gradient vector gives the direction in parameter space into which the learnt distribution should be modified to better fit the observed data [8]. The dimensionality of this vector depends only on the number of parameters in λ [5]. On the gradient vector, a kernel is defined in [4][5][6] in the inner product form as follows:

$$K(X,Y) = G_\lambda^X F_\lambda^{-1} G_\lambda^Y \tag{2}$$

where F_λ is the Fisher information matrix of p given by

$$F_\lambda = E_{x\sim p}\left[\nabla_\lambda \log p(x|\lambda) \nabla_\lambda \log p(x|\lambda)'\right] \ . \tag{3}$$

Through the Cholesky decomposition of $K(X,Y)$, a normalized gradient vector [5] is obtained as follows:

$$\varrho_\lambda^X = F_\lambda^{-1/2} \nabla_\lambda \log p(X|\lambda) \tag{4}$$

The normalized gradient is referred to as the Fisher vector of X [5] which is the gradient of the sample's likelihood with respect to the parameters of the underlying distribution, scaled by the inverse square root of the Fisher information matrix [8].

2.2 Music-Similarity Computation Based on the Fisher vector

As shown in Fig. 2, we first extract the low-level spectral features from an input audio. An audio signal is split into overlapping segments (called frames) of length L with 50% overlap (in our system, $L = 1024$ at a sampling frequency of 22050

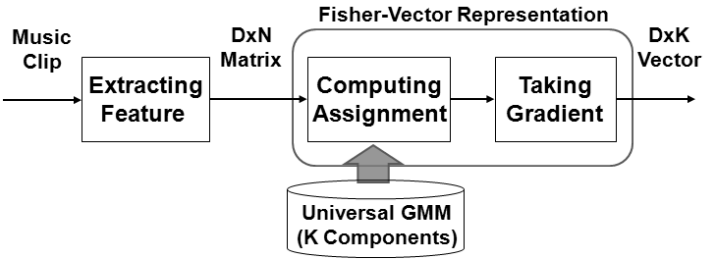

Fig. 2. Extraction of the Fisher vector from a music clip.

Hz). Each frame is windowed by a Hamming window of length L and transformed into the frequency domain. From each frame, we extract the low-level spectral features. We consider the D-order MFCC (in this paper, $D = 19$) as the low-level spectral feature as in [1]. Assuming that there are N frames in a music clip, the set of MFCC vectors from each frame is given by $X = \{x_0, x_1, \cdots, x_{N-1}\}$. We choose the GMM as a underlying distribution p for the feature space since the GMM has been used to represent the MFCC space in [2][3]. We denote the distribution p as a sum of mixtures by $p(x) = \sum_{k=0}^{K-1} w_k N(x|m_k, \Sigma_k)$ where the mixture weight w_k, mean vector m_k, and covariance matrix Σ_k are the parameters λ. In order to simplify the representation, as in [5], the covariance matrix is constrained to be diagonal with variance vector σ_k^2. We only consider the Fisher vector with respect to the mean and the standard deviation since that with respect to the weight carries little information [6]. Based on the assumption that x_n's are generated independently from p [6], the Fisher vector with respect to a parameter λ is given by

$$G_\lambda^X = \frac{1}{N} \sum_{n=0}^{N-1} \nabla_\lambda \log p(x_n|\lambda) \quad . \tag{5}$$

A closed-form expression of the Fisher information matrix of a GMM was derived in [5]. Using the derived Fisher information matrix, the Fisher vector ϱ_μ^X for the mean and ϱ_σ^X for the standard deviation are simplified in [5][6][7] as follows:

$$\varrho_\mu^X[kD + d] = \frac{1}{N\sqrt{w_k}} \sum_{n=0}^{N-1} \gamma_{nk} \left(\frac{x_n[d] - \mu_n[d]}{\sigma_n[d]} \right) \tag{6}$$

$$\varrho_\sigma^X[kD + d] = \frac{1}{N\sqrt{2w_k}} \sum_{n=0}^{N-1} \gamma_{nk} \left[\left(\frac{x_n[d] - \mu_n[d]}{\sigma_n[d]} \right)^2 - 1 \right] \tag{7}$$

where d denotes the d-th dimension of the feature vector x_n (in our case, $d = 0, 1, 2, \cdots, D-1$, and $k = 0, 1, 2, \cdots, K-1$), and γ_{nk} is the soft alignment (posterior probability) of feature vector x_n to the k-th Gaussian component of

the GMM given by

$$\gamma_{nk} = \frac{w_k N(x_n|m_k, \sigma_k)}{\sum_{j=0}^{K-1} N(x_n|m_j, \sigma_j)} \quad . \tag{8}$$

As shown in the Fig. 2, the Fisher kernel transforms an incoming variable-size (in our case, $D \times N$) set of independent features into a fixed-size (in our case, $D \times K$) vector representation, assuming that the features follow a parametric generative model estimated on a training set [8].

3 Experimental Results

Evaluating a music similarity function is intricate since the ground truth of the music similarity is difficult to obtain. Thus, in the previous works [1][2], it was assumed that the songs of the same genre or singer are perceptually more similar than those of the different genre or singer. With the same assumption, we evaluate the validity of the Fisher vector for music similarity on the genre and the singer datasets. The genre dataset is made by George Tzanetakis for his work [9] and consists of 1000 songs over ten different genres: blues, classical, country, disco, hiphop, jazz, metal, pop, reggae, and rock. The singer dataset is made by the authors and consists of 680 songs (20 songs per each singer) over 34 singers. For each query song in the dataset, we calculate the distances with the other songs in the dataset and examine the closest 5, 10, and 20 songs among which we count the number of songs in the same category (genre or singer) as the query song. The Fisher-vector based music similarity is compared to the Logan's music similarity function [1], where the MFCC vectors extracted from a song are modeled by the k-means clusters, and the clusters from two songs are compared each other using the EMD [1].

Each song in the datasets was converted to mono at a sampling frequency of 22050 Hz and then divided into frames of 46.4 ms ($L = 1024$) overlapped by 23.2 ms. We computed the 19-order MFCC of each frame as a low-level feature ($D = 19$). When extracting the Fisher vector, we considered three different GMMs as the underlying feature distribution with the number of mixture components in the GMM as 4, 8, and 16. The GMM was trained on 156 songs of various genres which are not overlapping with the test datasets. For each song, we calculated the Fisher vector with respect to the mean and the standard deviation as in (6) and (7) respectively. Table 1 is the result of the genre dataset and shows the average number of closest songs with the same genre as the query song. Table 2 is the result of the singer dataset and shows the average number of closest songs by the same singer as the query song. In obtaining the results in Table 1 and 2, each song in the dataset was used as a query, and the closest 5, 10, and 20 songs to each query were scrutinized. On the genre dataset (10 genres), the expected number of songs with the same genre as the query song among the closest 5 songs is $0.5 (= 5 \times 1/10)$ for random selection (assuming the identical and independent trials). In case of the singer dataset (34 singers),

Table 1. Average number of closest songs with the same genre as the seed song. The MFV and the SFV denote the Fisher vector with respect to the mean and the standard deviation respectively.

Types of Signatures	Distance Measure	Average number of songs in the same genre		
		Closest 5	Closest 10	Closest 20
MFV ϱ_{μ}^{X} $(K = 4)$	Euclidean	2.492	4.325	7.501
	Cosine	2.052	3.754	6.740
SFV ϱ_{σ}^{X} $(K = 4)$	Euclidean	1.054	1.979	3.73
	Cosine	1.569	2.846	5.16
MFV ϱ_{μ}^{X} $(K = 8)$	Euclidean	2.483	4.329	7.368
	Cosine	2.314	4.106	7.436
SFV ϱ_{σ}^{X} $(K = 8)$	Euclidean	1.463	2.728	5.012
	Cosine	1.643	2.965	5.391
MFV ϱ_{μ}^{X} $(K = 16)$	Euclidean	2.482	4.352	7.484
	Cosine	**2.545**	**4.557**	**8.014**
SFV ϱ_{σ}^{X} $(K = 16)$	Euclidean	1.581	2.929	5.358
	Cosine	1.640	2.994	5.487
Logan's Method [1]	EMD	**2.743**	**4.801**	**8.384**
Random Selection		0.5	1.0	2.0

Table 2. Average number of closest songs by the same singer as the seed song. The MFV and the SFV denote the Fisher vector with respect to the mean and the standard deviation respectively.

Types of Signatures	Distance Measure	Average number of songs by the same singer		
		Closest 5	Closest 10	Closest 20
MFV ϱ_{μ}^{X} $(K = 4)$	Euclidean	1.663	2.749	4.118
	Cosine	0.726	1.229	2.116
SFV ϱ_{σ}^{X} $(K = 4)$	Euclidean	0.319	0.602	1.096
	Cosine	0.559	0.929	1.497
MFV ϱ_{μ}^{X} $(K = 8)$	Euclidean	1.713	2.756	4.126
	Cosine	1.326	2.290	3.631
SFV ϱ_{σ}^{X} $(K = 8)$	Euclidean	0.476	0.804	1.410
	Cosine	0.547	0.929	1.531
MFV ϱ_{μ}^{X} $(K = 16)$	Euclidean	1.790	2.912	4.313
	Cosine	**1.919**	**3.121**	**4.800**
SFV ϱ_{σ}^{X} $(K = 16)$	Euclidean	0.618	1.091	1.890
	Cosine	0.656	1.151	1.999
Logan's Method [1]	EMD	**1.743**	**2.776**	**4.044**
Random Selection		0.147	0.294	0.588

the expected number of songs by the same singer as the query song among the closest 5 songs is $0.147(= 5 \times 1/34)$ for random selection. These indicate that the feature-based music similarity could provide a playlist which is much more meaningful than the random shuffling. In both Table 1 and 2, the Fisher vector with respect to the mean outperformed that with respect to the standard deviation. As the number of GMM components got larger (i.e. the dimensionality of the Fisher vector increased), the retrieval performance improved gradually. However, the performance gain was not quite notable. The retrieval performance of the Fisher-vector representation was more or less similar to that of the Logan's method [1] for both datasets. We note that the Fisher-vector representation has several merits over the Logan's method as stated in the Section 1. Moreover, the Fisher-vector representation is in vector form where many kinds of the distance measures can be easily incorporated. Although the Euclidean and the Cosine distance are considered in this paper, other distance measures can also be employed for the Fisher vector to boost the retrieval performance further. We leave it as a future work. We note that the scope of the experimental results in this paper is limited to the objective relevance with respect to the genre and the singer criterion. Each person's basis of the music similarity is multifarious depending on the personal preference and familiarity to a certain type of music [10]. Since designing and performing a subjective test on the music similarity is quite intricate in practice [10][11], we focus on the comparison between the proposed and the Logan's approach [1] with two objective criterions: the genre and the singer metadata. Further investigations of the proposed music similarity function are necessary with a subjective criterion by the empirical ratings of the human listeners to complement the experimental results reported in this paper.

4 Summary

In this paper, we apply the Fisher-vector representation of the spectral features to the content-based music similarity computation. The distance between the Fisher vectors of two songs is used as the similarity of the two songs. Compared with the previous mixture model representation, the Fisher-vector representation could provide a simplified alternative framework for music similarity computation. Experimental results show that the Fisher-vector representation can match the retrieval performance of the more complex ones.

References

1. Logan, B., Salomon, A.: A Music Similarity Function Based on Signal Analysis. In: IEEE International Conference on Multimedia and Expo, pp. 745–748. Tokyo (2001)
2. Aucouturier, J.-J., Pachet, F.: Improving Timbre Similarity : How High's the Sky?. Journal of Negative Results in Speech and Audio Sciences. 1, (2004)
3. Mandel, M., Ellis, D.: Song-Level Features and Support Vector Machines for Music Classification. In: International Conference on Music Info. Retrieval, London (2005)

4. Jaakkola, T., Haussler, D.: Exploiting generative models in discriminative classiers. In: Advances in Neural Information Processing Systems, pp. 487–493. Vancouver (1999)
5. Perronnin, F., Dance, C.: Fisher kernels on visual vocabularies for image categorization. In: IEEE Computer Vision and Pattern Recognition, pp. 1–8. Minneapolis (2007)
6. Perronnin, F., Sanchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: European Conference on Computer Vision, pp. 143–156. Crete (2010)
7. Perronnin, F., Liu, Y., Sanchez, J., Poirier, H.: Large-scale image retrieval with compressed Fisher vectors. In: IEEE Computer Vision and Pattern Recognition, pp. 3384–3391. San Francisco (2010)
8. Jegou, H., Douze, M., Schmid, C., Perez, P.: Aggregating local descriptors into a compact image representation. In: IEEE Computer Vision and Pattern Recognition, pp. 3304–3311. San Francisco (2010)
9. Tzanetakis, G., Cook, P.: Musical genre classification of audio signals. IEEE Transaction on Speech and Audio Processing. 10, 293–302 (2002)
10. Lee, J.H.: How similar is too similar?: Exploring users' perceptions of similarity in playlist evaluation. In: International Conference on Music Info. Retrieval, Miami (2011)
11. Bogdanov, D., Herrera, P.: How much metadata do we need in music recommendation? A subjective evaluation using preference sets. In: International Conference on Music Info. Retrieval, Miami (2011)