IDENTIFYING WORDS THAT ARE MUSICALLY MEANINGFUL

David Torres[1], **Douglas Turnbull**[1], **Luke Barrington**[2] **Gert Lanckriet**[2]
Dept. of Computer Science and Engineering[1]
Dept. of Electrical and Computer Engineering[2]
University of California, San Diego

ABSTRACT

A musically meaningful vocabulary is one of the keystones in building a computer audition system that can model the semantics of audio content. If a word in the vocabulary is inconsistently used by human annotators, or the word is not clearly represented by the underlying acoustic representation, the word can be considered as *noisy* and should be removed from the vocabulary to denoise the modeling process. This paper proposes an approach to construct a vocabulary of predictive semantic concepts based on *sparse canonical component analysis* (sparse CCA) . Experimental results illustrate that, by identifying musically meaningful words, we can improve the performance of a previously proposed computer audition system for music annotation and retrieval.

1 INTRODUCTION

Over the past two years we have been developing a computer audition system that can annotate songs with semantically meaningful words and retrieve relevant songs based on a text query. This system learns a joint probabilistic model between a vocabulary of words and acoustic feature vectors using a heterogeneous data set of song and song annotations. However, if a specific word is inconsistently used when annotating songs or is not represented well by the acoustic features, the system will not be able to model this *noisy* word well. In this paper, we explore the problem of *vocabulary selection* for semantic music annotation and retrieval. We will consider two concepts, *human agreement* and *acoustic correlation*, as indicators for picking candidate words.

Previously, we collected semantic annotations of music using various methods: text-mining song reviews [15], conducting a human survey [16], and exploring the use of a human computation game [17, 19]. In all cases, we are forced to choose a vocabulary using ad-hoc methods. For example, text-mining the song reviews resulted in a list of over 1,000 candidate words which the authors manually pruned if there was a general consensus that a word was not 'musically-relevant'. To collect the survey and game data, we built, a priori, a two-level hierarchical vocabulary by first considering a set of high-level semantic categories ('Instrumentation', 'Emotion', 'Vocal Characteristic', 'Genre') and then listing low-level words ('Electric Guitar', 'Happy', 'Breathy', 'Bebop') for each semantic category. In both cases, a vocabulary required manual construction and included some *noisy* words that degraded the performance of our computer audition system.

In this paper, we highlight two potential reasons why a word causes problems for our system. The first is related to the notion that aspects of music are subjective. That is, two individual listeners will use different words to describe the same piece of music. For example, a pre-teen girl might consider a Backstreet Boys song to be 'touching and powerful' whereas a dj at an indie radio station may consider it 'abrasive and pathetic'. If we consider them as one population, the annotations will be in conflict with one another. To address this issue we introduce in Section 2 a measure of *human agreement* to evaluate how consistently our population uses a word to label a large set of songs.

A second reason a word may be hard to model involves the expressive power of our chosen audio feature representation. For example, if we are interested in words related to long-term music structure (e.g., '12-bar blues') and we only represent the audio using short-term (< 1 sec) audio feature vectors, we may be unable to model such concepts. Another example is words that relate to a geographical association (e.g., 'British Invasion', 'Woodstock') which may have strong cultural roots, but are poorly represented in the audio content.

Given an audio feature representation, we would like to identify the words that are represented well by the audio content before we try to model such words. To do this we propose the use of a method based on *canonical correlation analysis* (CCA) to measure *acoustic correlation*.

CCA is a method of exploring dependencies across two different, but related, vector spaces and has been used in applications dealing with multi-language text analysis [18], learning a semantic representations between images and text [4], and localizing pixels which are correlated with audio from a video stream [6]. Similar to how principal component analysis (PCA) finds informative directions in one feature space by maximizing the variance of projected data, CCA finds directions (projections of the data) across multiple spaces that maximize correlation.

Given music data represented in both a semantic feature space and an acoustic feature space, we propose that

these directions of high correlation can be used to find words that are strongly characterized by an audio representation. We do so by imposing constraints on CCA that explicitly turn it into a vocabulary selection mechanism. This CCA variant is called *sparse CCA*.

2 HUMAN AGREEMENT

Recently, we collected the Computer Audition Lab 500 (CAL500) data set [16]: 500 songs by 500 unique artists each of which has been annotated according to a 173-word vocabulary by a minimum of three individuals. Most of the participants were paid, American, undergraduate students and the testing was conducted in a computer laboratory at UC San Diego. We purposely collected multiple annotations for songs so that we could gauge how consistently a population of college students label music.

Using this data set, we can calculate a statistic we refer to as *human agreement* for each word in our vocabulary. The agreement of a word-song pair (w, s) is:

$$A_{w,s} = \frac{\#(\text{positive associations})_{w,s}}{\#(\text{ annotations})_s}. \quad (1)$$

For example, if 3 out of 4 students label Elvis Presley's 'Heartbreak Hotel' as being a 'blues' songs then $A_{\text{'blues', 'heartbreak hotel'}} = 0.75$. We calculate the human agreement for a word by averaging over all the songs in which at least one subject has used the word to describe the song. This can be written as

$$A_w = \frac{\sum_s A_{w,s}}{\sum_s I[A_{w,s} > 0]} \quad (2)$$

where I is an indicator function that is 1 if $A_{w,s}$ is greater then zero, and 0 otherwise. That is, all word-song pairs are valid except the word-song pair that nobody associates with one another. We expect human agreement to be close to 1 for more 'objective' words such as words associated with instrumentation ('cow bell'), and close to 0 for words that are more 'subjective' such as those that related to song usages ('driving music').

3 ACOUSTIC CORRELATION WITH CCA

Canonical Correlation Analysis, or CCA, is a method of exploring dependencies between data which are represented in two different, but related, vector spaces. For example, consider a set of songs where each song is represented by both a *semantic annotation vector* and an *audio feature vector*. An annotation vector for a song is a real-valued (or binary) vector where each element represents the strength of association (e.g., Equation 1) between the song and a word from our vocabulary. An audio feature vector is a real-valued vector of statistics calculated from the digital audio signal. It is assumed that the two spaces share some joint information which can be captured in the form of correlations between the music data that live in these spaces. CCA finds a one-dimensional projection of the data in each space such that the correlations between the projections is maximized.

More formally, consider two data matrices, A and S, from two different feature spaces. The rows of A contain music data represented in the audio feature space $\mathcal{A}$. The corresponding rows of S contain the music data represented in the semantic annotation space $\mathcal{S}$ (e.g., annotation vectors). CCA seeks to optimize

$$\max_{\mathbf{w_a} \in \mathcal{A}, \mathbf{w_s} \in \mathcal{S}} \quad \mathbf{w_a'} \mathbf{A'} \mathbf{S} \mathbf{w_s} \quad (3)$$
$$\text{s.t.} \quad \mathbf{w_a'} \mathbf{A'} \mathbf{A} \mathbf{w_a} = 1$$
$$\mathbf{w_s'} \mathbf{S'} \mathbf{S} \mathbf{w_s} = 1.$$

The objective in Problem 3 is the dot product between projections of data points. By itself, the objective function is unbounded since we can scale the $\mathbf{w}$ terms arbitrarily. Thus, we add the constraints to bound the length of the $\mathbf{w}$ terms and ensure the result is proportional to a correlation score.

By analyzing the Lagrangian dual function of Problem 3, we find that it is equivalent to a pair of maximum eigenvalue problems,

$$\mathbf{S_{ss}^{-1} S_{sa} S_{aa}^{-1} S_{as} w_s} = \lambda^2 \mathbf{w_s} \quad (4)$$
$$\mathbf{S_{aa}^{-1} S_{as} S_{ss}^{-1} S_{sa} w_a} = \lambda^2 \mathbf{w_a} \quad (5)$$

where $\begin{pmatrix} \mathbf{S_{aa}} & \mathbf{S_{as}} \\ \mathbf{S_{sa}} & \mathbf{S_{ss}} \end{pmatrix} = \begin{pmatrix} \mathbf{A'A} & \mathbf{A'S} \\ \mathbf{S'A} & \mathbf{S'S} \end{pmatrix}$ and λ is the maximum of Problem 3.

Note that the solution vector $\mathbf{w_s}$ can be interpreted as a linear combination of words, learned from the music data, which are highly correlated with the audio representation. In the next section we modify Problem 3 so that a subset of words in our vocabulary is explicitly selected.

3.1 Sparse CCA

The solution vectors, $\mathbf{w_a}$ and $\mathbf{w_s}$, in Problem 3 can be considered dense since most of the elements of each vector will be non-zero. In many applications it may be of interest to limit the number of non-zero elements in the $\mathbf{w}$ terms. This may aid in the interpretability of the result, particularly when the coordinate axes of a vector space have a direct meaning. For example, in bioinformatics experiments, the input space may contain thousands of coordinate axes corresponding to individual genes. We may wish to perform sparse analysis if we suspect that some phenomenon is dependent on a handful of these genes. In this paper, we impose sparsity on the solution vector $\mathbf{w_s}$, corresponding to the semantic space where each coordinate axis describes a word. Our goal is to find a subset of words in a vocabulary that are highly correlated to audio. We expect that these words may be more objective than others in the vocabulary since they are potentially characterized by correlations with the underlying audio signal, and thus, using these words may improve the performance of semantic music analysis systems.

Sparsity has been well studied in the fields of statistics and machine learning [22, 2, 14]. Imposing sparsity

is theoretically achieved by constraining the zero norm of a solution vector $||\mathbf{w}||_0$, which is the number of non-zero elements in $\mathbf{w}$ (this is technically an abuse of terminology as the zero norm is not a true mathematical norm). Constraining the zero-norm of a solution, because it is a non-convex constraint, renders the problem intractable, NP hard in fact. Instead, most sparse methods relax the cardinality constraints by approximating the zero-norm with a more mathematically tractable (i.e., convex) term such as the one-norm, $||\mathbf{w}||_1 = \sum_i |w_i|$ [22] [2]. In this paper we approximate $||\mathbf{w}||_0$ by $\sum_i \log(\epsilon + |w_i|)$, where $0 < \epsilon \ll 1$, avoids problems when one of the w_i is zero. This approximation has shown superior performance in sparse methods literature and it can be shown that the cardinality that results from this approximation is only $O(\frac{1}{\log \epsilon})$ greater than the zero-norm constrained solution [14] [20]. In practice we assume that ϵ is equal to machine precision and set it to zero.

We impose sparsity on the semantic space $\mathcal{S}$ by penalizing the cardinality of the solution vector $\mathbf{w_s}$ in Eq. 4. This is equivalent to solving the problem

$$\max_{\mathbf{w_s} \in \mathcal{S}} \quad \mathbf{S_{ss}^{-1} S_{sa} S_{aa}^{-1} S_{as} w_s} - \rho_s \sum_i \log |w_{s,i}| \quad (6)$$

$$\text{s.t.} \quad \mathbf{w_s'} \mathbf{w_s} = 1$$

If the log term in the objective is removed, the problem is simply the variational characterization of the eigenvalue problem in Eq. 4. The addition of the log term penalizes the cardinality of the solution vector if it becomes too high. ρ_s mitigates how harsh that penalty is, so by setting ρ_s one can control the sparsity of the solution.

The non-zero elements of the sparse solution vector $\mathbf{w_s}$ can be interpreted as those words which have a high correlation with the audio representation. Thus, in the experiments that follow, setting values of ρ_s and solving Problem 6 reduces to a vocabulary selection technique.

Problem 6 is still difficult to solve because it requires *maximizing* a convex objective. In other words, the problem does not have a guaranteed global maximum. However local solutions can be found by gradient descent or, as is done in our experiments, by solving a sequence of linear approximations [14].

4 REPRESENTING AUDIO AND SEMANTIC DATA

In this section we describe the audio and semantic representations, as well as describe the CAL500 [16] and Web2131 [15] annotated music corpora that are used in our experiments. In both cases, the semantic information will be represented using a single annotation vector $\mathbf{s}$ with dimension equal to the size of the vocabulary. The audio content will be represented as multiple feature vectors $\{\mathbf{a}_1, ..., \mathbf{a}_T\}$, where T depends on the length of the song.

The construction of the matrices A and S to run sparse CCA follows: Each feature vector in the music corpus is associated with the label for its song. For example, for a given song, we duplicate its annotation vector s for a total of T times so that the song-label pair may be represented as $\{(\mathbf{s}, \mathbf{a}_1), ..., (\mathbf{s}, \mathbf{a}_T)\}$. To construct A we stack the feature vectors for all songs in the corpus into one matrix. S is constructed by stacking all the corresponding annotation vectors into one matrix. If each song has approximately 600 feature vectors and we have 500 hundred songs, then both A and S will have about 30,000 rows.

4.1 Audio Representation

Each song is represented as a *bag-of-feature-vectors*: we extract an unordered set of feature vectors for every song, by extracting one feature vector for each short-time segment of audio data. Specifically, we compute dynamic Mel-frequency cepstral coefficients (dMFCC) from each half-overlapping, medium-time ($\sim$743 msec) segment of audio content [9].

Mel-frequency cepstral coefficients (MFCC) describe the spectral shape of a short-time audio frame in a concise and perceptually meaningful way and are popular features for speech recognition and music classification (e.g., [11, 7, 13]). We calculate 13 MFCC coefficients for each short-time (23 msec) frame of audio. For each of the 13 MFCCs, we take a discrete Fourier transform (DFT) over the time series of 64 frames, normalize by the DC value (to remove the effect of volume) and summarize the resulting spectrum by integrating across 4 modulation frequency bands: (unnormalized) DC, 1-2Hz, 3-15Hz and 20-43Hz. Thus, we create a 52-dimensional features vector (4 features for each of the 13 MFCCs) for every 3/4 segment of audio content. For a five minute song, this results in about 800 52-dimensional feature vectors.

We have also explored a number of alternative feature representations, These include auditory filterbank temporal envelope [9], MFCCs (with and without instantaneous derivatives) [16], chroma features [3], and fluctuation patterns [10]. For our experiments we chose a DMFCC representation since it is compact compared with raw MFCC feature representations and shows good performance on the task of semantic music annotation and retrieval compared with these other representations.

4.2 Semantic Representation

The CAL500 is an annotated music corpus of 500 western popular songs by 500 unique artists. Each song has been annotated by a minimum of 3 individuals using a vocabulary of 174 words. We paid 66 undergraduate music students to annotate our music corpus with semantic concepts. We collected a set of semantic labels created specifically for a music annotation task. We considered 135 musically-relevant concepts spanning six semantic categories: 29 instruments were annotated as present in the song or not; 22 vocal characteristics were annotated as relevant to the singer or not; 36 genres, a subset of the Codaich genre list [8], were annotated as relevant to the song or not; 18 emotions, found by Skowronek et al. [12] to be both important and easy to identify, were rated on

Top 3 words by semantic category		
	Agreement	*Acoustic Correlation*
overall	male lead vocals, drum set, female lead vocals	rapping, at a party, hip-hop/rap
emotion	not angry/agressive, not weird, not tender/soft	arousing/awakening, exciting/thrilling, sad
genre	hip-hop/rap, electronica, world	hip-hop/rap, electronica, funk
instrument	male lead vocals, drum set, female lead vocals	drum machine, samples, synthesizer
general	electric texture, not danceable, high energy	heavy beat, very danceable, synthesized texture
usage	driving, at a party, going to sleep	at a party, exercising, getting ready to go out
vocals	rapping, emotional, strong	rapping, strong, altered with effects
Bottom 3 words by semantic category		
	Agreement	*Acoustic Correlation*
overall	at work, with the family, waking up	not weird, not arousing, not angry/agressive
emotion	not powerful/strong, not emotional, weird	not weird, not arousing, not angry/agressive
genre	contemporary blues, roots rock, alternative folk	classic rock, bebop, alternative folk
instrument	trombone, tamborine, organ	female lead vocals, drum set, acoustic guitar
general	changing energy level, minor key tonality, low song quality	constant energy level, changing energy level, not catchy
usage	at work, with the family, waking up	going to sleep, cleaning the house, at work
vocals	falsetto, spoken, monotone	high pitches, falsetto, emotional

Table 1. Top and bottom 3 words by semantic category as calculated by agreement and acoustic correlation.

a scale from one to three (e.g., "not happy", "neutral", "happy"); 15 song concepts describing the acoustic qualities of the song, artist and recording (e.g., tempo, energy, sound quality); and 15 usage terms from [5], (e.g., "I would listen to this song while *driving, sleeping, etc.*"). The 135 concepts are converted to the 174-word vocabulary by first mapping bi-polar concepts to multiple word labels ('Energy Level' maps to 'low energy' and 'high energy'). Then we prune all words that are represented in five or fewer songs to remove under-represented words. Lastly, we construct a real-valued 174-dimensional annotation vector by averaging the label frequencies of the individual annotators. Details of the summarization process can be found in [16]. In general, each element in the annotation vector contains a real-valued scalar indicating the strength of association.

The Web2131 is an annotated collection of 2131 songs and accompanying expert song reviews mined from a web-accessible music database [1] [15]. Exactly 363 songs from Web2131 overlap with the CAL500 songs. The vocabulary consists of 317 words that were hand picked from a list of the common words found in the corpus of song reviews. Common stop words are removed and the resulting words are preprocesses with a custom stemming algorithm. We represent a song review as a binary 317-dimensional annotation vector. The element of a vector is 1 if the corresponding word appears in the song review and 0 otherwise.

5 EXPERIMENTS

Both human agreement and acoustic correlation may be used to discover words that are musically meaningful and useful in the context of semantic music annotation and retrieval. In this section, we conduct three experiments to highlight potential uses.

[1] AMG All Music Guide www.allmusic.com

Human Agr.	**Acoustic Cor.**
emotion 53.5 (26.7)	emotion 127.9 (54.9)
instrument 53.9 (39.5)	vocals 146.2 (50.0)
vocals 88.2 (40.0)	instrument 154.5 (39.9)
genre 118.6 (42.4)	genre 156.7 (41.2)
usage 152.3 (21.9)	usage 162.5 (37.9)

Table 2. Average rank of words in a semantic category when ranked by human agreement and acoustic correlation: Columns are sorted downward in increasing average rank. The average rank of the category and std. dev. are shown (lower is better). Note that the order of the categories closely match across both columns.

5.1 Qualitative Analysis

Table 2 shows the average rank of words in a semantic category when words are ranked by human agreement and acoustic correlation. For human agreement, words are ranked by their agreement score. For acoustic correlation, words are ranked by how long they are kept by sparse CCA as the vocabulary size is reduced. This experiment was run on the CAL500 data set.

A good rank in the human agreement metric suggests that a word is less subjective. This is true by definition since a good human agreement score means that people used that word consistently to describe music. Not surprisingly, we found that more objective categories such as instrumentation are highly ranked on this list and subjective categories such as usage are ranked at the bottom.

vocab.sz.	488	249	203	149	103	50
# CAL500 words	173	118	101	85	65	39
# Web2131 words	315	131	102	64	38	11
%Web2131	**.64**	**.52**	**.50**	**.42**	**.36**	**.22**

Table 3. The fraction of noisy web-mined words in a vocabulary as vocabulary size is reduced: As the size shrinks sparse CCA prunes noisy words and the web-mined words are eliminated over higher quality CAL500 words.

Interestingly, the ranks of semantic categories for the acoustic correlation metric matched closely with human agreement. This indicates that acoustic correlation may be honing in on objective words too. This could be explained if acoustic correlation picks up on the same structure in the audio that is being used by humans to make semantic judgements.

For a closer inspection at "musically relevant" words given by our methods, Table 1 shows the top 3 and bottom 3 ranked words for all semantic categories.

5.2 Vocabulary Pruning using Sparse CCA

Sparse CCA can be used to perform vocabulary selection where the goal is to prune noisy words from a large vocabulary. To test this hypothesis we combined the vocabularies from the CAL500 and Web2131 data sets and consider the subset of 363 songs that are found in both data sets.

Based on our own informal user study, we found that the Web2131 annotations are noisy as compared to the CAL500 annotations. We showed subjects 10 words from each data set and asked them which set of words were relevant to a song. The Web2131 annotations were not much better than selecting words randomly to from the vocabulary, whereas CAL500 words were mostly considered relevant.

Because Web2131 was found to be noisier than CAL500, we expect sparse CCA to filter out more of the Web2131 words. That is, given progressively smaller vocabulary sizes, Web2131 should comprise a progressively smaller proportion of the vocabulary.

Table 3 shows the results of this experiment. The first column of data reflects the vocabulary at full size. The vocabulary size is 488 and the Web2131 vocabulary initially out numbers the CAL500 words nearly two to one. Subsequent columns show the state of the vocabulary when vocabulary size is reduced from 203 words down to 50 words. Notice that the percentage of the Web2131 words that comprise a vocabulary does, in fact, decrease. Noisy words are being removed. If sparse CCA had no preference for either Web2131 or CAL500 we would see a constant proportion in Table 3.

5.3 Vocabulary Selection for Music Retrieval

Human agreement and acoustic relevance can also be used to prune noisy words from a vocabulary in order to improve the performance of semantic music analysis systems. If a word is not well represented by the underlying acoustic and semantic representation, then attempting to model such a word will be fruitless.

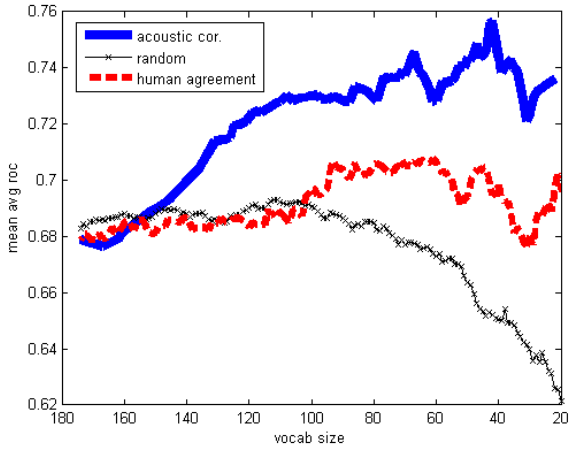

Figure 1. Comparison of vocabulary selection techniques: We compare vocabulary selection using human agreement, acoustic correlation, and a random baseline, as it effects retrieval performance.

In previous work, we have built a music retrieval system that can rank-order songs for a given word based on the audio content [16]. At a high level, this system builds Gaussian mixture models of audio feature vectors for songs associated with a given word. These mixtures can then be applied to the task of music retrieval.

One useful evaluation metric for this task is the area under the ROC (AROC) curve. (A ROC curve is a plot of the true positive rate as a function of the false positive rate as we move down a ranked list of songs.) For each word, the AROC ranges between 0.5 for a random ranking and 1.0 for a perfect ranking. Average AROC is found by averaging AROC over all words in the vocabulary. Once a specific vocabulary is set, we train our supervised model with a training set of 450 of the 500 test songs from the CAL500 data set. Then we calculate average AROC performance using the 50 songs that were not used during training. The average AROC scores, as a function of varying vocabulary size, are shown in Figure 1.

We find that pruning a vocabulary both based on human agreement and acoustic correlation improve the retrieval performance of our system. The performance of acoustic correlation is markedly superior to a baseline method in which we randomly select words to be in a vocabulary. Based on experiment 5.2 these methods seem to remove noisy words which are difficult to model, and thus improve system performance.

6 DISCUSSION

We have presented human agreement and acoustic correlation as metrics by which we can automatically, and in an unsupervised fashion, construct a musically meaning-

ful vocabulary prior to building semantic models. Our results indicate that vocabulary selection via these metrics remove noisy words that are difficult to model and lead to improved performance of our semantic music analysis systems. In the absence of human labeled data where human agreement cannot be calculated, acoustic correlation using sparse CCA provides a principled approach for pruning noisy words as is shown in Section 5.2.

Whitman and Ellis have previously looked at vocabulary selection by training binary classifiers (e.g., Support Vector Machines) on a heterogeneous data set of web-documents related to artists and the audio content produced by these artists [21]. By evaluating the performance of each 'word' classifier, they are able to rank order words for the purpose of vocabulary selection. This idea is conceptually similar to rank ordering words by average AROC, as was shown in Section 5.3, in that both evaluate a vocabulary a posteriori. i.e., after using a supervised learning framework. Moreover, such analysis evaluates the relevance of every word independent of every other word.

In future research, we will investigate the impact of using a musically meaningful vocabulary for assessing song similarity through semantic distance. We are interested in developing a query-by-semantic-example system [1] for music which retrieves similar songs by first representing them in the semantic space and then rank-ordering them based on a distance metric in that semantic space. We expect that having a compact semantic representation, which can be found using sparse CCA, we will be able to improve retrieval performance.We also plan to explore the possibility of extracting meaningful semantic concepts from web-based documents through acoustic correlation.

7 REFERENCES

[1] Luke Barrington, Antoni Chan, Douglas Turnbull, and Gert Lanckriet. Audio information retrieval using semantic similarity. Technical report, 2007.

[2] A. d'Aspremont, L. El Ghaoui, M.I. Jordan, and G.R.G. Lanckriet. A direct formulation for sparse pca using semidefinite programming. *Accepted for publication in SIAM Review*, 2006.

[3] D. Ellis and G. Poliner. Identifying cover songs with chroma features and dynamic programming beat tracking. IEEE ICASSP, 2007.

[4] David R. Hardoon, Sandor Szedmak, and Jogn Shawe-Taylor. Canonical correlation analysis; an overview with application to learning methods. *Neural Computation*, 2004.

[5] Xiao Hu, J. Stephen Downie, and Andreas F. Ehmann. Exploiting recommended usage metadata: Exploratory analyses. *ISMIR*, 2006.

[6] Einat Kidron, Yoav Y. Schechner, and Michael Elad. Pixels that sound. In *IEEE Computer Vision and Pattern Recognition*, 2005.

[7] Beth Logan. Mel frequency cepstral coefficients for music modeling. *ISMIR*, 2000.

[8] Cory McKay, Daniel McEnnis, and Ichiro Fujinaga. A large publicly accessible prototype audio database for music research. *ISMIR*, 2006.

[9] M. F. McKinney and J. Breebaart. Features for audio and music classification. *ISMIR*, 2003.

[10] E. Pampalk. *Computational Models of Music Similarity and their Application in Music Information Retrieval*. PhD thesis, Vienna University of Technology, Vienna, Austria, 2006.

[11] L. Rabiner and B. H. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

[12] Janto Skowronek, Martin McKinney, and Steven ven de Par. Ground-truth for automatic music mood classification. *ISMIR*, 2006.

[13] M. Slaney. Semantic-audio retrieval. *IEEE ICASSP*, 2002.

[14] Bharath K. Sriperumbudur, David A. Torres, and Gert R. G. Lanckriet. Sparse eigen methods by d.c. programming. *To appear in International Conference on Machine Learning*, 2007.

[15] D. Turnbull, L. Barrington, and G. Lanckriet. Modelling music and words using a multi-class naïve bayes approach. *ISMIR*, 2006.

[16] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet. Towards musical query-by-semantic description using the CAL500 data set. In *To appear in SIGIR '07*, 2007.

[17] D. Turnbull, R. Liu, L. Barrington, D. Torres, and Gert Lanckriet. UCSD CAL technical report: Using games to collect semantic information about music. Technical report, 2007.

[18] Alexei Vinokourov, John Shawe-Taylor, and Nello Cristianini. Inferring a sementic representation of text via cross-language correlation analysis. *Advances in Neural Information Processing Systems*, 2003.

[19] Luis von Ahn. Games with a purpose. *IEEE Computer Magazine*, 39(6):92–94, 2006.

[20] J. Weston, A. Elisseeff, B. Schölkopf, and M. Tipping. Use of the zero-norm with linear models and kernel methods. *Journal of machine learning research*, 2003.

[21] B. Whitman and D. Ellis. Automatic record reviews. *ISMIR*, 2004.

[22] Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 2004.