Unsupervised accuracy improvement for cover song detection using Spectral Connectivity Network

Mathieu Lagrange, Joan Serra

▶ **To cite this version:**

Mathieu Lagrange, Joan Serra. Unsupervised accuracy improvement for cover song detection using Spectral Connectivity Network. International Conference on Music Information Retrieval, 2010, Utrecht, Netherlands. pp.1-1. hal-01161246

HAL Id: hal-01161246
https://hal.science/hal-01161246

Submitted on 8 Jun 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNSUPERVISED ACCURACY IMPROVEMENT FOR COVER SONG DETECTION USING SPECTRAL CONNECTIVITY NETWORK

Mathieu Lagrange
Analysis-Synthesis team,
IRCAM-CNRS UMR 9912,
1 place Igor Stravinsky, 75004 Paris, France
mathieu.lagrange@ircam.fr

Joan Serrà
Music Technology Group,
Universitat Pompeu Fabra,
Roc Boronat 138, 08018 Barcelona, Spain
joan.serraj@upf.edu

ABSTRACT

This paper introduces a new method for improving the accuracy in medium scale music similarity problems. Recently, it has been shown that the raw accuracy of query by example systems can be enhanced by considering priors about the distribution of its output or the structure of the music collection being considered. The proposed approach focuses on reducing the dependency to those priors by considering an eigenvalue decomposition of the aforementioned system's output. Experiments carried out in the framework of cover song detection show that the proposed approach has good performance for enhancing a high accuracy system. Furthermore, it maintains the accuracy level for lower performing systems.

1. INTRODUCTION

Expressing the similarity between music streams is of interest for many multimedia applications [3]. Though, in many tasks in music information retrieval (MIR), one can observe a glass ceiling in the performance achieved by current methods and algorithms [5]. Several research directions can be considered for tackling this issue. In this paper, we focus on the cover song detection task, but most of the argumentation may be transferred to more general similarity tasks involving a query by example (QBE) system.

One option to boost the accuracy of current QBE systems is to use an enhanced description of the musical stream using the segregation principle [2]. Intuitively, a lot can be gained if an audio signal is available for each instrument. This way, one can easily focus on the stream of interest for each MIR task. In this line, Foucard et al. [8] show that considering a dominant melody removal algorithm as a pre-processing step is a promising approach for observing more robustly the harmonic progression and, in this way, achieve a better accuracy in the cover song detection task. However, it may be a long way until such pre-processing based on segregation will be beneficial for managing medium to large scale musical collections.

An efficient alternative is to consider post-processing approaches exploiting the regularities found in the results of a QBE system for a given music collection. Indeed, music collections are usually organized and structured at multiple scales. In the case of cover detection, songs naturally cluster into so-called cover sets [17]. Therefore, if those cover sets can be approximately estimated, one can gain significant retrieval accuracy, as evidenced by Serrà et al. [17] and Egorov & Linetsky [6]. A different and very interesting post-processing alternative is the general classification scheme proposed by Ravuri & Ellis in [15], where they employ the output of different cover song detection algorithms and a z-score normalization scheme to classify pairs of songs.

Unsupervised post-processing methods that have been introduced so far are rooted on (a) the knowledge of an experimental similarity threshold defining whether two songs are covers or not [17], or (b) the potential number of or cardinality of clusters of the dataset being considered [6]. Thus, these methods are either algorithm or data-dependant. The scheme in [15] is a supervised system trained on different algorithms outputs for some ground truth data. Therefore, it might potentially fail into one or both of the aforementioned dependencies [1].

In this paper, we focus on improving the output of a single QBE system in an unsupervised way. In contrast with the aforementioned references, we propose to consider "more global" approaches in order to alleviate their needs and in order to advance towards unsupervised parameter-free post-processing steps for QBE systems. To this extent we introduce spectral connectivity network (SCN). In addition, we focus on the benefits this technique might provide if the raw accuracy of the QBE system is rather low. This could be the case of a particularly difficult dataset, of a more simple and efficient system (or merely a suboptimal one), or a combination of both cases.

The remaining of the paper is organized as follows: after a presentation of previous work in Sec. 2, we introduce our new accuracy improvement scheme in Sec. 3. In this section, the algorithm is motivated and illustrated on

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page.
© 2010 International Society for Music Information Retrieval.

[1] Furthermore, issues could arise with the employed z-score normalization for some intricate data structures or algorithm outputs (e.g., binomially distributed classifier inputs).

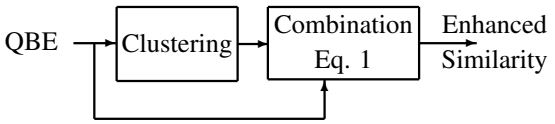

Figure 1. *Combination scheme used for clustering based systems.*

generic artificial datasets. In Sec. 4 we use the evaluation methodology considered in [17] to show the potential of the proposed approach.

2. PREVIOUS WORK

There exist different proposals for the unsupervised post-processing of the output of a single QBE cover song detection system [6, 17]. Most promising strategies so far consist in first estimating the cover sets and use this clustering information in order to increase the overall accuracy as shown in Fig. 1. This can be achieved by considering a classical agglomerative hierarchical clustering algorithm such as the well-known group average linkage (UPGMA) method [10, 19] or alternatively the Community Clustering method (CC) presented in [17], which looks for connected components in a complex network built upon the results of the considered QBE system. Once a clustering solution is obtained, the output distance for a couple of song entries (e_i, e_j) given by a QBE system can be modified to increase the overall accuracy [17]:

$$d'_{i,j} = \begin{cases} \frac{d(e_i,e_j)}{\max(d)} & \text{if } e_i, e_j \in E_k, \\ \frac{d(e_i,e_j)}{\max(d)} + \beta & \text{otherwise.} \end{cases} \quad (1)$$

We denote $d_{i,j}$ as the raw dissimilarity output of the QBE system between two songs e_i and e_j, E_k represents a given cluster, and $\beta > 1$.

Both UPGMA and CC depend on the setting of a threshold similarity value that overall discriminates between cover and non-cover song pairs. This parameter is usually algorithm-dependent. Therefore, for different music collections analyzed through the same QBE system, one should expect similar values for the similarity threshold. That seems to be the case for the algorithm presented in [16] when analyzing different datasets [2] (Fig. 2).

At a first glance one could screen Fig. 2 and set a dissimilarity threshold for roughly separating between covers and not covers. In the present case this threshold could be around 0.6 (or below, if we want to have less false positives). The threshold then would provide the necessary information to the post-processing clustering stage. However, this dissimilarity threshold might not directly correspond to what the clustering algorithm is using internally (e.g., intra-cluster cophenetic distances [10, 19]). In the end one might better perform a grid search for the involved parameter.

In a more general scenario, one might not always be sure about the data or algorithm dependencies of the problem. So, to be on the safe side, some data exploration, algorithm analysis, and/or parameter optimization needs to be done. To avoid those tedious steps is what motivates us to consider unsupervised parameter-free post-processing strategies.

[2] We notice however that both datasets have some similarities, e.g., in terms of genres.

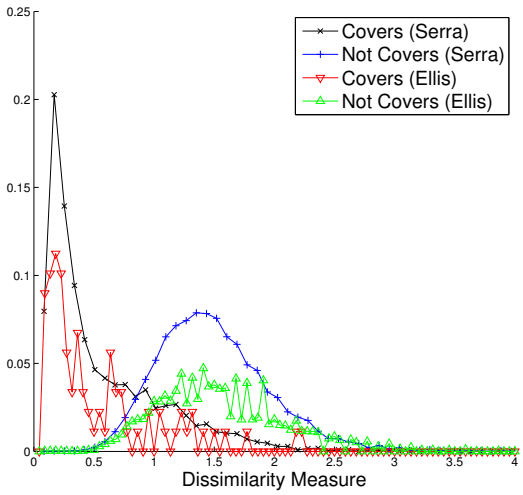

Figure 2. *Normalized histograms for the dissimilarity measure [16] on the music collection of [17] (lines with crosses) and on the "covers80" dataset [7] (lines with triangles).*

3. SPECTRAL CONNECTIVITY NETWORK (SCN)

Without any a priori knowledge about the problem at hand, one needs to root the method on a statistical analysis that is able to identify the underlying structure of the observation, being in our case the output of a QBE system over a large music collection.

Spectral graph clustering has gained popularity in many information retrieval areas, specially in gene, web, image, and audio processing [1, 11, 18]. The interested reader may be referred to [12] for a tutorial introduction.

If S is a square matrix encoding the similarities of all the entries e_i of our music collection E, it can be shown [14] that the eigenvectors of the corresponding Laplacian are relevant clustering indicators for determining the k disjoint set of clusters $E_1, ..., E_k$ (see Fig. 3). We propose to consider this property in order to increase the overall accuracy of QBE systems using the processing scheme shown in Fig. 4. Each of the steps are further detailed in the remaining of this section.

3.1 Similarity Computation

As most QBE systems output a dissimilarity value $d_{i,j}$ measuring how "far" a given couple of entries (e_i, e_j) are, one needs to convert this distance into a similarity value $s_{i,j}$. This is performed using the traditional radial basis function

$$s_{i,j} = e^{\left(\frac{-d^2(e_i,e_j)}{\sigma^2}\right)}, \quad (2)$$

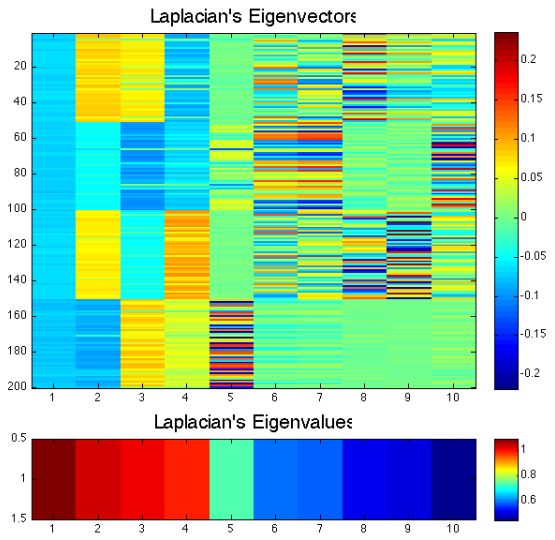

Figure 3. *Eigenvalues and eigenvectors of the Laplacian graph corresponding to a dataset made of 4 bi-dimensional sets of 50 components with low overlapping.*

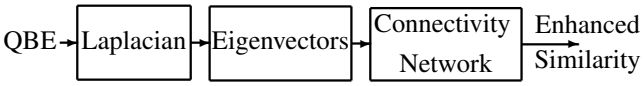

Figure 4. *Processing scheme used for the proposed system based on SCN.*

with σ determined using the local scaling procedure proposed in [20]. The similarities $s_{i,j}$ lead to a matrix S, which is further normalized as [18]:

$$S_d = \mathcal{D}^{-1/2} S \mathcal{D}^{-1/2}, \tag{3}$$

where $\mathcal{D}$ is a diagonal matrix with the degrees $\lceil_1, ..., \lceil_n$ along the diagonal, $\lceil_k = \sum_j s_{k,j}$.

3.2 Eigenvalue Decomposition

As proposed in [14] and illustrated in Fig. 3, the eigenvectors corresponding to the k highest eigenvalues of S_d can be considered as cluster indicators. For that purpose, the contribution of each eigenvector is first normalized with respect to each of the entries, (i.e., per rows).

For a clustering task, any traditional clustering algorithm may then be considered. The k-means algorithm is usually considered in the literature. Though, in the case of cover set detection, the number of clusters is high and their cardinality is low, which makes the algorithm rather slow and highly sensitive to the random initialization. In pre-analysis, it was found more suitable to use the aforementioned UPGMA algorithm. However, in this scenario, one still needs to perform the clustering decision based on a prior, be it the number of clusters or the similarity threshold and consider Eq. 1 for accuracy improvement.

3.3 Connectivity Network

An alternative approach is to consider the Connectivity Network (CN) as our enhanced dissimilarity $d'(i,j)$ by using the projection matrix of the normalized eigenvectors:

$$P = \sum_{k=1}^{N_q} q_k q_k^T, \tag{4}$$

where q_k is the eigenvector corresponding to the k highest eigenvalue λ_k and N_q is the number of eigenvectors to consider. This principle has been originally used for correspondence analysis of contingency tables [9] and reintroduced later in the context of spectral clustering [4].

The usual procedure is to set $N_p = k$ in order to retain only the relevant eigenvectors. If k cannot be considered as a prior (which is the case for cover set detection), one has to consider a method that can robustly estimate k. Unfortunately, no standard estimation procedure gave satisfying results both in terms of accuracy and complexity.

However, notice that in Fig. 3 the eigenvalues are high for the first k eigenvalues and lower afterwards. Considering the eigenvalues as weights in the computation leads us to the so-called Green's function

$$G = \sum_{k=2}^{N_q} q_k \lambda_k q_k^T, \tag{5}$$

where N_g can more safely be set to a high value. An alternate formulation was proposed in [4]:

$$SPCA = D \sum_{k=2}^{N_q} q_k \lambda_k q_k^T D. \tag{6}$$

In the experiments reported in this paper, the Green's function outperformed significantly the two others in the case of unknown k, *i.e.* when N_g is set to the total number of eigenvectors. Since we are interested in a parameter-free system, only the results obtained using this function are reported. Fig. 5 illustrates the use of the Green's function while considering a dataset made of four bi-dimensional Gaussian clusters with significant overlap. Fig. 5(b) is obtained by a bi-dimensional scaling of the Green's function.

4. RESULTS

We split our results into two parts. The first part concerns accuracy improvements related to QBE systems expected to have already a good accuracy and the second part relates to what might happen to systems with worse raw accuracies before the post-processing stages applied in this paper.

4.1 High accuracy QBE systems

In this subsection we attempt to improve a QBE system with quite high raw accuracy. We exactly use the same methodology and input data as in [17].

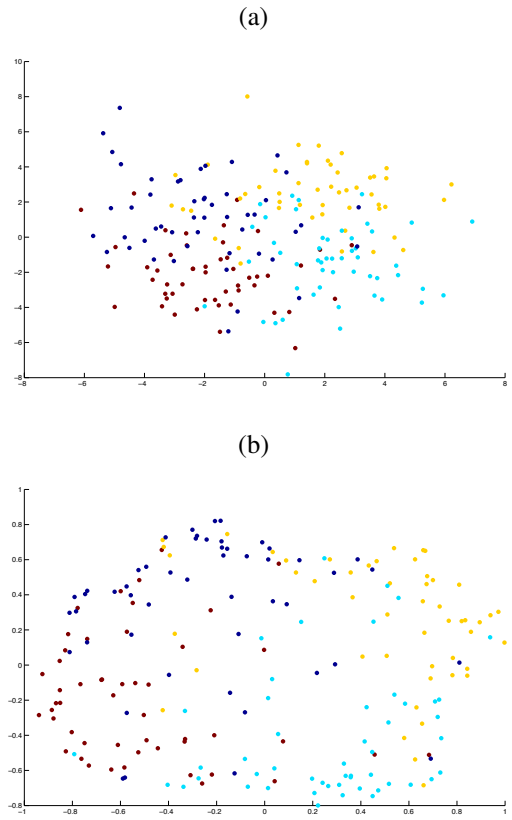

Figure 5. *Four bi-dimensional Gaussian clusters with significant overlap (a) and bi-dimensional scaling plot of the corresponding Green's function (b).*

4.1.1 Methodology

In order to replicate those experiments we use both the same synthetic and real data. Synthetic data is generated by considering a Gaussian noise font $\mathcal{N}(0, 0.25)$ with zero mean and 0.25 standard deviation. A dissimilarity measure between songs i and j is then defined as:

$$d_{i,j} = \begin{cases} 0 & \text{if } i = j, \\ |\mathcal{N}(0, 0.25)| & \text{if } i \text{ and } j \text{ are covers,} \\ 1 - |\mathcal{N}(0, 0.25)| & \text{otherwise.} \end{cases} \quad (7)$$

Real data is provided by the $Q_{\max}$ measure presented in [16] and sampled from the 2125 song collection of [17].

We also employ the same data setups as in [17] (Table 1, where var. means that cover sets have a variable cardinality). Different number of cover sets (n_C) and cardinalities are considered, as well as the fact of adding a different number of noise songs (n_N). For setups 1.1 to 2.4 we repeat the experiments 20 times.

To evaluate QBE systems we employ the mean of average precisions (MAP) over all queries. The MAP is routinely employed in a wide variety of tasks in the IR [13] and MIR communities, including the MIREX cover song identification task [5]. The average precision (AP) for a query i is calculated from the retrieved answer A_i as

Setup	Parameters			
	n_C	Card.	n_N	Trials
1.1	25	4	0	20
1.2	25	var.	0	20
1.3	25	4	100	20
1.4	25	var.	100	20
2.1	125	4	0	20
2.2	125	var.	0	20
2.3	125	4	400	20
2.4	125	var.	400	20
3	525	var.	0	1

Table 1. *Setup summary.*

$$\mathrm{AP}_i = \frac{1}{C_i} \sum_{r=1}^{N} \mathrm{P}_i(r)\mathrm{I}_i(r), \quad (8)$$

where C_i is the total number of covers for the i-th query, N is the total number of songs in the dataset, P_i is the precision of the sorted list A_i at rank r,

$$\mathrm{P}_i(r) = \frac{1}{r} \sum_{l=1}^{r} \mathrm{I}_i(l), \quad (9)$$

and I_i is a relevance function such that $\mathrm{I}_i(z) = 1$ if the song with rank z in A_i is a cover of the i-th song, $\mathrm{I}_i(z) = 0$ otherwise. A relative MAP increase is then computed just dividing the post-processed MAP by the raw one, subtracting 1, and multiplying by 100. For further details about methodology we resort to [17]. In the case of UPGMA and CC we report results with the optimal threshold found, independently for each data source.

4.1.2 Results

As it can be seen in Table 2, a significant accuracy improvement can be gained over the synthetic dataset. UPGMA performs best, followed by SCN which is handicapped by the cluster size variability (setups 2.2 and 2.4).

On the real dataset, UPGMA and CC perform equally well (Table 3). SCN achieves lower performance, probably due to the fact that real data has less intrinsic regularity

	UPGMA	CC	SCN
1.1	10.17	5.49	6.17
1.2	9.76	4.31	4.08
1.3	10.01	3.88	10.20
1.4	9.54	3.73	3.27
2.1	20.95	5.33	20.00
2.2	20.70	4.95	5.98
2.3	21.54	4.62	25.20
2.4	20.35	5.08	10.90

Table 2. *Accuracy improvement (expressed as relative MAP-improvement %) for the synthetic dataset processed using the QBE proposed in [17] as input.*

	UPGMA	CC	SCN
1.1	5.49	4.91	3.55
1.2	4.31	4.00	3.15
1.3	3.88	3.97	3.26
1.4	3.73	4.05	3.45
2.1	5.33	6.44	2.82
2.2	4.95	5.02	2.47
2.3	4.62	6.08	2.43
2.4	4.77	5.06	1.70
3	5.08	5.57	1.14

Table 3. *Accuracy improvement (expressed as relative MAP-improvement %) for the real dataset processed using the QBE proposed in [17] as input.*

than the synthetic one. Actually, no post-processing improves more than 5-6%. This may be explained by the fact that the MAP achieved by the considered system over this concrete dataset is rather high. As a consequence, setting a threshold distance can be done reliably (recall Fig. 2). Therefore, one can speculate that the best MAP that can be achieved given this configuration is in that range.

As a conclusion, it seems that approaches focusing on locality (UPGMA and CC) are more relevant than global approaches (SCN) for improving the performance of a QBE system with rather high raw accuracy provided that their clustering threshold can be set reliably.

4.2 Lower Accuracy QBE systems

In light of the previous results, we are interested in seeing how these clustering schemes perform on lower accuracy systems. Motivations for that could be that we either do not have a good, high performing QBE system for a given task, but a more modest one, or either that we are using a faster and more efficient version of the original system. Furthermore, we could be dealing with a particularly difficult dataset where our (otherwise reliable) QBE system performs more poorly.

In these cases, the accuracy improvement provided by the post-processing steps outlined in this paper could be more significant than with the original high accuracy system. It could even be the case that, with a (in principle) lower performing QBE system, we reached the same (or a higher) final MAP.

For lower accuracy systems it is theoretically relevant to consider more global approaches, as setting a dissimilarity threshold is more difficult due to the noise level. However, the overall structure of the dataset might not be completely lost, and therefore we can still take benefit of this fact by using a method like SCN. This can be asserted by comparing the MAP increase achieved by the studied methods when considering as input a lower accuracy system [7] (Table 4).

4.2.1 Methodology

We propose to further verify the previous assertion by simulating a QBE system with a controllable accuracy. For

	MAP	UPGMA	CC	SCN
Serrà et al. [16]	0.73	4.01	1.27	1.14
Ellis & Cotton [7]	0.42	8.06	3.04	19.70

Table 4. *MAP and MAP increase (%) for two QBE systems over the "covers80" dataset [7]. UPGMA and CC thresholds were specifically optimized for this dataset (however no significant difference was observed, c.f. Sec. 2).*

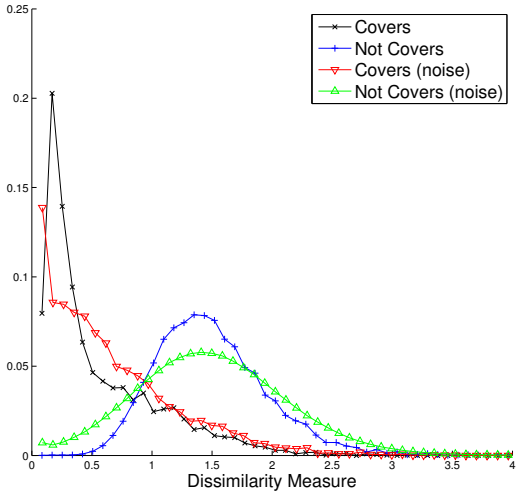

Figure 6. *Normalized histograms for real data with no noise (lines with crosses) and with $\sigma = 0.45$ (lines with triangles).*

that purpose, noise is added to our real data $d_{i,j}$ (the output of the high accuracy reference QBE system) such that

$$\tilde{d}_{i,j} = |d_{i,j} + \mathcal{N}(0, \sigma d_{\text{mx}})|\,, \qquad (10)$$

where σ is the noise level and d_{mx} is a normalization factor set to the maximal dissimilarity found (see Fig. 6 for the corresponding histograms).

4.2.2 Results

As it can be seen in Fig. 7, CC does not maintain its initial MAP increase when the noise level raises up. In contrast, UPGMA maintains or slightly increases its relative MAP. We finally see that SCN really boosts the MAP increase as more noise is added. This confirms our hypothesis and leads us to speculate that these methods are more robust for low accuracy QBE systems.

5. CONCLUSION

We proposed a global approach for improving the accuracy of query-by-example (QBE) systems based on spectral connectivity network. Contrasting with other state-of-the-art approaches, it does not rely on any parameter setting such as a dissimilarity threshold or the expected number of or cardinality of clusters within the data.

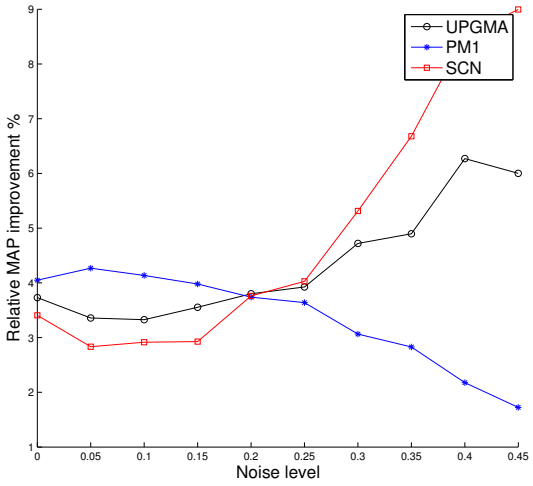

Figure 7. *Relative accuracy increase as a function of the noise level for setup 2.4 using Serra's data/QBE combination as input.*

The experiments showed that the proposed approach exhibits comparable results for improving high accuracy QBE systems and becomes highly competitive for improving lower accuracy QBE systems. Future research will include a more in depth study upon the selection of the relevant eigenvectors (a problem closely linked to the estimation of the number of clusters in a dataset).

6. ACKNOWLEDGEMENTS

M.L. has been partially funded by the Quaero project within the task 6.4: "Music Search by Similarity". J.S. has been partially funded by the Music 3.0 project TSI-070100-2008-318 of the Spanish Ministry of Industry, Tourism, and Trade.

7. REFERENCES

[1] F. Bach and M. I. Jordan. Learning spectral clustering, with application to speech separation. *Journal of Machine Learning Research*, 7:1963–2001, 2006.

[2] A. S. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound*. The MIT Press, 1990.

[3] M. Casey, R. C. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney. Content-based music information retrieval: current directions and future challenges. *Proceedings of the IEEE*, 96(4):668–696, April 2008.

[4] C. Ding, X. He, H. Zha, and H. Simon. Unsupervised learning: Self-aggregation in scaled principal component space. In *European Conference on Principles and Practice of Knowledge Discovery in Databases*, 2002.

[5] J. S. Downie. The music information retrieval evaluation exchange (2005–2007): a window into music information retrieval research. *Acoustical Science and Technology*, 29(4):247–255, 2008.

[6] A. Egorov and G. Linetsky. Cover song identification with IF-F0 pitch class profiles. *MIREX extended abstract*, September 2008.

[7] D. P. W. Ellis and C. Cotton. The 2007 LabROSA cover song detection system. *MIREX extended abstract*, September 2007.

[8] R. Foucard, J.-L. Durrieu, M. Lagrange, and G. Richard. Multimodal similarity between musical streams for cover version detection. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP'10)*, Dallas, Texas, USA, March 2010.

[9] M. J. Greenacre. *Theory and Applications of Correspondence Analysis*. Academic Press, 1984.

[10] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323, September 1999.

[11] M. Lagrange, L. G. Martins, J. Murdoch, and G. Tzanetakis. Normalized cuts for predominant melodic source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):278–290, Feb. 2008.

[12] U. Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17:395–416, 2007.

[13] C. D. Manning, R. Prabhakar, and H. Schutze. *An introduction to Information Retrieval*. Cambridge University Press, 2008.

[14] M. Meila and J. Shi. Learning segmentation by random walks. In *Advance on Neural Information Processing Systems*, 2000.

[15] S. Ravuri and D. P. W. Ellis. Cover song detection: from high scores to general classification. *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 55–68, March 2010.

[16] J. Serrà, X. Serra, and R. G. Andrzejak. Cross recurrence quantification for cover song identification. *New Journal of Physics*, 11:093017, September 2009.

[17] J. Serra, M. Zanin, C. Laurier, and M. Sordo. Unsupervised detection of cover song sets: Accuracy improvement and original identification. In *International Society for Music Information Retrieval Conference*, 2009.

[18] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:888–905, 1997.

[19] R. Xu and D. C. Wunsch. *Clustering*. IEEE Press, 2009.

[20] L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. In *Annual Conference on Neural Information Processing Systems*, 2004.