Unifying Low-level and High-level Music Similarity Measures

Dmitry Bogdanov, Joan Serrà, Nicolas Wack, Perfecto Herrera, and Xavier Serra

***Abstract*—Measuring music similarity is essential for multimedia retrieval. For music items, this task can be regarded as obtaining a suitable distance measurement between songs defined on a certain feature space. In this paper, we propose three of such distance measures based on the audio content. First, a low-level measure based on tempo-related description. Second, a high-level semantic measure based on the inference of different musical dimensions by support vector machines. These dimensions include genre, culture, moods, instruments, rhythm, and tempo annotations. Third, a hybrid measure which combines the above-mentioned distance measures with two existing low-level measures: a Euclidean distance based on principal component analysis of timbral, temporal, and tonal descriptors, and a timbral distance based on single Gaussian MFCC modeling. We evaluate our proposed measures against a number of baseline measures. We do this objectively based on a comprehensive set of music collections, and subjectively based on listeners' ratings. Results show that the proposed methods achieve accuracies comparable to the baseline approaches in the case of the tempo and classifier-based measures. The highest accuracies are obtained by the hybrid distance. Furthermore, the proposed classifier-based approach opens up the possibility to explore distance measures that are based on semantic notions.**

***Index Terms*—Music, Information retrieval, Distance measurement, Knowledge acquisition, Multimedia databases, Multimedia computing**

I. Introduction

RAPID development of digital technologies, the Internet, and the multimedia industry have provoked a huge excess of information. An increasingly growing amount of multimedia data complicates search, retrieval, and recommendation of relevant information. For example, in the digital music industry, major Internet stores such as the iTunes Store contain up to 14 million songs[1], adding thousands of new songs every month. In such circumstances, fast and efficient retrieval approaches operating on large-scale multimedia databases are necessary [1]. Specifically, similarity search is a challenging scientific problem, which helps to facilitate advances in multimedia knowledge, organization, and recommendation. Therefore, it can serve the user's needs and satisfaction within educative, explorative, social, and entertainment multimedia applications.

Studying the ways to search and recommend music to a user is a central task within the music information retrieval (MIR) community [2]. From a simplistic point of view, this task can be regarded as obtaining a suitable distance[2] measurement between a query song and a set of potential candidates. This way, one maps these songs to a certain feature space where a dissimilarity measure can be computed. Currently, researchers and practitioners fill in this feature space with information extracted from the audio content[3], context, or both. Contextual information, in the form of user ratings [3] and social tags [4], is a powerful source for measuring music similarity. However, it becomes problematic to obtain such data in a long-tail [5]. General lack of user ratings and social tags for unpopular multimedia items complicate their sufficient characterization, as multimedia consumption is biased towards popular items. Alternatively, information extracted from the audio content can help to overcome this problem [6].

The present work deals with content-based approaches to music similarity. We organize this paper into three parts, dealing with the state-of-the-art, the proposal of two simple distance measurements, and the proposal of a hybrid (non-simple) distance measurement, respectively.

In the first part (Sec. II), we review related state of the art, including current approaches to music similarity (Sec. II-A) and low-level audio descriptors available to our research (Sec. II-B). Furthermore, we briefly explain a number of existing simple approaches, which we use as a baseline for evaluating our proposed methods. Throughout the paper, we assume simple approaches to be those which are not constituted by a number of distances[4]. More concretely, as baseline approaches we consider Euclidean distances defined on sets of timbral, rhythmic, and tonal descriptors (Secs. II-C1 and II-C2) and Kullback-Leibler divergence defined on Gaussian mixture models (GMMs) of Mel-frequency cepstral coefficients (MFCCs, Sec. II-C3).

In the second part, which we partially presented in [7], we compare the aforementioned baseline approaches against two novel distance measures (Sec. III). The first idea we explore consists of the use of tempo-related musical aspects. We pro-

D. Bogdanov, J. Serrà, N. Wack, P. Herrera, and X. Serra are with the Music Technology Group, Universitat Pompeu Fabra, Roc Boronat, 138, 08018 Barcelona, Spain, tel: +34 935422000, +34 935422164, fax: +34 935422517, email: {dmitry.bogdanov, joan.serraj, nicolas.wack, perfecto.herrera, xavier.serra}@upf.edu.

Manuscript received Month 00, 0000; revised Month 00, 0000.

[1] http://en.wikipedia.org/wiki/ITunes_Store

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

[3] We pragmatically use the term "content" to refer to any information extracted from the audio signal.

[4] We have opted for the term "simple" instead of other appropriate terms, such as "non-hybrid" and "homogeneous".

pose a distance based on two low-level rhythmic descriptors, namely beats per minute and onset rate (Sec. III-A). The second idea we explore shifts the problem to a more high-level (semantic) domain as we propose to use high-level semantic dimensions, including information about genre and musical culture, moods and instruments, and rhythm and tempo. With regard to this aspect, we continue the research of [8]–[10] but, more in the line of [10], we investigate the possibility of benefiting from results obtained in different classification tasks and transferring this acquired knowledge to the context of music similarity (Sec. III-B). More specifically, as our first main technical contribution, we infer different groups of musical dimensions by using support vector machines. and use a high-level modular distance which combines these dimensions. Among the qualities of this classifier-based distance we strive for high modularity, being able to easily append additional dimensions. Moreover, we strive for descriptiveness, being able to explain similarity to a user.

We evaluate all the considered simple approaches with a uniform methodological basis, including an objective evaluation on several comprehensive ground truth music collections (Sec. IV-A) and a subjective evaluation based on ratings given by real listeners (Sec. IV-C). We show that, in spite of being conceptually different, the proposed methods achieve comparable or even higher accuracies than the considered baseline approaches (Secs. IV-B and IV-D). Finally, we illustrate the benefits of the proposed classifier-based distance for music similarity justification to a user (Sec. V). In addition, we demonstrate an example of possible semantic explanation of similarity between songs.

In the third part, we explore the possibility of creating a hybrid approach, based on the considered simple approaches as potential components. As our second main technical contribution, we propose a new distance measure that combines a low-level Euclidean distance based on principal component analysis (PCA), a timbral distance based on single Gaussian MFCC modeling, and our proposed tempo-based and semantic classifier-based distances (Sec. VI). These choices are motivated by the results obtained in the subjective evaluation of simple approaches performed in the second part of the paper. We hypothesize that such combination of conceptually different approaches, covering timbral, rhythmic, and semantic aspects of music similarity, is more appropriate from the point of view of music cognition [11] and, thus, it could lead to a better performance from the point of view of the listener. Indeed, a number of works support this idea though being limited by combining only timbral and rhythmic aspects into a hybrid distance [12]–[17], and, alternatively, timbral and tonal, or timbral and semantic ones [18]. To the best of the authors' knowledge, more extended combinations of timbral, rhythmic, tonal and semantic dimensions, providing a single hybrid distance, have not yet been studied.

We evaluate the hybrid approach against its component approaches objectively, performing a cross-collection out-of-sample test on two large-scale music collections (Sec. VII-A), and subjectively, based on ratings of 21 real listeners (Sec. VII-C). We find that the proposed hybrid method reaches a better performance than all considered approaches, both objectively (Sec. VII-B) and subjectively (Sec. VII-D). We subjectively evaluate our classifier-based and hybrid approaches against a number of state-of-the-art distance measures within the bounds of an international evaluation framework (Sec. VIII-A). Notably, our hybrid approach is found to be one of the best performing participants (Sec. VIII-B). We finally state general conclusions and discuss the possibility of further improvements (Sec. IX).

II. Scientific background

A. Music similarity

Focusing on audio content-based similarity, there exist a wide variety of approaches for providing a distance measurement between songs. These approaches comprise both the selection of audio descriptors and the choice of an appropriate distance function. Representing the songs as points in a feature space with an L_p metric is a straightforward approach. Cano et al. [19] demonstrate such an approach using a Euclidean metric after a PCA transformation of a preliminary selected combination of timbral, temporal, and tonal descriptors. Similarly, Slaney et al. [20] apply a Euclidean metric on loudness and temporal descriptors, and use a number of algorithms to improve performance. These algorithms include whitening transformation, linear discriminant analysis (LDA), relevant component analysis (RCA) [21], neighbourhood components analysis, and large-margin nearest neighbour classification [22].

As well, specific timbral representations exist, the most prominent one being modeling the songs as clouds of vectors of MFCCs, calculated on a frame basis. Logan and Salomon, [23] represent such clouds as cluster models, comparing them with the Earth mover's distance. Mandel and Ellis [24] compare means and covariances of MFCCs applying the Mahalanobis distance. Furthermore, GMMs can be used to represent the clouds as probability distributions, and then these distributions can be compared by the symmetrized Kullback-Leibler divergence. However, in practice, approximations are required for the case of several Gaussian components in a mixture. To this end, Aucouturier et al. [25], [26] compare GMMs by means of Monte Carlo sampling. In contrast, Mandel and Ellis [24] and Flexer et al. [27] simplify the models to single Gaussian representations, for which a closed form of the Kullback-Leibler divergence exists. Pampalk [13] gives a global overview of these approaches. As well, Jensen et al. [28] provide an evaluation of different GMM configurations. Besides MFCCs, more descriptors can be used for timbral distance measurement. For example, Li and Ogihara [29] apply a Euclidean metric on a set of descriptors, including Daubechies wavelet coefficient histograms.

Temporal (or rhythmic) representation of the songs is another important aspect. A number of works propose specific temporal distances in combination with timbral ones. For example, Pampalk et al. [12], [13] exploit fluctuation patterns, which describe spectral fluctuations over time, together with several derivative descriptors, modeling overall tempo and fluctuation information at specific frequencies. A hybrid distance is then defined as a linear combination of a Euclidean distance on fluctuation patterns together with a timbral distance, based on GMMs of MFCCs. Pohle et al. [14] follow this idea, but propose using a cosine similarity distance for fluctuation patterns together with a specific distance measure related to cosine similarity for GMMs of MFCCs. Furthermore, they propose an alternative temporal descriptor set, including a modification of fluctuation patterns (onset patterns and onset coefficients), and additional timbral descriptors (spectral contrast coefficients, harmonicness, and attackness) along with MFCCs for single Gaussian modeling [15], [16]. Song and Zhang [17] present a hybrid distance measure, combining a timbral Earth mover's distance on MFCC cluster models, a timbral Euclidean distance on spectrum histograms, and a temporal Euclidean distance on fluctuation patterns.

Finally, some attempts to exploit tonal representation of songs exist. Ellis and Poliner [30], Marolt [31] and Serrà et al. [32], present specific melodic and tonality distance measurements, not addressed to the task of music similarity, but to version (cover) identification. In principle, their approaches are based on matching sequences of pitch class profiles, or chroma feature vectors, representing the pitch class distributions (including the melody) for different songs.

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 semantic explanation [33]. In contrast, research on computing high-level semantic features from low-level audio descriptors exists. In particular, in the context of MIR classification problems, genre classification [34], mood detection [35], [36], and artist identification [24] have gathered much research attention.

Starting from the relative success of this research, we hypothesize that the combination of classification problem outputs can be a relevant step to overcome the so-called semantic gap [33] between human judgements and low-level machine learning inferences, specifically in the case of content-based music similarity. A number of works support this hypothesis. Berenzweig et al. [9] propose to infer high-level semantic dimensions, such as genres and "canonical" artists, from low-level timbral descriptors, such as MFCCs, by means of neural networks. The inference is done on a frame basis, and the resulting clouds in high-level feature space are compared by centroids with a Euclidean distance. Barrington et al. [8] train GMMs of MFCCs for a number of semantic concepts, such as genres, moods, instrumentation, vocals, and rhythm. Thereafter, high-level descriptors can be obtained by computing the probabilities of each concept on a frame basis. The resulting semantic clouds of songs can be represented by GMMs, and compared with Kullback-Leibler divergence. McFree and Lanckriet [18] propose a hybrid low-dimensional feature transformation embedding musical artists into Euclidean space subject to a partial order, based on a set of manually annotated artist similarity triplets, over pairwise low-level and semantic distances. As such, the authors consider low-level timbral distance, based on MFCCs, tonal distance, based on chroma descriptors, and the above-mentioned semantic distance [8]. The evaluation includes the embeddings, which merge timbral and tonal distances, and, alternatively, timbral and semantic distances. West and Lamere [10] apply classifiers to infer semantic features of the songs. In their experiment, Mel-frequency spectral irregularities are used as an input for a genre classifier. The output class probabilities form a new high-level feature space, and are compared with a Euclidean distance. The authors propose to use classification and regression trees or LDA for classification.

In spite of having a variety of potential content-based approaches to music similarity, still there exist certain open issues. The distances, operating solely on low-level audio descriptors, lack semantic explanation of similarity on a level which human judgements operate. The majority of approaches, both low-level and high-level, focus mostly on timbral descriptors, whereas other types of low-level descriptors, such as temporal and tonal, are potentially useful as well. Furthermore, comparative evaluations are necessary, especially those carried out comprehensively and uniformly on large music collections. In existing research, there is a lack of such comparative evaluations, taking into consideration different approaches. Objective evaluation criteria of music similarity are generally reduced to co-occurrences of genre, album, and artist labels, being tested on relatively small ground truth collections. In turn, subjective evaluations with human raters are not common. We will focus on filling in these open issues, employing comprehensive music collections, objective criteria for similarity, and human listeners for subjective evaluations. As existing approaches still perform relatively poorly, we hypothesize that better performance may be achieved by combining conceptually different distance measurements, which will help to jointly exploit different aspects of music similarity.

B. Musical descriptors

In the present work, we characterize each song using an in-house audio analysis tool[5]. From this tool we use 59 descriptor classes in total, characterizing global properties of songs, and covering timbral, temporal, and tonal aspects of musical audio. The majority of these descriptors are extracted on a frame-by-frame basis with a 46 ms frame size, and 23 ms hop size, and then summarized by their means and variances across these frames. In the case of multidimensional descriptors, covariances between components are also considered (e.g. with MFCCs). Since it is not the objective of this paper to review existing methods for descriptor extraction, we just provide a brief overview of the classes we use in Table I. The interested reader is referred to the cited literature for further details.

C. Baseline simple approaches

In this work, we consider a number of conceptually different simple approaches to music similarity. Among them we indicate several baselines, which will be used in objective and subjective evaluations, and moreover will be regarded as potential components of the hybrid approach.

[5] http://mtg.upf.edu/technologies/essentia

TABLE I
OVERVIEW OF MUSICAL DESCRIPTORS.

Descriptor group	Descriptor class
Timbral	Bark bands [35], [37]
	MFCCs [13], [35], [37], [38]
	Pitch [39], pitch centroid [40]
	Spectral centroid, spread, kurtosis, rolloff, decrease, skewness [35], [37], [41]
	High-frequency content [39], [41]
	Spectral complexity [35]
	Spectral crest, flatness, flux [37], [41]
	Spectral energy, energy bands, strong peak, tristimulus [41]
	Inharmonicity, odd to even harmonic energy ratio [37]
Rhythmic	BPM, onset rate [35], [39], [41]
	Beats loudness, beats loudness bass [40]
Tonal	Transposed and untransposed harmonic pitch class profiles, key strength [35], [42]
	Tuning frequency [42]
	Dissonance [35], [43]
	Chord change rate [35]
	Chords histogram, equal tempered deviations, non-tempered/tempered energy ratio, diatonic strength [40]
Miscellaneous	Average loudness [37]
	Zero-crossing rate [13], [37]

1) Euclidean distance based on principal component analysis (L_2-PCA): As a starting point, we follow the ideas proposed by Cano et al. [19], and apply an unweighted Euclidean metric on a manually selected subset of the descriptors outlined above[6]. This subset includes bark bands, pitch, spectral centroid, spread, kurtosis, rolloff, decrease, skewness, high-frequency content, spectral complexity, spectral crest, flatness, flux, spectral energy, energy bands, strong peak, tristimulus, inharmonicity, odd to even harmonic energy ratio, beats loudness, beats loudness bass, untransposed harmonic pitch class profiles, key strength, average loudness, and zero-crossing rate.

Preliminary steps include descriptor normalization in the interval $[0, 1]$ and principal component analysis (PCA) [44] to reduce the dimension of the descriptor space to 25 variables. The choice of the number of target variables is conditioned by a trade-off between target descriptiveness and the curse of high-dimensionality [45]–[47], typical for L_p metrics, and is supported by research work on dimension reduction for music similarity [48]. Nevertheless, through our PCA dimensionality reduction, an average of 78% of the information variance was preserved on our music collections, reducing the number of 201 native descriptors by a factor of 8.

2) Euclidean distance based on relevant component analysis (L_2-RCA-1 and L_2-RCA-2): Along with the previous measure, we consider more possibilities of descriptor selection. In particular, we perform relevant component analysis (RCA) [21]. Similar to 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. Having such training data, the transformation reduces irrelevant variability in the data while amplifying relevant variability. As in the L_2-PCA approach, the output dimensionality is chosen to be 25. We consider both the descriptor subset used in L_2-PCA and the full descriptor set of Table I (L_2-RCA-1 and L_2-RCA-2, respectively).

3) Kullback-Leibler divergence based on GMM of MFCCs (1G-MFCC): Alternatively, we consider timbre modeling with GMM as another baseline approach [26]. We implement the simplification of this timbre model using single Gaussian with full covariance matrix [24], [27], [49]. Comparative research of timbre distance measures using GMMs indicates that such simplification can be used without significantly decreasing performance while being computationally less complex [13], [28]. 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,

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

where μ_X and μ_Y are MFCC means, Σ_X and Σ_Y are MFCC covariance matrices, and N_{MFCC} is the dimensionality of the MFCCs. This dimensionality can vary from 10 to 20 [28], [35], [50]. To preserve robustness against different audio encodings, the first 13 MFCC coefficients are taken [51].

III. PROPOSED SIMPLE APPROACHES

Concerning simple approaches to music similarity, here we propose two novel distance measures that are conceptually different than what has been reviewed. We regard both approaches as potential components of the hybrid approach.

A. Tempo-based distance (TEMPO)

The first approach we propose is related to the exploitation of tempo-related musical aspects with a simple distance

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

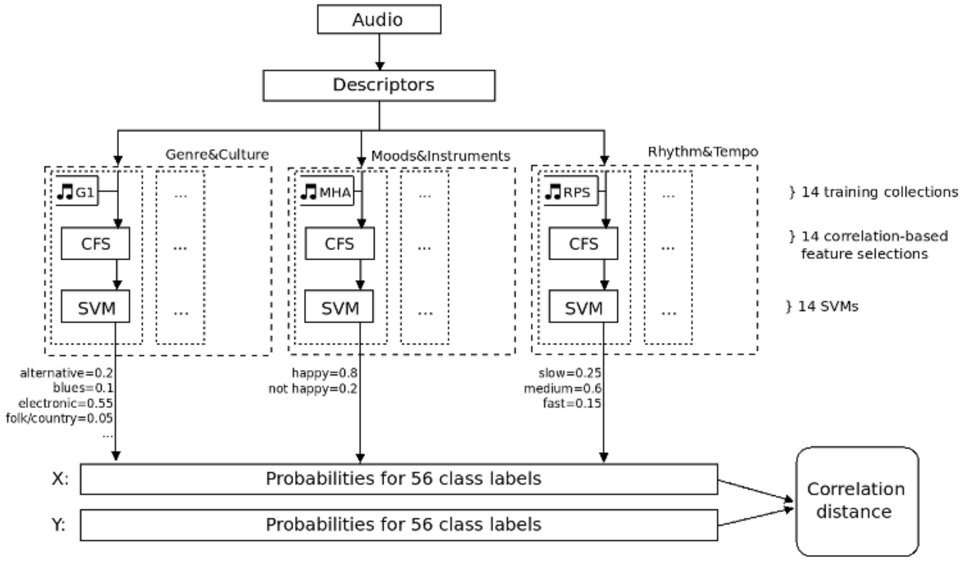


Fig. 1. General schema of CLAS distance. Given two songs X and Y, low-level audio descriptors are extracted, a number of SVM classifications are run based on ground truth music collections, and high-level representations, containing probabilities of classes for each classifier, are obtained. A distance between X and Y is calculated with correlation distances such as Pearson correlation distance.

measure. This measure is based on two descriptors, beats per minute (BPM) and onset rate (OR), the latter representing the number of onsets per second. These descriptors are fundamental for the temporal description of music. Among different implementations, we opted for BPM and OR estimation algorithms presented in [39].

For two songs X and Y with BPMs X_{BPM} and Y_{BPM}, and ORs X_{OR} and Y_{OR}, respectively, we determine a distance measure by a linear combination of two separate distance functions,

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

defined for BPM as

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

and for OR as

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

where $X_{\text{BPM}}, Y_{\text{BPM}}, X_{\text{OR}}, Y_{\text{OR}} > 0$, $\alpha_{\text{BPM}}, \alpha_{\text{OR}} \geq 1$. 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 multiples of the BPM (OR), e.g. $X_{\text{BPM}} = iY_{\text{BPM}}$, are more similar than songs with non-multiple BPMs (ORs). For example, the songs X and Y with $X_{\text{BPM}} = 140$ and $Y_{\text{BPM}} = 70$ should have a closer distance than the songs X and Z with $Z_{\text{BPM}} = 100$. Our assumption is motivated by research on the perceptual effects of double or half tempo [52]. The strength of this assumption depends on the parameter α_{BPM} (α_{OR}). Moreover, such a distance can be helpful in relation to the common problem of tempo duplication (or halving) in automated tempo estimation [53], [54]. In the case of $\alpha_{\text{BPM}} = 1$, all multiple BPMs are treated equally, while in the case of $\alpha_{\text{BPM}} > 1$, preference inversely decreases with i. In practice we use $i = 1, 2, 4, 6$.

Eqs. 2, 3, and 4 formulate the proposed distance in the general case. In a parameter-tuning phase we performed a grid search with one of the ground truth music collections (RBL) under the objective evaluation criterion described in Sec. IV-A. Using this collection, which is focused on rhythmic aspects and contains songs with various rhythmic patterns, we found $w_{\text{BPM}} = w_{\text{OR}} = 0.5$ and $\alpha_{\text{BPM}} = \alpha_{\text{OR}} = 30$ to be the most plausible parameter configuration. Such values reveal the fact that in reality both components are equally meaningful and that mainly a one-to-one relation of BPMs (ORs) is relevant for for the music collection and descriptors we used to evaluate such rhythmic similarity. When our BPM (OR) estimator has increased duplicity errors (e.g. a BPM of 80 was estimated as 160), we should expect lower α values.

B. Classifier-based distance (CLAS)

The second approach we propose derives a distance measure from diverse classification tasks. In contrast to 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. A schema of the approach is presented in Fig. 1

For the first step we choose standard multi-class support vector machines (SVMs) [44], which are shown to be an effective tool for different classification tasks in MIR [24], [35], [36], [55], [56]. We apply SVMs to infer different groups of musical dimensions such as (i) genre and musical culture, (ii) moods and instruments, and (iii) rhythm and tempo. To this end, 14 classification tasks are run according to all available ground truth collections presented in Table II. More concretely, we train one SVM per each ground truth collection, providing its annotated songs as a training input. For each collection and the corresponding SVM, a preliminary correlation-based feature selection (CFS) [44] over all available $[0, 1]$-normalized descriptors (Sec. II-B) is performed to optimize the descriptor selection for this particular classification task. As an output, the classifier provides probability values of classes on which it was trained. For example, a classifier using the G1 collection is trained on an optimized descriptor space, according to the collection's classes and the CFS process, and returns genre probabilities for the labels "alternative", "blues", "electronic", "folk/country", etc. Altogether, the classification results form a high-level descriptor space, which contains the probability values of each class for each SVM. Based on results in [35], we decided to use the libSVM[7] implementation with the C-SVC method and a radial basis function kernel with default parameters.

For the second step, namely defining a distance operating on a formed high-level semantic space (i.e. the one of the label probabilities), we consider different measures frequently used in collaborative filtering systems. Among the standard ones, we select the cosine distance (CLAS-Cos), Pearson correlation distance (CLAS-Pears) [5], [57], and Spearman's rho correlation distance (CLAS-Spear) [58]. Moreover, we consider a number of more sophisticated measures. In particular, the adjusted cosine distance (CLAS-Cos-A) [5], [57] is computed by taking into account the average probability for each class, i.e. compensating distinction between classifiers with different numbers of classes. Weighted cosine distance (CLAS-Cos-W) [59] and weighted Pearson correlation distance (CLAS-Pears-W) [60] are both weighted manually (W_M) and also based on classification accuracy (W_A). For W_M, we split the collections into 3 groups of musical dimensions, namely genre and musical culture, moods and instruments, and rhythm and tempo. We empirically assign weights 0.50, 0.30, and 0.20 respectively. Our choice is supported by research on the effect of genre in terms of music perception [11], [61] and the fact that genre is the most common aspect of similarity used to evaluate distance measures in the MIR community [12]. For W_A, we evaluate the accuracy of each classifier, and assign proportional weights which sum to 1.

With this setup, the problem of content-based music similarity can be seen as a collaborative filtering problem of item-to-item similarity [57]. Such a problem can generally be solved by calculating a correlation distance between rows of a song/user rating matrix with the underlying idea that similar items should have similar ratings by certain users. Transferring this idea to our context, we can state that similar songs should have similar probabilities of certain classifier labels. To this extent, we compute song similarity on a song/user rating matrix 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.

IV. Evaluation of simple approaches

We evaluated all considered approaches with a uniform 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.

A. Objective evaluation methodology

In our evaluations 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 both full songs and excerpts) were employed for that purpose, and are presented in Table II. For some dimensions we used existing collections in the MIR field [34], [36], [55], [62]–[64], while for other dimensions we created manually labeled in-house collections. For each collection, we considered songs from the same class to be similar and songs from different classes to be dissimilar, and assessed the relevance of the songs' rankings returned by each approach.

To assess the relevance of the songs' rankings, we used the mean average precision (MAP) measure [65]. The MAP is a standard information retrieval measure used in the evaluation of many query-by-example tasks. For each approach and music collection, MAP was computed from the corresponding full distance matrix. The average precision (AP) [65] was computed for each matrix row (for each song query) and the mean was calculated across queries (columns).

For consistency, we applied the same procedure to each of the considered distances, whether they required training or not: the results for RAND, L_2-PCA, L_2-RCA-1, L_2-RCA-2, 1G-MFCC, TEMPO, and CLAS-based distances were averaged over 5 iterations of 3-fold cross-validation. On each iteration, all 17 ground truth collections were split into training and testing sets. For each testing set, the CLAS-based distances were provided with 14 out of 17 training sets. The G3, ART, and ALB collections were not included as training sets due to the insufficient size of their class samples. In contrast, for each testing set, L_2-RCA-1, and L_2-RCA-2 were provided with a single complementary training set belonging to the same collection.

B. Objective evaluation results

The average MAP results are presented in Fig. 2 and Table III. Additionally, the approaches with statistically non-significant difference in MAP performance according to the independent two-sample t-tests are presented in Table IV. These t-tests were conducted to separately compare the performances for each music collection. In the cases that are not reported in Table IV, we found statistically significant differences in MAP performance ($p < 0.05$).

We first see that all considered distances outperform the random baseline (RAND) for most of the music collections. When comparing baseline approaches (L_2-PCA, L_2-RCA-1, L_2-RCA-2, 1G-MFCC), we find 1G-MFCC to perform best on average. Still, L_2-PCA performs similarly (MHA, MSA, MRE, and MEL) or slightly better for some collections (MAC and RPS). With respect to tempo-related collections, TEMPO

[7] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

TABLE II
GROUND TRUTH MUSIC COLLECTIONS EMPLOYED FOR OBJECTIVE EVALUATION OF THE SIMPLE APPROACHES. ALL PRESENTED COLLECTIONS ARE USED FOR TRAINING CLAS-BASED DISTANCES, EXCEPT G3, ART, AND ALB COLLECTIONS DUE TO INSUFFICIENT SIZE OF THEIR CLASS SAMPLES.

Acronym	Category	Classes (musical dimensions)	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	[62]
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	[63]
G4	Genre & Culture	Blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, rock	993 song excerpts, 100 per genre	[34]
CUL	Genre & Culture	Western, non-western	1640 song excerpts, 1132/508 per class	[55]
MHA	Moods & Instruments	Happy, non-happy	302 full songs + excerpts, 139/163 per class	[36] + in-house
MSA	Moods & Instruments	Sad, non-sad	230 full songs + excerpts, 96/134 per class	[36] + in-house
MAG	Moods & Instruments	Aggressive, non-aggressive	280 full songs + excerpts, 133/147 per class	[36] + in-house
MRE	Moods & Instruments	Relaxed, non-relaxed	446 full songs + excerpts, 145/301 per class	[36] + 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	[36] + in-house
MEL	Moods & Instruments	Electronic, non-electronic	332 full songs + excerpts, 164/168 per class	[36] + 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	[64]

TABLE III
OBJECTIVE EVALUATION RESULTS (MAP) OF THE SIMPLE APPROACHES FOR THE DIFFERENT MUSIC COLLECTIONS CONSIDERED. N.C. STANDS FOR "NOT COMPUTED" DUE TO TECHNICAL DIFFICULTIES. FOR EACH COLLECTION, THE MAPS OF THE APPROACHES, WHICH PERFORM BEST WITHOUT STATISTICALLY SIGNIFICANT DIFFERENCE BETWEEN THEM, ARE MARKED IN BOLD.

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

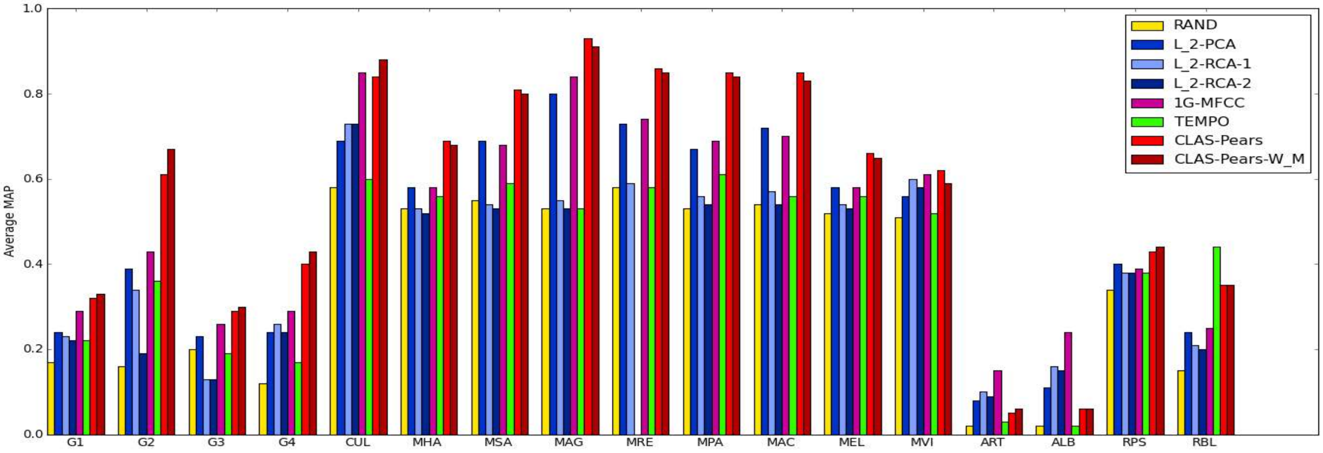


Fig. 2. Objective evaluation results (MAP) of the simple approaches for the different music collections considered.

performs similarly (RPS) or significantly better (RBL) than baseline approaches. Indeed, 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. In contrast, no statistically significant difference was found in comparison with the random baseline for the G3, MAG, MRE, and ALB collections. Finally, we saw that classifier-based distances achieved the best accuracies for the majority of the collections. Since all CLAS-based distances (CLAS-Cos, CLAS-Pears, CLAS-Spear, CLAS-Cos-W, CLAS-Pears-W, CLAS-Cos-A) showed comparable accuracies, we only report two examples (CLAS-Pears, CLAS-Pears-W$_M$). In particular, CLAS-based distances achieved large accuracy improvements with the G2, G4, MPA, MSA, and MAC collections. In contrast, no improvement was achieved with the ART, ALB, and RBL collections. The distance 1G-MFCC performed best for the ART and ALB collections. We hypothesize that the success of 1G-MFCC for the ART and ALB collections might be due to the well known "album effect" [24]. This effect implies that, due to production process, songs from the same album share much more timbral characteristics than songs from different albums of the same artist, and, moreover, different artists.

TABLE IV

THE APPROACHES WITH STATISTICALLY NON-SIGNIFICANT DIFFERENCE IN MAP PERFORMANCE ACCORDING TO THE INDEPENDENT TWO-SAMPLE T-TESTS. THE L_2-RCA-2 APPROACH WAS EXCLUDED FROM THE ANALYSIS DUE TO TECHNICAL DIFFICULTIES.

Collection	Compared approaches	P-value
G3	RAND, TEMPO	0.40
MHA	RAND, L_2-RCA-1	1.00
	L_2-PCA, 1G-MFCC	1.00
	CLAS-Pears, CLAS-Pears-W$_M$	0.37
MSA	L_2-PCA, 1G-MFCC	0.37
	CLAS-Pears, CLAS-Pears-W$_M$	0.50
MAG	RAND, TEMPO	1.00
MRE	RAND, TEMPO	0.33
	L_2-PCA, 1G-MFCC	0.09
	CLAS-Pears, CLAS-Pears-W$_M$	0.37
MPA	CLAS-Pears, CLAS-Pears-W$_M$	0.50
MAC	CLAS-Pears, CLAS-Pears-W$_M$	0.08
MEL	L_2-PCA, 1G-MFCC	1.00
	CLAS-Pears, CLAS-Pears-W$_M$	0.37
ALB	RAND, TEMPO	0.33
	CLAS-Pears, CLAS-Pears-W$_M$	0.33
RPS	L_2-RCA-1, TEMPO	1.00

C. *Subjective evaluation methodology*

In the light of the results of the objective evaluation (Sec. IV-B), we selected 4 conceptually different approaches (L_2-PCA, 1G-MFCC, TEMPO, and CLAS-Pears-W$_M$) together with the random baseline (RAND) for the listeners' subjective evaluation. We designed a web-based survey where registered listeners performed a number of iterations blindly voting for the considered distance measures, assessing the quality of how each distance reflects perceived music similarity. In particular, we evaluated the resulting sets of most similar songs produced by the selected approaches, hereafter referred as "playlists". Such a scenario is a popular way to assess the quality of music similarity measures [3], [6]. It increases discrimination between approaches in comparison with a pairwise song-to-song evaluation. Moreover, it reflects the common applied context of music similarity measurement, which consists of playlist generation.

During each iteration, the listener was presented with 5 different playlists (one for each measure) generated from the same seed song (Fig. 3). 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). No playlist contained more than one song from the same artist.

Independently for each playlist, we asked the listeners to provide (i) a playlist similarity rating and (ii) a playlist inconsistency boolean answer. For playlist similarity ratings we used a 6-point Likert-type scale (0 corresponding to the lowest similarity, 5 to the highest) to evaluate the appropriateness of the playlist with respect to the seed. Likert-type scales [66] are bipolar scales used as tools-of-the-trade in many disciplines to capture subjective information, such as opinions, agreements, or disagreements with respect to a given issue or question. The two opposing positions occupy the extreme ends of the scale (in our case, low-high similarity of the playlist to the seed), and several ratings are allocated for intermediate positions. We explicitly avoided a "neutral" point in order to increase the discrimination between positive and negative opinions. We did not present examples of playlist inconsistency but they might comprise of speech mixed with music, extremely different tempos, completely opposite feelings or emotions, distant musical genres, etc.

We divided the test into two phases: in the first, 12 seeds and corresponding playlists were shared between all listeners; in the second one the seeds for each listener (up to a maximum of 21) were randomly selected. Listeners were never informed of this distinction. Additionally, we asked each listener about his musical background, which included musicianship and listening expertise information (each measured in 3 levels). Altogether we collected playlist similarity ratings, playlist inconsistency indicators, and background information from 12 listeners[8].

D. *Subjective evaluation results*

In any experimental situation such as our subjective evaluation, analysis of variance (ANOVA) is the usual methodology employed to assess the effects of one variable (like the similarity computation approach) on another one (such as the similarity rating obtained from listeners). ANOVA provides a statistical test of whether or not the means of several groups (in our case, the ratings obtained using a specific similarity

[8] Due to confidential reasons, the survey was conducted on a limited closed set of participants, and was unavailable to general public.

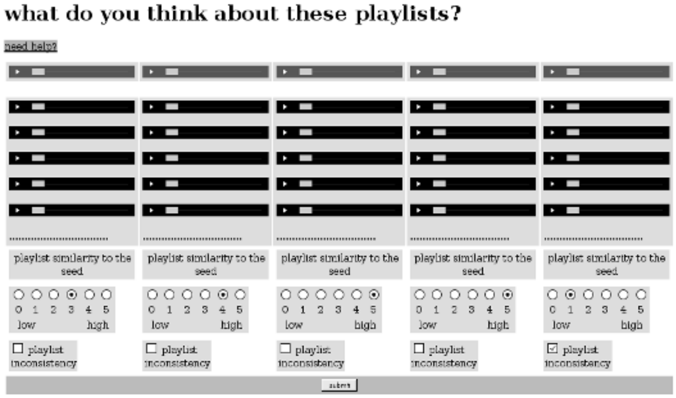


Fig. 3. A screenshot of the subjective evaluation web-based survey.

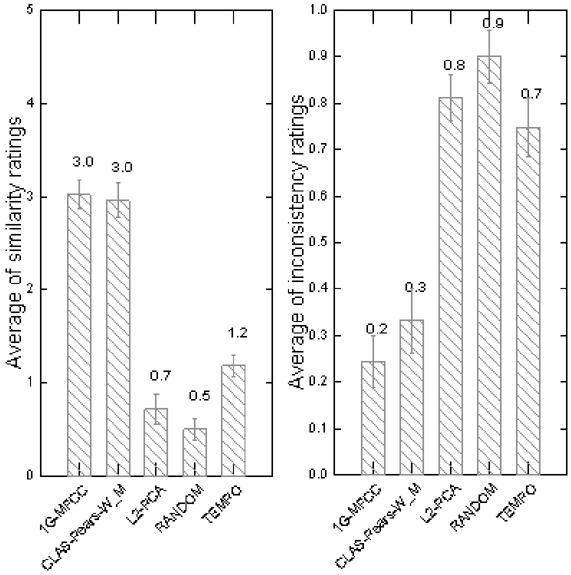


Fig. 4. Average playlist similarity rating and proportion of inconsistent playlists for the subjective evaluation of the simple approaches. Error bars indicate 1 Standard Error of the Mean.

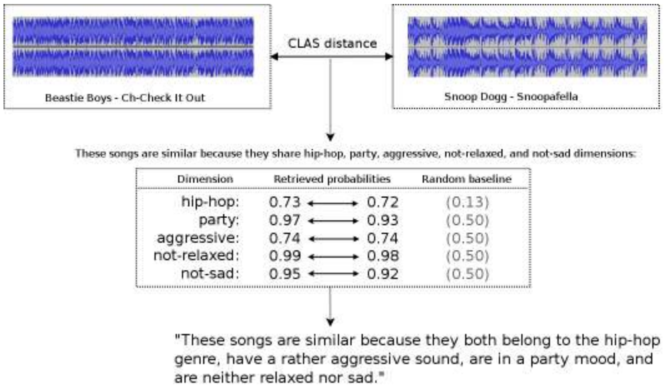
computation approach) are equal. In addition to the effect of the different similarity computation methods, in our evaluation we wanted to know the possible effect of the musicianship and listening experience of the participants. Furthermore, we also wanted to know the effect produced by the two consecutive testing phases used (one presenting the same songs to all the listeners, and the other using different songs for each of them). Therefore a mixed-design ANOVA with two between-subjects factors (musicianship and listening expertise) and two within-subjects factors (similarity computation approach and testing phase) was required. Results from this analysis revealed that the effect of the similarity computation method on the similarity ratings was statistically significant (Wilks Lambda $=$ 0.005, $F(4,2) = 93.943$, $p < 0.05$) and that it separated the methods in 3 different groups: RANDOM and L_2-PCA (which yielded the lowest similarity ratings) versus TEMPO versus 1G-MFCC and CLAS-Pears-W$_M$ (which yielded the highest similarity ratings). The same pattern was obtained for the effects on the inconsistency ratings. The effect of the testing phase, also found to be significant, reveals that ratings yielded slightly lower values in the second phase. This could be due to the "tuning" of the similarity ratings experienced by each subject as the experiment proceeded. Fortunately, the impact of phase was uniform and did not depend on or interact with any other factor. Hence, the similarity ratings are only made "finer" or more "selective" as the experiment progresses, but irrespective of the similarity computation approach. On the other hand, the potential effects of musicianship and listening expertise revealed no impact on the similarity ratings. Overall, we conclude that the L_2-PCA and TEMPO distances, along with a random baseline, revealed poor performance, tending to provide disruptive examples of playlist inconsistency. Contrastingly, CLAS-Pears-W$_M$ and 1G-MFCC revealed acceptable performance with slightly positive user satisfaction. We have omitted for clarity the specific results of the statistical tests which validated our concluding statements.

V. Semantic explanation of music similarity

Here we give some thoughts concerning the proposed CLAS distance and its semantic application. An interesting aspect of this proposed approach is the ability to provide a user of the final system with a concrete motivation for the retrieved songs starting from a purely audio content-based analysis. To the best of the authors' knowledge, this aspect is very rare among other music content-processing systems [67]. However, there is evidence that retrieval or recommendation results perceived as transparent (getting an explanation of why a particular retrieval or recommendation was made) are preferred by users, increasing there confidence in a system [68].

Remarkably, the proposed classifier-based distance gives the possibility of providing high-level semantic descriptions for the similarity between a pair of songs along with the distance value itself. In a final system, such annotations can be presented in terms of probability values of the considered dimensions that can be understood by a user. Alternatively, automatic text generation can be employed to present the songs' qualities in a textual way. For a brief justification of similarity, a subset of dimensions with the highest impact on overall similarity can be selected. A simple use-case example is shown in Fig 5. For a pair of songs and the CLAS-Pears-W$_M$ distance measure, a subset of 15 dimensions was determined iteratively by greedy distance minimization. In each step the best candidate for elimination was selected from different dimensions, and its weight was zeroed. Thereafter, the residual dimension probabilities that exceeded corresponding random baselines[9] can be presented to a user. Notice however that as random baselines differ for different dimensions depending on the number of output classes of the corresponding classifier, the significance of dimension probabilities cannot be treated equally. For example, the 0.40 probability of a dimension regressed by an 8-class classifier is considerably more significant than the 0.125 random baseline. Though not presented, the dimensions with probabilities below random baselines also have an impact on the distance measurement. Still, such

[9] Under the assumptions of the normal distribution of each classifier's labels for a music collection.



These songs are similar because they share hip-hop, party, aggressive, not-relaxed, and not-sad dimensions:

Dimension	Retrieved probabilities	Random baseline
hip-hop:	0.73 ⟷ 0.72	(0.13)
party:	0.97 ⟷ 0.93	(0.50)
aggressive:	0.74 ⟷ 0.74	(0.50)
not-relaxed:	0.99 ⟷ 0.98	(0.50)
not-sad:	0.95 ⟷ 0.92	(0.50)

"These songs are similar because they both belong to the hip-hop genre, have a rather aggressive sound, are in a party mood, and are neither relaxed nor sad."

Fig. 5. A real example of a semantic explanation of the similarity between two songs retrieved from our music collection for the classifier-based distance.

negative statements (in the sense of a low probability of a regressed dimension) are probably less suitable than positive ones for justification of music similarity to a user.

VI. Proposed hybrid approach (HYBRID)

Finally, we hypothesize that an important performance gain can be achieved by combining conceptually different approaches, covering timbral, rhythmic, and semantic aspects of music similarity. We propose a hybrid distance measure, consisting of a subset of the simple measures described above. We define the distance as a weighted linear combination of L_2-PCA, 1G-MFCC, TEMPO, and CLAS-Pears-W$_M$ distances. We select these 4 conceptually different approaches relying on the results of the objective evaluation of potential components (Sec. IV-B). For each selected component, we apply score normalization, following ideas in [69], [70]. More concretely, each original distance variable d_i is equalized to a new variable $\overline{d_i} = E_i(d_i)$, uniformly distributed in $[0, 1]$. The equalizing function E_i is given by the cumulative distribution function of d_i, which can be obtained from a distance matrix on a given representative music collection. As such, we use an aggregate collection of 16K full songs and music excerpts, composed from the ground truth collections previously used for objective evaluation of simple approaches (Table II). The final hybrid distance is obtained by a weighted linear combination of component distances. The weights are based on the results of the subjective evaluation (Sec. IV-D) and are set as follows: 0.7 for L_2-PCA, 3.0 for 1G-MFCC, 1.2 for TEMPO, and 3.0 for CLAS-Pears-W$_M$ distances. Hence, for each component a weight corresponds to an average playlist similarity rating given by listeners.

VII. Evaluation of hybrid approach

A. Objective evaluation methodology

Here we followed a different evaluation strategy than with the simple approaches. This strategy comes from the fact that the ground truth music collections available to our evaluation, both in-house and public, can have different biases (due to different collection creators, music availability, audio formats, covered musical dimensions, how the collection was formed, etc.). Therefore, in order to minimize these effects, we carried out a large-scale cross-collection evaluation of the hybrid approach against its component approaches, namely L_2-PCA, 1G-MFCC, TEMPO, and CLAS-Pears-W$_M$, together with the random baseline (RAND). Cross-collection comparison implies that the queries and their answers belong to different music collections (out-of-sample results), thus making evaluation results more robust to possible biases.

Solely the genre musical dimension was covered in this experiment. Two large in-house ground truth music collections were employed for that purpose: (i) a collection of 299K music excerpts (30 sec.) (G-C1), and (ii) a collection of 73K full songs (G-C2). Both collections had a genre label associated with every song. In total, 218 genres and subgenres were covered. The size of these music collections is considerably large, which makes evaluation conditions closer to a real world scenario. As queries, we randomly selected songs from the 10 most common genres from both collections G-C1 and G-C2. The distribution of the selected genres among the collections is presented in Table V. More concretely, for each genre, 790 songs from collection G-C1 were randomly selected as queries. The number of queries per genre corresponds to a minimum number of genre occurrences among the selected genres.

Each query was applied to the collection G-C2, forming a full row in a distance matrix. As with the objective evaluation of simple approaches (Sec. IV-A), MAP was used as an evaluation measure, but was calculated with a cutoff (similarly to pooling techniques in text retrieval [71]–[73]) equal to the 10 closest matches due to the large dimensionality of the resulting distance matrix. The evaluation results were averaged over 5 iterations. In the same manner, a reverse experiment was carried out, using songs from the G-C2 collection as queries, and applied to the collection G-C1. As the evaluation was completely out-of-sample, the full ground truth collections were used to train the CLAS approach.

B. Objective evaluation results

The results are presented in Table VI. In addition, we analyzed the obtained MAPs with a series of independent two-sample t-tests. All the approaches were found to perform with statistically significant difference ($p < 0.001$).

We see that all considered distances outperform the random baseline (RAND). We found 1G-MFCC and CLAS-Pears-W$_M$ to have comparable performance, being the best among the simple approaches. As well, the TEMPO distance was found to perform similarly or slightly better than L_2-PCA. Overall, the results for simple approaches conform with our previous objective evaluation. Meanwhile, our proposed HYBRID distance achieved the best accuracy in the cross-collection evaluation in both directions.

C. Subjective evaluation methodology

We repeated the listening experiment, conducted for simple approaches (Sec. IV-C) to evaluate the hybrid approach against its component approaches. The same music collection of 300K music excerpts (30 sec.) by 60K artists (5 songs/artist) was used for that purpose. Each listener was presented with a series of 24 iterations, which, according to the separation of

TABLE V
NUMBER OF OCCURRENCES OF 10 MOST FREQUENT GENRES, COMMON FOR COLLECTIONS G-C1 AND G-C2.

Genre	G-C1	G-C2
Reggae	2991	790
New Age	4294	1034
Blues	6229	2397
Country	8388	1699
Folk	10367	1774
Pop	15796	4523
Electronic	16050	4038
Jazz	22227	5440
Classical	43761	4802
Rock	49369	11486

TABLE VI
OBJECTIVE CROSS-COLLECTION EVALUATION RESULTS (MAP WITH CUTOFF AT 10) AVERAGED OVER 5 ITERATIONS.

Distance	G-C1 → G-C2	G-C2 → G-C1
RANDOM	0.07	0.08
L_2-PCA	0.09	0.11
1G-MFCC	0.23	0.22
TEMPO	0.11	0.12
CLAS-Pears-W$_M$	0.21	0.23
HYBRID	**0.25**	**0.28**

the experiment into two phases, included 12 iterations with seeds and corresponding playlists shared between all listeners, and 12 iterations with randomly selected seeds, different for each listener. In total, we collected playlist similarity ratings, playlist inconsistency indicators, and background information about musicianship and listening expertise from 21 listeners.

D. *Subjective evaluation results*

An ANOVA with two between-subjects factors (musicianship and listening expertise) and two within-subjects factors (similarity computation approach and testing phase) was used to test their effects on the similarity ratings and on the inconsistency ratings given by the listeners (Fig. 6). The only clearly significant factor explaining the observed variance in the similarity ratings was the similarity computation approach (Wilks lambda $= 0.43$, $F(4, 11) = 9.158$, $p < 0.005$). The specific pattern of significant differences between the tested computation approaches makes the HYBRID metric to clearly stand out from the rest, while L_2-PCA and TEMPO score low (but without statistical differences between them), and CLAS-Pears-W$_M$ and 1G-MFCC (again without statistically significant differences between them) score between the two extremes. As we did not find any significant effect of musicianship and listening expertise on the similarity ratings, it seems clear that the differences in similarity ratings can be attributed only to the differences in the similarity computation approaches.

The same pattern and meaning was also found for the inconsistency ratings: they were dependent on the similarity computation approach, and most of them were generated by the L_2-PCA and TEMPO methods, whereas the HYBRID method provided significantly lower inconsistency ratings. No other factor or interaction between factors was found to be statistically significant, but a marginal interaction effect of similarity computation approach and testing phase was found. This effect means that some similarity computation methods (but not all) lowered the ratings as the evaluation progressed. The same pattern was obtained for the inconsistency ratings. In conclusion, we found a similarity computation method (HYBRID) that was clearly preferred over the rest and no effect other than the computation method was responsible for that preference.

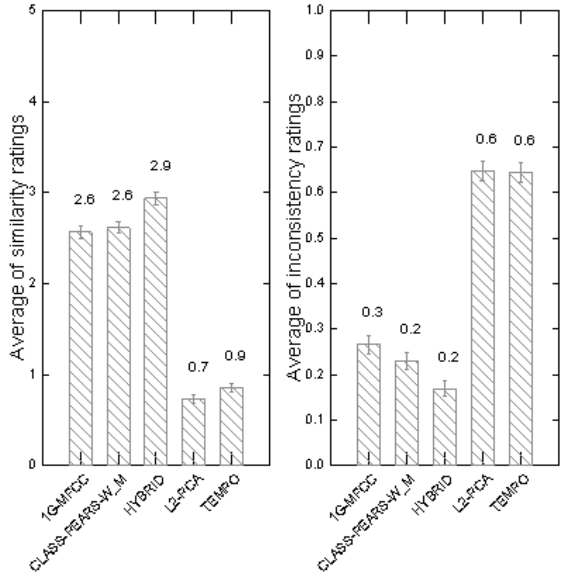


Fig. 6. Average playlist similarity rating and proportion of inconsistent playlists for the subjective evaluation of the hybrid approach. Error bars indicate 1 Standard Error of the Mean.

VIII. MIREX 2009 EVALUATION

A. *Methodology*

We submitted the HYBRID and CLAS-Pears-W$_M$ systems to the Music Information Retrieval Evaluation eXchange (MIREX). MIREX is an international community-based framework for the formal evaluation of MIR systems and algorithms [74], [75]. Among other tasks, MIREX allows for the comparison of different algorithms for artist identification, genre classification, or music transcription. In particular, MIREX allows for a subjective human assessment of the accuracy of different approaches to music similarity by community members, this being a central task within the framework. For that purpose, participants can submit their algorithms as binary executables and the MIREX organizers determine and publish the algorithms' accuracies and runtimes. The underlying music collections are never published or disclosed to the participants, neither before or after the contest. Therefore, participants cannot tune their algorithms to the music collections used in the evaluation process.

In the MIREX'2009 edition, the evaluation of each submitted approach was performed on a music collection of

7000 songs (30 sec. excerpts), which were chosen from IMIRSEL's[10] collections [75] and pertained to 10 different genres. For each participant's approach, a 7000×7000 distance matrix was calculated. A query set of 100 songs was randomly selected from the music collection, representing each of the 10 genres (10 songs per genre). For each query and participant approach, the 5 nearest-to-the-query songs out of the 7000 were chosen as candidates (after filtering out the query itself and all songs of the same artist). All candidates were evaluated by human graders using the Evalutron 6000 grading system [76]. For each query, a single grader was assigned to evaluate the derived candidates from all approaches. Thereby, the uniformity of scoring within each query was ensured. For each query/candidate pair, a grader provided (i) a categorical broad score in the set $\{0, 1, 2\}$ (corresponding to "not similar", "somewhat similar", and "very similar" categories), and (ii) a fine score in the range from 0 (failure) to 10 (perfection). The listening experiment was conducted with 50 graders, and each one of them evaluated 2 queries. As this evaluation was completely out-of-sample, our submitted systems were trained on the full ground truth collections required for the CLAS distance.

B. Results

The overall evaluation results are reproduced in Table VII[11]. Our measures are noted as BSWH1 for CLAS-Pears-W$_M$, and BSWH2 for HYBRID. The results of the Friedman test against the summary data of fine scores are presented in Fig. 7. First, and most importantly, we found the HYBRID measure to be one of the best performing distances in the MIREX 2009 audio music similarity task. HYBRID was very close to PS1, but worse than the leading PS2 distance [15]. However, no statistically significant difference between PS2, PS1 and our HYBRID measure was found in the Friedman test. Second, the CLAS-Pears-W$_M$ measure revealed satisfactory average performance comparing to other distances with no statistically significant difference to the majority of the participant approaches. Nevertheless, CLAS-Pears-W$_M$ outperformed a large group of poor performing distances with a statistically significant difference. Finally, we state that despite the fact that we do not observe examples of stable excellent performance among all participant distances, up to above-average user satisfaction was achieved by the majority of the approaches, including our HYBRID and CLAS-Pears-W$_M$ distances.

IX. Conclusions

In the current work we presented, studied, and comprehensively evaluated, both objectively and subjectively, new and existing content-based distance measures for music similarity. We studied a number of simple approaches, each of which apply a uniform distance measure for overall similarity. We considered 5 baseline distances, including a random one. We explored the potential of two new conceptually different distances not strictly operating on the often exclusively used musical timbre aspects. More concretely, we presented a simple tempo-based distance which can be especially useful for expressing music similarity in collections where rhythm aspects are predominant. Using only two low-level temporal descriptors, BPM and OR, this distance is computationally inexpensive, yet effective for such collections. As well, our subjective evaluation experiments revealed a slight preference by listeners of tempo-based distance over a generic euclidean distance.

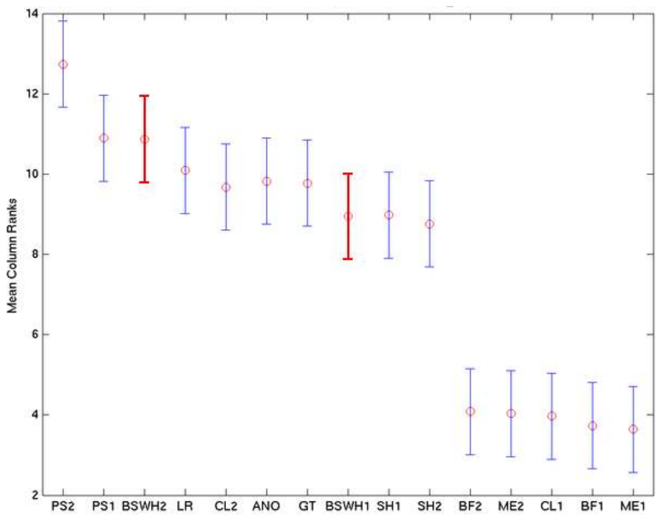


Fig. 7. MIREX 2009 Friedman's test (fine scores). Figure obtained from the official results webpage for MIREX'2009.

In addition, we investigated the possibility of benefiting from the results of classification problems and transferring this gained knowledge to the context of music similarity. To this end, we presented a classifier-based distance which makes use of high-level semantic descriptors inferred from low-level ones. This distance covers diverse groups of musical dimensions such as genre and musical culture, moods and instruments, and rhythm and tempo. The classifier-based distance outperformed all the considered simple approaches in most of the ground truth music collections used for objective evaluation. Contrastingly, this performance improvement was not seen in the subjective evaluation when compared with the best performing baseline distance considered. However, they were found to perform at the same level and, therefore, no statistically significant differences were found between them. In general, the classifier-based distance represents a semantically rich approach to music similarity. Thus, in spite of being based solely on audio content information, this approach can overcome the so-called "semantic gap" in content-based music similarity and provide a semantic explanation to justify the retrieval results to a user.

We explored the possibility of creating a hybrid approach, based on the studied simple approaches as potential components. We presented a new distance measure, which combines a low-level Euclidean distance based on principal component analysis (PCA), a timbral distance based on single Gaussian MFCC modeling, our tempo-based distance, and a high-level

[10] http://www.music-ir.org/evaluation/

[11] Detailed results can be found on the official results webpage for MIREX'2009: http://www.music-ir.org/mirex/2009/index.php/Audio_Music_Similarity_and_Retrieval_Results

TABLE VII
MIREX 2009 OVERALL SUMMARY RESULTS SORTED BY AVERAGE FINE SCORE. THE PROPOSED APPROACHES CLAS AND HYBRID ARE HIGHLIGHTED IN GRAY (BSWH1 AND BSWH2, RESPECTIVELY).

Acronym	Authors (measure)	Average fine score	Average broad score
PS2	Tim Pohle, Dominik Schnitzer (2009)	6.458	1.448
PS1	Tim Pohle, Dominik Schnitzer (2007)	5.751	1.262
BSWH2	Dmitry Bogdanov, Joan Serrà, Nicolas Wack, and Perfecto Herrera (HYBRID)	5.734	1.232
LR	Thomas Lidy, Andreas Rauber	5.470	1.148
CL2	Chuan Cao, Ming Li	5.392	1.164
ANO	Anonymous	5.391	1.126
GT	George Tzanetakis	5.343	1.126
BSWH1	Dmitry Bogdanov, Joan Serrà, Nicolas Wack, and Perfecto Herrera (CLAS-Pears-W$_M$)	5.137	1.094
SH1	Stephan Hübler	5.042	1.012
SH2	Stephan Hübler	4.932	1.040
BF2	Benjamin Fields (mfcc10)	2.587	0.410
ME2	François Maillet, Douglas Eck (sda)	2.585	0.418
CL1	Chuan Cao, Ming Li	2.525	0.476
BF1	Benjamin Fields (chr12)	2.401	0.416
ME1	François Maillet, Douglas Eck (mlp)	2.331	0.356

semantic classifier-based distance. This distance outperformed all previously considered approaches in an objective large-scale cross-collection evaluation, and revealed the best performance for listeners in a subjective evaluation. Moreover, we participated in a subjective evaluation against a number of state-of-the-art distance measures, within the bounds of the MIREX'2009 audio music similarity and retrieval task. The results revealed high performance of our hybrid measure, with no statistically significant difference from the best performing method submitted. In general, the hybrid distance represents a combinative approach, benefiting from timbral, rhythmic, and high-level semantic aspects of music similarity.

Further research will be devoted to improving the classifier-based distance with the addition of classifiers dealing with 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 particular, contextual dimensions, in the form of user ratings or social tags, can be added to make possible a fusion with collaborative filtering approaches. As well, to improve the classifier-based distance itself, we will consider a better combination of classifiers' output probabilities. Additionally, an enhancement of the tempo-based distance component of the proposed hybrid approach is possible by using a richer representation for rhythm, such as the fluctuation patterns.

ACKNOWLEDGMENTS

The authors would like to thank Jordi Funollet for technical support, Owen Meyers for technical support and proofreading, and all participants of the subjective evaluation. This research has been partially funded by the FI Grant of Generalitat de Catalunya (AGAUR), the Music 3.0 project of the Spanish Ministry of Industry, Tourism, and Trade (Avanza Contenidos, TSI-070100-2008-318), and the Buscamedia project (CEN-20091026).

REFERENCES

[1] G. Lu, "Techniques and data structures for efficient multimedia retrieval based on similarity," *IEEE Transactions on Multimedia*, vol. 4, no. 3, p. 372–384, 2002.

[2] 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*, vol. 96, no. 4, pp. 668–696, 2008.

[3] M. Slaney and W. White, "Similarity based on rating data," in *International Symposium on Music Information Retrieval (ISMIR'07)*, 2007.

[4] M. Levy and M. Sandler, "Music information retrieval using social tags and audio," *IEEE Transactions on Multimedia*, vol. 11, no. 3, pp. 383–395, 2009.

[5] O. Celma, "Music recommendation and discovery in the long tail," Ph.D. dissertation, UPF, Barcelona, Spain, 2008.

[6] L. Barrington, R. Oda, and G. Lanckriet, "Smarter than genius? human evaluation of music recommender systems," in *International Society for Music Information Retrieval Conference (ISMIR'09)*, 2009, pp. 357–362.

[7] D. Bogdanov, J. Serrà, N. Wack, and P. Herrera, "From low-level to high-level: Comparative study of music similarity measures," in *IEEE International Symposium on Multimedia (ISM'09). International Workshop on Advances in Music Information Research (AdMIRe'09)*, 2009, pp. 453–458.

[8] L. Barrington, D. Turnbull, D. Torres, and G. Lanckriet, "Semantic similarity for music retrieval," in *Music Information Retrieval Evaluation Exchange (MIREX'07)*, 2007, http://www.music-ir.org/mirex/abstracts/2007/AS_barrington.pdf.

[9] 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)*, vol. 1, 2003, pp. 29–32.

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

[11] G. C. Cupchik, M. Rickert, and J. Mendelson, "Similarity and preference judgments of musical stimuli," *Scandinavian Journal of Psychology*, vol. 23, no. 4, pp. 273–282, 1982.

[12] 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)*, 2005, pp. 628–633.

[13] E. Pampalk, "Computational models of music similarity and their application in music information retrieval," Ph.D. dissertation, Vienna University of Technology, Mar. 2006.

[14] 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.

[15] ——, "Submission to MIREX 2009 audio similarity task," in *Music Information Retrieval Evaluation Exchange (MIREX'09)*, 2009, http://music-ir.org/mirex/2009/results/abs/PS.pdf.

[16] T. Pohle, D. Schnitzer, M. Schedl, P. Knees, and G. Widmer, "On rhythm and general music similarity," in *International Society for Music Information Retrieval Conference (ISMIR'09)*, 2009, pp. 525–530.

[17] Y. Song and C. Zhang, "Content-Based information fusion for Semi-Supervised music genre classification," *IEEE Transactions on Multimedia*, vol. 10, no. 1, pp. 145–152, 2008.

[18] B. McFee and G. Lanckriet, "Heterogeneous embedding for subjective artist similarity," in *International Conference on Music Information Retrieval (ISMIR'09)*, 2009.

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

[20] M. Slaney, K. Weinberger, and W. White, "Learning a metric for music similarity," in *International Symposium on Music Information Retrieval (ISMIR'08)*, 2008, pp. 313–318.

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

[22] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *The Journal of Machine Learning Research*, vol. 10, p. 207–244, 2009.

[23] B. Logan and A. Salomon, "A music similarity function based on signal analysis," in *IEEE International Conference on Multimedia and Expo (ICME'01)*, 2001, p. 190.

[24] 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)*, 2005, pp. 594–599.

[25] J. J. Aucouturier and F. Pachet, "Music similarity measures: What's the use," in *Proceedings of the ISMIR*, 2002, p. 157–163.

[26] 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*, vol. 7, no. 6, pp. 1028–1035, 2005.

[27] 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)*, 2008, pp. 173–178.

[28] 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*, vol. 17, pp. 693–703, 2009.

[29] T. Li and M. Ogihara, "Toward intelligent music information retrieval," *IEEE Transactions on Multimedia*, vol. 8, no. 3, pp. 564–574, 2006.

[30] D. P. Ellis and G. E. Poliner, "Identifying 'cover songs' with chroma features and dynamic programming beat tracking," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'07)*, 2007, pp. IV–1429–1432.

[31] M. Marolt, "A Mid-Level representation for Melody-Based retrieval in audio collections," *IEEE Transactions on Multimedia*, vol. 10, no. 8, pp. 1617–1625, 2008.

[32] J. Serrà, X. Serra, and R. G. Andrzejak, "Cross recurrence quantification for cover song identification," *New Journal of Physics*, vol. 11, no. 9, p. 093017, 2009.

[33] 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.

[34] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on speech and audio processing*, vol. 10, no. 5, pp. 293–302, 2002.

[35] C. Laurier, O. Meyers, J. Serrà, M. Blech, P. Herrera, and X. Serra, "Indexing music by mood: design and integration of an automatic content-based annotator," *Multimedia Tools and Applications*, 2009.

[36] 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.

[37] 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/.

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

[39] P. M. Brossier, "Automatic annotation of musical audio for interactive applications," Ph.D. dissertation, QMUL, London, UK, 2007.

[40] E. Gómez, P. Herrera, P. Cano, J. Janer, J. Serrà, J. Bonada, S. El-Hajj, T. Aussenac, and G. Holmberg, "Music similarity systems and methods using descriptors," patent. WO 2009/001202, published December 31, 2008.

[41] 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," Ph.D. dissertation, UPF, Barcelona, Spain, 2005.

[42] E. Gómez, "Tonal description of music audio signals," Ph.D. dissertation, UPF, Barcelona, Spain, 2006.

[43] W. A. Sethares, *Tuning, timbre, spectrum, scale*. Springer Verlag, 2005.

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

[45] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is "Nearest neighbor" meaningful?" in *Database Theory — ICDT'99*, 1999, pp. 217–235.

[46] F. Korn, B. Pagel, and C. Faloutsos, "On the 'dimensionality curse' and the 'self-similarity blessing'," *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, no. 1, pp. 96–111, 2001.

[47] C. C. Aggarwal, "On k-anonymity and the curse of dimensionality," in *International Conference on Very Large Data Bases (VLDB'05)*. Trondheim, Norway: VLDB Endowment, 2005, pp. 901–909.

[48] N. Wack, P. Cano, B. de Jong, and R. Marxer, "A comparative study of dimensionality reduction methods: The case of music similarity," 2006.

[49] 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)*, 2006, pp. 66–75.

[50] E. Pampalk, S. Dixon, and G. Widmer, "On the evaluation of perceptual similarity measures for music," in *6th International Conference on Digital Audio Effects (DAFx'03)*, London, UK, 2003, p. 7–12.

[51] S. Sigurdsson, K. B. Petersen, and T. Lehn-Schiøler, "Mel frequency cepstral coefficients: An evaluation of robustness of mp3 encoded music," in *International Conference on Music Information Retrieval (ISMIR'07)*, 2006, p. 286–289.

[52] M. F. McKinney and D. Moelants, "Ambiguity in tempo perception: What draws listeners to different metrical levels?" *Music Perception*, vol. 24, no. 2, pp. 155–166, 2006.

[53] F. Gouyon, A. Klapuri, S. Dixon, M. Alonso, G. Tzanetakis, C. Uhle, and P. Cano, "An experimental comparison of audio tempo induction algorithms," *IEEE Transactions on Speech and Audio Processing*, vol. 14, no. 5, p. 1832–1844, 2006.

[54] L. M. Smith, "Beat critic: Beat tracking octave error identification by metrical profile analysis," in *International Society for Music Information Retrieval Conference (ISMIR'10)*, 2010.

[55] E. Gómez and P. Herrera, "Comparative analysis of music recordings from western and Non-Western traditions by automatic tonal feature extraction," *Empirical Musicology Review*, vol. 3, no. 3, pp. 140–156, 2008.

[56] 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)*, 2003, pp. 429–432.

[57] B. Sarwar, G. Karypis, J. Konstan, and J. Reidl, "Item-based collaborative filtering recommendation algorithms," in *International Conference on World Wide Web (WWW'01)*, 2001, pp. 285–295.

[58] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl, "Evaluating collaborative filtering recommender systems," *ACM Transactions on Information Systems*, vol. 22, no. 1, pp. 5–53, 2004.

[59] A. Cripps, C. Pettey, and N. Nguyen, "Improving the performance of FLN by using similarity measures and evolutionary algorithms," in *IEEE International Conference on Fuzzy Systems*, 2006, p. 323–330.

[60] M. B. Abdullah, "On a robust correlation coefficient," *Journal of the Royal Statistical Society. Series D (The Statistician)*, vol. 39, no. 4, pp. 455–460, 1990.

[61] A. Novello, M. F. McKinney, and A. Kohlrausch, "Perceptual evaluation of music similarity," in *International Conference on Music Information Retrieval (ISMIR'06)*, 2006.

[62] 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)*, 2005, pp. 528–531.

[63] 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*, vol. 84, pp. 1236–1256, 2003.

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

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

[66] W. E. Saris and I. N. Gallhofer, *Design, evaluation, and analysis of questionnaires for survey research*. Wiley-Interscience, Jul. 2007.

[67] F. Maillet, D. Eck, G. Desjardins, and P. Lamere, "Steerable playlist generation by learning song similarity from radio station playlists," in *International Conference on Music Information Retrieval (ISMIR'09)*, 2009.

[68] R. Sinha and K. Swearingen, "The role of transparency in recommender systems," in *CHI'02 extended abstracts on Human Factors in Computing Systems*, 2002, p. 831.

[69] M. Fernández, D. Vallet, and P. Castells, "Probabilistic score normalization for rank aggregation," in *Advances in Information Retrieval*, 2006, pp. 553–556.

[70] M. Arevalillo-Herráez, J. Domingo, and F. J. Ferri, "Combining similarity measures in content-based image retrieval," *Pattern Recognition Letters*, vol. 29, no. 16, pp. 2174–2181, Dec. 2008.

[71] A. Turpin and F. Scholer, "User performance versus precision measures for simple search tasks," in *International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'06)*, 2006, p. 11–18.

[72] W. B. Croft, D. Metzler, and T. Strohman, *Search engines: information retrieval in practice*. Addison-Wesley, 2010.

[73] F. Radlinski and N. Craswell, "Comparing the sensitivity of information retrieval metrics," in *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, 2010, p. 667–674.

[74] J. S. Downie, "The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research," *Acoustical Science and Technology*, vol. 29, no. 4, p. 247–255, 2008.

[75] J. Downie, A. Ehmann, M. Bay, and M. Jones, "The music information retrieval evaluation eXchange: some observations and insights," in *Advances in Music Information Retrieval*, 2010, pp. 93–115.

[76] A. A. Gruzd, J. S. Downie, M. C. Jones, and J. H. Lee, "Evalutron 6000: collecting music relevance judgments," in *ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'07)*. ACM, 2007, pp. 507–507.

Dmitry Bogdanov received a degree in Applied Mathematics and Informatics at the Lomonosov Moscow State University, Moscow, Russia, in 2006. There he participated in several research projects devoted to network security and intrusion detection. In 2007 he became a member of the Music Technology Group at the Universitat Pompeu Fabra (UPF), Barcelona, Spain, as a researcher, and at the present time he is a PhD candidate. His current research interests include music information retrieval, music similarity and recommendation, music content description and classification, user modeling, and preference elicitation.

Joan Serrà obtained both the degrees of Telecommunications and Electronics at Enginyeria La Salle, Universitat Ramón Llull, Barcelona, Spain, in 2002 and 2004, respectively. After working from 2005 to 2006 at the research and development department of Music Intelligence Solutions Inc, he joined the Music Technology Group of Universitat Pompeu Fabra (UPF), Barcelona, where he received the MSc in Information, Communication and Audiovisual Media Technologies in 2007. He is currently a PhD candidate with the Music Technology Group of the UPF. He is also a part-time associate professor with the Dept. of Information and Communication Technologies of the same university. In 2010 he was a guest scientist with the Research Group on Nonlinear Dynamics and Time Series Analysis of the Max Planck Institute for the Physics of Complex Systems, Dresden, Germany. His research interests include music retrieval and understanding, signal processing, time series analysis, complex networks, complex systems, information retrieval and music perception and cognition.

Nicolas Wack received his Telecommunication Engineer degree from Telecom ParisTech, Paris, France, in 2003. Since then, he has been involved with the Music Technology Group (UPF) in various European projects dealing with sound and music classification and similarity. He spearheaded the development of the Essentia and Gaia technologies, which respectively extract audio characteristics from sounds/songs and perform similarity queries on them. His main interests are in software design, efficient audio analysis and working with very large databases.

Perfecto Herrera received a Degree in Psychology from the University of Barcelona, Barcelona, Spain, in 1987. He was with the University of Barcelona as a Software Developer and as Assistant Professor. His further studies focused on sound engineering, audio postproduction, and computer music. He has been working with the Music Technology Group, (UPF) since its inception in 1996, first as the person responsible for the sound laboratory/studio, then as a Researcher. Now he is finishing his PhD on Music Content Description in the Universitat Pompeu Fabra (UPF), Barcelona. He worked in the MPEG-7 standardization initiative from 1999 to 2001. Then, he collaborated in the EU-IST-funded CUIDADO project, contributing to the research and development of tools for indexing and retrieving music and sound collections. This work continued and was expanded as Scientific Coordinator for the Semantic Interaction with Music Audio Contents (SIMAC) project, again funded by the EU-IST. He is currently the Head of the Department of Sonology, Higher Music School of Catalonia (ESMUC), where he teaches music technology and psychoacoustics. His main research interests are music content analysis, description and classification, and music perception and cognition.

Xavier Serra is Associate Professor of the Department of Information and Communication Technologies and Director of the Music Technology Group at the Universitat Pompeu Fabra in Barcelona, Spain. After a multidisciplinary academic education he obtained a PhD in Computer Music from Stanford University in 1989 with a dissertation on the spectral processing of musical sounds that is considered a key reference in the field. His research interests cover the understanding, modeling and generation of musical signals by computational means, with a balance between basic and applied research and approaches from both scientific/technological and humanistic/artistic disciplines. Dr. Serra is very active in promoting initiatives in the field of Sound and Music Computing at the local and international levels, being editor and reviewer of a number of journals, conferences and research programs of the European Commission, and also giving lectures on current and future challenges of the field. He is the principal investigator of more than 15 major research projects funded by public and private institutions, the author of 31 patents and of more than 75 research publications.