MINING THE CORRELATION BETWEEN LYRICAL AND AUDIO FEATURES AND THE EMERGENCE OF MOOD

Matt McVicar
Intelligent Systems Lab,
University of Bristol
matt.mcvicar@bris.ac.uk

Tim Freeman
Engineering Mathematics
University of Bristol
tf7960@bris.ac.uk

Tijl De Bie
Intelligent Systems Lab,
University of Bristol
tijl.debie@gmail.com

ABSTRACT

Understanding the mood of music holds great potential for recommendation and genre identification problems. Unfortunately, hand-annotating music with mood tags is usually an expensive, time-consuming and subjective process, to such an extent that automatic mood recognition methods are required. In this paper we present a new unsupervised learning approach for mood recognition, based on the lyrics and the audio of a song. Our system thus eliminates the need for ground truth mood annotations, even for training the system.

We hypothesize that lyrics and audio are both partially determined by the mood, and that there are no other strong common effects affecting these aspects of music. Based on this assumption, mood can be detected by performing a multi-modal analysis, identifying what lyrics and audio have in common. We demonstrate the effectiveness of this using Canonical Correlation Analysis, and confirm our hypothesis in a subsequent analysis of the results.

1. INTRODUCTION

Detecting the mood evoked by a musical piece is a task which is relatively easy for human listeners to perform. The ability to automate this process would be of use for music search, retrieval and recommendation, and for these reasons automatic techniques that recognize emotion in music have been an active topic of research in the past few years (e.g. [5, 8, 10, 17]).

The most common method of quantifying a mood state is by associating it with a point in a 2-dimensional space with valence (attractiveness/aversiveness) and arousal (energy) as dimensions, a concept first proposed by Russell [14]. High valence values correspond to positive moods such as 'pleased' or 'satisfied', with negative examples being emotions such as 'frustrated' or 'miserable'. Arousal can range from negative values ('sleepy') to positive ('excited'). This domain is known as the *valence-arousal* space (see Figure 1). Thus, automatic methods for mood recognition would map a song onto a point in this 2-dimensional space. However, also other ways of quantifying mood have been considered (e.g. [13]).

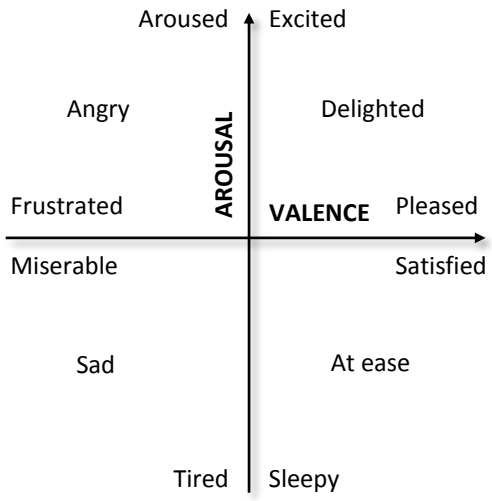

Figure 1. The 2−dimensional valence-arousal space, showing a range of emotions on a attractiveness/energy scale.

A major problem with evaluating (and—for machine learning methods— training) such algorithms is that high-quality ground truth mood annotations are hard to come by. Ideally these would be obtained by questioning a range of people on which emotions (and to which degree) they experience when listening to a range of songs in many styles. Such studies are expensive and time-consuming and clearly do not scale to the quantity of music required to tackle realistic research problems. A further confounding factor is that the emotion or mood associated with a song is a subjective and often personal feature.

1.1 Contributions

In this paper, we conduct a bi-modal analysis of music, simultaneously studying the audio and the lyrics of songs. Our goal is to extract factors that simultaneously underly aspects of the audio and the lyrics of popular music, at least statistically. In other words, we ask the question: "What do the audio and the lyrics of songs have in common?"

Our hypothesis is that answering this question is likely to resolve the problems faced in developing and assessing the quality of mood recognition systems, both those that are based on audio and those based on lyrics (or both). Indeed, we assume that the intended mood of a song will inspire the songwriter to use certain timbres, harmony, and rhythmic features, in turn affecting the choice of lyrics as well. A further hypoth-

esis is that factors unrelated to mood typically do not simultaneously influence the audio and the lyrics. If these hypotheses hold, uncovering what lyrics and audio share is equivalent to uncovering the mood of a song.

As a partial verification our hypotheses, below we first describe an exploratory analysis investigating if audio features correlate with valence and arousal, as predicted by a naive mood recognition algorithm based on lyrical information only.

The main result in this paper is the application of Canonical Correlation Analysis (CCA) [6] between paired representations of a song's audio and its lyrics. This is an unsupervised learning method that is independent of human experiments, able to extract common factors affecting both modes under study. We illustrate results which intuitively seem to coincide remarkably well with a notions of valence, and with another notion that is different but seems related to arousal.

1.2 Related work

Previous work in the area of multi-mode (text and audio) mood recognition has been focused on combining lyrics and audio into combined features for classification [7, 8]. This however still depends on the availability of good quality mood annotations for a large number of songs. Most strongly related to our current work is the investigation of correlations between social (non-lyrical) tags and audio [16]. Note that it is far less obvious that lyrics contain information about mood than in social tags. However, lyrics are easy to obtain, less subject to spamming, and objective. Thus, our work combines the benefits of the two types of prior work.

During the final stages of our study, the MusiXmatch lyrics database that is paired with the Million Song dataset was released [4]. Our study here is conducted on lyrics gathered by ourselves, the size of which is smaller but of similar order of magnitude as the MusiXmatch database. The approach presented in the current paper can directly be used as a blueprint for future research into the relationship between lyrics and audio based on this larger set of data.

1.3 Outline

The remainder of this paper is organised as follows. In Section 2 we outline our general approach and hypotheses. In Section 3 we describe the set of audio and lyric features used in this paper. A simple experiment is conducted in Section 4 exploring correlations between lyrics and audio. Section 5 contains our main result on CCA analysis and we conclude our findings in Section 6.

2. MOOD: THE SYNERGY OF LYRICS & AUDIO?

Since 2007, the Music Information Retrieval Evaluation eXchange (MIREX) has run a task on audio mood classification. The task is to 'tag' audio clips with an emotional label. Here, the ground truth is provided by users of the musical radio site `www.last.fm`. There are generally three approaches to tackling mood classification in these tasks and we summarise them here to highlight the interplay between text and audio.

2.1 Classification based on Audio Features

The most common method for classification is based on harmonic and spectral features of the audio [8]. Commonly used features include low level indicators such as spectral centroid, rolloff, flux, slope, skewness and kurtosis [3], harmonic features such as MFCCs [12] and those based on Short Time Fourier Transforms [15]. In many cases Support Vector Machines are used to discriminate between features and have proved to be successful in this setting [9].

2.2 Classification based on Lyrical Features

Other approaches are based on lyrical content only. Bag-Of-Words (BOW) representations have recently been successful in identifying mood, as well as higher-order statistics such as combinations of unigrams, bigrams and trigrams [5].

2.3 Classification using both Audio and Lyrics

More complex approaches simultaneously exploit lyrical and audio features. Such approaches generally achieve higher classification accuracy than those methods presented in Subsections 2.1 and 2.2 (see for example [11, 17]).

A recent analysis by Hu et. al. [8] showed that lyrical features typically outperform audio when used as a classifier, although they note that in their study audio was more useful in determining emotions in the 3^{rd} quadrant of the valence-arousal space in Figure 1 (i.e. 'sad', 'depressed' etc.).

2.4 Framework

In this paper, we will search for correlations between a set of features from audio and from the lyrics, under the assumption that the causal factor of any such correlations is the mood, i.e. that emotion is the unique facet that lyrics and audio share. Of course, such patterns may be subtle and they will be present only 'on average', such that they cannot be reliably detected on small samples. For this reason, we study such patterns on a large scale, allowing even subtle correlations to emerge as statistically significant.

Informally speaking, if $x_a \in \mathbb{R}^{d_a}$ is a d_a-dimensional audio-based feature vector for a given song, and $x_l \in \mathbb{R}^{d_l}$ is a d_l-dimensional lyrical feature vector for the same composition, we seek real-valued functions f_a and f_l such that for many songs and to a good approximation:

$$f_a(x_a) \approx f_l(x_l). \tag{1}$$

A core assumption is that if such functions f_a and f_l can be found, they must be capturing some notion of mood of an audio piece. Due to variability in style, genre, instrumentation and potential use of irony (i.e. different mood exhibited by the lyrics and the audio), we do not expect to find this approximate equality to be very strong, or to be valid for many songs, but the size of the data used (see below) should nevertheless allow us to find statistically significant relation.

Our strategy differs from previous ones in that it does not need a training set of songs with ground truth mood annotations. Rather than supervising the learning process using ground truth labels, we simultaneously train two mood recognizers, one based on lyrics and one on audio, which supervise each other's learning.

3. THE DATA: SONG CORPUS AND FEATURES

Below we describe the feature representations of the lyrics and audio modes of songs we used in this paper, as well as the corpus of songs used.

3.1 Lyrics feature representation

We used the **Term Frequency-Inverse Document Frequency** (TF-IDF) measure to represent the lyrics in a song. The TF-IDF representation of a document is a reweighted version of a BOW account, accounting for how rare a word is with respect to a document and the overall collection. Consider the i^{th} word in the j^{th} lyric. Then the term frequency is the number of times word i appears in document j, normalised by the document's length:

$$TF_{i,j} = \frac{|\text{word } i \text{ appears in lyric } j|}{|\text{lyric } j|}$$

The inverse document frequency is a measure of the general importance of the word in the lyric database:

$$IDF_i = \log \frac{\text{total number of lyrics}}{|\text{lyrics containing word } i|}$$

The TF-IDF for word i in lyric j is then the product

$$TFIDF_{i,j} = TF_{i,j} \times IDF_i$$

3.2 Audio Feature Extraction

We used the Echonest API [1] to extract features from our audio and thus obtained 65 spectral, percussive, harmonic and structural features, which are summarised in Table 1.

Field	Feature
1	Tempo
2	Tempo Confidence
3-7	Time Signature
8	Time Signature Confidence
9	Mode
10	Mode Confidence
11	Number of Sections
12	Energy
13	Danceability
14-25	Mean Chroma Pitches
26-37	Standard Deviation Chroma Pitches
38-49	Timbre Mean
50-61	Timbre Standard Deviations
62	Loudness Start Mean
63	Loudness Start Standard Deviations
64	Loudness Max Mean
65	Loudness Max Standard Deviations

Table 1. Audio features extracted from Echonest.

Note that some of these features (e.g. the Mean Chroma Pitches) are unlikely to be relevant for mood recognition. Still, we have included them in our experiments to validate our approach.

[1] `http://developer.echonest.com/docs/v4/`

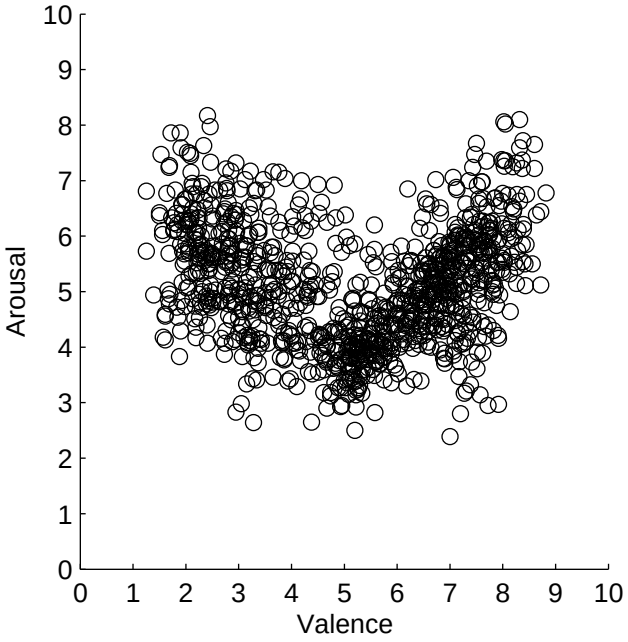

Figure 2. Valence and arousal for the ANEW database.

3.3 The song corpus

Using a simple web-scraper, we obtained lyrics from the popular lyrical database website `www.lyricsmode.com`, which contains over 800,000 song entries. We also obtained audio features using the Echonest API and found the intersection of these two datasets to be $119,664$ lyric/audio pairs. We are not aware of any other lyrical/audio combined studies carried out on this scale.

4. EXPLORING MOOD, AUDIO, AND LYRICS RELATIONS

In a first exploratory study, we build a simple mood recognition system based on lyrics, and we verify which (if any) audio features are correlated with this mood estimate. This is to confirm our basic hypothesis that on average both lyrics and audio reflect the mood of a song. To this end we implemented a simple method for estimating mood from lyrics based on the valence/arousal space described in Sec. 1.

4.1 Valence/Arousal Estimation

One method of analysing emotive content of lyrics is to measure the average valence or arousal over a song, picking out particular words from a dictionary where the valence/arousal scores are known. We chose the Affective Norms for English Words (ANEW) as our dictionary, which contains ratings of 1030 words on pleasure, arousal and dominance collected by psycholinguistic experiments [2]. The words within were chosen to cover a wide range of the valence-arousal space [10] and we show their means (taken over participants) in Fig. 2.

Let $l^i = (w_1, w_2 \dots w_{n^i})$ be the i^{th} lyric, comprised of n^i words and let $\mathcal{L} = \{l_1, l_2, \dots l_m\}$ be the complete collection of lyrics. We then estimate the valence v^i and arousal a^i of

lyric i via

$$v^i = \frac{1}{n^i}\sum_{j=1}^{n^i} V(w_{n^i}), \quad a^i = \frac{1}{n^i}\sum_{j=1}^{n^i} A(w_{n^i}), \quad i = 1 \ldots m.$$

V and A are functions that return the mean valence/arousal if word w_{n^i} is in the ANEW dictionary and zero otherwise.

This is obviously a crude mood recognition system. Note however that our goal here is to use a simple and transparent system, only to verify our hypothesis that audio and lyrics share a common cause.

4.2 Correlations between audio features and mood estimates based on lyrics

Given our simple mood recognition system based on lyrics, we computed Pearson's correlation coefficient between each of the audio features and our valence/arousal estimate based on lyrics. We found many of the correlations to be extremely statistically significant, but below 0.2 in absolute value. For illustration, in Table 2 we show the audio features that are correlated with p-value numerically equal to 0, and from those only the 5 highest correlations by absolute value.

Audio Feature	Lyrical Feature	Correlation
12	Valence	-0.1943
62	Valence	-0.1939
38	Valence	-0.1897
64	Valence	-0.1818
61	Valence	0.1739
57	Arousal	-0.0591
59	Arousal	-0.0553
39	Arousal	0.0511
17	Arousal	0.0462
24	Arousal	0.0434

Table 2. Top correlations with valence and arousal with p-value numerically 0 (audio feature indices refer to Table 1).

The strongest relationship is valence against energy, with a correlation of -0.1943. This suggests than an increase in 'lyrical positiveness' corresponds to a decrease in energy, and is perhaps caused by love ballads, which typically will contain many positive words ('love','heart' etc.) along with gentle audio. Several other audio features strongly correlated with valence are loudness (62,64).

The correlations with arousal are more difficult to interpret. The top three correlations relate to timbre, and seem plausible. The features 17 and 24 are mean chroma values over the song, and their apparent significance to mood seems counter-intuitive. However, the magnitude of the correlations is very small when compared to the valence correlations, and we suspect that these correlations are due to artefacts (e.g., mean chroma values may not be independent of certain loudness features). Unfortunately, this is hard to verify, as the exact mechanism of how they are computed is unknown to us (they were obtained through the echonest API).

The overall conclusion that can be drawn is that a correlation between valence/arousal is present and significant, which confirms our hypothesis that, to some extent, mood is indeed simultaneously related to both lyrics and audio. However, the correlations are not very strong. We suggest two possible explanations for this. Firstly, the mood recognition method based on lyrics is simple and imperfect. More crucially, probably none of the audio features by themselves relate strongly to mood—probably that a combination of them is more relevant (in different combinations for valence and arousal) than each of the features individually.

In the next Section, we will demonstrate a method that is immune to both these problems. We will simultaneously learn linear combinations of the features in the lyrics and audio representations, so as to maximize the correlation between the resulting linear combinations. In this way, we avoid our dependency on an initial method for mood recognition based on lyrics such as the one introduced in Sec. 4.1. Furthermore, by considering linear combinations of features, we expect to find much stronger and more meaningful relations.

5. CANONICAL CORRELATION ANALYSIS

We will first discuss the theory of CCA before presenting our findings (see e.g. [1] for a more in depth treatment).

5.1 Background

CCA is a technique that can be used to find information that is consistent in two datasets by revealing linear correlations between them, and is particularly useful in high-dimensional datasets such as ours.

Given two datasets $X \in \mathbb{R}^{n\times d_x}$ and $Y \in \mathbb{R}^{n\times d_y}$, the objective of CCA is to find weightings $w_x \in \mathbb{R}^{d_x}$ and $w_y \in \mathbb{R}^{d_y}$ that maximise the correlation between the projections of X and Y, Xw_x and Xw_y. Thinking of these projections as directions through the data spaces, CCA looks for a projection which will minimise the angle $\angle$ between Xw_x and Xw_y. Mathematically, this optimization problem is written:

$$\begin{aligned}
\{w_x^*, w_y^*\} &= \underset{w_x,w_y}{\operatorname{argmin}} \angle(Xw_x, Yw_y), \\
&= \underset{w_x,w_y}{\operatorname{argmax}} \cos(\angle(Xw_x, Yw_y)), \\
&= \underset{w_x,w_y}{\operatorname{argmax}} \frac{(Xw_x)'(Yw_y)}{\sqrt{(Xw_x)'(Xw_x)}\sqrt{(Yw_y)'(Yw_y)}}, \\
&= \underset{w_x,w_y}{\operatorname{argmax}} \frac{w_x'X'Yw_y}{\sqrt{w_x'X'Xw_x}\sqrt{w_y'Y'Yw_y}}.
\end{aligned}$$

It is known that this optimization problem can be solved by solving the following generalized eigenvalue problem (see e.g. [1] for a derivation):

$$\begin{pmatrix} 0 & X'Y \\ Y'X & 0 \end{pmatrix} \begin{pmatrix} w_x \\ w_y \end{pmatrix} = \lambda \begin{pmatrix} X'X & 0 \\ 0 & Y'Y \end{pmatrix} \begin{pmatrix} w_x \\ w_y \end{pmatrix}. \quad (2)$$

The eigenvalue λ in Eq. (2) is equal to the achieved correlation between the projections of X and Y on their respective weight vectors w_x and w_y. Thus, the eigenvector corresponding to the largest eigenvalue is of greatest interest, with successive ones of decreasing importance. An additional property of CCA is that projections on successive components are independent, such that each of the eigenvectors capture uncorrelated information.

5.2 Experiments

In our setting, the data X and Y refer to audio and lyrical features. For lyrical features independent of mood, we used the TF-IDF measure described in Subsection 3.1.

To prevent overfitting of the method we performed 100-fold cross validation. I.e., we split the set of $119,664$ songs into 100 disjoint subsets and apply CCA on the union of 99 of them, after which we compute the correlation between the projections of the remaining subset on the obtained weight vectors as a validation. This is repeated 100 times, leaving out each of the 100 subsets in turn. The mean training and testing correlations over the folds are shown in Figure 3.

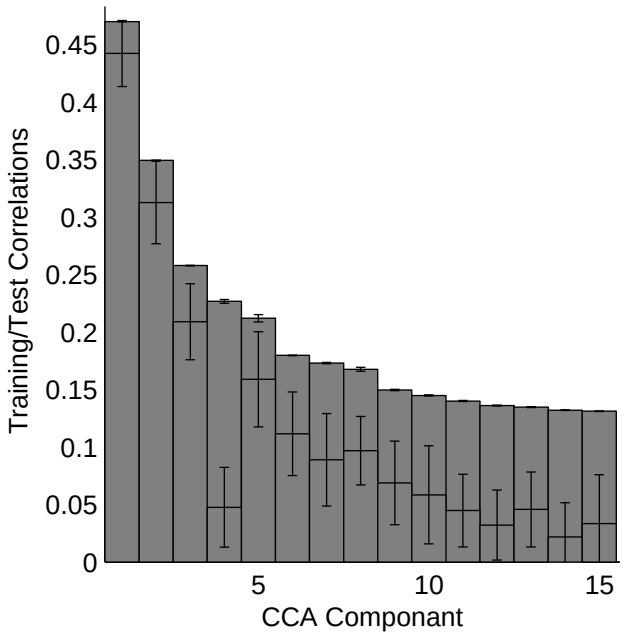

Figure 3. Training/Testing (upper/lower bars) correlations of the CCA components, with Error bars of 1 standard deviation.

It can be seen that training and test correlations are quite close, especially in the first two components (suggesting the data is not significantly overfitted). Correlations on the training set are likely to always be higher than on the test set, but it appears not significantly so, as the error bars on the test set overlap those for the training data in these cases.

Confident that the CCA algorithm was not overfitting the training data, we proceeded to train the weights on all of the training data, and tested on the complete set. The first component is shown in detail in Table 3.

Inspecting Table 3, the first component seems to closely correspond to valence—even though this was not imposed by the algorithm. Low weights are associated with strongly negative emotions/words, which would lie in the 4^{th} quadrant of the valence-arousal space (see Fig. 1). In contrast, the words with high weights appear to correspond to positive moods (1^{st} quadrant), although there are some outliers in the 3^{rd} and 4^{th} columns. In the audio domain the features most negatively weighted in the CCA components were all related to Timbre, the most positive to Loudness.

To verify that the first component relates to valence, we correlated the weights which resulted from the CCA output to the valences from the ANEW database. The resulting correlation was -0.3519, with a $p-$value numerically equal to 0. This is an important result, as it shows we have successfully reconstructed words which carry the meaning of 'positive/negative' emotions without the need for expensive human interventions. It shows that valence is the aspect of mood most dominantly affecting both lyrics and audio.

Lowest		Highest	
Word	**Lyrical Weight**	**Word**	**Lyrical Weight**
Death	-0.075996	Love	0.1248
Dead	-0.064387	Baby	0.049397
Hate	-0.054789	Heart	0.047417
Pain	-0.047474	Hay	0.029812
Evil	-0.04673	Home	0.028472
Life	-0.042257	Lonely	0.027777
Stench	-0.040415	Good	0.027413
Hell	-0.038346	Blue	0.026954
War	-0.037502	Sin	0.026194
Destroy	-0.036671	Loved	0.026123
Feature	**Audio Weight**	**Feature**	**Audio Weight**
38	-0.61774	64	0.3919
50	-0.22214	62	0.28949
42	-0.15033	65	0.19222

Table 3. First component of the CCA analysis, which appears to relate to valence. The 10 most negatively and positively weighted words and 3 most weighted audio features are shown, along with their associated weights.

Lowest		Highest	
Word	**Lyrical Weight**	**Word**	**Lyrical Weight**
Heart	-0.024301	Baby	0.02641
Love	-0.019733	Man	0.021014
Lost	-0.018202	Hit	0.020528
World	-0.015552	Money	0.020241
Moment	-0.015103	Rock	0.019736
Fall	-0.015003	Party	0.018319
Lonely	-0.014069	Girl	0.017076
Dream	-0.013675	Mad	0.015997
Hope	-0.013444	Kick	0.015813
Sun	-0.012514	Fat	0.012571
Feature	**Audio Weight**	**Feature**	**Audio Weight**
38	-0.77382	64	0.49949
12	-0.10808	62	0.26838
43	-0.080392	5	0.092167

Table 4. Second component of the CCA analysis, which we postulate relates to arousal.

The second component is shown in Table 4, and is more difficult to interpret, although there seems to be a relation with arousal. Words in the first column ('dream', 'heart') are generally calming and restful, whilst those in the third column are more energetic ('kick','party'). Audio features with significant weight relate to Timbre/Energy and Loudness.

5.3 Discussion

It is remarkable that our CCA analysis automatically detects aspects of mood that appear to align with Russell's model for human perception of emotion [14], without any dependence on human trials or mood annotations. We should point out that further components (not shown here due to space constraints) are harder to interpret in terms of aspects of mood we are aware of. However, given the encouraging results for the dominant components we believe they are likely to be helpful in a multi-dimensional characterization of mood in audio and in lyrics. As such they may be helpful in applications such as music classification and recommendation in particular.

Interestingly, our approach also opens up possibilities of detecting more high-level properties in music, such as irony and sarcasm. The ability to recognize strongly correlated aspects of mood from both audio and lyrics also allows us to identify songs where there is a discrepancy or tension between the mood in the audio and the mood in the lyrics, violating the global pattern of correlation.

6. CONCLUSIONS

In this paper we investigated the correlation between audio and lyrics, demonstrating that there exist weak but highly significant correlations between lyrical and audio features. Following this, we used Canonical Component Analysis to uncover strong correlations between linear combinations of lyrical and audio features which, at least in part, appear to correspond to known aspects of mood and valence and arousal.

In further work we intend to rerun our experiments including also the MusiXmatch dataset [4]. Furthermore, we intend to use more features such as images, video, social tags and $n-$gram features in the lyrical domain.

7. REFERENCES

[1] T. De Bie, N. Cristianini, and R. Rosipal. Eigenproblems in pattern recognition. In E. Bayro-Corrochano, editor, *Handbook of Computational Geometry for Pattern Recognition, Computer Vision, Neurocomputing and Robotics*. Springer-Verlag, 2004.

[2] M.M. Bradley and P.J. Lang. Affective norms for english words (anew): Instruction manual and affective ratings. *University of Florida: The Center for Research in Psychophysiology*, 1999.

[3] J.J. Burred, M. Ramona, F. Cornu, and G. Peeters. Mirex-2010 single-label and multi-label classification tasks: ircamclassification09 submission. *MIREX 2010*, 2010.

[4] The Echo Nest Corp. The million song dataset gets lyrics, too. http://blog.echonest.com/post/4578901170/the-million-song-dataset-gets-lyrics-too, May 2011.

[5] H. He, J. Jin, Y. Xiong, B. Chen, W. Sun, and L. Zhao. Language feature mining for music emotion classification via supervised learning from lyrics. *Advances in Computation and Intelligence*, pages 426–435, 2008.

[6] H. Hotelling. Relations between two sets of variables. *Biometrika*, 28:321–377, 1936.

[7] X. Hu and J.S. Downie. Improving mood classification in music digital libraries by combining lyrics and audio. In *Proceedings of the 10th annual joint conference on Digital libraries*, pages 159–168. ACM, 2010.

[8] X. Hu and J.S. Downie. When lyrics outperform audio for music mood classification: a feature analysis. In *ISMIR*, pages 1–6, 2010.

[9] X. Hu, J.S. Downie, C. Laurier, M. Bay, and A.F. Ehmann. The 2007 mirex audio mood classification task: Lessons learned. In *Proceedings of ISMIR*, pages 462–467, 2008.

[10] Y. Hu, X. Chen, and D. Yang. Lyric-based song emotion detection with affective lexicon and fuzzy clustering method. In *Proceedings of ISMIR*, 2009.

[11] C. Laurier, J. Grivolla, and P. Herrera. Multimodal music mood classification using audio and lyrics. In *Machine Learning and Applications, 2008. ICMLA'08. Seventh International Conference on*, pages 688–693. IEEE, 2008.

[12] M.I. Mandel. Svm-based audio classification, tagging, and similarity submissions. *MIREX 2010*, 2010.

[13] A. Pepe and J. Bolle. Between conjecture and memento: shaping a collective emotional perception of the future. In *AAAI Spring Symposium on Emotion, Personality, and Social Behavior*, 2008.

[14] J.A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[15] K. Seyerlehner, M. Schedl, T. Pohle, and P. Knees. Using block-level features for genre classification, tag classification and music similarity estimation. *MIREX 2010*, 2010.

[16] D. Torres, D. Turnbull, L. Barrington, and G. Lanckriet. Identifying words that are musically meaningful. *Proc. ISMIR07*, pages 405–410, 2007.

[17] Y.H. Yang, Y.C. Lin, H.T. Cheng, I.B. Liao, Y.C. Ho, and H. Chen. Toward multi-modal music emotion classification. *Advances in Multimedia Information Processing-PCM 2008*, pages 70–79, 2008.