2009 11th IEEE International Symposium on Multimedia

From Low-level to High-level: Comparative Study of Music Similarity Measures

Dmitry Bogdanov, Joan Serrà, Nicolas Wack, and Perfecto Herrera
Music Technology Group
Universitat Pompeu Fabra
Roc Boronat, 138, 08018 Barcelona, Spain
{dmitry.bogdanov,joan.serraj,nicolas.wack,perfecto.herrera}@upf.edu

Abstract

Studying the ways to recommend music to a user is a central task within the music information research community. From a content-based point of view, this task can be regarded as obtaining a suitable distance measurement between songs defined on a certain feature space. We propose two such distance measures. First, a low-level measure based on tempo-related aspects, and second, a high-level semantic measure based on regression by support vector machines of different groups of musical dimensions such as genre and culture, moods and instruments, or rhythm and tempo. We evaluate these distance measures against a number of state-of-the-art measures objectively, based on 17 ground truth musical collections, and subjectively, based on 12 listeners' ratings. Results show that, in spite of being conceptually different, the proposed methods achieve comparable or even higher performance than the considered baseline approaches. Furthermore, they open up the possibility to explore distance metrics that are based on truly semantic notions.

1. Introduction

Studying the ways to recommend music to a user is a central task within the music information research (MIR) community [7]. From a simplistic point of view, in addition to generating personal preference profiles, this task can be regarded as obtaining a suitable distance[1] measurement between a "preferred" song and a set of potential "to-be-liked" candidates defined in a certain feature space. Currently, researchers and practitioners fill in this feature space with information extracted from the audio content, context, or both. Focusing on audio content-based MIR, there exist a wide variety of approaches for providing such a distance measurement. Examples include applying an L_p metric after a preliminary selection of audio descriptors [6], comparing Gaussian mixture models (GMM) of mel-frequency cepstral coefficients (MFCCs) [1], or more elaborated approaches [2, 3, 19, 20, 21, 22, 27].

Though common approaches for content-based music similarity may include a variety of perceptually relevant descriptors related to different musical aspects, such descriptors are, in general, relatively low-level and not directly associated with a high-level semantic explanation [8]. In contrast, research on computing high-level semantic features from low-level audio descriptors exists. Moreover, in the context of MIR classification problems, this research has yielded remarkable results [15, 17, 26]. Starting from this relative success, we hypothesize that the combination of classification problems for distance-based music recommendation could be a relevant step to overcome the so-called semantic gap [8].

The present work deals with content-based approaches for music similarity. Using state-of-the-art low-level audio descriptors (Sec. 2), we compare several baseline approaches and explore two basic ideas to create novel distance measures (Sec. 3). More concretely, as baseline approaches we consider Euclidean distances defined on descriptor subsets (Secs. 3.1 and 3.2) and Kullback-Leibler divergence defined on GMMs of MFCCs (Sec. 3.3). The first idea we explore consists of the use of tempo-related musical aspects. To this extent, we propose a simple distance based on two low-level descriptors, namely beats per minute (BPM) and onset rate (OR) (Sec. 3.4). The second idea we explore shifts the problem to a more high-level (semantic) domain. To this extent, we continue the research of [2, 3, 27] but, more in the line of [27], we investigate the possibility of benefiting from results obtained in different classification tasks and transferring this gained knowledge to the context of music recommendation (Sec. 3.5). We evaluate all the considered approaches with a unique methodological basis, including an objective evaluation on several comprehensive ground truth music collec-

[1] We here pragmatically use the term distance to refer to any dissimilarity measurement between songs.

978-0-7695-3890-7/09 $26.00 © 2009 IEEE
DOI 10.1109/ISM.2009.72

tions (Sec. 4.1) and a subjective evaluation based on ratings given by real listeners (Sec. 4.2). We show that, in spite of being conceptually different, the proposed methods achieve comparable or even higher performance than the considered baseline approaches (Sec. 5). Finally, we state general conclusions and discuss the possibility of further improvements (Sec. 6).

2. Musical descriptors

We characterize each song using an in-house audio analysis tool. This tool provides over 60 descriptor classes in total, characterizing global properties of songs. The majority of these descriptors are extracted on a frame-by-frame basis and then summarized by (at least) their means and variances across frames. In the case of multidimensional descriptors, covariances between components are also considered (e.g. with MFCCs). Extracted descriptor classes include inharmonicity, odd to even harmonic energy ratio, tristimulus, spectral centroid, spread, skewness, kurtosis, decrease, flatness, crest, and roll-off factors [20], MFCCs [16], spectral energy bands, zero-crossing rate [10], spectral and tonal complexities [25], transposed and untransposed harmonic pitch class profiles, key strength, tuning, chords [11], BPM, and onsets [4].

3. Studied approaches

3.1. Euclidean distance based on principal component analysis (L_2-PCA)

As a starting point we follow the ideas proposed in [6] and apply an unweighted Euclidean metric on a manually selected subset of the descriptors outlined above[2]. Preliminary steps include descriptor normalization in the interval $[0, 1]$ and principal component analysis (PCA) [28] to reduce the dimension of the descriptor space to 25 variables.

3.2. Euclidean distance based on relevant component analysis (L_2-RCA-1 and L_2-RCA-2)

Along with the L_2-PCA measure, we consider more possibilities of descriptor selection. To this extent, instead of PCA, we perform relevant component analysis (RCA) [24]. As well as PCA, RCA gives a rescaling linear transformation of a descriptor space but is based on preliminary training on a number of groups of similar songs. In the objective evaluation (Sec. 4.1) for each collection we supply the algorithm with part of the ground truth information. As in the L_2-PCA approach, the output dimensionality is chosen to be 25. In addition to the descriptor subset used in L_2-PCA, the overall set of descriptors is analyzed (L_2-RCA-1 and L_2-RCA-2, respectively).

[2] Specific details not included in the cited reference were consulted with P. Cano in personal communication.

3.3. Kullback-Leibler divergence based on GMM MFCC modeling (1G-MFCC)

Alternatively, we consider timbre modeling with GMM as another baseline approach [1]. We implement the simplification of this timbre model using single Gaussian with full covariance matrix [9, 17]. Comparative research of timbre distance measures using GMMs indicates that such simplification can be used without significantly decreasing performance while being computationally less complex [14]. As a distance measure between single Gaussian models for songs X and Y we use a closed form symmetric approximation of the Kullback-Leibler divergence,

$$\begin{aligned} d(X,Y) = \\ & Tr(\Sigma_X^{-1}\Sigma_Y) + Tr(\Sigma_Y^{-1}\Sigma_X) + \\ & Tr((\Sigma_X^{-1} + \Sigma_Y^{-1})(\mu_X - \mu_Y)(\mu_X - \mu_Y)^T) - \\ & 2N_{MFCC}, \end{aligned} \quad (1)$$

where μ_X and μ_Y are MFCC means, Σ_X and Σ_Y are MFCC covariance matrices, and $N_{MFCC} = 13$ is the number of used MFCCs.

3.4. Tempo-based distance (TEMPO)

The first approach we propose is related to the exploitation of tempo-related musical aspects with a simple distance measure based on BPM and OR. For two songs X and Y with BPMs X_{BPM} and Y_{BPM}, and ORs X_{OR} and Y_{OR}, we determine this measure as a linear combination of two separate distance functions,

$$d(X,Y) = w_{BPM} d_{BPM}(X,Y) + w_{OR} d_{OR}(X,Y), \quad (2)$$

defined for BPM as

$$d_{BPM}(X,Y) = \min_{i \in \mathbb{N}} \alpha_{BPM}^{i-1} \left| \frac{max(X_{BPM}, Y_{BPM})}{min(X_{BPM}, Y_{BPM})} - i \right|, \quad (3)$$

and for OR as

$$d_{OR}(X,Y) = \min_{i \in \mathbb{N}} \alpha_{OR}^{i-1} \left| \frac{max(X_{OR}, Y_{OR})}{min(X_{OR}, Y_{OR})} - i \right|, \quad (4)$$

where $X_{BPM}, Y_{BPM}, X_{OR}, Y_{OR} > 0$, $\alpha_{BPM}, \alpha_{OR} \geq 1$.

The parameters w_{BPM} and w_{OR} of Eq. 2 define the weights for each distance component. Eq. 3 (Eq. 4) is based on the assumption that songs with the same BPMs (ORs) or multiple ones (e.g. $X_{BPM} = iY_{BPM}$) are more similar

than songs with non-multiple BPMs (ORs). For example, the songs X and Y with $X_{BPM} = 140$ and $Y_{BPM} = 70$ should have a closer distance than the songs X and Z with $Z_{BPM} = 100$. The strength of this assumption depends on the parameter α_{BPM} (α_{OR}). In the case of $\alpha_{BPM} = 1$, all multiple BPMs are treated equally, while in the case of $\alpha_{BPM} > 1$, preference inversely decreases with i. In practice we use $i = 1, 2, 4, 6$.

In pre-analysis we performed a grid search with one of the ground truth music collections (Sec. 4.1) and we found $w_{BPM} = w_{OR} = 0.5$ and $\alpha_{BPM} = \alpha_{OR} = 30$ to be the best parameter configuration. Such values reveal the fact that actually both components are equally meaningful and that mainly a 1-to-1 relation of BPMs (ORs) is relevant for the overall song similarity, respectively. When our BPM (OR) estimator has more duplicity errors (e.g. a BPM of 80 was estimated as 160), we should expect lower α values.

3.5. Classifier-based distance (CLAS)

The second approach we propose derives a distance measure from diverse classification tasks. In distinction from the aforementioned methods, which directly operate on a low-level descriptor space, we first infer high-level semantic descriptors using suitably trained classifiers and then define a distance measure operating on this newly formed high-level semantic space.

For the first step we choose standard multi-class support vector machines (SVMs) [28], which are shown to be an effective tool for different classification tasks in MIR [12, 15, 17, 29]. We apply an SVM regression to different musical dimensions such as genre and culture, moods and instruments, or rhythm and tempo. More concretely, 14 classification tasks are run according to all available ground truth collections[3] (Sec. 4.1). For each ground truth collection, one SVM is trained with a preliminary correlation-based feature selection (CFS) [28] over all $[0, 1]$-normalized descriptors (Sec. 2). The resulting high-level descriptor space is formed by the probability values of each class for each SVM. In pre-analysis we compared several SVM models and we finally decided to use the libSVM[4] implementation with the C-SVC method and a radial basis function kernel with default parameters.

For the second step we consider different measures frequently used in collaborative filtering systems: cosine distance (CLAS-Cos), Pearson correlation distance (CLAS-Pears), Spearman's rho correlation distance (CLAS-Spear), weighted cosine distance (CLAS-Cos-W), weighted Pearson correlation distance (CLAS-Pears-W), and adjusted cosine distance (CLAS-Cos-A). Adjusted cosine distance is computed by taking into account the average probability for each class. Weighting is done both manually (W_M) and based on classification accuracy (W_A). For W_M, we split the collections into 3 musical dimensions, namely genre and culture, moods and instruments, and rhythm and tempo, and empirically assign weights 0.50, 0.30, and 0.20 respectively. For W_A, we evaluate the accuracy of each classifier, and assign directly proportional weights which sum to 1.

From this perspective, the problem of content-based music recommendation can be seen as a collaborative filtering problem with class labels playing the role of users and probabilities playing the role of user ratings, so that each N-class classifier corresponds to N users.

4. Evaluation Methodology

We evaluated all considered approaches with a unique methodological basis, including an objective evaluation on comprehensive ground truths and a subjective evaluation based on ratings given by real listeners. As an initial benchmark for the comparison of the considered approaches we used a random distance (RAND), i.e. we selected a random number from the standard uniform distribution as the distance between two songs.

4.1. Objective evaluation

We covered different musical dimensions such as genre, mood, artist, album, culture, rhythm, or presence or absence of voice. A number of ground truth music collections (including full songs and excerpts) were employed for that purpose (Table 1). For some dimensions we used already existing collections in the MIR field [5, 12, 13, 15, 23, 26], while for other dimensions we created different manually labeled in-house collections.

For our evaluation measure, we used the mean average precision (MAP) [18]. For each approach and music collection, MAP was computed from the corresponding full distance matrix. The average precision (AP) [18] was computed for each matrix row (for each song query) and the mean was calculated. The results were averaged over 5 iterations of 3-fold cross-validation.

4.2. Subjective evaluation

Starting from the results of the objective evaluation (Sec. 5.1), we selected 4 conceptually different approaches (L_2-PCA, 1G-MFCC, TEMPO, and CLAS-Pears-W_M) together with the random baseline (RAND) for subjective evaluation. To this extent, we designed a web-based survey where registered listeners performed a number of iterations blindly voting for the considered distance measures. During one iteration each listener was presented with 5 different playlists (one for each measure) generated from the

[3] We ignored music collections with insufficient size of class samples.

[4] http://www.csie.ntu.edu.tw/∼cjlin/libsvm/

Acronym	Musical dimension	Classes	Size	Source
G1	Genre & Culture	Alternative, blues, electronic, folk/country, funk/soul/rnb, jazz, pop, rap/hiphop, rock	1820 song excerpts, 46 - 490 per genre	[13]
G2	Genre & Culture	Classical, dance, hip-hop, jazz, pop, rhythm'n'blues, rock, speech	400 full songs, 50 per genre	In-house
G3	Genre & Culture	Alternative, blues, classical, country, electronica, folk, funk, heavy metal, hip-hop, jazz, pop, religious, rock, soul	140 full songs, 10 per genre	[23]
G4	Genre & Culture	Blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, rock	993 song excerpts, 100 per genre	[26]
CUL	Genre & Culture	Western, non-western	1640 song excerpts, 1132/508 per class	[12]
MHA	Moods & Instruments	Happy, non-happy	302 full songs + excerpts, 139/163 per class	[15] + in-house
MSA	Moods & Instruments	Sad, non-sad	230 full songs + excerpts, 96/134 per class	[15] + in-house
MAG	Moods & Instruments	Aggressive, non-aggressive	280 full songs + excerpts, 133/147 per class	[15] + in-house
MRE	Moods & Instruments	Relaxed, non-relaxed	446 full songs + excerpts, 145/301 per class	[15] + in-house
MPA	Moods & Instruments	Party, non-party	349 full songs + excerpts, 198/151 per class	In-house
MAC	Moods & Instruments	Acoustic, non-acoustic	321 full songs + excerpts, 193/128 per class	[15] + in-house
MEL	Moods & Instruments	Electronic, non-electronic	332 full songs + excerpts, 164/168 per class	[15] + in-house
MVI	Moods & Instruments	Voice, instrumental	1000 song excerpts, 500 per class	In-house
ART	Artist	200 different artist names	2000 song excerpts, 10 per artist	In-house
ALB	Album	200 different album titles	2000 song excerpts, 10 per album	In-house
RPS	Rhythm & Tempo	Perceptual speed: slow, medium, fast	3000 full songs, 1000 per class	In-house
RBL	Rhythm & Tempo	Chachacha, jive, quickstep, rumba, samba, tango, viennese waltz, waltz	683 song excerpts, 60 - 110 per class	[5]

Table 1. Objective evaluation ground truth music collections.

same seed song[5]. Each playlist consisted of the 5 nearest-to-the-seed songs. The entire process used an in-house collection of 300K music excerpts (30 sec.) by 60K artists (5 songs/artist) covering a wide range of musical dimensions (different genres, styles, arrangements, geographic locations, and epochs). Independently for each playlist, we asked listeners to provide (i) a playlist similarity rating (appropriateness of the playlist with respect to the seed) using a 6-point Likert-type scale (0 corresponding to the lowest similarity, 5 to the highest) and (ii) a playlist inconsistency boolean answer. We did not present examples of inconsistency but they might comprise of speech mixed with music, extremely different tempos, completely opposite feelings or emotions, distant musical genres, etc. The first 12 seeds and corresponding playlists were shared between all listeners, while the remaining iteration seeds (up to a maximum of 21) were different for each listener as the seeds were randomly selected. Altogether we collected playlist similarity ratings, playlist inconsistency indicators, and background information about listening and musical expertise (each measured in 3 levels) from 12 listeners.

5. Results and discussion

5.1. Objective evaluation

We first show that the considered distances outperform the random baseline (RAND) for most of the music collections (Table 2). When comparing baseline approaches (L_2-PCA, L_2-RCA-1, L_2-RCA-2, 1G-MFCC), we found 1G-MFCC to perform best on average. Still, L_2-PCA performed similarly or slightly better for some collections (e.g. MAC or RPS). With respect to tempo-related collections, TEMPO performs similarly (RPS) or significantly better (RBL) than baseline approaches. Furthermore, it is the best performing distance for the RBL collection. Surprisingly, TEMPO yielded accuracies which are comparable to some of the baseline approaches for music collections not strictly related to rhythm or tempo such as G2, MHA, and MEL. Finally, we see that classifier-based distances achieved the best accuracies for the large majority of the collections. Due to space reasons and since all CLAS-based distances (CLAS-Cos, CLAS-Pears, CLAS-Spear, CLAS-Cos-W, CLAS-Pears-W, CLAS-Cos-A) showed equal accuracies, we only report two examples of them. In particular, CLAS-based distances achieved significant accuracy improvements with the G2, G4, MPA, MSA, and MAC collections. In contrast, no improvement was achieved with the ART, ALB, and RBL collections: 1G-MFCC performed best for ART and ALB collections, while TEMPO had the highest accuracy for RBL. We hypothesize that the success of 1G-MFCC for ART and ALB collections might be due to the well known "album effect" [17].

[5] A screenshot of the survey can be accessed online: http://www.iua.upf.edu/∼perfe/misc/simsubjeval.png

Method	G1	G2	G3	G4	CUL	MHA	MSA	MAG	MRE	MPA	MAC	MEL	MVI	ART	ALB	RPS	RBL
RAND	0.17	0.16	0.20	0.12	0.58	0.53	0.55	0.53	0.58	0.53	0.54	0.52	0.51	0.02	0.02	0.34	0.15
L_2-PCA	0.24	0.39	0.23	0.24	0.69	0.58	0.69	0.80	0.73	0.67	0.72	0.58	0.56	0.08	0.11	0.40	0.24
L_2-RCA-1	0.23	0.34	0.13	0.26	0.73	0.53	0.54	0.55	0.59	0.56	0.57	0.54	0.60	0.10	0.16	0.38	0.21
L_2-RCA-2	0.22	0.19	0.13	0.24	0.73	0.52	0.53	0.53	N.C.	0.54	0.54	0.53	0.58	0.09	0.15	0.38	0.20
1G-MFCC	0.29	0.43	0.26	0.29	0.85	0.58	0.68	0.84	0.74	0.69	0.70	0.58	0.61	**0.15**	**0.24**	0.39	0.25
TEMPO	0.22	0.36	0.19	0.17	0.60	0.56	0.59	0.53	0.58	0.61	0.56	0.56	0.52	0.03	0.02	0.38	**0.44**
CLAS-Pears	0.32	0.61	0.29	0.40	0.84	**0.69**	**0.81**	**0.93**	**0.86**	**0.85**	**0.85**	**0.66**	**0.62**	0.05	0.06	0.43	0.35
CLAS-Pears-W$_M$	**0.33**	**0.67**	**0.30**	**0.43**	**0.88**	0.68	0.80	0.91	0.85	0.84	0.83	0.65	0.59	0.06	0.06	**0.44**	0.35

Table 2. Objective evaluation results (MAP) for the different music collections considered. N.C. stands for "not computed" due to technical difficulties.

5.2. Subjective evaluation

A one-way within-subjects ANOVA test using the entire set of subjective ratings was carried out. The test checks the effect of subjects as a main source of the variance found in the ratings. As this effect was found to be non-significant, we had green light to check the effect of the considered distances, which proved to be significant ($F(4, 44) = 276.310$, $p = 0.000$). Furthermore, post-hoc tests revealed no significant difference between CLAS-Pears-W$_M$ and 1G-MFCC, and no significant differences between L2-PCA, RANDOM and TEMPO (Fig. 1). In contrast, significant differences between the methods of these two groups were found. A final ANOVA test, using only the shared data, revealed the same pattern, which points to the conclusion that the different similarities captured by the different methods are quickly grasped (and easily assessed) by listeners. The proportion of playlists considered to be inconsistent followed the same pattern of differences and significance as the similarity ratings. An additional Spearman test can be done by computing the correlation between subject's ratings. This test reveals that the correlations are high. The observed range is $0.452 - 0.869$, and the average between-raters correlation is 0.772 ($sd = 0.089$). In total, we can be confident on the reliability of the ratings.

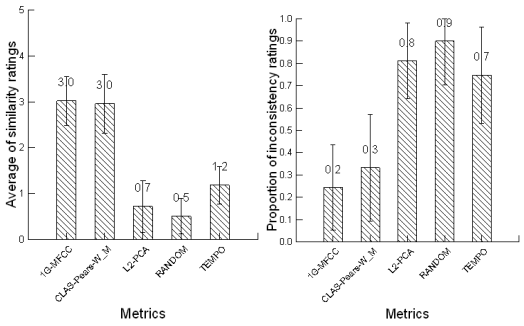

Figure 1. Average playlist similarity rating and proportion of inconsistent playlists for the subjective evaluation.

6. Conclusions

In the present work we study and comprehensively evaluate, both objectively and subjectively, the accuracy of different content-based distance measures for music recommendation. We consider 4 baseline distances and a random-based one. Furthermore, we explore the potential of two new conceptually different distances not strictly operating on musical timbre aspects. More concretely, we present a simple tempo-based distance which can be especially useful for expressing music similarity in collections where rhythm aspects are predominant. In addition, we investigate the possibility of benefiting from classification problems' results and transferring this gained knowledge to the context of music recommendation. To this extent, we present a classifier-based distance which makes use of high-level semantic descriptors inferred from low-level ones. This distance covers diverse musical dimensions such as genre and culture, moods and instruments, and rhythm and tempo, and outperforms all the considered approaches in most of the ground truth music collections used for objective evaluation. Contrastingly, this performance improvement is not seen in the subjective evaluation when comparing with the best performing baseline distance considered. However, no statistically significant differences are found between them.

Further research will be devoted to improving the classifier-based distance with more musical dimensions such as tonality or instrument information. Given that several separate dimensions can be straightforwardly combined with this distance, additional improvements are feasible and potentially beneficial. In general, the classifier-based distance represents a semantically rich approach to recommending music. Thus, in spite of being based solely on audio content information, this approach can overcome the so-called "semantic gap" in content-based music recommendations and provide a semantic explanation to justify the recomendations to a user.

7. Acknowledgments

The authors would like to thank Jordi Funollet and Owen Meyers for technical support and all participants of the subjective evaluation. This work was partially funded by the EU-IP project PHAROS IST-2006-045035, and the FI Grant of Generalitat de Catalunya (AGAUR).

References

[1] J. J. Aucouturier, F. Pachet, and M. Sandler. "The way it sounds": timbre models for analysis and retrieval of music signals. *IEEE Transactions on Multimedia*, 7(6):1028–1035, 2005.

[2] L. Barrington, A. Chan, D. Turnbull, and G. Lanckriet. Audio information retrieval using semantic similarity. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'07)*, volume 2, pages 725–728, 2007.

[3] A. Berenzweig, D. P. W. Ellis, and S. Lawrence. Anchor space for classification and similarity measurement of music. In *International Conference on Multimedia and Expo (ICME'03)*, volume 1, pages 29–32, 2003.

[4] P. M. Brossier. *Automatic Annotation of Musical Audio for Interactive Applications*. PhD thesis, QMUL, London, UK, 2007. http://aubio.org/phdthesis/.

[5] P. Cano, E. Gómez, F. Gouyon, P. Herrera, M. Koppenberger, B. Ong, X. Serra, S. Streich, and N. Wack. ISMIR 2004 audio description contest. Technical report, 2006. http://mtg.upf.edu/node/461.

[6] P. Cano, M. Koppenberger, and N. Wack. Content-based music audio recommendation. In *ACM International Conference on Multimedia (ACMMM'05)*, pages 211–212, 2005.

[7] M. A. Casey, R. 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, 2008.

[8] O. Celma, P. Herrera, and X. Serra. Bridging the music semantic gap. In *ESWC 2006 Workshop on Mastering the Gap: From Information Extraction to Semantic Representation*, 2006. http://mtg.upf.edu/node/874.

[9] A. Flexer, D. Schnitzer, M. Gasser, and G. Widmer. Playlist generation using start and end songs. In *International Symposium on Music Information Retrieval (ISMIR'08)*, pages 173–178, 2008.

[10] F. Gouyon. *A computational approach to rhythm description: Audio features for the computation of rhythm periodicity functions and their use in tempo induction and music content processing*. PhD thesis, UPF, Barcelona, Spain, 2005. http://www.iua.upf.es/∼fgouyon/thesis/.

[11] E. Gómez. *Tonal Description of Music Audio Signals*. PhD thesis, UPF, Barcelona, Spain, 2006. http://www.iua.upf.es/∼egomez/thesis/.

[12] E. Gómez and P. Herrera. Comparative analysis of music recordings from western and Non-Western traditions by automatic tonal feature extraction. *Empirical Musicology Review*, 3(3):140–156, 2008.

[13] H. Homburg, I. Mierswa, B. Möller, K. Morik, and M. Wurst. A benchmark dataset for audio classification and clustering. In *International Conference on Music Information Retrieval (ISMIR'05)*, pages 528–531, 2005.

[14] J. H. Jensen, M. G. Christensen, D. P. W. Ellis, and S. H. Jensen. Quantitative analysis of a common audio similarity measure. *IEEE Transactions on Audio, Speech, and Language Processing*, 17:693–703, 2009.

[15] C. Laurier, O. Meyers, J. Serrà, M. Blech, and P. Herrera. Music mood annotator design and integration. In *International Workshop on Content-Based Multimedia Indexing (CBMI'2009)*, 2009. http://mtg.upf.edu/node/1260.

[16] B. Logan. Mel frequency cepstral coefficients for music modeling. In *International Symposium on Music Information Retrieval (ISMIR'00)*, 2000.

[17] M. I. Mandel and D. P. Ellis. Song-level features and support vector machines for music classification. In *International Conference on Music Information Retrieval (ISMIR'05)*, pages 594–599, 2005.

[18] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*. Cambridge University Press, 2008.

[19] E. Pampalk, A. Flexer, and G. Widmer. Improvements of audio-based music similarity and genre classification. In *International Conference on Music Information Retrieval (ISMIR'05)*, pages 628–633, 2005.

[20] G. Peeters. A large set of audio features for sound description (similarity and classification) in the CUIDADO project. *CUIDADO Project Report*, 2004. http://recherche.ircam.fr/equipes/analyse-synthese/peeters/ARTICLES/.

[21] T. Pohle, P. Knees, M. Schedl, and G. Widmer. Automatically adapting the structure of audio similarity spaces. In *Workshop on Learning the Semantics of Audio Signals (LSAS'06)*, pages 66–75, 2006.

[22] T. Pohle and D. Schnitzer. Striving for an improved audio similarity measure. *Music Information Retrieval Evaluation Exchange (MIREX'07)*, 2007. http://www.music-ir.org/mirex/2007/abs/AS_pohle.pdf.

[23] P. J. Rentfrow and S. D. Gosling. The do re mi's of everyday life: The structure and personality correlates of music preferences. *Journal of Personality and Social Psychology*, 84:1236–1256, 2003.

[24] N. Shental, T. Hertz, D. Weinshall, and M. Pavel. Adjustment learning and relevant component analysis. *Lecture Notes In Computer Science*, pages 776–792, 2002.

[25] S. Streich. *Music complexity: a multi-faceted description of audio content*. PhD thesis, UPF, Barcelona, Spain, 2007. http://www.tesisenxarxa.net/TDX-0124108-174625/.

[26] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, 10(5):293–302, 2002.

[27] K. West and P. Lamere. A model-based approach to constructing music similarity functions. *EURASIP Journal on Advances in Signal Processing*, 2007:149–149, 2007.

[28] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2005.

[29] C. Xu, N. C. Maddage, X. Shao, F. Cao, and Q. Tian. Musical genre classification using support vector machines. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03)*, pages 429–432, 2003.