A SEMANTIC SPACE FOR MUSIC DERIVED FROM SOCIAL TAGS

Mark Levy
Centre for Digital Music
Queen Mary, University of London
Mile End Road, London E1 4NS
mark.levy@elec.qmul.ac.uk

Mark Sandler
Centre for Digital Music
Queen Mary, University of London
Mile End Road, London E1 4NS
mark.sandler@elec.qmul.ac.uk

ABSTRACT

In this paper we investigate social tags as a novel high-volume source of semantic metadata for music, using techniques from the fields of information retrieval and multivariate data analysis. We show that, despite the ad hoc and informal language of tagging, tags define a low-dimensional semantic space that is extremely well-behaved at the track level, in particular being highly organised by artist and musical genre. We introduce the use of Correspondence Analysis to visualise this semantic space, and show how it can be applied to create a browse-by-mood interface for a psychologically-motivated two-dimensional subspace representing musical emotion.

1 INTRODUCTION

Social tags are of interest as a potential high-volume source of descriptive metadata for music. Such metadata can either be used directly to drive search applications, as already happens to some extent in the commercial domain, or as a source of groundtruth to train audio content-based classification and search engines. In the academic literature, comparable text metadata for music has previously been found by mining web-pages such as blogs and music reviews [2, 19, 9]. Although some interesting preliminary results have been reported, two significant problems are associated with this approach. Firstly, text retrieved from the web is often noisy, i.e. it unavoidably contains a great deal of irrelevant content. Secondly, for computational reasons, and because the noise problem becomes insuperable, text has to be mined on a per-artist rather than per-track basis: as a result it offers only low-quality groundtruth for learning the characteristics of audio content. Social tags as applied to individual tracks appear to offer a solution to both of these issues. At the time of writing there is no previous relevant academic literature on social tags for music.

The tags discussed here were all applied to individual tracks. They were aggregated from the last.fm [1] and MyStrands [2] web services during January and February 2007.

[1] http://ws.audioscrobbler.com
[2] https://www.musicstrands.com

Music in general has no semantics, in the strict sense of representing or being 'about' something, and, perhaps as a result, tags for music are often discursive. Of some 45,000 distinct tags in our dataset, over a third of consist of three or more words, while over 10% contain 5 or more words: these are frequently complete phrases. In our experiments we therefore treat tags as regular text, tokenizing them with a standard stop-list (to remove common words such as 'it', 'and', 'the', etc.). We then create a conventional document-term matrix, tabulating the number of occurrences of each word in tags applied to each track. We do not use a stemmer, because of the idiosyncratic vocabulary of social tagging and the large number of words used as proper nouns (particularly artist names). Working with words rather than tags nonetheless goes some way towards capturing the common meaning of alternate forms such as 'female vocalist', 'female vocals', 'good female vocals', 'sexy female vocals', 'lovely female vocals', etc.

The language of tags for music is ad hoc and often highly informal, as shown by the following few tags selected at random: 'all my hope is gone', 'oregon trips', 'my favourite muse songs', 'french-canadian', 'Tool Mix', 'comp1', 'ragga rhythm', 'Dave Brubeck Quartet', 'american wedding', 'fora do mundo', 'space trucking', 'right in two', 'desert island songs - songs which keep me alive or otherwise enraged', 'heard on 96wave', 'put on mikey cds'. We might indeed question whether the collaborative tagging model can be applied successfully to music: beyond standard metadata like artist or title, which are already likely to be known in any real application, it may be far from obvious which tags are appropriate for any particular track.

This study provides evidence, however, that despite the vagaries of individual tags, patterns of co-occurrences of words in tags can reveal terms or combinations of terms which are significantly grounded in the music they describe (rather than expressing arbitrary personal reactions) and generalisable across tracks. In particular we show that tags define a vector space with highly attractive properties for music retrieval, and which appears to have genuine semantics.

Table 1. Top terms describing Portishead

Tags	Web-mined text
trip-hop	cynical
electronic	produced
portishead	smooth
female vocalists	dark
downtempo	particular
alternative	loud
mellow	amazing
chillout	vocal
sad	unique
90s	simple

2 THE SEMANTIC SPACE OF TAGS

2.1 Tags vs web-mined text

The only comparable source of high-volume metadata for music explored to date is web-mined text. This is typically retrieved by searching for pages that appear to be relevant to a particular artist, and then attempting to retain only terms that relate to their music [2, 19]. The resulting text is inherently noisy on two levels. Firstly, the pages retrieved by any automated system are not guaranteed to be relevant (in particular when an artist's name has other meanings), and come from a variety of kinds of source, each with its own characteristic vocabulary. Secondly, in general only a small unknown part of the content of each page will refer directly to music of interest. One consequence of the inevitable inclusion of irrelevant terms is that the vocabulary size explodes. A typical web crawl reported in [9] found over 200,000 terms for a set of 200 well-known artists. In contrast, we found less than 13,500 distinct tags for tracks by the same set of artists. Such a comparison is necessarily informal, because of the difficulty of comparing the sizes of the input data sources (50 web pages vs tags from the order of 100 different users for each artist). More importantly, however, web-mining appears to be impractical as a source of metadata at the track level, as the problems of noise multiply still further.

The vocabulary of tags is different from web-mined text not only in size, but also in character, as illustrated in Table 1, which compares the ten most widely applied tags for the group Portishead with the top web-mined adjectives given in [18]. We observe that, in contrast to the tags, as many as half of the web-mined adjectives ('cynical', 'produced', 'particular', 'amazing', 'unique') are very unlikely to be grounded in the music of this particular group.

2.2 Catalogue organisation

A natural question to ask of any new representation for collections of music is to what extent it respects a traditional recording catalogue organisation, in which tracks are grouped by artist and genre. A good deal of work has been devoted to addressing this issue in relation to low-level audio features, with the problem cast, perhaps somewhat unhelpfully, as a pair of classification tasks (see [14] for a recent review). The conclusion, after several years of research, is that current low-level feature sets lead to a representation that is only weakly structured by artist and genre [1, 12].

While individual genre tags attached to tracks are not reliable in general, we can reasonably ask whether the semantic space defined by co-occurrences of terms does capture concepts of artist and genre, and, if so, what dimensionality is required to represent these concepts effectively in a vector space.

2.3 Retrieval experiments

The results reported here are based on 236,974 tags collected for 5,722 tracks drawn from all of the mainstream genres. The total vocabulary size, after tokenizing with a standard stop-list, was 24,160 distinct words, of which 13,312 were applied to 2 or more tracks, and 3,992 to 10 or more. The choice of tracks was seeded with a set of artists balanced across the mainstream musical genres, and an influential list of words associated with musical expression from [8], expanded with synonyms from their WordNet synsets [6]. The scale of the dataset was chosen to give reasonable coverage across tracks and terms without becoming computationally intractable.

To investigate the organisation of the tag space, and to give a quantitative comparison with web-mined text for retrieval applications, we replicated the experimental set-up used in [9], in which similarities were calculated for a set of 224 well-known artists split equally over 14 genres. From our dataset, we found tags for 1196 tracks by 223 of the 224 artists, with between 4 and 12 tracks for each artist. We measured retrieval performance over this dataset, using each track in turn as a query.

We created a document-term matrix $\mathbf{X} = \{x_{ij}\}$ with simple tf-idf values

$$x_{ij} = \mathrm{tf}_{ij} \log \frac{N}{\mathrm{df}_j} \quad (1)$$

where tf_{ij} is the number of times that term j appears in the tags for track i, df_j is the number of tracks whose tags contain term j, and N is the total number of tracks. We then used cosine distance to compare the term vectors for each track.

We compared three different approaches when calculating the term frequencies tf_{ij}: weighting them to reflect the number of users who had applied the term to the track in question; ignoring the number of users; and using the qtag [3] part-of-speech (POS) tagger to restrict the terms considered to adjectives only. The weights in the first approach are based on unexplained 'counts' published by last.fm, and should therefore be considered ad hoc: we use them only to get an idea of the potential value of including such information.

We extended our experiment by using Latent Semantic Analysis (LSA) [5] to reduce the dimensionality of the

[3] http://www.english.bham.ac.uk/staff/omason/software/qtag.html

feature vector for each track: we calculated the rank-k Singular Value Decomposition of the document-term matrix $\tilde{\mathbf{X}}_\mathbf{k} = \mathbf{U_k S_k V_k^T}$ for a range of ranks, and based our similarities on the reduced vectors $\mathbf{U_k S_k}$. The SVD was calculated over our full dataset.

2.4 Results

We show per-word mean Average Precision (AP), over the sets of artist and genre labels, in Fig. 2. The AP for a particular query is calculated as

$$\mathrm{AP} = \frac{\sum_{r=1}^{N} P(r)\mathrm{rel}(r)}{R} \qquad (2)$$

where $P(r)$ is the precision at rank r, $\mathrm{rel}(r)$ is 1 if the document at rank r is relevant and 0 otherwise, R is the total number of relevant documents, and N is the total number of documents in the collection. AP therefore measures the average precision over the ranks at which each relevant track is retrieved. The per-word mean AP for a particular genre or artist label is the mean AP over all queries labelled with that term. Besides being a standard IR performance metric, mean AP rewards the retrieval of relevant tracks ahead of irrelevant ones, and is consequently an extremely good indicator of how our vector space is organised.

Vectors based on term frequencies using all terms applied to these tracks clearly perform better than those based on adjectives only. The benefit of taking user weights into account is somewhat less clear, improving genre precision at all ranks, but having a negligible effect on artist precision above rank 60. Using the weights has the effect of emphasizing the 'majority view' for the relevance of a particular term to any given track, and a possible interpretation of the results is that genre precision improves artificially as minority opinions are discounted.

Using the full term vectors, the genre precision reaches 80%, and the artist precision 61%. For historical reasons, [9] gives genre performance as a Leave One Out 1-nearest neighbour classification rate (effectively showing precision at rank 2) of 87%. Using our full term vectors, the LOO genre classification rate was 95%. The rate using the nearest track by a different artist to the query was 83%. Using LSA at ranks 30 and above consistently improves the genre precision, and with the weighted counts the maximum is over 82% at rank 20. LSA improves artist precision at ranks 200 and above, with a maximum of 63% at rank 300.

There is, of course, no 'right answer' for the precision that we would hope for when doing retrieval in a vector space for tracks, because songs by other artists or from different genres can quite reasonably be considered very similar to any given query. On the other hand, organisation by artist and genre is well understood by music lovers, and the lack of such organisation in low-level feature representations appears to be a major barrier to their acceptance in practical applications. Our view is that a reduced-dimension semantic space defined by tags may be an ideal

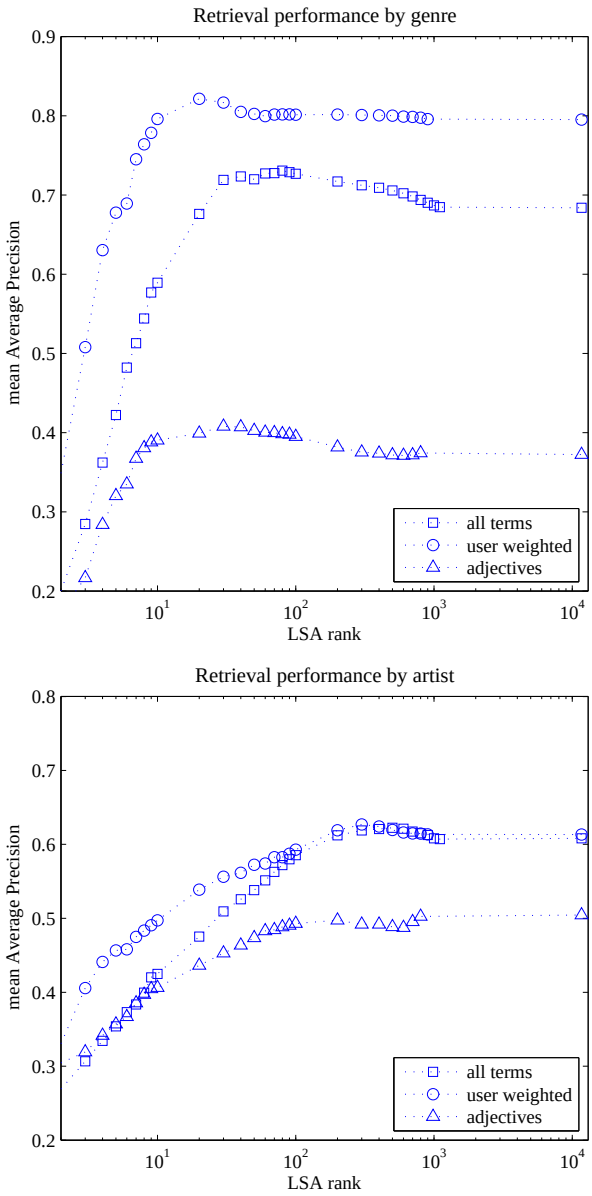

Figure 1. Retrieval performance of tag term vectors

representation, capturing rich descriptions for each track based on a very large vocabulary, but also respecting traditional catalogue organisation with high precision.

3 THE SUB-SPACE OF MUSICAL EMOTION

3.1 The dimensional representation of emotion

A significant psychological literature has investigated a so-called 'dimensional' approach to the representation of emotion in general [15, 11, 13, 16], and with particular regard to emotional responses to music [8, 17]. A focus of these studies has been to map relevant terms onto low-dimensional spaces with named axes, intended to correspond to internal human representations of emotion. This has led to a widely-accepted emotional space with 2 primary significant dimensions, most frequently referred to

as *valence* (from pleasant to unpleasant) and *arousal* (from mild to intense). A further two secondary dimensions have been identified in some studies, but are generally regarded as less significant [4].

Social tags provide a unique source of high-volume, non-invasive data from which to study emotional responses to music. We investigate elsewhere the extent to which tagging conforms to, or departs from, established models for emotion in music [10], noting in particular that the vocabulary for mood and emotion arising organically from the user community in tags differs significantly from that commonly used in controlled psychological experiments.

Just as we can compare tracks by co-occurrence of terms, we can compare terms by co-occurrence over tracks: we simply use the columns of the document-term matrix to create track vectors representing terms. Applying LSA to the matrix as in Section 2.3, the dimensionally-reduced term vectors are given by $\mathbf{VS}$. We selected all the words in our dataset that were applied to at least 50 tracks, and which appear to relate to mood or musical expression, resulting in a list of 57 emotion words. We then trained a Self-Organising Map on the track vectors for these words, using LSA at rank 40, and mapped each word onto its best-matching unit in the trained SOM. The resulting configuration of terms is shown in Table 2, and gives an impression of the organisation of emotion words in our semantic space. This shows some relationship to the traditional arousal-valence axes, with valence increasing broadly from left to right and arousal from top to bottom.

3.2 Correspondence Analysis for visualisation

Correspondence Analysis (CA) is a well-established technique of dimension reduction used primarily for visualising multivariate categorical data [3, 7]. It has two properties that make it extremely attractive for our purposes:

1. it enables the visualisation of two sets of cross-tabulated variables (in our case tracks and semantic terms) in the same low-dimensional space;
2. Euclidean distances in the visualisation represent distributional (χ^2) distances in the data.

CA is a generalised form of Principal Component Analysis suitable for application to an M by N table of co-occurrence data $\mathbf{F}$, where $\mathbf{F}$ has been normalised to have total sum 1. CA finds a low-dimensional projection of $\mathbf{F}$ which optimally preserves χ^2-distances between row and column *profiles*

$$\mathbf{f}^{c|r=i} = \left(\frac{f_{i1}}{f_i}, ..., \frac{f_{iN}}{f_i}\right)$$
$$\mathbf{f}^{r|c=j} = \left(\frac{f_{1j}}{f_j}, ..., \frac{f_{Mj}}{f_j}\right)$$

where f_i, f_j are the row and column sums respectively, i.e. $f_i = \sum_{j=1}^{N} f_{ij}$ and $f_j = \sum_{i=1}^{M} f_{ij}$.

The χ^2-metric between row profiles is a weighted Euclidean distance where the weight for each column is given by $\frac{1}{f_j}$; the metric between column profiles is weighted similarly by $\frac{1}{f_i}$. The χ^2-metric has the desirable property that distances between columns (tag words) do not change if columns (tracks) with identical profiles (normalised term vectors) are amalgamated, and vice versa.

We compute a generalised SVD of $\mathbf{F}$

$$\tilde{\mathbf{F}} = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}' \tag{3}$$

where $\boldsymbol{\Delta}$ is a diagonal matrix, and $\mathbf{U}$ and $\mathbf{V}$ satisfy

$$\mathbf{U}'(\mathbf{F^r})^{-1}\mathbf{U} = \mathbf{V}'(\mathbf{F^c})^{-1}\mathbf{V} = \mathbf{I} \tag{4}$$

where $\mathbf{F_r}$ and $\mathbf{F_c}$ are diagonal matrices of the row and column sums respectively. Co-ordinates $\mathbf{S}$ of row profiles onto axes $\mathbf{U}$ are then given by

$$\mathbf{f}^{c|r} = \mathbf{US} \tag{5}$$

where

$$\mathbf{S} = \boldsymbol{\Delta}\mathbf{V}'(\mathbf{F^c})^{-1} \tag{6}$$

Co-ordinates $\mathbf{T}$ of column profiles onto axes $\mathbf{V}$ are given similarly by

$$\mathbf{f}^{r|c} = \mathbf{VT} \tag{7}$$

where

$$\mathbf{T} = \boldsymbol{\Delta}\mathbf{U}'(\mathbf{F^r})^{-1} \tag{8}$$

Row and column profiles can then be plotted in the same d-dimensional space, taking only the first d co-ordinates of $\mathbf{S}$ and $\mathbf{T}$. Although it is not meaningful in general to interpret row-column distances in this visualisation, it does show the relative distances of a single row (track) to all the columns (emotion words), and vice versa.

This suggests a natural application of CA with $d = 2$ to create a browse-by-mood interface to a collection of tracks, using a normalised portion of our document-term matrix, with row profiles representing tracks and columns restricted to mood terms. The resulting plot of tracks and terms shows mood words in a meaningful relationship, while tracks in any particular region of the space should be well described by nearby words.

3.3 Evaluation

We tested this approach on a small list of 14 mood words, consisting of the subset of terms from the classic list of musical emotions given in [8] which were applied to at least 50 tracks in our dataset, and the subset of 3176 tracks tagged with at least one of these words. In Figure 2 we show the resulting positions of the terms and tracks. We evaluate the organisation of the plot by calculating the mean AP for each mood word, where we consider a track to be relevant to its closest mood word in the plot if it has been tagged with it.

To comply with the allowable interpretation of distances in CA, we take the mean AP for each term only over tracks which are closer to it in the CA space than they are to any other term (so each track in the dataset gets considered exactly once). The results are given in Table 3, showing

Table 2. Emotion terms mapped onto a SOM

soft mellow		chill	relax chillout	sweet		summer happy
love	romantic	relaxing	smooth dreamy	downtempo		fun
beautiful	melancholic calm	soothing		melodic	feelgood	upbeat catchy
slow sad quiet	sleep	pretty	lovely bittersweet	uplifting nice	fast	funky
melancholy emotional		night		singalong	heavy	cool
moody depressing		haunting	intense	energetic clean		sexy sex
dark	experimental atmospheric	ethereal	silent intensity	angry	psychedelic	party

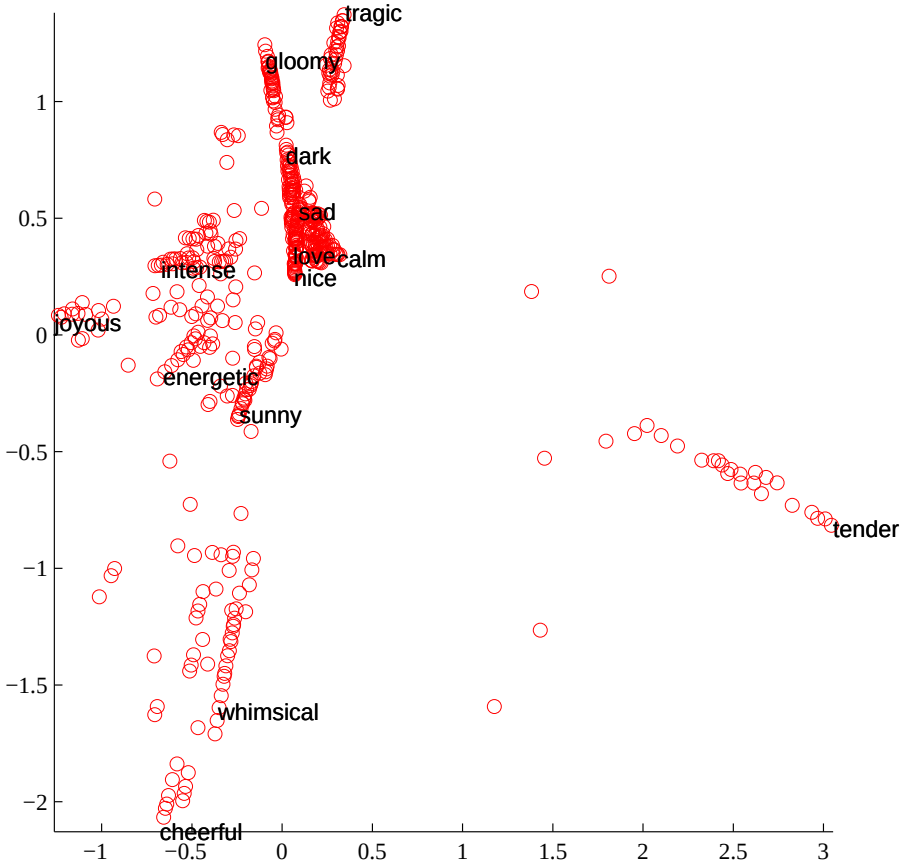

Figure 2. CA joint plot of mood words and tracks

Table 3. Mean Average Precision for mood words

Mood	mean AP
calm	0.998
cheerful	1.000
dark	0.947
energetic	0.925
gloomy	0.987
intense	0.924
joyous	1.000
love	1.000
nice	0.939
sad	0.965
sunny	0.942
tender	1.000
tragic	1.000
whimsical	0.919

that the plot partitions the space almost perfectly by this measure, although it is important to note that precision is measured here against words found in tags themselves, not a verifiable external source of information.

4 CONCLUSIONS

Despite the ad hoc and informal usage typical of social tagging, tags are highly effective in capturing music similarity. Although they are often discursive, tags for music appear to capture sensible attributes grounded in individual tracks, defining a well-behaved similarity space with an effective dimensionality of around 10^2. Given these encouraging results as to the usefulness of tags as music metadata, and the low dimensionality of an effective feature space for music similarity, future work includes the use of tags as groundtruth for joint feature-annotation models for music.

5 ACKNOWLEDGMENTS

The authors would like to thank Gunter Kreutz for illuminating discussions about psychological approaches to representing musical emotion.

This research was supported by EPSRC grants GR/S84750/01 (Hierarchical Segmentation and Semantic Markup of Musical Signals) and EP/E017614/1 (Online Music Recognition And Searching).

6 REFERENCES

[1] J.-J. Aucouturier. *Ten experiments on the modelling of polyphonic timbre*. PhD thesis, University of Paris 6, 2006.

[2] S. Baumann and O. Hummel. Using cultural metadata for artist recommendations. In *Proc. Wedelmusic*, 2003.

[3] J.-P. Benzécri. Histoire et préhistoire de l'analyse des données. *Cahiers de l'Analyse des Données*, 2:9–40, 1977.

[4] G. L. Collier. Beyond valence and activity in the emotional connotations of music. *Psychology of Music*, 35(1):110–131, 2007.

[5] S. Deerwester, S. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the Society for Information Science*, 1990.

[6] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[7] M. J. Greenacre. *Theory and applications of correspondence analysis*. Academic Press, 1984.

[8] K. Hevner. Experimental studies of the elements of expression in music. *American Journal of Psychology*, 48:246–68, 1936.

[9] P. Knees, E. Pampalk, and G. Widmer. Artist classification with web-based data. In *Proc. ISMIR*, 2004.

[10] G. Kreutz and M. Levy. Emotion annotations for popular music in internet communities. In preparation.

[11] C. E. Osgood, G. J. Succi, and P. H. Tannenbaum. *The measurement of meaning*. University of Illinois Press, 1957.

[12] E. Pampalk. *Computational models of music similarity and their application to music information retrieval*. PhD thesis, Vienna University of Technology, 2006.

[13] J. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–78, 1980.

[14] N. Scaringella, G. Zoia, and D. Mlynek. Automatic genre classification of music content: a survey. *IEEE Signal Processing Magazine*, 23(2):133–141, 2006.

[15] H. Schlosberg. Three dimensions of emotion. *Psychological Review*, 61(2):81–8, March 1954.

[16] P. Shaver, J. Schwartz, D. Kirson, and C. O'Connor. Emotion knowledge: further exploration of a prototype approach. *Journal of Personality and Social Psychology*, 52(6):1061–86, 1987.

[17] L. Wedin. Dimension analysis of emotional expression in music. *Swedish Journal of Musicology*, 51:119–140, 1969.

[18] B. Whitman. Semantic rank reduction of music audio. In *Proc. IEEE WASPAA*, 2003.

[19] B. Whitman. *Learning the Meaning of Music*. PhD thesis, MIT, 2005.