IEICE TRANS. INF. & SYST., VOL.E101–D, NO.1 JANUARY 2018

LETTER *Special Section on Enriched Multimedia — Potential and Possibility of Multimedia Contents for the Future —*

A Local Feature Aggregation Method for Music Retrieval

Jin S. SEO[†a)], ***Member***

SUMMARY The song-level feature summarization is an essential building block for browsing, retrieval, and indexing of digital music. This paper proposes a local pooling method to aggregate the feature vectors of a song over the universal background model. Two types of local activation patterns of feature vectors are derived; one representation is derived in the form of histogram, and the other is given by a binary vector. Experiments over three publicly-available music datasets show that the proposed local aggregation of the auditory features is promising for music-similarity computation.

key words: *music retrieval, music search, music information retrieval, supervector*

1. Introduction

For a practical query-by-example music retrieval system, a compact statistical representation of auditory features with a computationally efficient similarity measure is indispensable. In [1], [2], the low-level spectral features, such as mel-frequency cepstral coefficients (MFCC), extracted from a song are been modeled by the k-means cluster or Gaussian mixture model (GMM). Despite their excellent performance, there are several issues in applying them in practice. The construction of the song-level representations is based on an iterative process, and their distance measures are computationally expensive. As an attempt to mitigate these problems, supervector concept has been applied to model auditory features by adapting the *universal background model* (UBM) [3]. Typically the UBM is a Gaussian mixture distribution estimated from a number of training songs. Instead of modeling each test song separately, the parameters of the UBM are adapted by the feature vectors of the test song. The supervector is given by concatenating the mean vectors of the adapted GMM. The GMM supervector of the MFCC has been applied to various applications including the speaker verification [4] and the music retrieval [5]. One noticeable limitation of the GMM supervector is that only one mode can be represented as a supervector for each mixture component. To incorporate more diverse local feature distribution of a song, this paper employs one more layer of feature coding over each mixture component of the UBM inspired from the success of the multi-way local pooling [6] in image classification. First, we train a supervector codebook for each mixture component from a number of music supervectors. For a song in the music repository, we compute the posteriori probability that each feature vector in the song is from a supervector code and construct the histogram of the posteriori probabilities of the feature vectors for each mixture component. The histograms obtained from all the mixture components are concatenated to form a song-level vector representation of the song, which is referred to as the local activation histogram vector (LAHV) in this paper. The LAHVs of two songs can be compared with any types of the histogram-distance measures, such as the histogram intersection or the chi-squared distance. When more compact form of music representation is needed, the LAHV can be further binarized to form the local activation binary vector (LABV) by utilizing the sparsity of the LAHV. This paper considers both LAHV and LABV for query-by-example music retrieval.

The main contributions of our work are summarized as follows. 1) The LAHV and the LABV can be easily indexed and incorporated with computationally-efficient distance measures, which is practically important in searching over large-size music archives. 2) Diverse local feature distribution can be accommodated by adding one more layer of feature coding. 3) The proposed LAHV and LABV can be directly derived over a trained codebook without iteration.

2. Proposed Song-Level Music Representation

In this section, we propose the LAHV and LABV as an extension and an alternative to the GMM supervector for the song-level music representation.

2.1 GMM Supervector for Music Representation

The followings are the introduction to the GMM supervector as was proposed in [4], [5]. Let $\mathbf{x}$ be a sample vector whose generation process can be modeled by a probability density function $p(\mathbf{x})$ of the GMM as UBM given by

$$p(\mathbf{x}) = \sum_{k=1}^{K} w_k N(\mathbf{x}|\mathbf{m}_k, \Sigma_k) \tag{1}$$

where the mixture weights w_k, mean vectors $\mathbf{m}_k$, and covariance matrices Σ_k are the parameters. In order to simplify the representation, as in [4], the covariance matrix is often constrained to be diagonal with variance vector σ_k^2. As shown

Manuscript received March 15, 2017.
Manuscript revised August 10, 2017.
Manuscript publicized October 16, 2017.

[†]The author is with the Department of Electrical Engineering, Gangneung-Wonju National University, Gangneung, Rep. of Korea.
a) E-mail: jsseo@gwnu.ac.kr

DOI: 10.1587/transinf.2017MUL0001

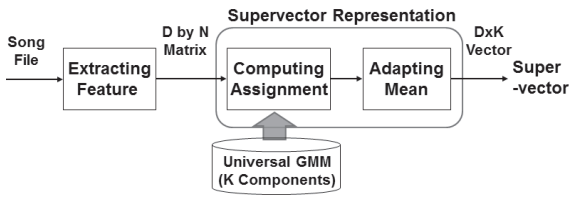

Fig. 1 The extraction of the supervector from a song file.

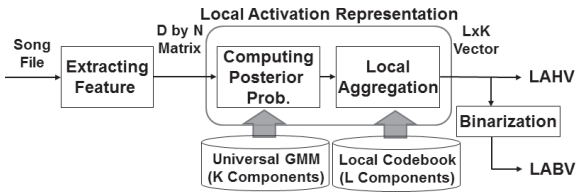

Fig. 2 The extraction of the proposed LAHV and LABV from a song file.

in Fig. 1, we first extract the low-level spectral features from an input audio. An audio signal is split into overlapping segments (called frames). 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

$$\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}. \tag{2}$$

Given the set of feature vectors of a music clip, the supervector is obtained by the maximum a posteriori adaptation [3], [5] of the mean vectors $\mathbf{m}_k$ as stated here after. For the t-th frame feature vector $\mathbf{x}_t$, the posteriori membership probability, $Pr(k|\mathbf{x}_t)$, of the k-th Gaussian component of the UBM is given by

$$Pr(k|\mathbf{x}_t) = \frac{w_k N(\mathbf{x}_t|\mathbf{m}_k, \Sigma_k)}{\sum_{j=1}^{K} w_j N(\mathbf{x}_t|\mathbf{m}_j, \Sigma_j)} \quad . \tag{3}$$

From the $Pr(k|\mathbf{x}_t)$, we can compute the adaptation parameters for the weight and the mean vector as follows:

$$n_k = \sum_{t=1}^{N} Pr(k|\mathbf{x}_t) \tag{4}$$

$$E_k = \frac{1}{n_k} \sum_{t=1}^{N} Pr(k|\mathbf{x}_t)\mathbf{x}_t \quad . \tag{5}$$

The mean adaptation is performed by the weighted update as follows:

$$\tilde{\mathbf{m}}_k = \alpha_k E_k + (1 - \alpha_k)\mathbf{m}_k \tag{6}$$

where the weight factor α_k is given by a fixed relevance factor r as

$$\alpha_k = \frac{n_k}{n_k + r} \quad . \tag{7}$$

After replacing $\mathbf{m}_k$ with the adapted $\tilde{\mathbf{m}}_k$, the above process is iterated. The finally obtained mean vectors are further normalized by the weight and the covariance matrix of GMM as follows:

$$\bar{\mathbf{m}}_k = \sqrt{w_k}\Sigma_k^{-1/2}\tilde{\mathbf{m}}_k \tag{8}$$

which is derived to approximate the Kullback-Leibler divergence by the simple Euclidean distance [4], [5]. Finally the adapted mean vectors are concatenated to form a DK-dimensional vector, which is called supervector [4]. The obtained GMM supervector can be thought of as a mapping between a music clip and a high-dimensional vector.

2.2 Local Activation Patterns Based on Supervector Codebook

The overview of the proposed LAHV and LABV extraction is shown in Fig. 2. To derive the LAHV of a song, we train the supervector codebook from a number of training songs. For the k-th component of UBM, we calculate the $\bar{\mathbf{m}}_k$ of each training song as in Sect. 2.1 and cluster them into L clusters. We denote the cluster centers as the codebook $\mathbf{V}_k = \{\mathbf{v}_{k1}, \mathbf{v}_{k2}, \cdots, \mathbf{v}_{kL}\}$. In total, we have K codebooks, $\mathbf{V}_k$, from $k = 1$ to $k = K$. For the t-th frame feature vector $\mathbf{x}_t$ of the input music clip, the cluster-membership index z_{kt} of the k-th component is given by

$$z_{kt} = \arg\min_{1 \le l \le L} \|\bar{\mathbf{x}}_t^k - \mathbf{v}_{kl}\|^2 \tag{9}$$

where $\|\cdot\|$ denotes the L^2 norm of the vector space, and $\bar{\mathbf{x}}_t^k = \sqrt{w_k}\Sigma_k^{-1/2}\mathbf{x}_t$. Since each cluster center $\mathbf{v}_{kl}$ might represent a specific local timber group, the cluster-membership index z_{ki} bears corresponding timbral characteristics. For an music clip with the feature vectors $\mathbf{X}$ in (2), the cluster-membership set C_{kl} is denoted by

$$C_{kl} = \{t | z_{kt} = l, \quad 1 \le t \le N\} \quad . \tag{10}$$

We add all the posteriori membership probabilities corresponding to each cluster-membership set C_{kl} by

$$g_{kl} = \sum_{t \in C_{kl}} Pr(k|\mathbf{x}_t) \quad . \tag{11}$$

For each mixture component k, we derive the histogram vector h_k by normalizing g_{kl} to sum to one as follows:

$$h_k[l] = \frac{g_{kl}}{\sum_{j=1}^{L} g_{kj}} \quad . \tag{12}$$

Finally, the LK-dimensional LAHV h is defined by concatenating L-dimensional histogram vector h_k for all k from 1 to K.

In this paper, the musical difference between two music clips A and B is represented by the histogram distance between the two LAHVs h_A and h_B obtained from A and B respectively. Among various histogram distance measures, this paper employs the following distance measure D_{HI} based on the computationally-efficient histogram intersection:

$$D_{HI}^{(A,B)} = K - \sum_{i=1}^{LK} \min(h_A[i], h_B[i]) \quad . \tag{13}$$

The LAHV of a song is sparse by nature, which means that only several codewords for each UBM mixture component are actively engaged for the song. By thresholding the LAHV, we obtain the binary LABV p of the music-signal spectrum given by

$$p[i] = \begin{cases} 1 & \text{if } h[i] > thr \\ 0 & \text{if } h[i] \le thr \end{cases} \tag{14}$$

where the binarization threshold thr is a predetermined constant (typically $1/(2L)$). The LABV is given as a binary vector, which is easily indexed with the Hamming distance.

3. Experimental Results

Evaluating a music similarity function is intricate since the ground truth of the music similarity is difficult to obtain. Each person's basis of the music similarity is multifarious depending on the personal preference and familiarity to a certain type of music [7]. Since designing and performing a subjective test on the music similarity is quite intricate in practice [7], [8], the objective relevances have been employed [1], [2], [9]. In this paper, we use two criterions, the genre match and the tag similarity, in evaluating the retrieval performance of the proposed method. For the genre-match criterion, it was assumed that the songs of the same genre are perceptually more similar than those of the different genre. Two publicly-available musical genre datasets were used in the evaluation. For the tag-similarity criterion, the retrieval performance of LAHV and LABV was evaluated with the human tag-based distance of the CAL500 dataset [10] as a ground truth of music similarity.

The first genre dataset (abbreviated as GTZAN) was made by George Tzanetakis for his work [11] and consists of 1000 songs over ten different types of genres. The second genre dataset (abbreviated as ISMIR2004) is the one from the ISMIR 2004 genre classification contest in which there are 1458 songs over the six different types of genres. The CAL500 dataset is composed of 500 songs, where each song is manually annotated with 174 tags. 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 overlapped by 23.2 ms where the 19-order MFCC was computed as a low-level feature. The supervector of each song was obtained by adapting the mean vector of UBM iteratively using the MFCC vectors of the frames in the song as in Sect. 2.1. The GMM was used for UBM, and the number K of mixture components in the UBM was 12. The LAHV and LABV were computed from the posteriori probabilities of the MFCC vectors of a song over the trained codebook as in Sect. 2.2, where we considered various size L of codebook ranging from 16 to 24 in the experiments. From ten thousand training songs, which are strictly separated with the songs in the evaluation datasets, we generate the UBM five times and construct the supervector codebook five times for each UBM. Thus all the evaluations of the LAHV and LABV were performed 25 times, and the experimental results in this Section are the average performance of the 25 trials. We note that the standard deviation of the results from the 25 trials was small; the coefficient of variation (ratio of the standard deviation and the mean) was below 0.01 which means that the experimental results are quite stable regardless of the construction of UBM and codebook. The performance of the LAHV and LABV was compared experimentally to that of the GMM supervector. In Tables 1, 2, and 3, the SV and the USV denote the GMM supervector and the UBM-normalized supervector respectively. The HD and the ED denote the Hamming and the Euclidean distance respectively.

Tables 1 and 2 show the average number of the clos-

Table 1 Average number of closest songs correctly retrieved with the criterion of the same genre on the GTZAN dataset.

Types of model	Distance	Avg. # of correctly-retrieved songs		
		Closest5	Closest10	Closest20
LAHV ($L = 16$)	D_{HI}	2.865	5.248	9.427
LAHV ($L = 19$)	D_{HI}	2.889	5.295	9.527
LAHV ($L = 24$)	D_{HI}	2.917	5.326	9.591
LABV ($L = 16$)	HD	2.680	4.923	8.908
LABV ($L = 19$)	HD	2.717	4.989	8.995
LABV ($L = 24$)	HD	2.776	5.083	9.127
SV [4]	ED	2.651	4.724	8.126
USV [5]	ED	2.904	5.212	9.085
Random		0.5	1.0	2.0

Table 2 Average number of closest songs correctly retrieved with the criterion of the same genre on the ISMIR2004 dataset.

Types of model	Distance	Avg. # of correctly-retrieved songs		
		Closest5	Closest10	Closest20
LAHV ($L = 16$)	D_{HI}	3.514	6.717	12.766
LAHV ($L = 19$)	D_{HI}	3.534	6.756	12.834
LAHV ($L = 24$)	D_{HI}	3.563	6.817	12.927
LABV ($L = 16$)	HD	3.397	6.481	12.294
LABV ($L = 19$)	HD	3.426	6.536	12.373
LABV ($L = 24$)	HD	3.461	6.599	12.473
SV [4]	ED	3.518	6.674	12.457
USV [5]	ED	3.616	6.797	12.723
Random		0.833	1.667	3.333

Table 3 Quartiles of the tag-based distance of the closest songs retrieved by the feature-based distance on the CAL500 dataset.

Types of model	Distance	Quartiles of tag-based dist.		
		Q1	Q2	Q3
LAHV ($L = 16$)	D_{HI}	0.1508	0.1793	0.2126
LAHV ($L = 19$)	D_{HI}	0.1499	0.1782	0.2126
LAHV ($L = 24$)	D_{HI}	0.1499	0.1784	0.2126
LABV ($L = 16$)	HD	0.1503	0.1793	0.2126
LABV ($L = 19$)	HD	0.1494	0.1786	0.2122
LABV ($L = 24$)	HD	0.1497	0.1782	0.2111
SV [4]	ED	0.1552	0.1839	0.2144
USV [5]	ED	0.1552	0.1793	0.2115

est songs with the same genre as the query song on GTZAN and ISMIR2004 respectively. 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 GTZAN 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 ISMIR2004 (6 genres), the expected number of songs with the same genre as the query song among the closest 5 songs is 0.833 ($= 5 \times 1/6$) for random selection. These indicate that the content-based music similarity could provide a playlist which is much more meaningful than the random shuffling. As the size of codebook (L) got larger (i.e. the dimensionality of LAHV and LABV increased), the retrieval performance improved gradually. However, the performance gain was not quite notable when L is greater than 24. For both genre datasets, the performance of the proposed LAHV and LABV was better than that of the conventional supervector and more or less similar to the UBM-normalized supervector [5]. We note that the distance measures associated with LAHV and LABV are computationally-efficient without requiring any multiplication, which is practically important when dealing with large-size music repositories.

The performance of the proposed feature representation was compared with that of the human-annotated tags on the CAL500 dataset [10]. All songs in the CAL500 dataset were annotated by predefined 174 tags and represented by a 174-dimensional binary vector, where each element is assigned to 1 if 80% of the human annotators label the song with the corresponding tag and assigned to 0 otherwise. The tag-based distance of the CAL500 dataset was obtained by the Hamming distance between the binary tag vectors and used as a ground truth in comparing the feature-based distances using LAHV, LABV, and supervector. We first listed the closest 10 songs to a query song with respect to each feature-based distance. Then we computed the first, the second, and the third quartiles (abbreviated as Q1, Q2, and Q3 respectively) of the tag-based distances of closest songs to all query songs, which is shown in Table 3. Although the performance difference was little, the LAHV and LABV resembled the human tag-based distance more closely than the supervector especially at the first and the second quartiles.

4. Conclusion

In this paper, a local feature aggregation method over the UBM is proposed for music retrieval. The local feature aggregation leads to two simplified song-level music representations; LAHV and LABV. Compared with the supervector, the LAHV and the LABV, which encode the local feature distribution over each mixture component of UBM, can be obtained without iteration and incorporated with computationally-efficient distance measures. Experimental results show that the proposed local feature representations yield similar or better performance compared with the supervector.

Acknowledgments

This research project was supported by Ministry of Culture, Sports and Tourism (MCST) and from Korea Copyright Commission in 2017 [Development of predictive detection technology for the search for the related works and the prevention of copyright infringement].

References

[1] B. Logan and A. Salomon, "A music similarity function based on signal analysis," Proc. ICME-2001, pp.745–748, 2001.
[2] J.J. Aucouturier and F. Pachet, "Improving timbre similarity: How high's the sky?," Journal of Negative Results in Speech and Audio Sciences, vol.1, no.1, pp.1–13, 2004.
[3] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker verification using adapted Gaussian mixture models," Digital Signal Processing, vol.10, no.1-3, pp.19–41, January 2000.
[4] W.M. Campbell, D.E. Sturim, and D.A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," IEEE Signal Process. Lett., vol.13, no.5, pp.308–311, 2006.
[5] C. Charbuillet, D. Tardieu, and G. Peeters, "GMM supervector for content based music similarity," Proc. DAFX-2011, 2011.
[6] Y.-L. Boureau, N.L. Roux, F. Bach, J. Ponce, and Y. LeCun, "Ask the locals: Multi-way local pooling for image recognition," Proc. ICCV-2011, pp.2651–2658, 2011.
[7] J. Lee, "How similar is too similar?: Exploring users' perceptions of similarity in playlist evaluation," Proc. ISMIR-2011, 2011.
[8] D. Bogdanov and P. Herrera, "How much metadata do we need in music recommendation? A subjective evaluation using preference sets," Proc. ISMIR-2011, 2011.
[9] J.S. Seo, "A music similarity function based on the centroid model," IEICE Trans. Inf. & Syst., vol.E97-D, no.7, pp.1573–1576, July 2013.
[10] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Towards musical query-by-semantic-description using the CAL500 data set," Proc. SIGIR-2007, pp.439–446, 2007.
[11] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," IEEE Trans. Speech Audio Process., vol.10, no.5, pp.293–302, July 2002.