EURASIP Journal on
Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

Semantic structures of timbre emerging from social and acoustic descriptions of music

Rafael Ferrer* and Tuomas Eerola

Abstract

The perceptual attributes of timbre have inspired a considerable amount of multidisciplinary research, but because of the complexity of the phenomena, the approach has traditionally been confined to laboratory conditions, much to the detriment of its ecological validity. In this study, we present a purely bottom-up approach for mapping the concepts that emerge from sound qualities. A social media (http://www.last.fm) is used to obtain a wide sample of verbal descriptions of music (in the form of tags) that go beyond the commonly studied concept of genre, and from this the underlying semantic structure of this sample is extracted. The structure that is thereby obtained is then evaluated through a careful investigation of the acoustic features that characterize it. The results outline the degree to which such structures in music (connected to affects, instrumentation and performance characteristics) have particular timbral characteristics. Samples representing these semantic structures were then submitted to a similarity rating experiment to validate the findings. The outcome of this experiment strengthened the discovered links between the semantic structures and their perceived timbral qualities. The findings of both the computational and behavioural parts of the experiment imply that it is therefore possible to derive useful and meaningful structures from free verbal descriptions of music, that transcend musical genres, and that such descriptions can be linked to a set of acoustic features. This approach not only provides insights into the definition of timbre from an ecological perspective, but could also be implemented to develop applications in music information research that organize music collections according to both semantic and sound qualities.

Keywords: timbre, natural language processing, vector-based semantic analysis, music information retrieval, social media

1 Introduction

In this study, we have taken a purely bottom-up approach for mapping sound qualities to the conceptual meanings that emerge. We have used a social media (http://www.last.fm) for obtaining as wide a sample of music as possible, together with the free verbal descriptions made of music in this sample, to determine an underlying semantic structure. We then empirically evaluated the validity of the structure obtained, by investigating the acoustic features that corresponded to the semantic categories that had emerged. This was done through an experiment where participants were asked to rate the perceived similarity between acoustic examples of prototypical semantic categories. In this way, we were attempting to recover the correspondences between semantic and acoustic features that are ecologically relevant in the perceptual domain. This aim also meant that the study was designed to be more exploratory than confirmative. We applied the appropriate and recommended techniques for clustering, acoustic feature extraction and comparisons of similarities; but this was only after assessing the alternatives. But, the main focus of this study has been to demonstrate the elusive link that exists between the semantic, perceptual and physical properties of timbre.

1.1 The perception of timbre

Even short bursts of sound are enough to evoke mental imagery, memories and emotions, and thus provoke immediate reactions, such as the sensation of pleasure or fear. Attempts to craft a bridge between such acoustic features and the subjective sensations they provoke [1] have usually started with describing instrument

* Correspondence: rafael.ferrer-flores@jyu.fi
Finnish Centre of Excellence in Interdisciplinary Music Research, University of Jyväskylä, Jyväskylä, Finland

sounds via adjectives on a bipolar scale (e.g. bright-dark, static-dynamic) and matching these with more precise acoustic descriptors (such as the *envelope* shape, or high-frequency energy content) [2,3]. However, it has been difficult to compare these studies when such different patterns between acoustic features and listeners' evaluations have emerged [4]. These differences may be attributed to the cross-study variations in context effects, as well as the choice of terms, stimuli and rating scales used. It has also been challenging to link the findings of such studies to the context of actual music [5], when one considers that real music consists of a complex combination of sound. A promising approach has been obtained to evaluate short excerpts of recorded music with a combination of bipolar scales and acoustic analysis [6]. However, even this approach may well omit certain sounds and concepts that are important for the majority of people, since the music and scales have usually been chosen by the researcher, not the listeners.

1.2 Social tagging

Social tagging is a way of labelling items of interest, such as songs, images or links as a part of the normal use of popular online services, so that the tags then become a form of categorization in themselves. Tags are usually semantic representations of abstract concepts created essentially for mnemonic purposes and used typically to organize items [7,8]. Within the theory of *information foraging* [9], tagging behaviour is one example of a transition from internalized to externalized forms of knowledge where, using *transactional memory*, people no longer have to know everything, but can use other people's knowledge [10]. What is most evident in the social context is that what escapes one individual's perception can be captured by another, thus transforming tags into memory or knowledge cues for the undisclosed transaction [11].

Social tags are usually thought to have an underlying *ontology* [12] defined simply by people interested in the matter, but with no institutional or uniform direction. These characteristics make the vocabulary and implicit relations among the terms considerably richer and more complex than in formal taxonomies where a hierarchical structure and set of rules are designed *apriori* (cf. folksonomy versus taxonomy in [13]). When comparing ontologies based on social tagging and the classification by experts, it is presumed that there is an underlying organization of musical knowledge hidden among the tags. But, as raised by Celma and Serra [1]), this should perhaps not to be taken for granted. For this reason, Section 2 addresses the uncovering of an ontology from the tags [14] in an unsupervised form, to investigate whether such an ontology is not an imposed construction. Because a latent structure has been assumed, we use a technique called *vector-based semantic analysis*, which is a generalization of *Latent Semantic Analysis* [15] and similar to the methods used in latent semantic mapping [16] and latent perceptual indexing [17]. Thus, although some of the terminology is borrowed from these areas, our method is also different in several crucial respects. While ours is designed to explore emergent structures in the semantic space (i.e. clusters of musical descriptions), the other methods are designed primarily to improve information retrieval by reducing the dimensionality of the space [18]. In our method, the reduction is not part of the analytical step, but rather implemented as a pre-filtering stage (see Appendix sections A.1 and A.2). The indexing of documents (songs in our case) is also treated separately in Section 2.2 which presents our solution based on the Euclidean distances of clusters profiles in a vector space. The reasons outlined above show that tags, and the structures that can be derived from them, impart crucial cues about how people organize and make sense of their experiences, which in this case is music and in particular its timbre.

2 Emergent structure of timbre from social tags

To find a semantic structure for timbre analysis based on social tags, a sample of music and its associated tags were taken. The tags were then filtered, first in terms of their statistical relevancy and then according to their semantic categories. This filtering left us with five such categories, namely *adjectives*, *nouns*, *instruments*, *temporal references* and *verbs* (see Appendix A for a detailed explanation of the filtering process). Finally, the relations between different combinations of tags were analysed by means of distance calculations and hybrid clustering.

The initial database of music consisted of a collection of 6372 songs [19], from a total of 15 musical genres (with approximately 400 examples for each genre), namely, *Alternative*, *Blues*, *Classical*, *Electronic*, *Folk*, *Gospel*, *Heavy*, *Hip-Hop*, *Iskelmä*, *Jazz*, *Pop*, *Rock*, *Soul*, *Soundtrack* and *World*. Except for some songs in the *Iskelmä* and *World* genres (which were taken from another corpus of music), all of the songs that were eventually chosen in November 2008 from each of these genres could already be found on the musical social network (http://www.last.fm), and they were usually among the "top tracks" for each genre (i.e. the most played songs tagged with that genre on the Internet radio). Although larger sample sizes exist in the literature (e.g. [20,21]), this kind of sample ensured that (1) typicality and diversity were optimized; while (2) the sample could still be carefully examined and manually verified. These musical genres were used to maximize musical variety in the collection, and to ensure that the sample was

compatible with a host of other music preference studies (e.g. [22,23]), as these studies have also provided lists of between 13 and 15 broad musical genres that are relevant to most Western adult listeners.

All the tags related to each of the songs in the sample were then retrieved in March 2009 from the millions of users of the mentioned social media using a dedicated *application programming interface* called *Pylast* (http://code.google.com/p/pylast/). As expected, not quite all (91.41%) of the songs in the collection could be found; those not found were probably culturally less familiar songs for the average Western listener (e.g., from the *Iskelmä* and *World* music genres). The retrieved *corpus* now consisted of 5825 lists of tags, with a mean length of 62.27 tags. As each list referred to a particular song, the song's title was also used as a label, and together these were considered as a document in the *Natural Language Processing* (NLP) context (see the preprocessing section of Appendix A). In addition to this textual data, numerical data for each list were obtained that showed the number of times a tag had been used (*index of usage*) up to the point when the tags were retrieved. The corpus contained a total of 362,732 tags, of which 77,537 were distinct and distributed over 323 frequency classes (in other words, the shape of the spectrum of rank frequencies), and this is reported here to illustrate the prevalence of hapax legomena–tags that appear only once in the corpus–in Table 1 (cf. [24]). The tags usually consisted of one or more words (M = 2.48, SD = 1.86), with only a small proportion containing long sentences (6% with five words or more). Previous studies have *tokenized* [20,25] and *stemmed* [26] the tags to remove common words and normalize the data. In this study however, a tag is considered as a holistic unit representing an element of the *vocabulary* (cf. [27]), disregarding the number of words that compose it. Treating tags as *collocations* (i.e. words that are frequently placed together for a combined effect)–rather than as separate, single keywords–has the advantage of keeping the link between the music and its description a priority, rather than the words themselves. This approach shifts the focus from data processing to concept processing [28], where the tags function as conceptual expressions [29] instead of purely words or phrases. Furthermore, this treatment (collocated versus separated) does not distort the underlying nature of the corpus, given that the distribution of the sorted frequencies of the vocabulary still exhibits a Zipfian curve. Such a distribution suggests that tagging behaviour is also governed by the principle of least effort [30], which is an essential underlying feature of human languages in general [27].

Table 1 Frequency classes of tags

Class	*N*	Cumulative (%)
1 (hapaxes)	46 727	60.26
2	11 724	75.38
3	5512	82.49
4	2938	86.28
5	2020	88.89
6	1420	90.72
7	1055	92.08
8	838	93.16
9	674	94.03
10+	4094	100

2.1 Exposing the structure via cluster analysis

The tag structure was obtained via a vector-based semantic analysis that consisted of three stages: (1) the construction of a Term-Document Matrix, (2) the calculation of similarity coefficients and (3) cluster analysis.

The *Term Document Matrix* $\mathbf{X} = \{x_{ij}\}$ was constructed so that each song i corresponded to a "Document" and each unique tag (or item of the vocabulary) j to a "Term". The result was a binary matrix $\mathbf{X}(0, 1)$ containing information about the presence or absence of a particular tag to describe a given song.

$$x_{ij} = \begin{cases} 1, & \text{if } j \in i \\ 0, & \text{if } j \notin i \end{cases} \tag{1}$$

The similarity matrix $n \times n$ $\mathbf{D}$ with elements $d_{ij}$where d_{ii}= 0 was created by computing similarity indices between tag vectors $x_{i^*j}$of $\mathbf{X}$ with:

$$d_{ij} = \frac{ad}{\sqrt{(a+b)(a+c)(d+b)(d+c)}} \tag{2}$$

where a is the number of (1,1) matches, b = (1,0), c = (0,1) and d = (0,0). A choice then had to be made between the several methods available to compute similarity coefficients between binary vectors [31]. The coefficient (2) corresponding to the 13th coefficient of Gower and Legendre was selected because of its *symmetric* quality. This effectively means that it considers double absence (0,0) as equally important as double presence (1,1), which is a feature that has been observed to have a positive impact in ecological applications [31]. Using Walesiak and Dudek algorithm [32], we then compared its performance with nine alternative similarity measures used for binary vectors, in conjunction with five distinct clustering methods. The outcome of this comparison was that the coefficient we had originally chosen was indeed best suited to create an intuitive and visually appealing result in terms of *dendrograms* (i.e. visualizations of hierarchical clustering).

The last step was to find meaningful clusters of tags. This was done using a hierarchical clustering algorithm that transformed the similarity matrix into a sequence of nested partitions. The aim was to find the most compact, spherical clusters, hence Ward's minimum variance method [33] was chosen due to its advantages in general [34], but also in this particular respect, when compared to other methods (i.e. single, centroid, median, McQuitty and complete linkage).

After obtaining a hierarchical structure in the form of a dendrogram, the clusters were then extracted by "pruning" the branches with another algorithm that combines a "partitioning around medioids" clustering method with the height of the branches [35]. The result of this first hybrid operation can be seen in the 19 clusters shown in Figure 1, shown as vertical-coloured stripes in the top section of the bottom panel. In addition, the typical tags related to each of these cluster medioids are shown in Table 2.

To increase the interpretability of these 19 clusters, a second operation was performed, consisted of repeating the hybrid pruning to increase the minimum amount of items per cluster (from 5 to 25), which thereby decreased the overall number of actual clusters. It resulted in five meta-clusters, shown in the lower section of stripes in Figure 1. These were labelled according to their contents as *Energetic* (I), *Intimate* (II), *Classical* (III), *Mellow* (IV) and *Cheerful* (V).

In both the above operations, the size of the clusters varied considerably. This was most noticeable for the first cluster in both, which was significantly larger than the rest. We interpreted this to be due to the fact that these first clusters might be capturing tags with weak relations. Indeed, for practical purposes, the first in both solutions was not as well defined and clean-cut in the semantic domain as the rest of the clusters. This was probably due to the fact that the majority of tags used in them was highly polysemic (i.e. using words that have different, and sometimes unrelated senses).

2.2 From clustered tags to music

This section explains how the original database, of 6372 songs, was then reorganized according to their closeness to each tag cluster in the semantic space. In other words, the 19 clusters from the analysis were now considered as prototypical descriptions of 19 ways that music shares similar characteristics. These prototypical descriptions were referred to as "clusters profiles" in the vector space, containing sets of between 5 and 334 tags in common (to a particular concept). Songs were then described in terms of a comparable ranked list of tags, varying in length from 1 to 96. The aim was then to measure (in terms of Euclidean distance) how close each song's ranked list of tags was to each prototypical description's set of tags. The result of this would tell us how similar each song was to each prototypical description.

An $m \times n$ *Term Document Matrix* $\mathbf{Y} = \{y_{ij}\}$ was therefore constructed to define the cluster profiles in the vector space. In this matrix, the lists of tags

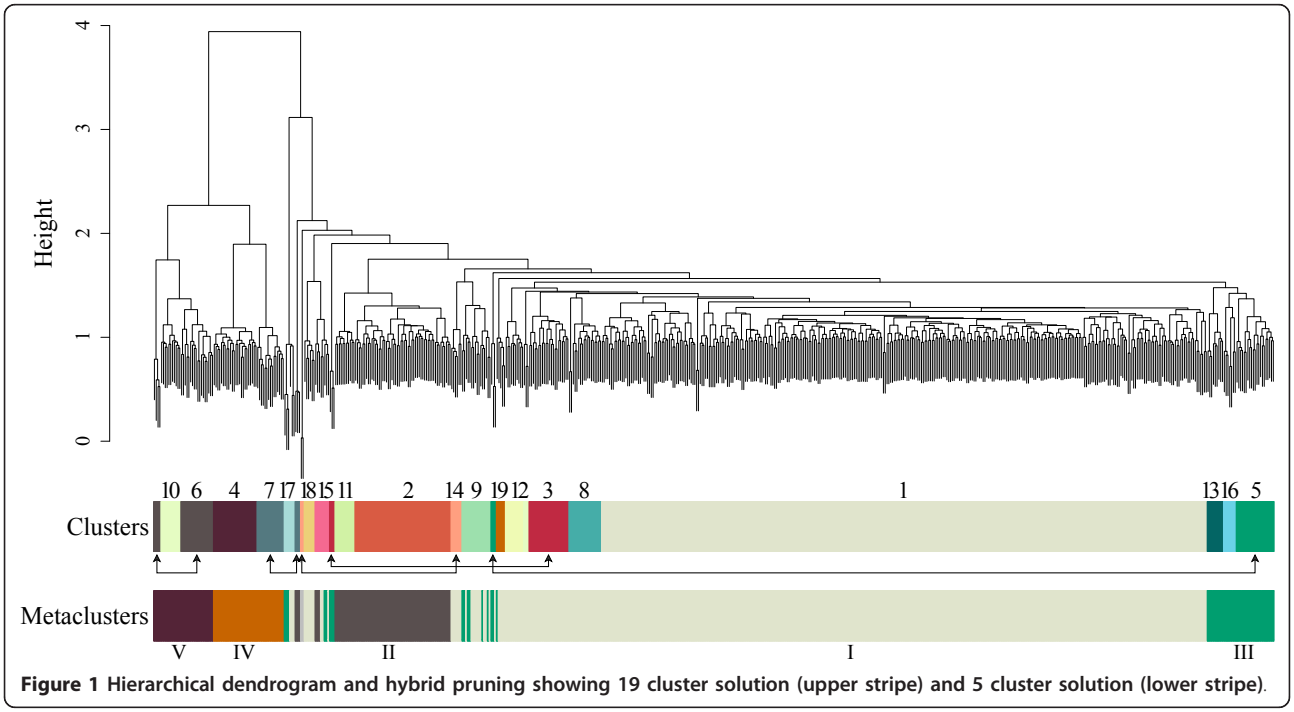

Figure 1 Hierarchical dendrogram and hybrid pruning showing 19 cluster solution (upper stripe) and 5 cluster solution (lower stripe).

Table 2 Most representative tags and corresponding artists for each of the 19 clusters

ID	Tags closest to cluster centroids	Top artists in the cluster
1	*energetic, powerful, hot*	Amy Adams, Fred Astaire, Kelly Clarkson
2	*dreamy, chill out, sleep*	Nick Drake, Radiohead, Massive Attack
3	*sardonic, sarcastic, cynical*	Alabama 3, Yann Tiersen, Tom Waits
4	*awesome, amazing, great*	Guns N' Roses, U2, Metallica
5	*cello, piano, cello rock*	Camille Saint-Saëns, Tarja Turunen, Franz Schubert
6	*00s, sexy, catchy*	Fergie, Lily Allen, Amy Winehouse
7	*mellow, beautiful, sad*	Katie Melua, Phil Collins, Coldplay
8	*hard, angry, aggressive*	System of a Down, Black Sabbath, Metallica
9	*60s, 70s, legendary*	Simon & Garfunkel, Janis Joplin, The Four Tops
10	*feelgood, summer, cheerful*	Mika, Goo Goo Dolls, Shekinah Glory Ministry
11	*wistful, intimate, reflective*	Soulsavers, Feist, Leonard Cohen
12	*high school, 90's, essential*	Fool's Garden, The Cardigans, No Doubt
13	*50s, saxophone, trumpet*	Miles Davis, Thelonious Monk, Charles Mingus
14	*1980s, eighties, voci maschili*	Ray Parker Jr., Alphaville, Michael Jackson
15	*affirming, lyricism, life song*	Lisa Stansfield, KT Tunstall, Katie Melua
16	*choral, a capella, medieval*	Mediæval Bæbes, Alison Krauss, Blackmore's Night
17	*voce femminile, donna, bella topolina*	Avril Lavigne, The Cranberries, Diana Krall
18	*tangy, coy, sleek*	Kylie Minogue, Ace of Base, Solange
19	*rousing, exuberant, passionate*	James Brown, Does It Offend You, Yeah?, Tchaikovsky

attributed to a particular song (i.e. the song descriptions) are represented as m, and n represents the 618 tags left after the filtering stage (i.e. the preselected tags). Each list of tags (i) is represented as a finite set {1, ..., k}, where $1 \leq k \leq 96$ (with a mean of 29 tags per song). Finally, each element of the matrix contains a value of the normalized rank of a tag if found on a list, and it is defined by:

$$y_{ij} = \left(\frac{r_k}{k}\right)^{-1} \tag{3}$$

where $r_k$is the cardinal rank of the tag j if found in i, and k is the total length of the list. Next, the mean rank of the tag across $\mathbf{Y}$ is calculated with:

$$\bar{r}_j = \frac{\sum_{i=1}^{m} y_{ij}}{m} \tag{4}$$

And the cluster profile or *mean ranks vector* is defined by:

$$\mathrm{p}_l = \bar{r}_{j \in C_l} \tag{5}$$

$C_l$denotes a given cluster l where $1 \leq l \leq 19$, and $\mathbf{p}$ is a vector {5, ..., k}, where $5 \leq k \leq 334$ (5 is the minimum number of tags in one cluster, and 334 is the maximum in another).

The next step was to obtain, for each cluster profile, a list of songs ranked in order according to their closeness to the profile. This consisted in calculating the Euclidean distance $d_i$between each song's rank vector $y_{i,j \in C_l}$ and each cluster profile $\mathbf{p}l$ with:

$$d_i = \sqrt{\sum_{j \in C_l} (y_{ij} - \mathrm{p}_l)^2} \tag{6}$$

Examples of the results can be seen in Table 2, where top artists are displayed beside the central tags for each cluster, while Figure 2 shows more graphically how the closeness to cluster profiles was calculated for this ranking scheme. In it are shown three artificial and partly overlapping clusters (I, II and III). In each cluster, the centroid $\mathbf{p}_l$has been calculated, together with the Euclidean distance from it to each song, as formally explained in Equations 3-6. This distance is graphically represented by the length of each line from centroid to the songs (a, b, c, ...), and the boxes next to each cluster

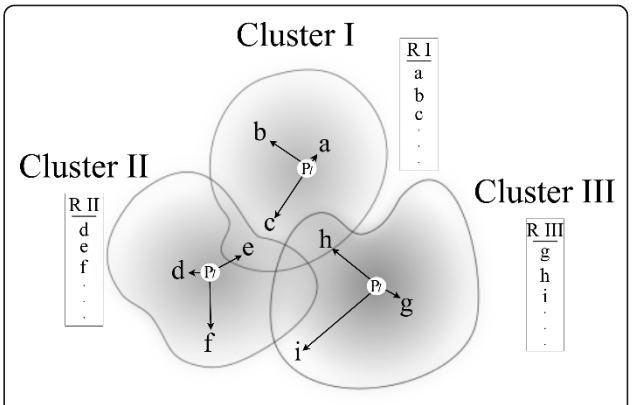

Figure 2 Visual example of the ranking of the songs based on their closeness to each cluster profile.

show their ranking (the boxes with R I, R II, R III) accordingly. Furthermore, this method allows for systematic comparisons of the clusters to be made when sampling and analysing the musical material in different ways, which is the topic of the following section.

3 Determining the acoustic qualities of each cluster

Previous research on explaining the semantic qualities of music in terms of its acoustic features has taken many forms: genre discrimination tasks [36,37], the description of soundscapes [5], bipolar ratings encompassing a set of musical examples [6] and the prediction of musical tags from acoustic features [21,38-40]. A common approach in these studies has been to extract a range of features, often low-level ones such as timbre, dynamics, articulation, Mel-frequency cepstral coefficients (MFCC) and subject them to further analysis. The parameters of the actual feature extraction are dependent on the goals of the particular study; some focus on shorter musical elements, particularly the MFCC and its derivatives [21,39,40]; while others utilize more high-level concepts, such as harmonic progression [41-43].

In this study, the aim was to characterize the semantic structures with a combined set of non-redundant, robust low-level acoustic and musical features suitable for this particular set of data. These requirements meant that we employed various data reduction operations to provide a stable and compact list of acoustic features suitable for this particular dataset [44]. Initially, we considered a large number of acoustic and musical features divided into the following categories: dynamics (e.g. root mean square energy); rhythm (e.g. fluctuation [45] and attack slope [46]); spectral (e.g. brightness, roll-off [47,48], spectral regularity [49] and roughness [50]); spectro-temporal (e.g. spectral flux [51]) and tonal features (e.g. key clarity [52] and harmonic change [53]). By considering the mean and variance of these features across 5-s samples of the excerpts (details given in the following section), we were initially presented with 50 possible features. However, these features contained significant redundancy, which limits the feasibility of constructing predictive classification or regression models and also hinders the interpretation of the results [54]. For this reason, we did not include MFCC, since they are particularly problematic in terms of redundancy and interpretation [6].

The features were extracted with the *MIRtoolbox* [52] using a frame-based approach [55] with analysis frames of 50-ms using a 50% overlap for the dynamic, rhythmic, spectral and spectro-temporal features and 100-ms with an overlap of 87.5% for the remaining tonal features.

The original list of 50 features was then reduced by applying two criteria. Firstly, the most stable features were selected by computing the Pearson's correlation between two random sets taken from the 19 clusters. For each set, 5-s sound examples were extracted randomly from each one of the top 25 ranked songs representing each of the 19 clusters. More precisely: $P(t)$ for $0.25T \leq t \leq 0.75T$, where T represents the total duration of a song. This amounted to 475 samples in each set, which were then tested for correlations between sets. Those features correlating above $r = 0.5$ between two sets were retained, leaving 36 features at this stage. Secondly, highly collinear features were discarded using a *variance inflation factor* $(\hat{\beta}_i < 10)$ [56]. This reduction procedure resulted in a final list of 20 features, which are listed in Table 3.

3.1 Classification of the clusters based on acoustic features

To investigate whether they differed in their acoustic qualities, four test sets were prepared to represent the clusters. For each cluster, the 50 most representative songs were selected using the ranking operation defined in Section 2.2. This number was chosen because an analysis of the rankings within clusters showed that the top 50 songs per cluster remained predominantly within the target cluster alone (89%), whereas this discriminative property became less clear with larger sets (100 songs at 80%, 150 songs at 71% and so on). From these

Table 3 Selected 20 acoustic features

Domain	Name	Σ	*MDA*
Rhythm	Attack time	M	0.23
		SD	0.08
	Fluctuation centr.	M	0.63
	Fluctuation peak	M	0.58
Spectral	Brightness	SD	0.39
	Entropy	SD	0.66
	Flatness	SD	0.60
	Regularity	M	0.33
		SD	0.26
	Roll-off	SD	0.06
	Roughness	M	0.75
	Spread	M	0.54
Spectro-Temporal	Spectral flux	M	1.20
		SD	0.44
Tonal	Chromagram centr.	M	0.98
		SD	0.35
	Chromagram peak	M	0.60
	Harmonic change	M	0.50
		SD	0.61
	Key clarity	M	0.07

Σ stands for the summary measure, where M = mean and SD = standard deviation. MDA is the *Mean Decrease Accuracy* in classification of the five meta-clusters by the acoustic features using RF.

candidates, two random 5-s excerpts were then extracted to establish two sets, to train and test each clustering, respectively. For 19 clusters, this resulted in 950 excerpts per set; and for the 5 meta-clusters, it resulted in 250 excerpts per set. After this, classification was carried out using Random Forest (RF) analysis [57]. RF is a recent variant of the regression tree approach, which constructs classification rules by recursively partitioning the observations into smaller groups based on a single variable at a time. These splits are created to maximize the between groups sum of squares. Being a non-parametric method, regression trees are thereby able to uncover structures in observations which are hierarchical, and yet allow interactions and nonlinearity between the predictors [58]. RF is designed to overcome the problem of overfitting; bootstrapped samples are drawn to construct multiple trees (typically 500 to 1000), which have randomized subsets of predictors. Out-of-bag samples are used to estimate error rate and variable importance, hence, eliminating the need for cross-validation, although in this particular case we still resorted to validation with a test set. Another advantage of RF is that the output is dependent only on one input variable, namely, the number of predictors chosen randomly at each node, heuristically set to 4 in this study. Most applications of RF have demonstrated that this technique has improved accuracy in comparison to other supervised learning methods.

For 19 clusters, a mere 9.1% of the test set could correctly be classified using all 20 acoustic features. Although this is nearly twice the chance level (5.2%), clearly the large number of target categories and their apparent acoustic similarities degrade the classification accuracy. For the meta-clusters however, the task was more feasible and the classification accuracy was significantly higher: 54.8% for the prediction per test set (with a chance level of 20%). Interestingly, the meta-clusters were found to differ quite widely in their classification accuracy: Energetic (I, 34%), Intimate (II, 66%), Classical (III, 52%), Mellow (IV, 50%) and Cheerful (V, 72%). As mentioned in Section 2.1, the poor classification accuracy of meta-cluster I is understandable, since that cluster contained the largest number of tags and was also considered to contain the weakest links between the tags (see Figure 1). However, the main confusions for meta-cluster I were with clusters III and IV, suggesting that labelling it as "Energetic" may have been premature (see Table 4). The advantage of the RF approach is the identification of critical features for classification using the Mean Decrease Accuracy [59].

Another reason for RF classification chosen was that it uses relatively unbiased estimates based on out-of-bag samples and the permutation of classification trees. The mean decrease in accuracy (MDA) is the average of such estimates (for equations and a fuller explanation, see [57,60]). These are reported in Table 3, and the normalized distributions of the three most critical features are shown in Figure 3. Spectral flux clearly distinguishes the meta-clusters II from III and IV from V, in terms of the amount of change within the spectra of the sounds used. Differences in the dominant registers also distinguish meta-clusters I from II and III from V, and these are reflected in differences in the estimated mean centroid of the chromagram for each, and roughness, the remaining critical feature, partially isolates cluster IV (Mellow, Awesome, Great) from the other clusters.

Table 4 Confusion matrix for five meta-clusters (showing 54.8% success in RF classification)

		Predicted				
		I Energetic	II Intimate	III Classical	IV Mellow	V Cheerful
Actual	I Energetic	17	5	3	2	5
	II Intimate	9	33	10	11	2
	III Classical	8	4	26	5	3
	IV Mellow	13	5	3	25	4
	V Cheerful	3	3	8	7	36

The classification results imply that the acoustic correlates of the clusters can be established if we are looking only at the broadest semantic level (meta-clusters). Even then, however, some of the meta-clusters were not adequately discriminated by their acoustical properties. This and the analysis with all 19 clusters suggest that many of the pairs of clusters have similar acoustic contents and are thus indistinguishable in terms of classification analysis. However, there remains the possibility that the overall structure of the cluster solution is nevertheless distributed in terms of the acoustic features along dimensions of the cluster space. The cluster space itself will therefore be explored in more detail next.

3.2 Acoustic characteristics of the cluster space

As classifying the clusters according to their acoustic features was not hugely accurate at the most detailed cluster level, another approach was taken to define the differences between the clusters in terms of their mutual distances. This approach examined in more detail their underlying acoustic properties; in other words, whether there were any salient acoustic markers delineating the concepts of cluster 19 ("Rousing, Exuberant, Confident, Playful, Passionate") from the "Mellow, Beautiful, Chillout, Chill, Sad" tags of cluster 7, even though the actual boundaries between the clusters were blurred.

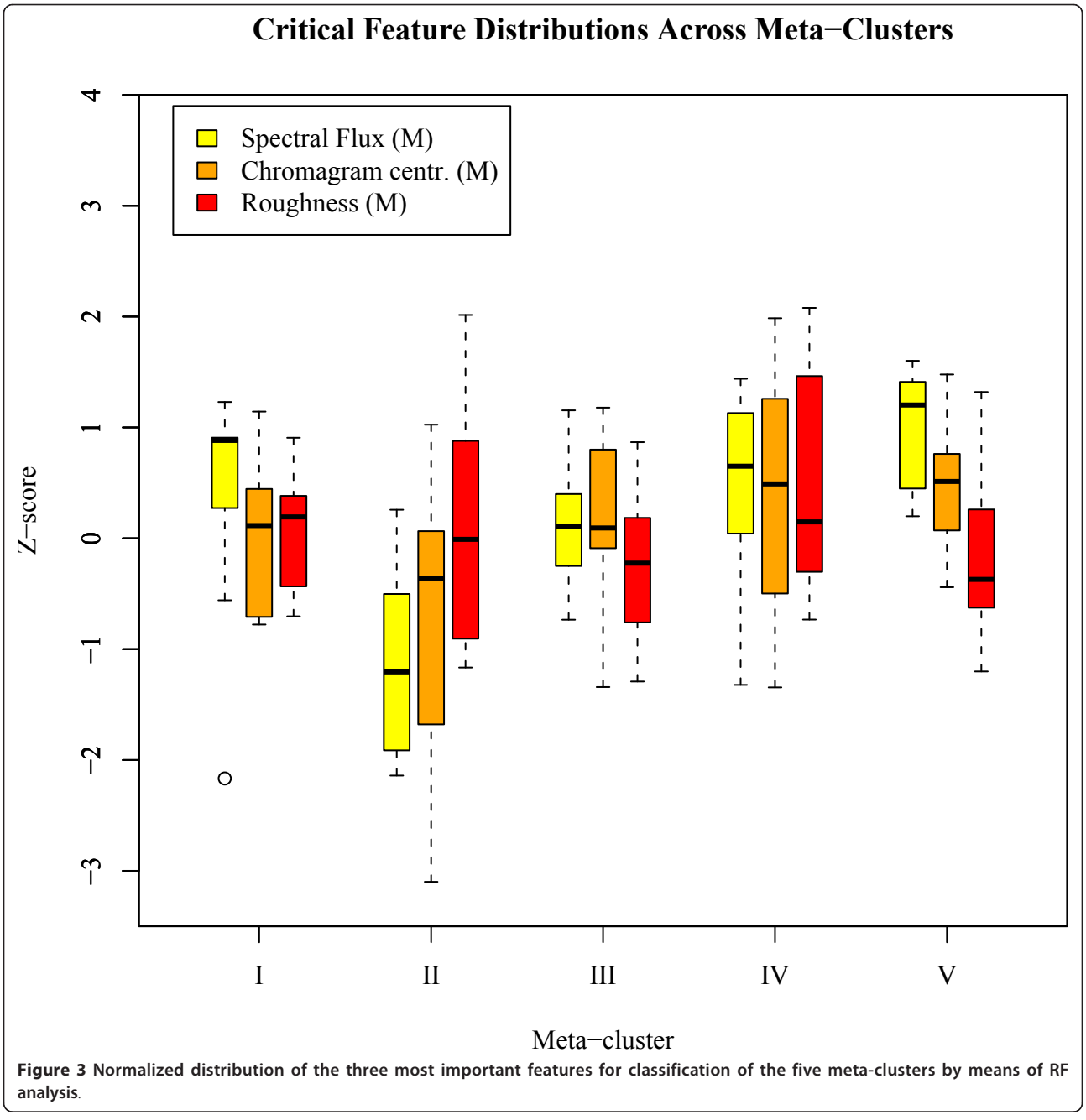

Figure 3 Normalized distribution of the three most important features for classification of the five meta-clusters by means of RF analysis.

To explore this idea fully, the intercluster distances were first obtained by computing the closest Euclidean distance between two tags belonging to two separate clusters [61]:

$$\text{dist}(C_i, C_j) = \min\{d(x, y) : x \in C_i, y \in C_j\} \quad (7)$$

where $C_i$and $C_j$represent a pair of clusters and x and y two different tags.

Nevertheless, before settling on this method of single linkage, we checked three other intercluster distance measures (Hausdorff, complete and average) for the purposes of comparison. Single linkage was finally chosen due to its intuitive and discriminative performance in this material and in general (cf. [61]).

The resulting distance matrix was then processed with classical metrical *Multidimensional Scaling* (MDS) analysis [62]. We then wanted to calculate the minimum number of dimensions that were required to approximate the original distances in a lower dimensional space. One way to do this is to estimate the proportion of variation explained:

$$\frac{\sum_{i=1}^{p}\lambda_i}{\sum(\text{positive eigenvalues})} \quad (8)$$

where p is the number of dimensions and $\lambda_i$represents the eigenvalues sorted in decreasing order [63].

However, the results of this procedure suggested that considering only a reduced number of dimensions would not satisfactorily reflect the original space, so we instead opted for an exploratory approach (cf. [64]). An exploration of the space meant that we could investigate whether any of the 18 dimensions correlated with the previously selected set of acoustic features, which had been extracted from the top 25 ranked examples of the 19 clusters. This analysis yielded statistically significant correlations for dimensions 1, 3 and 14 of the MDS solution with the acoustic features that are shown in Table 5. For the purpose of illustration, Figure 4 shows the relationship, in the inter-cluster space, between four of these acoustic features (shown in the labels for each axis) and two of these dimensions (1 and 3 in this case). If we look at clusters 14 and 16, we can see that they both contain tags related with the human voice (*Voci maschili* and *Choral*, respectively), and they are situated around the mean of the *X*-axis. However, this is in spite of a large difference in sound character, which can best be described in terms of their perceptual dissonance (e.g. spectral roughness), hence their positions at either end of the *Y* -axis. Another example of tags relating to the human voice, concerns clusters 17 and 4 (*Voce femminile* and *Male Vocalist*, respectively), but this time they are situated around the mean of the *Y* -axis, and it is in terms of the shape of the spectrum (e.g. spectral spread) that they differ most, hence their positions at the end of the *X*-axis. In sum, despite the modest classification accuracy of the clusters according to their acoustic features, the underlying semantic structure embedded into tags could nonetheless be more clearly explained in terms of their relative positions to each other within the cluster space. The dimensions yielded intuitively interpretable patterns of correlation, which seem to adequately pinpoint the essence of what musically characterize the concepts under investigation in this study (i.e. adjectives, nouns, instruments, temporal references and verbs). However, although these semantic structures could be distinguished sufficiently by their acoustic profiles at the generic, meta-cluster level; this was not the case at the level of the 19 individual clusters. Nevertheless, the organization of the individual clusters across the semantic space could be connected by their acoustic features. Whether the acoustic substrates that musically characterize these tags is what truly distinguishes them for a listener is an open question that will be explored more fully next.

4 Similarity rating experiment

In order to explore whether the obtained clusters were perceptually meaningful, and to further understand what kinds of acoustic and musical attributes they actually consisted of, new empirical data about the clusters needed to be gathered. For this purpose, a similarity rating experiment was designed, which assessed the timbral qualities of songs from each of the tag clusters. We chose to focus on the low-level, non-structural qualities of music, since we wanted to minimize the possible confounding factor of association, caused by recognition of lyrics, songs or artists. The stimuli for the experiment therefore consisted of semi-randomly spliced [37,65], brief excerpts. These stimuli, together with other details of the experiment, will be explained more fully in the remaining parts of this section.

4.1 Experiment details

4.1.1 Stimuli

Five-second excerpts were randomly taken from a middle part ($P(t)$ for $0.25T \leq t \leq 0.75T$, where T represents the total duration of a song) of each of the 25 top ranked songs from each cluster (see the ranking procedure detailed in Section 2.2). However, when splicing the excerpts together for similarity rating, we wanted to minimize the confounds that were caused by disrupting the onsets (i.e. bursts of energy). Therefore, the exact temporal position of the onsets for each excerpt was detected with the aid of the MIRToolbox [52]. This

Table 5 Correlations between acoustic features and the inter-item distances between the clusters

Dimension 1		Dimension 3		Dimension 14	
Acoustic feature	***r***	**Acoustic feature**	***r***	**Acoustic feature**	***r***
Fluctuation centroid (M)	0.53*	Regularity (SD)	-0.51*	Chromagram centroid (M)	0.60**
Spread (M)	0.51*	Harmonic change (SD)	-0.50*	Flatness (SD)	0.54*
Entropy (SD)	0.50*	Roughness (M)	0.50*	Attack time (M)	-0.51*
Brightness (SD)	0.49*	Harmonic change (M)	-0.50*	Regularity (M)	-0.51*
Flatness (SD)	0.49*	Chromagram centroid (SD)	-0.45*	Attack time (SD)	-0.48*
Flux (SD)	0.49*	Flux (SD)	-0.45*	Chromagram peak (M)	-0.46*

* p <0.05, ** p <0.01, df = 17

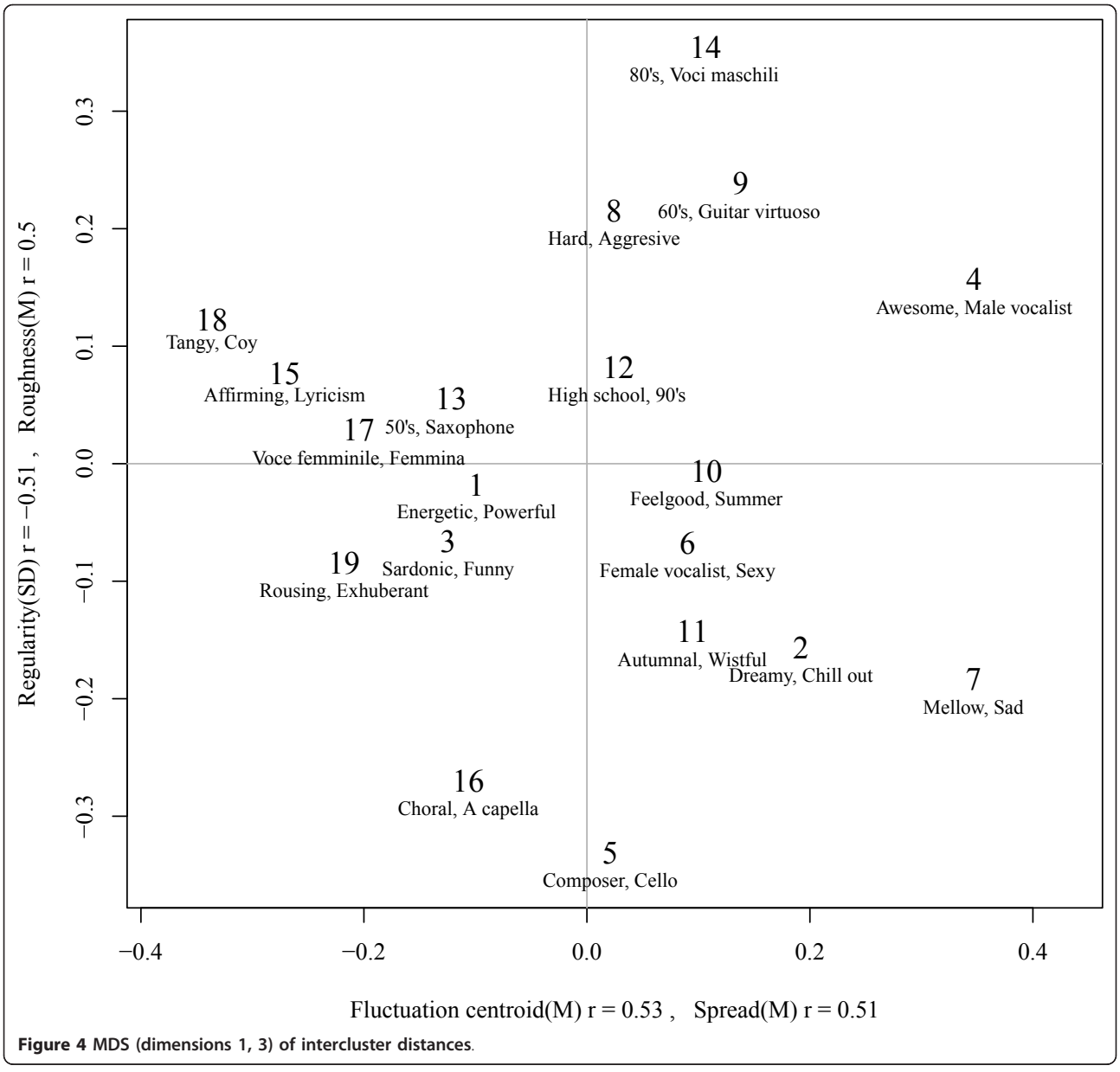

Figure 4 MDS (dimensions 1, 3) of intercluster distances.

process consisted of computing the *spectral flux* within each excerpt by focussing on the increase in energy in successive frames. It produced a temporal curve from which the highest peak was selected as the reference point for taking a slice, providing that this point was not too close to the end of the signal ($t \leq 4500$ ms).

Slices of random length ($150 \leq t \leq 250$ ms) were then taken from a point that was 10 ms before the peak onset for each excerpt that was being used to represent a tag cluster. The slices were then equalized in loudness, and finally mixed together using a fade in/out of 50 ms and an overlap window of 100 ms. This resulted in 19 stimuli (examples of the spliced stimuli can be found at http://www.jyu.fi/music/coe/materials/splicedstimuli) of variable length, each corresponding to a cluster, and each of which was finally trimmed to 1750 ms (with a fade in/out of 100 ms). To finally prepare these 19 stimuli for a similarity rating experiment, the resulting 171 paired combinations were mixed with a silence of 600 ms between them.

4.1.2 Participants

Twelve females and nine males were participated in this experiment (age M = 26.8, SD = 4.15). Nine of them had at least 1 year of musical training. Twelve reported listening to music attentively between 1 and 10 h/week, and 19 of the subjects listened to music while doing another activity (63% $1 \leq t \leq 10$, 26% $11 \cdot \leq t \leq 20$, 11% $t \leq 21$ h/week).

4.1.3 Procedure

Participants were presented with pairs of sound excerpts in random order using a computer interface and high-quality headphones. Their task was to rate the similarity of sounds on a 9-level Likert scale, the extremes of which were labelled as *dissimilar* and *similar*. Before the actual experimental trials, the participants were also given instructions and some practice to familiarize themselves with the task.

4.2 Results of experiment

The level at which participants' ratings agreed with each other was estimated with Cronbach's method (α = 0.94), and the similarity matrices derived from their ratings were used to make a representation of the perceptual space. Individual responses were thus aggregated by computing a mean similarity matrix, and this was subjected to a classical metric MDS analysis. With Cox and Cox's [63] method (8) we estimated that four dimensions were enough to represent the original space since these can explain 70% of the variance.

4.2.1 Perceptual distances

As would be hoped, the arrangement of clusters, as represented by their spliced acoustic samples, illustrates a clear organization according to an underlying semantic structure. This perceptual distance can be seen in Figure 5 where, for example, *Aggressive* and *Chill out* are in opposite corners of the psychological space. There is also a clear acoustical organization of the excerpts, as cluster number 5 (*Composer*, *Cello*) is depicted as being high in roughness and high in spectral regularity, with a well-defined set of harmonics, and those clusters that have similar overall descriptors, such as 15 (*Affirming*, *Lyricism*), 7 (*Mellow*, *Sad*) and 11 (*Autumnal*, *Wistful*), are located within proximity of each other. Noticeably, cluster number 1 is located at the centre of the MDS solution, which could be expected from a cluster that worked as a trap for tags with weak relations.

4.2.2 Acoustic attributes of the similarities between stimuli

Acoustic features were extracted from the stimuli in a similar fashion to that described in Section 3, but the list of features was consolidated again by trimming it down to a robust minimal set. Trimming consisted of creating another random set of stimuli and correlating their acoustic features with the stimuli used in the experiment. Those features which performed poorly (r <0.5, df = 17) were removed from the list. After this, the coordinates of the resulting 4-dimensional space were correlated with the set of acoustic features extracted from the stimuli to show the perceptual distances of the stimuli from one another. Only dimensions 1 and 2 had statistically significant linear correlations with the acoustic features, the other dimensions having only low correlations ($|r| \leq 0.5$, or p >0.05, df = 17). The final selection of both acoustic features and dimensions is displayed in Table 6.

The first dimension correlates with features related to the organization of pitch and harmonics, as revealed by the mean chromagram peaks (r = 0.82) and the degree of variation between successive peaks in the spectrum (mean spectral regularity r = 0.72). There is also correlation with the variance of the energy distribution (standard deviation of the spectral roll-off at 95% r = 0.7); the distance between the spectrum of successive frames (mean spectral flux r = -0.7); and to a lesser degree with the shape of the spectrum in terms of its "width" (mean spectral spread r = -0.61). The second dimension correlates significantly with the perceived dissonance (mean roughness r = -0.74); pitch salience (chromagram centroid r = -0.72); and also captures the mean spectral spread (r = 0.65), although in an inverse fashion. Table 6 provides a more detailed summary of this.

4.2.3 Comparing a semantic structure based on social tags, to one based on perceptual similarities

As we have now explored the emergent structure from tags using a direct acoustic analysis of the best exemplars in each cluster, and probed this semantic space further in a perceptual experiment, the question remains as to whether the two approaches bear any similarities. The most direct way to examine this is to look at the pattern of correlations between both: i.e. to compare tables 5 and 6. Although the lists of features vary slightly, due to the difference in redundancy and robustness criteria applied to each set of data, convergent patterns can still be found. An important shared feature is the variation in brightness, which is both present in dimension 1 of the direct cluster analysis, and in the perceptual space depicting the spliced stimuli (from the same 19 clusters). In the first case, it takes the form of "brigthenss SD", and in the second, it is "roll-off SD" (virtually identical). In addition, the second dimension in both solutions is characterized by roughness, although the underlying polarities of the space have been flipped in each. Of course, one major reason for differences between the two sets of data must be due to the effects of splicing, conducted in the perceptual experiment but not in the other. However, there were nevertheless analogies between the two perspectives of the semantic structure that could be detected in the acoustic substrates. They have been used here to highlight such features that are little affected by form, harmony, lyrics and other high-level musical (and extra-musical) characteristics. From this perspective, a tentative convergence between the two approaches was successfully obtained.

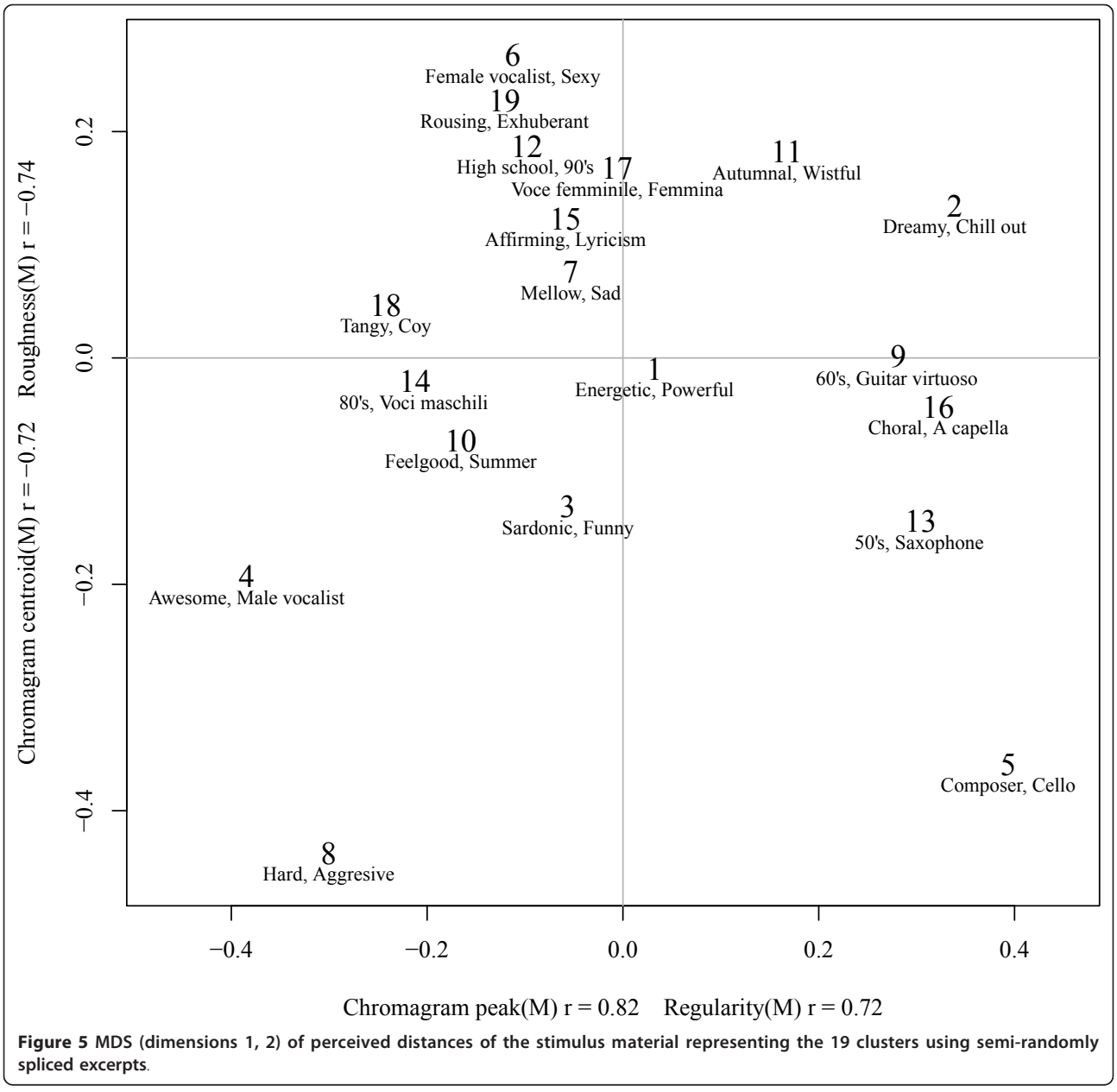

Figure 5 MDS (dimensions 1, 2) of perceived distances of the stimulus material representing the 19 clusters using semi-randomly spliced excerpts.

5 Discussion and conclusions

Semantic structures within music have been extracted from the social media previously [20,25] but the main difference between the prior genre-based studies and this study is that we focussed more on the way people describe music in terms of how it sounds in conceptual expressions. We argue that these expressions are more stable than musical genres, which have previously proven to be of a transient nature and a source of disagreement (cf. [37]), despite important arguments vindicating their value for classification systems [66]. Perhaps the biggest problem with expert classifications (such as genre) is that the result may not reach the same level of ecological validity in describing how music sounds, as a semantic structure derived from social tags. This is a very important reason to examine tag-based semantic structures further, in spite of their inherent weaknesses as pointed out by Lamere [7].

A second way in which this study differs from those previous lies in the careful filtering of the retrieved tags using manual and automatic methods before the actual analysis of the semantic structures was conducted. Not only that, but a prudent trimming of the acoustic features was done to avoid overfitting and any possible increases in model complexity. Finally, a perceptual exploration of the semantic structure found was carried

Table 6 Correlations between MDS solutions (dimensions 1 and 2) and acoustic features for the experiment

Domain	Name	Σ	Dim 1	Dim 2
Spectral	Entropy	SD	0.36	0.46*
	Flatness	SD	-0.13	0.32
	Regularity	M	0.72***	0.10
	Roll-Off	SD	0.70***	0.14
	Roughness	M	-0.35	-0.74***
	Spread	M	-0.61**	0.65**
Spectro-temporal	Spectral flux	M	-0.70***	-0.16
Tonal	Chromagram centroid	M	-0.23	-0.72***
	Chromagram peak	M	0.82***	-0.28

* p <0.05, ** p <0.01, *** p <0.001, df = 17

out to assess whether the sound qualities alone would be sufficient to uncover the tag-based structure.

The whole design of this study offers a preliminary approach to the cognition of timbre in semantic terms. In other words, it uses verbal descriptions of music, expressed by the general population (in the form of social tags), as a window to study how a critical feature of music (timbre) is represented in the *semantic memory* [67]. It is however evident that if each major step of this study was treated separately, there would be plenty of room for refining their respective methodologies, namely, tag filtering, uncovering the semantic structure, acoustic summarization and conducting a perceptual experiment to examine the two empirical perspectives. This being said, we *did* consider some of the alternatives for these steps to avoid methodological pitfalls (particularly in the clustering and the distance measures). But even if each analytical step was optimized to enhance the solution to an isolated part of the problem, this would inevitably come at the expense of unbalancing the overall picture. Since this study relies on an exploratory approach, we chose mainly conventional techniques for each step, with the expectation that further research will be conducted to corroborate the findings and improve the techniques used here.

The usefulness of signal summarization based on the random spliced method [37] has been assessed for audio pattern recognition [65]. Our findings in the perceptual domain seem to vindicate the method where listeners rate sounds differing in timbral qualities, especially if the scope is the long-term non-structural qualities of music [68]. Such a focus is attained by cutting the slices in a way that preserves important aspects of music (onsets and sample lengths), while ensuring that they are from a wide cross section of timbrically related songs (i.e. belonging to the same semantic region or *timbral environment* [69] in the perceptual space).

In conclusion, this study provided a bottom-up approach for finding the semantic qualities of music descriptions, while capitalizing on the benefits of social media, NLP, similarity ratings and acoustic analysis to do so. We learned that when listeners are presented with brief and spliced excerpts taken from the clusters representing a tag-based categorization of the music, they are able to form coherent distinctions between them. Through an acoustic analysis of the excerpts, clear correlations between the dimensional and timbral qualities of music emerged. However, it should be emphasized that the high relevance of many timbral features is only natural since the timbral characteristics of the excerpts were preserved and structural aspects were masked by the semi-random splicing. Nevertheless, we are positively surprised at the level of coherence in regard to the listener ratings and their explanations in terms of the acoustic features; in spite of the limitations we imposed on the setting using a random splicing method, and the fact that we tested a large number of clusters.

The implications of the present findings relate to several open issues. The first is whether structural aspects of music are required to explain the semantic structures or whether low-level, timbral characteristics are sufficient, as was suggested by the present findings. Secondly, what new semantic layers (as indicated by the categories of tags) can meaningfully be connected with the acoustic properties of the music? Finally, if the timbral characteristics are indeed strongly connected with such semantic layers as *adjectives*, *nouns* and *verbs*, do these arise by means of learning and associations or are the underlying regularities connected with the emotional, functional and gestural cues of the sounds?

A natural continuation of this study would be to go deeper into the different layers of tags to explore which layers are more amenable to direct mapping by acoustic qualities, and which are mostly dependent on the functional associations and cultural conventions of the music.

A Preprocessing

Preprocessing is necessary in any text mining application because the retrieved data do not follow any particular set of rules, and there are no standard steps to follow [70]. Moreover, with the aid of *Natural Language Processing* (NLP) [71,72] methods, it is possible to explore the nature of the tags from statistical and lexicological perspectives. In the following sections, the rationale and explanation for each preprocessing step is given.

A.1 Filtering

Three filtering rules were applied to the corpus.

Remove *Hapax legomena* (i.e. tags that appear only once in the corpus), under the rationale of discarding unrelated data (see Table 1).

Capture the most prevalent tags by eliminating from the vocabulary those whose *index of usage* (see Section 2) is below the mean.

Discard tags composed by three or more words in order to prune short sentence-like descriptions from the corpus.

The subset resulting from such reductions represents 46.6% of the corpus (N = 169, 052, Vocabulary = 2029 tags).

A.2 Lexical categories for tags

At this point, the data had been de-noised but only in the quantitative domain. To extract a meaningful ontology from the tags, not only filtering, but semantic analysis of the tags was necessary. To do so in an effective fashion, a qualitative analysis was performed using a number of sources: the Brown Corpus [73] to identify parts of speech; the Wordnet database [74] to disambiguate words; and the online Urban Dictionary (http://www.urbandictionary.com) and http://www.Last.fm database for general reference. We were thus aiming for a balanced set of references; two sources were technical (the Brown and Wordnet), one vernacular (the Urban Dictionary) and one highly specialized in musical jargon (Last.fm's wiki pages). An underlying motivation for relying on this broad set of references, rather than exclusively on an English dictionary, was to recognize the multilingual nature of musical tags. Tag meanings were thus looked up and the selection of a category was decided case by case. The criteria applied in this process favoured categories of meaning closely related to music and the music industry, such as the genre, artist, instrument, form of music, and commercial entity. The next most important type of meaning looked for was adjectival, and finally other types of descriptor were considered. For instance, "Acid" is well known to be a corrosive substance, but it is also a term used extensively to describe certain musical genres, so this latter meaning took priority. Table 7 shows the aforementioned tag categories, examples of each, a definition of each, and their percentage of distribution in the sample.

The greatest percentage of tags refer to *musical genres*, but there are significant percentages in other categories. For instance, the second most commonly found tags are *adjectives*, followed by *nouns* which except for some particular contextual connotations, are used for the most part adjectivally to describe the general sound of a song (e.g. *mellow*, *beautiful* for adjectives and *memories* and *melancholy* for nouns).

The rest of the categories suggest that music is often tagged in terms of association, whether it be to known auditory objects (e.g. instruments and band names), specific circumstances (e.g. geographical locations and time of the day or season) or idiosyncratic things that only make sense at a personal level. This classification is mainly consistent with past efforts [7], although the vocabulary analysed is larger, and there are consequently more categories.

The result allowed for a finer discrimination of tags to be made, that might better uncover the semantic structure. Since one of the main motivations of this project was to obtain prototypical timbral descriptions, we focused on only a few of the categories: *adjectives*, *nouns*, *instruments*, *temporal references* and *verbs*, and this resulted in a vocabulary of 618 tags.

The rest of the tag categories were left for future analysis. Note that this meant discarding such commonly used descriptors as musical genres, which on the one hand provide an easy way to discriminate music [36] in

Table 7 Main categories of tags

Categories	%	Definition	Examples
Genre	36.72	Musical genre or style	Rock, Alternative, Pop
Adjective	12.17	General category of adjectives	Beautiful, Mellow, Awesome
Noun	9.41	General category of nouns	Love, Melancholy, Memories
Artist	8.67	Artists or group names	Coldplay, Radiohead, Queen
Locale	8.03	Geographic situation or locality	British, American, Finnish
Personal	6.80	Words used to manage personal collections	Seen Live, Favourites, My Radio
Instrument	4.83	Sound source	Female vocalists, Piano, Guitar
Unknown	3.79	Unclassifiable	aitch, prda, < 3
Temporal	2.41	Temporal circumstance	80's, 2000, Late Romantic
Form	2.22	Musical form or compositional technique	Ballad, Cover, Fusion
Commercial	1.72	Record label, radio station, etc.	Motown, Guitar Hero, Disney
Verb	1.63	General category of verbs	Chillout, Relax, Wake up
Content	1.03	Emphasis in the message or literary content	Political, Great lyrics, Love song
Expression	0.54	Exclamations	Wow, Yeah, lol

terms of fairly broad categories, but on the other hand makes them hard to adequately define by virtue of this very same quality [37]. This manuscript is devoted to exploring timbre and by extension the way people describe the general sound of a piece of music, hence the idea has been to explore the concepts that lie underneath the genre descriptions. For this reason, *genre* was utilized as the most significant semantic filter. The other discarded categories had their own reasons, for instance *Personal* and *Locale* contents are strongly centered in the individual's perspective, *Artist* contents are redundantly referring to the creator/performer of the music. The rest of the omissions concerned rare categories (e.g. *unknown terms*, *expressions*, *commercial branches or recording companies*) or not explicitly related with timbre (e.g. *musical form*, description of the *lyrics*); these were left out to simplify the results.

Acknowledgements
This study was supported by the Finnish Centre of Excellence in Interdisciplinary Music Research.

Competing interests
The authors declare that they have no competing interests.

Received: 25 March 2011 Accepted: 7 December 2011
Published: 7 December 2011

References

1. O Celma, X Serra, FOAFing the Music: Bridging the semantic gap in music recommendation. *Web Semantics. Science, Services and Agents on the World Wide Web*. **6**(4), 250–256 (2008). [Semantic Web Challenge 2006/2007]. doi:10.1016/j.websem.2008.09.004
2. J Grey, Multidimensional Perceptual Scaling of Musical Timbres. The Journal of the Acoustical Society of America. **61**(5), 1270–1277 (1977). doi:10.1121/1.381428
3. S McAdams, S Winsberg, S Donnadieu, G De Soete, J Krimphoff, Perceptual Scaling of Synthesized Musical Timbres: Common dimensions, specificities and latent subject classes. Psychological Research. **58**(3), 177–192 (1995). doi:10.1007/BF00419633
4. J Burgoyne, S McAdams, A Meta-analysis of Timbre Perception Using Nonlinear Extensions to CLAS-CAL. Computer Music Modeling and Retrieval. Sense of Sounds, 181–202 (2009)
5. JJ Aucouturier, F Pachet, M Sandler, The Way it Sounds: Timbre models for analysis and retrieval of music signals. Multimedia, IEEE Transactions on. **7**(6), 1028–1035 (2005)
6. V Alluri, P Toiviainen, Exploring Perceptual and Acoustical Correlates of Polyphonic Timbre. Music Perception. **27**(3), 223–242 (2010). doi:10.1525/mp.2010.27.3.223
7. P Lamere, Social Tagging and Music Information Retrieval. Journal of New Music Research. **37**(2), 101–114 (2008). doi:10.1080/09298210802479284
8. JJ Aucouturier, E Pampalk, Introduction-From Genres to Tags: A little epistemology of music information retrieval research. Journal of New Music Research. **37**(2), 87–92 (2008). doi:10.1080/09298210802479318
9. C Held, U Cress, Learning by Foraging: The impact of social tags on knowledge acquisition, in *Learning in the Synergy of Multiple Disciplines. 4th European Conference on Technology Enhanced Learning*, Nice, France, (2009)
10. F Hesse, Use and Acquisition of Externalized Knowledge, in *Learning in the Synergy of Multiple Disciplines. 4th European Conference on Technology Enhanced Learning*, (Nice, France: Springer, 2009), p. 5
11. E Chi, Augmented Social Cognition: Using social web technology to enhance the ability of groups to remember, think, and reason, in *Proceedings of the 35th SIGMOD International Conference on Management of Data, Providence*, Rhode Island, USA, (2009)
12. H Kim, S Decker, J Breslin, Representing and Sharing Folksonomies with Semantics. Journal of Information Science. **36**, 57–72 (2010). doi:10.1177/0165551509346785
13. A Mathes, Folksonomies-cooperative Classification and Communication through Shared Metadata. http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html (2004)
14. H Lin, J Davis, Y Zhou, An Integrated Approach to Extracting Ontological Structures from Folksonomies, in *Proceedings of the 6th European Semantic Web Conference on The Semantic Web. Research and Applications*, (Heraklion, Greece: Springer, 2009), p. 668
15. S Deerwester, S Dumais, G Furnas, T Landauer, R Harshman, Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science. **41**(6), 391–407 (1990). doi:10.1002/(SICI)1097-4571(199009)41:63.0.CO;2-9
16. J Bellegarda, Latent Semantic Mapping: Principles & applications. Synthesis Lectures on Speech and Audio Processing. **3**, 1–101 (2007). doi:10.2200/S00048ED1V01Y200609SAP003
17. S Sundaram, S Narayanan, Audio Retrieval by Latent Perceptual Indexing, in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, (IEEE, 2008), pp. 49–52
18. S Dumais, Latent Semantic Analysis. Annual Review of Information Science and Technology (ARIST). **38**, 189–230 (2004)
19. T Eerola, R Ferrer, Setting the Standards: Normative data on audio-based musical features for musical genres, in *Proceedings of the 7th Triennial Conference of European Society for the Cognitive Sciences of Music*, Jyväskylä, Finland, (2009)
20. M Levy, M Sandler, Learning Latent Semantic Models for Music from Social Tags. Journal of New Music Research. **37**(2), 137–150 (2008). doi:10.1080/09298210802479292
21. T Bertin-Mahieux, D Eck, F Maillet, P Lamere, Autotagger: A model for predicting social tags from acoustic features on large music databases. Journal of New Music Research. **37**(2), 115–135 (2008). doi:10.1080/09298210802479250
22. P Rentfrow, S Gosling, Message in a Ballad. Psychological Science. **17**(3), 236–242 (2006). doi:10.1111/j.1467-9280.2006.01691.x
23. M Delsing, T ter Bogt, R Engels, W Meeus, Adolescents' Music Preferences and Personality Characteristics. European Journal of Personality. **22**(2), 109–130 (2008). doi:10.1002/per.665
24. I Popescu, G Altmann, *Word Frequency Studies* (Berlin: Walter de Gruyter, 2009)
25. M Levy, M Sandler, A Semantic Space for Music Derived from Social Tags, in *Proceedings of the 8th International Society for Music Information Retrieval Conference*, vol. 1, ed. by Dixon S, Bainbridge D, Typke R (Vienna, Austria: Österreichische Computer Gesellschaft, 2007), p. 12
26. B Zhang, Q Xiang, H Lu, J Shen, Y Wang, Comprehensive Query-dependent Fusion Using Regression-on-folksonomies: A case study of multimodal music search, in *Proceedings of the seventeen ACM international conference on Multimedia*, Beijing, China: ACM, pp. 213–222 (2009)
27. H Halpin, V Robu, H Shepherd, The Complex Dynamics of Collaborative Tagging, in *Proceedings of the 16th international conference on World Wide Webal Conference on World Wide Web*, Banff, Alberta, Canada: ACM, p. 220 (2007)
28. J Brank, M Grobelnik, D Mladenic, Automatic Evaluation of Ontologies, in *Natural Language Processing and Text Mining*, ed. by Kao A, RPoteet S (USA: Springer, 2007)
29. J Siskind, Learning Word-to-meaning Mappings, in *Models of Language Acquisition. Inductive and deductive approaches*, USA: Oxford University Press, pp. 121–153 (2000)
30. G Zipf, *Human Behavior and the Principle of Least Effort*, *An introduction to human ecology* (addison-wesley press, 1949)
31. J Gower, P Legendre, Metric and Euclidean Properties of Dissimilarity Coefficients. Journal of Classification. **3**, 5–48 (1986). doi:10.1007/BF01896809
32. M Walesiak, A Dudek, *ClusterSim. Searching for optimal clustering procedure for a data set* (2011). http://CRAN.R-project.org/package=clusterSim [R package version 0.39-2]
33. R Development Core Team, *R A language and environment for statistical computing*, (R Foundation for Statistical Computing, Vienna, Austria, 2009). http://www.R-project.org [ISBN 3-900051-07-0]
34. A Jain, R Dubes, *Algorithms for Clustering Data* (Englewood Cliffs, NJ: Prentice Hall, 1988)

35. P Langfelder, B Zhang, S Horvath, *DynamicTreeCut. Methods for detection of clusters in hierarchical clustering dendrograms* (2009). http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/BranchCutting/ [R package version 1.20]
36. G Tzanetakis, P Cook, Musical Genre Classification of Audio Signals. IEEE Transactions on Speech and Audio Processing. **10**(5), 293–302 (2002). doi:10.1109/TSA.2002.800560
37. R Gjerdingen, D Perrott, Scanning the Dial: The rapid recognition of music genres. Journal of New Music Research. **37**(2), 93–100 (2008). doi:10.1080/09298210802479268
38. M Hoffman, D Blei, P Cook, Easy as CBA: A simple probabilistic model for tagging music, in *Proceedings of the 10th International Society for Music Information Retrieval Conference*, Kobe, Japan, (2009)
39. MI Mandel, DPW Ellis, A Web-based Game for Collecting Music Metadata. Journal of New Music Research. **37**(2), 151–165 (2008). doi:10.1080/09298210802479300
40. D Turnbull, L Barrington, D Torres, G Lanckriet, Towards Musical Query-by-semantic-description Using the CAL500 Data Set, in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, New York, NY, USA: ACM, pp. 439–446 (2007)
41. K Jacobson, M Sandler, B Fields, Using Audio Analysis and Network Structure to Identify Communities in On-line Social Networks of Artists, in *Proceedings of the 9th International Society for Music Information Retrieval Conference*, ed. by Bello JP, Chew E, Turnbull D (Philadelphia, USA, 2008), pp. 269–274
42. C Laurier, O Meyers, J Serrà, M Blech, P Herrera, X Serra, Indexing Music by Mood: Design and integration of an automatic content-based annotator. *Multimedia Tools and Applications*. **48**, 161–184 (2010). [Springerlink link: http://www.springerlink.com/content/jj01750u20267426]. doi:10.1007/s11042-009-0360-2
43. J Bello, J Pickens, A Robust Mid-level Representation for Harmonic Content in Music Signals, in *Proceedings of the 6th International Society for Music Information Retrieval Conference*, London, UK, pp. 304–311 (2005)
44. S Chu, S Narayanan, CC Kuo, Environmental Sound Recognition With Time-frequency Audio Features. Audio, Speech, and Language Processing, IEEE Transactions on. **17**(6), 1142–1158 (2009)
45. E Pampalk, A Rauber, D Merkl, Content-based Organization and Visualization of Music Archives, in *Proceedings of the tenth ACM international conference on Multimedia*, Juan les Pins, France: ACM, p. 579 (2002)
46. G Peeters, A Large Set of Audio Features for Sound Description (Similarity and Classification) in the CUIDADO Project. *CUIDADO IST Project Report* 1–25 (2004)
47. P Juslin, Cue Utilization in Communication of Emotion in Music Performance: Relating performance to perception. Journal of Experimental Psychology. Human perception and performance., 6 **26**, 1797–1812 (2000)
48. P Laukka, P Juslin, R Bresin, A Dimensional Approach to Vocal Expression of Emotion. Cognition & Emotion. **19**(5), 633–653 (2005). doi:10.1080/02699930441000445
49. K Jensen, *Timbre Models of Musical Sounds* (Department of Computer Science, University of Copenhagen, 1999)
50. W Sethares, *Tuning, Timbre, Spectrum, Scale* (Springer Verlag, 2005)
51. J Bello, C Duxbury, M Davies, M Sandler, On the use of Phase and Energy for Musical Onset Detection in the Complex Domain. Signal Processing Letters, IEEE. **11**(6), 553–556 (2004). doi:10.1109/LSP.2004.827951
52. O Lartillot, P Toiviainen, T Eerola, A Matlab Toolbox for Music Information Retrieval, in *Data Aalysis, Machine Learning and Applications*, ed. by Preisach C, Burkhardt H, Schmidt-Thieme L, Decker R (Berlin, Germany: Springer, 2008), pp. 261–268. Studies in Classification, Data Analysis, and Knowledge Organization
53. C Harte, M Sandler, M Gasser, Detecting Harmonic Change in Musical Audio, in *Proceedings of the 1st ACM Workshop on Audio and Music Computing Multimedia*, Santa Barbara, CA, USA: ACM, p. 26 (2006)
54. I Guyon, A Elisseeff, An Introduction to Variable and Feature Selection. Journal of Machine Learning Research. **3**, 1157–1182 (2003)
55. G Tzanetakis, P Cook, Manipulation, Analysis and Retrieval Systems for Audio Signals, PhD thesis, (Princeton University, Princeton, NJ, 2002)
56. J Fox, G Monette, Generalized Collinearity Diagnostics. Journal of the American Statistical Association. **87**(417), 178–183 (1992). doi:10.2307/2290467
57. L Breiman, Random Forests. Machine Learning. **45**, 5–32 (2001). doi:10.1023/A:1010933404324
58. B Ripley, *Pattern Recognition and Neural Networks*, (Cambridge: Cambridge University Press, 1996)
59. K Archer, R Kimes, Empirical Characterization of Random Forest Variable Importance Measures. Computational Statistics & Data analysis. **52**(4), 2249–2260 (2008). doi:10.1016/j.csda.2007.08.015
60. H Pang, A Lin, M Holford, B Enerson, B Lu, M Lawton, E Floyd, H Zhao, Pathway Analysis Using Random Forests Classification and Regression. Bioinformatics. **22**(16), 2028 (2006). doi:10.1093/bioinformatics/btl344
61. L Nieweglowski, *CLV: Cluster validation techniques* (2009). http://CRAN.R-project.org/package=clv [R package version 0.3-2]
62. J Gower, Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis. Biometrika. **53**(3-4), 325 (1966). doi:10.1093/biomet/53.3-4.325
63. M Cox, T Cox, *Multidimensional Scaling: Handbook of data visualization*, (USA: Chapman & Hall, 2001)
64. I Borg, P Groenen, *Modern Multidimensional Scaling: Theory and applications*, (Springer Verlag, 2005)
65. JJ Aucouturier, B Defreville, F Pachet, The Bag-of-frames Approach to Audio Pattern Recognition: A sufficient model for urban soundscapes but not for polyphonic music. The Journal of the Acoustical Society of America. **122**(2), 881–891 (2007). doi:10.1121/1.2750160
66. C McKay, I Fujinaga, Musical Genre Classification: Is it worth pursuing and how can it be improved. in *Proceedings of the 7th International Society for Music Information Retrieval Conference* 101–6 (2006)
67. D Balota, J Cohane, Semantic Memory, in *Learing and Memory: A comprehensive reference, Volume 2 Cognitive Pyshology of Memory*, ed. by Byrne JH, III HLR (Oxford, UK: Academic Press, 2008), pp. 511–534
68. G Sandell, Macrotimbre: Contribution of attack, steady state, and verbal attributes. The Journal of the Acoustical Society of America. **103**, 2966 (1998)
69. R Ferrer, Embodied Cognition Applied to Timbre and Musical Appreciation: Theoretical Foundation. British Postgraduate Musicology. **X,** http://www.bpmonline.org.uk/bpm10/ferrer_rafael-embodied_cognition_applied_to_timbre_and_musical_appreciation_theoretical_foundation.pdf (2009)
70. Kao A, Poteet SR (eds.), *Natural Language Processing and Text Mining* (Springer Verlag, 2006)
71. C Manning, H Schütze, *Foundations of Statistical Natural Language Processing* (MIT Press, 2002)
72. S Bird, E Klein, E Loper, *Natural Language Processing with Python* (Oreilly & Associates Inc, 2009)
73. W Francis, H Kucera, *Brown Corpus. A Standard Corpus of Present-Day Edited American English, for use with Digital Computers* (Department of Linguistics, Brown University, Providence, Rhode Island, USA, 1979)
74. Fellbaum C (ed.), *WordNet: An electronic lexical database* (Language, speech, and communication, Cambridge, Mass: MIT Press, 1998)

doi:10.1186/1687-4722-2011-11
Cite this article as: Ferrer and Eerola: **Semantic structures of timbre emerging from social and acoustic descriptions of music.** *EURASIP Journal on Audio, Speech, and Music Processing* 2011 **2011**:11.