GMM SUPERVECTOR FOR CONTENT BASED MUSIC SIMILARITY

Christophe Charbuillet, Damien Tardieu and Geoffroy Peeters, *

IRCAM-STMS-CNRS
Centre Pompidou
Paris, France
charbuillet, tardieu, peeters (at) ircam.fr

ABSTRACT

Timbral modeling is fundamental in content based music similarity systems. It is usually achieved by modeling the short term features by a Gaussian Model (GM) or Gaussian Mixture Models (GMM). In this article we propose to achieve this goal by using the GMM-supervector approach. This method allows to represent complex statistical models by an Euclidean vector. Experiments performed for the music similarity task showed that this model outperform state of the art approches. Moreover, it reduces the similarity search time by a factor of ≈ 100 compared to state of the art GM modeling. Furthermore, we propose a new supervector normalization which makes the GMM-supervector approach more preformant for the music similarity task. The proposed normalization can be applied to other Euclidean models.

1. INTRODUCTION

Exploring the wide world of music requires some navigation tools. To discover new tracks, one might consider several options. Specialized magazines or music expert friends can guide the user. In a more passive way, the user can wait for new music production by listening to his favorite radio or following the statistically made recommendation of online mp3 providers, based on user profiles and puchase analyses. But to explore several million of iTunes© music tracks, one may need to employ a content based similarity search system. The principle is quite simple. From a starting music and for a given similarity measure,the system provides the user a list of similar songs found in the entire database. If the user is not satisfied with the result, he/she can change or adapt the similarity measure according to his/her wishes. The system can also learn the user preferences using relevance feedback. One can also use the result of previous queries as starting point for a new search and, thereby, perform a step by step smart exploration of the music space.

Obviously, the relevance of the similarity measure is fundamental. A music track can be described in several ways. Using the mpeg-7 taxonomy, we distinguish the meta description (e.g.: music author or title) and the content description. Similarity systems based on content description mimic human perception of similarity. Timbral modeling is nowaday state of the art in such systems. It consists in statistical modeling of short term audio features, usually the Mel Frequency Cepstrum Coefficients (MFCC). The model used can be a Gaussian Mixture Model (GMM) as proposed in [1, 2], or a single Gaussian Model with full covariance matrix [3, 4] which provides similar performances. The measure used to compare the models is the Symmetrized Kullback-Leibler Divergence (SKLD) [5] or alternatively the Earth Mover's Distance based on the SKLD when models are GMMs [2].

We present here an application of the Gaussian Mixture Model using Universal Background Model (GMM-UBM) approach for content based music similarity. This method, initially developed in the field of speaker recognition [6] has been successfully applied for music genre classification and similarity [7]. The main idea is to build a generic Gaussian mixture model by using a large data set of representative signals, which are in our case extracted from a large set of music tracks. This model, named Universal Background Model, aims at modeling the overall data distribution and can be composed of hundred of Gaussians. The model for a specific track is then obtained by adapting the UBM model parameters by using the track data. The final model is composed of a subset of the GMM parameters, stacked into a vector, the so called supervector. This approach presents several advantages:

- it allows to build a complex model from a small amount of data,
- the final model can be embedded into the Euclidean space, which allows fast similarity search.

In this paper, a complete description of the GMM-UBM model is proposed in section 2. Then our main contribution it presented in section 3. It consists in a new supervector transformation, which provides a significant improvement of the similarity system. Experiments are detailed in section 4 and the perspectives of this work are presented in section 5.

2. GMM-UBM APPROACH

2.1. Universal Background Model

The Universal Background Model (UBM) aims at modeling the overall data distribution. It consists of a classical Gaussian Mixture Model. For a D-dimensional feature vector $\mathbf{x}$ the mixture density used for the likelihood function is defined as a weighted sum of unimodal Gaussian models :

$$p(\mathbf{x}|\lambda) = \sum_{i=1}^{M} \omega_i p_i(\mathbf{x}) \quad (1)$$

where M is the number of Gaussian components, $p_i = \mathcal{N}(\mu_i, \Sigma_i)$. λ represents the GMM parameters, where $\lambda_i = \{\omega_i, \mu_i, \Sigma_i\}$, $i = 1, \cdots, M$. $\mathbf{x}$ represents a feature vector, which in our case is a short term descriptor, usually an MFCC.

* This work was supported by the French Oseo project QUAERO

The UBM is usually composed of Gaussian models with diagonal covariance matrix. The loss of modeling ability due the diagonal covariance matrix can be compensated by increasing the number of Gaussian in the mixture [6]. The UBM is trained using a large and representative set of data by using the Expectation Maximization (EM) algorithm.

2.2. UBM adaptation

The UBM adaptation is the process of modifying the UBM parameters in order to fit a particular data distribution. In our application, this subset is the data extracted from a track to modelize.

This adaptation is made using the Maximum A Prosteriori (MAP) approach. The first step consists in determining the probabilistic alignement of the training vectors with the UBM Gaussian components. For a Gaussian i in the UBM we compute :

$$\mathbf{Pr}(i, \mathbf{x_t}) = \frac{\omega_i p_i(\mathbf{x_t})}{\sum_{j=1}^{M} \omega_j p_j(\mathbf{x_t})} \tag{2}$$

$$n_i = \sum_{t=1}^{T} \mathbf{Pr}(i, \mathbf{x_t}) \tag{3}$$

$$E_i(\mathbf{x}) = \frac{1}{n_i} \sum_{t=1}^{T} \mathbf{Pr}(i, \mathbf{x_t}) \mathbf{x_t} \tag{4}$$

These statistical values are then used for adapting the mean vector $\hat{\mu}$ of each Gaussian during the following iterative process:

$$\hat{\mu}_i^0 = \mu_i \tag{5}$$

$$\hat{\mu}_i^k = \alpha_i E_i(\mathbf{x}) + (1 - \alpha_i)\hat{\mu}_i^{k-1} \tag{6}$$

$$\alpha_i = \frac{n_i}{n_i + r} \tag{7}$$

where $\mathbf{x}_t$ represents the t^{th} feature vector of the music track to modelize and r is a fixed "relevance factor", usually set between 8 and 20. $k = 1, \cdots, K$ represents the iteration number.

2.3. GMM supervector

To summarize, a music track model is directly derived from a generic GMM, estimated using a large set of representative data (the so called UBM). During the adaptation process, only the mean vectors of the Gaussians are modified to fit the particular music track distribution. Consequently, all the the music track models have both the same covariance matrix and weight. Knowing the parameter of the UBM, a particular music model can be summarized by the mean vectors of its Gaussian mixture components:

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \cdots \\ \mu_N \end{pmatrix} \tag{8}$$

where μ, named the GMM supervector, is the concatenation of all the mean vectors of the N Gaussian components. In [8], the authors propose to approximate the Kullback-Leibler divergence between two models a and b by :

$$d(\mu^a, \mu^b) = \frac{1}{2} \sum_{i=1}^{N} \omega_i (\mu_i^a - \mu_i^b)^T \Sigma_i^{-1} (\mu_i^a - \mu_i^b) \tag{9}$$

where μ^a and μ^b are the GMM supervectors of the models a and b respectively, λ_i represents the mixture weights and Σ_i the covariance matrix of the i^{th} Gaussian component (which is common to the models a and b). From this representation, we can deduce the following natural normalization:

$$\bar{\mu}_i = \sqrt{\omega_i} \Sigma_i^{-1/2} \mu_i \quad i \in 1, \cdots, N \tag{10}$$

where N is the number of Gaussian components of the model. Then, the divergence presented in eq. 9 can be rewritten as the square Euclidean distance between the normalized supervectors:

$$d(\mu^a, \mu^b) = \frac{1}{2} \|\bar{\mu}^a - \bar{\mu}^b\|^2 \tag{11}$$

Finally, because of the monotony of the square function $(\cdot)^2$, one can directly use the Euclidean distance $\|\bar{\mu}^a - \bar{\mu}^b\|$ for music similarity retrieval, as proposed in [7].

3. SUPERVECTOR NORMALIZATION FOR MUSIC SIMILARITY

3.1. Hubs and orphans

Even if the statistical modeling of short term descriptors gives good results for music similarity, it usually tends to create false positive results which are usually the same songs. This songs, named hubs, are falsely close to all the tracks of the database. As well, some songs, named orphans, are falsely far from the rest of the database. J. Aucouturier *et al.*[9] showed that this phenomenon is "not a property of a given modeling strategy and tends to appear with any type of model".

For a better understanding of the problem we propose to modelize a set of music tracks and to study their distance distributions. The music database and the modeling process are fully described in section 4. From this set of supervectors we compute the distance matrix between all the supervectors. Figure 1 presents the distance distribution between the track supervectors and the rest of the database. We can observe that the distributions have a significant variability. For example, the distribution related to the first music track shows that this model is far from the rest of the database. Consequently, it will have a poor probability to appear within the results of the similarity search. This is a good example of an "orphan" song.

3.2. P-norm

To overcome this drawback, a distance normalization method was proposed by T. Pohle *et al.*in [4]. The key idea of this method is to transform the distances between two models by using their distance distribution according to a normalization set. After the normalization process, the histogram of the new "distance" between

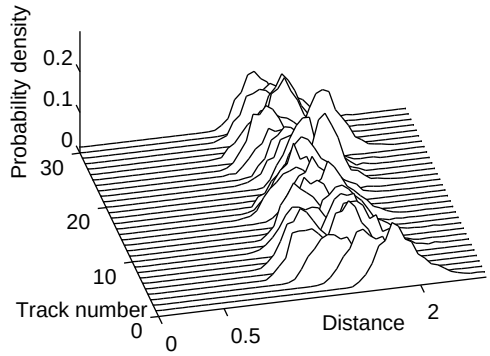

Figure 1: Distance distribution between supervectors. Each curve represents the histogram of the distance between a given supervector and the rest of the database.

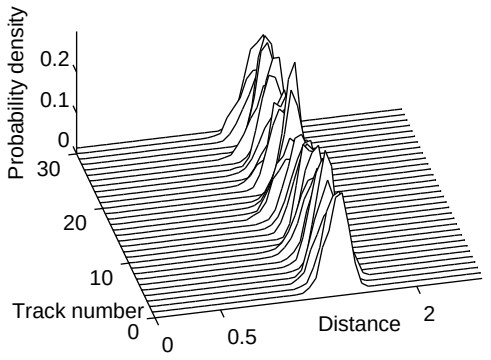

Figure 2: Distance distribution between supervectors normalized by the UCS-norm.

a model and the rest of the database must be a normal distribution $\mathcal{N}(0,1)$. This normalization is given by:

$$\text{P-norm}\left(d(a,b)\right) = \frac{1}{2}\left(\frac{d(a,b)-\mathring{\mu}_a}{\mathring{\sigma}_a} + \frac{d(a,b)-\mathring{\mu}_b}{\mathring{\sigma}_b}\right) \quad (12)$$

where $d(a,b)$ is the original distance between models a and b, $\mathring{\mu}_a, \mathring{\sigma}_a$ are the mean and standard deviation of the distances between the models a and the normalization set. For convenience, in the rest of this paper, we will refer to this method as the P-norm (Pohle-normalization). One can notice that this type of normalization is very close to the ZT-norm developed in the field of speaker verification [10].

3.3. UCS and MCS normalizations

An important benefit of the supervector approach is the ability to represent a complex statistical model as an Euclidean vector. It allows the use of efficient indexing algorithms for fast similarity search into very large databases like local sensitive hashing [11]. The use of the P-norm (which modifies the distances) transforms the original Euclidean space into a non-metric space, constraining the use of ad-hoc indexing methods which are usually slower. Therefore, a normalization which can be applied directly to the supervector is more suitable. Let us consider the supervector as a point into a high dimensional space and a large representative data set. To reproduce the benefit of the P-norm by a geometric transformation of the supervectors, the projected points must "see the world in a same way" i.e. the distance distribution between a point and the rest of the database must be the same for all the points. It is easy to show that a uniform data distribution on a hyper sphere satisfies this constraint. We propose two different methods to reach this goal:

1. project the supervectors on a unit sphere centered on the Universal Background Model,
2. project the supervectors on a unit sphere centered on the mean supervector of a representative data set (here we used the entire database).

For convenience, we named the first approach the *UBM Centered Spherical normalization* (UCS-norm) and the second one the *Mean Centered Spherical normalization* (MCS-norm). The following equations detail the implementation we used.

$$\bar{\mu}^a_{UCS} = \frac{\bar{\mu}^a - \bar{\mu}^{UBM}}{\|\bar{\mu}^a - \bar{\mu}^{UBM}\|} \quad (13)$$

$$\bar{\mu}^a_{MCS} = \frac{\bar{\mu}^a - \mathcal{M}}{\|\bar{\mu}^a - \mathcal{M}\|} \quad (14)$$

$$\text{with } \mathcal{M} = \frac{1}{N}\sum_{k=1}^{N}\bar{\mu}^k \ , \ \bar{\mu}^k \in \Omega \quad (15)$$

where $\bar{\mu}^a$ represents the supervector to normalize, $\bar{\mu}^{UBM}$ is the supervector of the UBM and $\mathcal{M}$ represents the mean supervector of a subset Ω composed of N supervectors. Figure 2 clearly shows that the UCS-norm allows to reduce the variability of the distance distributions. One can observe that the track number one is no more "orphan" still its distances from the rest of the database have been significantly reduced.

4. EXPERIMENTS

4.1. Data set

For our experiments, we used a music data set composed of 1304 tracks belonging to the following music genres: Country, Electronica, Folk, Gospel, Jazz, Latin, New Age, Pop/Rock, R&B, Rap, Reggae, World. These songs, originally encoded in mp3 32kHz stereo were down-sampled in 22050 kHz and turned into mono by summing the two channels.

4.2. Feature extraction

The short term feature vectors extracted are composed of 13 Mel Frequency Cepstrum Coefficients (MFCC) and 4 Spectral Flatness Measures (SFM). This extraction is made using a sliding window of 40 ms and a hop size of 20 ms.

4.3. Model

For this experiment, we used two types of models: the GMM supervector and a classical multivariate Gaussian Model (GM) with full covariance matrix. For the GMM-UBM approach, the whole data set was used for building a UBM composed of 64 Gaussian components with diagonal covariance matrix. This model was adapted for each song with 5 iterations of MAP using a relevance factor $r = 10$ (see 2.2). Normalized supervectors were extracted as described in section 2.3. The similarity is obtained by the Euclidean distance between supervectors. In the case of GM, the Symmetrized Kullback-Leibler divergence is used.

4.4. Evaluation metric

The evaluation metric used is the "average ratio of genre matches" in the top 1, 3 and 5 nearest neighbors after filtering the results belonging to the same artist as proposed in the MIREX Audio Music Similarity and Retrieval task [1].

4.5. Similarity search time cost

For the GM approach we used a fast implementation of the Symetrized Kullback-Leibler Divergence using its close form expression for multivariate Gaussain models. The covariance matrix inversion was computed off-line and stored into the model. With this system, the time cost for computing the full similarity matrix was of 16 s in a 3GHz 64bits computer which represents $\approx 9.4 \cdot 10^{-6} s$ by model comparison. Using the supervector approach, the duration of the entire similarity matrix process was of 0.13 s, which represents $\approx 7.6 \cdot 10^{-8} s$ by model comparison, representing a time improvement factor of 123.

4.6. Results

The obtained similarity results are presented in Table 1. First of all, we can observe that the supervector approach is slightly better than the standard Gaussian Model using the Kullback-Leibler divergence when no normalization is used. We can also notice the relevance of the P-norm. The UCS-norm and MCS-norm when applied for supervector normalization allows a significant performance improvement compare to the supervector whithout normalization. Moreover, the proposed normalizations methods perform slightly better in average than the P-norm. It is interesting to notice that the MCS-norm achives a better normalization than the UBM centered one. Furthermore, chaining the UCS-norm and the MCS-norm (SV + UCS-norm + MCS-norm) and using a sequence of all the normalization methods (SV + UCS-norm + MCS-norm + P-norm) significantly improve the results, showing that these normalization methods are complementary.

Table 1: Average ratio of artist-filtered genre matches in the top 1, 3 and 5 nearest neighbors. GM = Gaussian Model, SV = Supervector. The last column shows the type of distance related.

System	1NN	3NN	5NN	dist. type
GM	45.01	44.06	44.20	non eucl.
GM + P-norm	48.31	47.52	47.14	non eucl.
SV	46.93	45.67	45.07	euclidean
SV + P-norm	51.38	49.16	47.95	non eucl.
SV + UCS-norm	50.15	49.13	48.81	euclidean
SV + MCS-norm	50.92	49.80	49.09	euclidean
SV + UCS-norm + MCS-norm	51.08	50.13	49.45	euclidean
SV + UCS-norm + MCS-norm + P-norm	52.61	51.51	50.41	non eucl.

5. CONCLUSIONS

We have presented here an application of GMM supervector approach to the music similarity task. This modeling method allows to represent a complex statistical distribution into a Euclidean vector. We have proposed two new supervector projections suitable for the music similarity task. Experiments showed the relevance of our approach.

The improvement obtained by the MCS-norm is promising. Indeed, this normalization can be applied to all type of models which can be embedded into the Euclidean space. Our current research focuses on extending this normalization to other type of models.

[1] http://www.music-ir.org/mirex

6. REFERENCES

[1] J J Aucouturier and F Pachet, "Finding songs that sound the same," in *Proc. of IEEE Benelux Workshop on Model based Processing and Coding of Audio*, 2002, pp. 1–8.

[2] B. Logan and a. Salomon, "A music similarity function based on signal analysis," in *IEEE International Conference on Multimedia and Expo, 2001. ICME 2001.* 2001, vol. 00, pp. 745–748, Ieee.

[3] M. Mandel and D. Ellis, "Song-level features and support vector machines for music classification," in *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR'05), London, UK*, 2005.

[4] T. Pohle and D. Schnitzer, "Striving for an improved audio similarity measure," in *4th Annual Music Information Retrieval Evaluation Exchange*, 2007.

[5] S. Kullback, *Information theory and statistics*, Wiley Publication in Mathematical Statistics, 1959.

[6] DA Reynolds, TF Quatieri, and RB Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital signal processing*, 2000.

[7] C. Cao and M. Li, "Thinkits submissions for MIREX 2009 audio music classification and similarity tasks," in *MIREX abstracts, International Conference on Music Information Retrieval*, 2009.

[8] W M Campbell, D E Sturim, and D A Reynolds, "Support Vector Machines Using GMM Supervectors for Speaker Verification," *IEEE SIGNAL PROCESSING LETTERS*, vol. 13, no. 5, pp. 308, 2006.

[9] J.J. Aucouturier and F. Pachet, "A scale-free distribution of false positives for a large class of audio similarity measures," *Pattern Recognition*, vol. 41, no. 1, pp. 272–284, 2008.

[10] R Auckenthaler, M Carey, and H Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, 2000.

[11] Mayur Datar, Nicole Immorlica, Piotr Indyk, and Vahab S. Mirrokni, "Locality-sensitive hashing scheme based on p-stable distributions," in *SCG'04: Proceedings of the twentieth annual symposium on Computational geometry*, New York, NY, USA, 2004, pp. 253–262, ACM.