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

RESEARCH Open Access

Expressed music mood classification compared with valence and arousal ratings

Bert den Brinker[1*], Ralph van Dinther[1] and Janto Skowronek[2]

Abstract

Mood is an important aspect of music and knowledge of mood can be used as a basic feature in music recommender and retrieval systems. A listening experiment was carried out establishing ratings for various moods and a number of attributes, e.g., valence and arousal. The analysis of these data covers the issues of the number of basic dimensions in music mood, their relation to valence and arousal, the distribution of moods in the valence–arousal plane, distinctiveness of the labels, and appropriate (number of) labels for full coverage of the plane. It is also shown that subject-averaged valence and arousal ratings can be predicted from music features by a linear model.

Keywords: Music, Mood, Valence, Arousal

Introduction

Music recommendation and retrieval is of interest due to the increasing amount of audio data available to the average consumer. Experimental data on similarity in mood of different songs can be instrumental in defining musical distance measures [1,2] and would enable the definition of prototypical songs (or song features) for various moods. These latter can then be used as the so-called mood presets in music recommendation systems. With this in mind, we defined an experiment to collect the relevant data. In view of the mentioned applications, we are interested in the perceived song mood (not the induced mood), annotation per song (not per part of a song), and annotation by average users (as opposed to expert annotators). Furthermore, the test should be executed with a sufficient amount of participants as well as a good cross-section of music with clear moods covering the full range and, obviously, a proper set of mood labels (easy-to-use and discriminative). The data collected in earlier studies on music mood [3-12] only partially meet these requirements.

Part of the knowledge (mood labels, song selection, interface) used to define the experiment stems from earlier experience gained in this area [13-15]. Valence and arousal ratings were included since mood is assumed to be mainly governed by these two dimensions [1,2,16,17]. This article describes the experiment and the analysis of the collected data. The analysis comprises the fundamental mood dimensions [6], comparison of these dimensions to valence and arousal, coverage of the valence and arousal plane, comparison of mood labels in the valence–arousal plane and ratings for affect words [16,17] and the predictability of the valence and arousal ratings from a set of music features. The latter is of interest since predictability would imply the possibility of an automatic valence and arousal rating which presumably could be used as a basis for mood annotation. To study the predictability, we use music features determined from the audio signal. These include spectro-temporal features derived from Mel-frequency cepstral coefficients (MFCCs) as well as features based on the statistics of tonality, rhythm, and percussiveness [15].

Before describing the experiment and the analysis, we would like to comment on our terminology. In music research, it is common to categorize music according to mood. In our experiment, we also used the term mood. In emotion research, there is a clear tendency to distinguish between emotion and mood, where the former is associated with a shorter timescale than the latter. Such distinction is virtually absent in music research [2]. In view of the fact that we are looking for full song annotation and a full song has a somewhat larger time stretch, the term mood is probably the better option. We will therefore use the term mood for the music categorization throughout this article. Only in Section “Comparison with affect word scaling”,

*Correspondence: bert.den.brinker@philips.com
[1] Philips Research, High Tech Campus 36, NL-5656 AE Eindhoven, The Netherlands
Full list of author information is available at the end of the article

emotion scaling appears since there a comparison of our music mood rating with affect word rating is considered.

The article starts with a description of the music experiment in Section "Mood experiment". Next, the fundamental dimensions in music mood are determined and compared with the attribute ratings. The distribution of the different moods in the valence–arousal plane and the coverage of the plane is addressed in Section "Music moods in the valence–arousal plane". The last part of the analysis covers the predictability of the subject-averaged valence and arousal from music features in Section "Valence and arousal prediction". The article ends with a discussion and the conclusions.

Mood experiment

In a series of articles [13-15], the issue of creating a proper mood database was considered. The current rating experiment was inspired by the experiment conducted in [15] but differed in a number of ways. In particular, (1) a different questionnaire was used, i.e., moods were rated differently, and additional ratings were incorporated in the questionnaire, e.g., 'pleasant', 'energetic', 'tensed', and 'liking'; (2) participants were allowed to scroll through the entire song instead of a preselected piece of about 20 s (in view of full-song annotation); (3) a larger group of subjects participated (which resulted in more ratings per song).

The reason for incorporating additional ratings (i.e., next to mood) was to gain more insight into the basic dimensions determining the music mood. In particular, the relation between mood ratings and *valence, arousal*, and *tenseness* is addressed. In addition, the test considered a *liking, familiarity*, and *association* rating. The results of this last part of experiment are beyond the scope of the article. We only note here that the familiarity rating showed that most of the songs were unknown to the majority of the subjects.

Since music mood experiments are time-consuming, we opted for a minimal number of songs and participants such that the resulting data would accurately be enough for analysis purposes (e.g., robust to outliers in the data). Based on the experience, we estimated that ratings of eight different participants per song would yield relatively reliable estimates. For similar reasons, we targeted at least a dozen songs per mood category. Since the participants were free in their judgments of each category we doubled the amount to 24 songs per presumed mood category.

Mood categories

As argued in [13-15], there is a problem with expressed music mood because the subjectivity in the rating tends to be more prominent than for, e.g., genre. The considerations in the mentioned articles led to the following notions.

- There are 12 mood categories which are relatively consistent between subjects and easy-to-use as well;
- Mood categories are non-exclusive categories;
- Moods should be ranked as 'belonging to this class' or 'not-belonging to this class' (as opposed to working with antagonistic labels);
- Proper wording of the labels is important.

The set-up of the mood experiment was based on these findings. The 12 mood category labels are given in Table 1. In the figures, we will use either the alphabetical identifiers A–L or the shorthand labels instead of the full labels. For clarity, we note that the shorthand labels were never used in the experiment. They are only used here in this article for convenience.

Since the selection of the songs involves, e.g., balancing across moods and genres and thus heavily depends on previously gathered knowledge and data, we decided to use the mood categories we had been using so far and not to switch to another, e.g., the five mood categories used in the MIREX evaluations [1].

Participants

The target was a wide variety of participants in terms of gender, age, and experience in listening to music. Participation was rewarded with a shopping voucher of 20 euro.

In total, 36 volunteers accepted to participate in the experiment of which 32 completed the test. The latter group comprised 10 females and 22 males with ages ranging between 19 and 48 years (mean: 32, std: 9), consisted of 12 nationalities (mainly European). On average, participants listened to music for 12 h per week (std: 12) and the years of music practice ranged from 0 to 30 years (mean: 6, std: 9).

Table 1 The twelve mood labels used in the experiment and their shorthand notation

Identifier	Music mood label	Shorthand
A	Sad	Sad
B	Calming/soothing	Calming
C	Arousing/awakening	Arousing
D	Powerful/strong	Powerful
E	Tender/soft	Tender
F	Cheerful/festive	Cheerful
G	Carefree/lighthearted/light/playful	Carefree
H	Angry/furious/aggressive	Aggressive
I	Peaceful	Peaceful
J	Emotional/passionate/touching/moving	Emotional
K	Loving/romantic	Loving
L	Restless/jittery/nervous	Restless

Music selection

The tracks used in the experiment were selected from a large number (1,059) of music excerpts of a previous experiment [15], where participants labeled excerpts using 12 mood classes identified in [14].

For the current experiment, the third author selected in total 288 songs by reviewing the collection of 1,059 excerpts. Songs were selected when the earlier used excerpts were proper examples for the full song (e.g., chorus or verse, but not intros, etc.) and had a consistent mood rating according to the experiment in [15]. The 288 tracks were divided into sets of 64 tracks per participant in such a way that each excerpt was rated 8 times by different participants and each presumed mood was rated about 5 times per participant. The sets per participant were mutually overlapping.

The songs were drawn from different genres. The subdivision of songs over 12 different genres is given in Table 2. The song count for language of the lyrics is given in Table 3.

Rating experiment

The rating experiment was conducted over the Internet. This procedure was chosen in order to be able to include participants outside of our Lab and for the convenience of the participants. The convenience has several aspects. The participants were able to do the experiment at home and at a time which suited them best. Furthermore, it was also allowed to do the experiment in steps, i.e., the experiment could be stopped by closing the Internet browser window and continued at any time they wanted.

An instruction guide was distributed by email and clarifications of the instructions were offered on request. According to the instructions, ratings should be based on the mood that the song conveys or expresses, but not on other knowledge, e.g., on artist. It was advised to set the audio volume to a comfortable level. The participants were also instructed to ignore lyrical content in the judgment. One may doubt whether a participant is able to do so and, if yes, to what extent. This instruction was adopted nevertheless in order to bias the judgment as much as possible away from the lyrics and toward the music itself since lyrical content is not represented in our feature set. Most of the lyrics are in English (see Table 3) and presumably understandable by the majority of the participants.

Table 2 The subdivision of the 288 songs over 12 genres

Genre	Number of songs
Blues	24
Classical	24
Country	21
Folk	27
Hip-Hop	24
Jazz	24
Latin	24
Pop	24
R&B	24
Reggae	24
Rock	24
Electronic	24

Table 3 Number of songs with and without lyrics and languages of the lyrics

Lyrics and language	Number of songs
Instrumental	47
Lyrics	241
English	203
Spanish	22
German	8
Latin	5
French	1
Norwegian	1
Russian	1

Before starting the experiment, the participant was asked to complete a brief questionnaire, e.g., age, music preference, etc. After completing the questionnaire the experiment started with the assignment explained in the instructions.

For each song, the interface provided a screen divided into five parts. The first part, ‘Mood Rating Method 1’, consisted of a rating for each of the 12 moods from Table 1. The participant rated his/her agreement on a 7-point scale from strongly disagree to strongly agree by clicking on the buttons. The second part, ‘Mood Rating Method 2’, consisted of three ratings, where the participants were asked to judge in how far the music is pleasant, energetic, or tensed, again on a 7-point scale, from ‘unpleasant’ to ‘pleasant’, from ‘without energy’ to ‘full of energy’ and from ‘relaxed’ to ‘tensed’. The third part, ‘Liking’, consisted of three ratings in which participants were asked to judge in how far they liked the music, whether the music was known to them and what associations (bad–good) they had with the song. This part of the test was included for screening purposes in case of very unexpected outcomes and is actually not used in this article. In the fourth part, the participants were asked if they had any comments, in particular if they missed a mood category in the list of Method 1 for this particular music piece and, if so, to write it down in the text field. This part was built in as a safety net especially with regard to the wording of the labels. Lastly, participants had to press the accept button to go to the next song.

The song started by clicking on the play button at the top of the screen or it started automatically, depending on the web browser used by the participant. It was advised to scroll through the entire song and to spend at least 20 s before going to the next song. Because the rating experiment was not supervised by the experimenter, the test system ensured that the participant had to spend at least 20 s before he/she was able to continue with the next trial. On average the participants needed about 1.5–2 h of their time to complete the test.

All 7-point rating scales are represented by the numbers 0, . . . , 6.

Dimensions in music mood

Annotation

As a start in the analysis, we considered the number of basic dimensions determining the mood of a song. For this, we performed an eigenvalue decomposition of the covariance matrix (i.e., principal component analysis—PCA) of the 12 mood ratings (see Section "Number of dimensions"). This is a straightforward approach to get insight into the dominant (number of) dimensions underlying the experimental data. Next, we considered whether we could interpret the relevant dimensions (see Section "Axis interpretation"). This interpretation is validated by a model fit comparing the actually measured dimensions with the dominant axis according to the covariance analysis (see Section "Validation of axis interpretation").

We distinguish two approaches in the dimension analysis. The first one was building the covariance matrix using each rating separately. In a second approach, we first averaged the ratings per song. We refer to these approaches as trial-based and song-based, respectively.

The observation matrices are called S_t and S_s, respectively. The first one is a matrix of dimension 2131×12, the second one is 288×12 since 2131 is the number of ratings we had, 288 is the number of songs in the test, and 12 is the number of moods used in the test. From S_t and S_s, we determined the 12×12 covariance matrix on which an eigenvalue decomposition was performed. The fact that we have 2131 trials instead of $32 \times 64 = 2048$ (i.e., number of participants times the size of the song set) is due to the fact that we included the ratings of subjects that did not complete the full test. By including the partially completed forms, the number of songs with an equal number of ratings increases.

A first screening of the results was done to check for participants with clearly different judgments than the majority. Though there were some indications of systematically different scores for some participants, we decided not to discard any data. We checked whether removing data from these participants largely influenced the results, which is not the case. In fact, the analysis presented in the remainder of the article was repeated by excluding what was deemed as systematically different scores. Though this obviously gives different numbers than those presented in the plots and graphs of this article, the conclusions drawn from the full dataset remain valid. In line with [6], we found more consistency over subjects for arousal than for valence.

Number of dimensions

In Figure 1 the eigenvalues are plotted on a log scale as a function of index. We also approximated this curve as piece-wise linear curves (manual fit of the lines). We observe that the eigenvalues can be modeled as consisting of two linear parts. We interpret the straight lines at higher indices as a noise-driven part in the eigenvalue decomposition, and the first part as the essential part. In this view, there are only three dimensions which are the essential part of the covariance matrix.

Axis interpretation

For each of the three dominant eigenvectors, we try to give an interpretation. In Figure 2, the first eigenvector is plotted. The eigenvector has large positive values in the direction of indices 3, 8, and 12 corresponding to *arousing, aggressive*, and *restless* and large negative values for the moods *calming, tender*, and *peaceful* (indices 2, 5, and 9). We interpret this eigenvector as reflecting a state of arousal.

In Figure 3, the second eigenvector is plotted. The eigenvector has large positive values in the direction of indices 1 and 8 corresponding to *sad* and *aggressive*, and large negative values for the moods *cheerful* and *carefree* (indices 6 and 7). We interpret this eigenvector as reflecting a valence or pleasantness dimension.

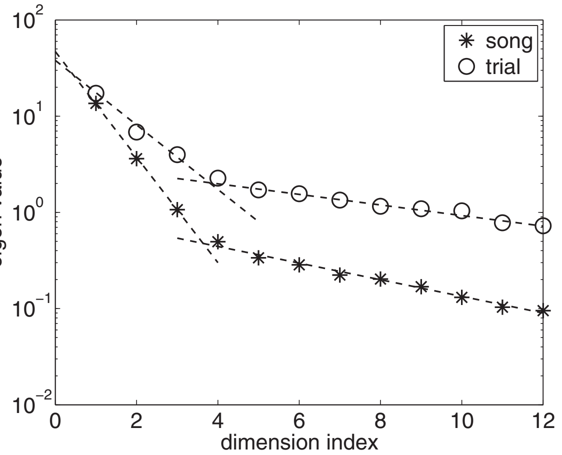

Figure 1 Eigenvalues as function of index from song and trial-based covariance analysis. The dashed lines are manual fits to the data.

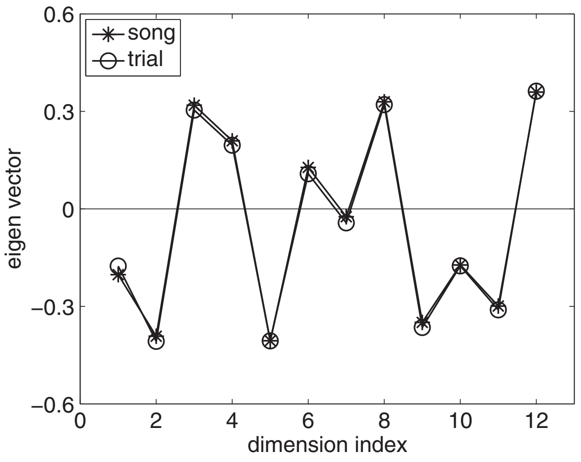

Figure 2 First eigenvector as function of mood label index.

In Figure 4, the third eigenvector is plotted. The eigenvector has large negative values for the moods *powerful* and *emotional* (indices 4 and 10). This dimension is reminiscent of dominance or Wedin's third dimension in music mood [2].

The above interpretations correspond well with the dominant mood dimensions known from literature [2,12,16,17]. Most common are mood interpretations in two dimensions: valence and arousal. The third dimension is typically weak and consensus is missing.

In conclusion, the number of basic dimensions and their character is very much in line with the expectations. These expectations were actually the basis for including the valence and arousal rating in our test. As a third dimension we incorporated upfront the attribute tenseness. There is however no clear indication in our analysis that the third dimension corresponds to tenseness.

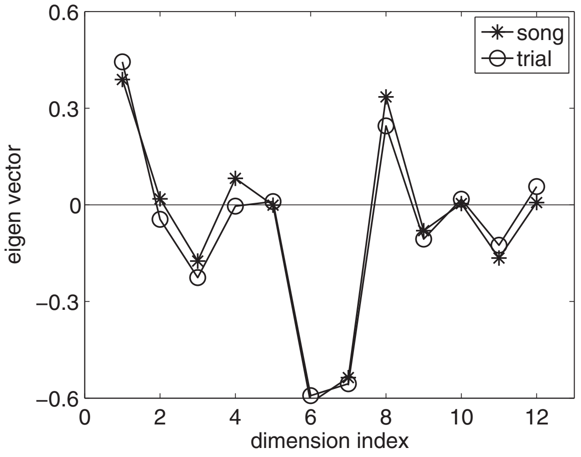

Figure 3 Second eigenvector as function of mood label index.

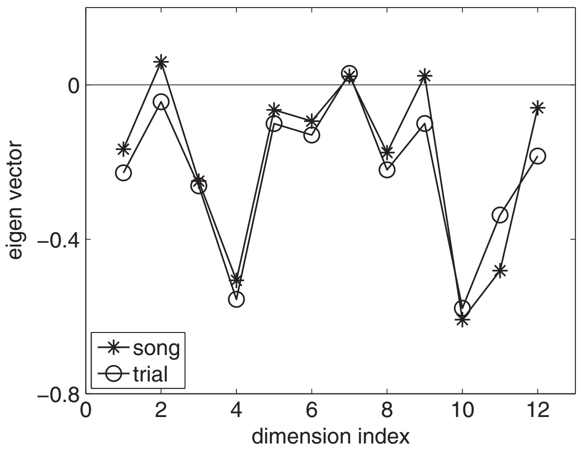

Figure 4 Third eigenvector as function of mood label index.

In the remainder of this article, we consider only the valence and arousal ratings and not the tension dimension rating. The reason to concentrate on the first two dimensions is threefold. To start with, these are the most significant directions as indicated by the eigenvalue analysis. Second, in contrast to the interpretation of the first two dimensions, the eigenvector interpretation for the third dimension is not clearly associated with the actually measured variable. Lastly, we considered the correlation coefficients between the three ratings, see Table 4. We infer that the rating along the tension axis is highly correlated with the other two ratings, while valence and arousal are almost uncorrelated. In other words, there is little independent information in the tension dimension rating. A similar finding is reported in [12].

Validation of axis interpretation

On the basis of the eigenvectors we argued that the two main dimensions in music mood are associated with valence and arousal. If this is the case then we should be able to estimate the experimental valence and arousal ratings from the (two) main dimensions found from the mood analysis. This issue is considered here in a qualitative sense.

In view of the applications, we are interested in tools for an average user. Therefore, we use S_s which is the 288 × 12 matrix containing the subject-averaged mood ratings

Table 4 Correlation coefficients between valence (V), arousal (A), and tenseness (T)

	Song-based			Trial-based		
	V	**A**	**T**	**V**	**A**	**T**
V	1	−0.27	−0.63	1	0.04	−0.47
A	−0.27	1	0.79	0.04	1	0.51
T	−0.63	0.79	1	−0.47	0.51	1

per song. From the covariance matrix, we determine its eigenvectors and eigenvalues and select the two dominant eigenvectors which we store in a matrix called V. Thus, V is a 12×2 matrix. We map the observations S_s onto a two-dimensional plane spanned by these eigenvectors by

$$[\hat{S}_1\ \hat{S}_2] = \hat{S}_s = S_s V, \tag{1}$$

with $\hat{S}_s$ being a 288×2 matrix containing the vectors $\hat{S}_1$ and $\hat{S}_2$. These vectors reflect the two dominant mood dimensions according to the eigenvalue decomposition (PCA).

Consider now the song-based valence or arousal ratings denoted as vectors r_v and r_a, respectively, both of length 288. If the dominant eigenvectors correspond to the dimensions of valence and arousal that were actually measured, we should be able to predict them with sufficient accuracy. The prediction should be a linear predictor, the accuracy is assessed by a χ^2 goodness-of-fit criterion.

When applying the matrix eigenvectors to the measured mood ratings, we obtain 12 orthogonal signals. We suggested in the previous section that the two dominant directions would equal the valence and arousal axis. However, the measured valence and arousal ratings are not completely orthogonal (see Table 4), while the two dominant directions from the covariance analysis are by definition orthogonal. This means that a straightforward identification of the first and second dominant dimensions from the mood ratings with the arousal and valence rating, respectively, is not strictly proper.

It is however still possible that the space spanned by the two dominant mood dimensions equals the space spanned by the valence and arousal rating. We consider this case where we also assume that the first dominant dimension from the mood covariance analysis corresponds to the arousal rating, thus the assumed linear relation is given by

$$r_a \approx \hat{r}_a = c_{a0} + \hat{S}_1 c_{a1}, \tag{2}$$

$$r_v \approx \hat{r}_v = c_{v0} + \hat{S}_1 c_{v1} + \hat{S}_2 c_{v2}. \tag{3}$$

In words, on the basis of the measured non-orthogonality, we adapt our interpretation that the first and second dimensions are arousal and valence, respectively, by the assumption that the first dimension is arousal but that valence depends not only on the second principal dimension, but also partly on the first one.

The optimal parameters c were determined using a least-squares error criterion, i.e., $\min \sum_k (r_z(k) - \hat{r}_z(k))^2$, where k denotes the song index and z is either a or v. The prediction and the actual data are shown in Figures 5 and 6 and suggest that there is a good correspondence, i.e., that the main dimensions as found from the mood ratings indeed lie in the plane spanned by valence and arousal.

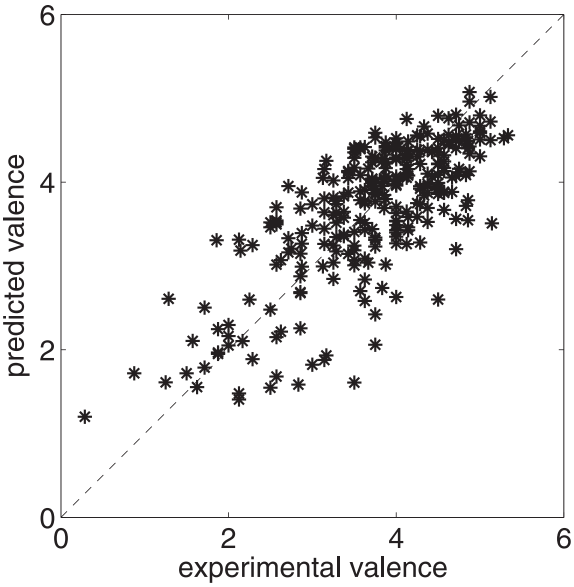

Figure 5 Valence rating per song and predicted valence rating based on the two dominant dimensions from the mood ratings.

In order to assess, in a quantitative way, whether there is a good correspondence, we consider a χ^2 goodness-of-fit test. The goodness of fit is evaluated by considering whether the sum of squared errors divided by the estimated variance of the mean is within a 95% confidence interval of the probability density function defined by a χ^2 distribution with D degrees of freedom [18]. The

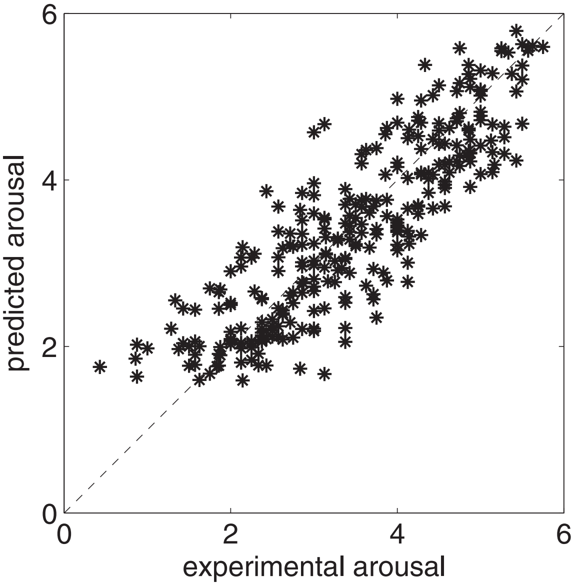

Figure 6 Arousal rating per song and predicted arousal rating based on the dominant dimension from the mood ratings.

Table 5 Goodness of fit evaluation

Rating	E	D	2.5%	97.5%
Valence	311	285	240	334
Arousal	288	286	241	335

Per rating, the weighted squared error sum E, the degrees of freedom D and the 2.5 and 97.5% confidence points of the χ^2 distribution are given.

outcome of the test is given in Table 5 where the weighted prediction error energies E_v and E_a are defined as

$$E_v = (r_v - \hat{r}_v)^T(r_v - \hat{r}_v)/q_v^2, \tag{4}$$

$$E_a = (r_a - \hat{r}_a)^T(r_a - \hat{r}_a)/q_a^2. \tag{5}$$

with q^2 the variance of the mean which we estimated from the distribution of the measured standard variances over the songs. The table also shows the 2.5 and 97.5% points of a χ^2 distribution with D degrees of freedom. From the table, it is clear that the error energies nicely agree with the expected values based on the measurement noise since the error E lies in the 95% confidence interval.

We conclude that the two principal dimensions in music mood correspond to the plane spanned by the actually measured valence and arousal ratings.

Music moods in the valence–arousal plane

Before considering locations of different moods in the valence–arousal plane, we first take a look at the distribution of the individual songs in this plane. In Figure 7, we plotted the (mean) valence and arousal ratings for each of the 288 songs. We note that these points are spanning roughly a triangular part of the plane; the lower-left triangular part is virtually empty. This might just mean that we did not include songs in our test with negative valence and arousal. We will return to this issue later after having inspected the locations of different moods.

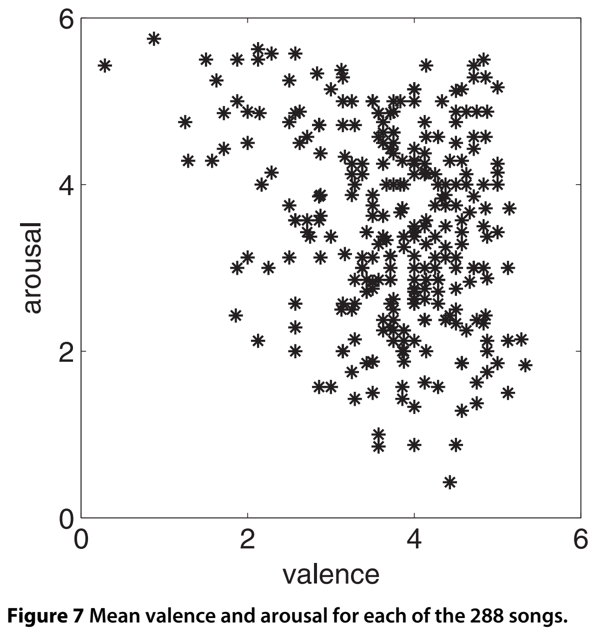

Figure 7 Mean valence and arousal for each of the 288 songs.

In order to establish relations between mood categories and valence and arousal ratings, we took the following approach. For a particular mood, we selected all trials which had an extreme rating, i.e., either 6 (definitely this mood) or 0 (definitely not this mood). Since we have 12 moods, this gives us 24 categories: 12 moods and 12 negated moods. On top of that, we added a *no mood* category. This was defined as all trials for which all mood ratings are in the mid range: 2–4.

First, we considered whether there was sufficient material within each of these categories. We simply counted the number of trials which gave for each mood a rating of 6 and a rating of 0. Similarly, we counted the number of records in the *no mood* category. The result is plotted in Figure 8. From these numbers, we infer that per category we have a sufficient number of ratings to allow processing by standard statistical tools.

For each of these categories, we calculated the means and covariance matrix from the ratings for valence and arousal. In the simplest view, this reflects a Gaussian probability density function in the valence–arousal plane. The ratings, however, have a limited range (0–6), while such a probability density function would range over two full real number axes. In order to stick to this interpretation, we therefore decided to use a logarithmic scaling of the valence and arousal ratings in the following way. First, we interpret the ratings $0, \ldots, 6$ as mid points of category bins of size 1, meaning that the full scale size F equals 7 and its center C equals 3. We map this linearly to the range $[-1, 1]$ by

$$z = 2\,\frac{r - C}{F}, \tag{6}$$

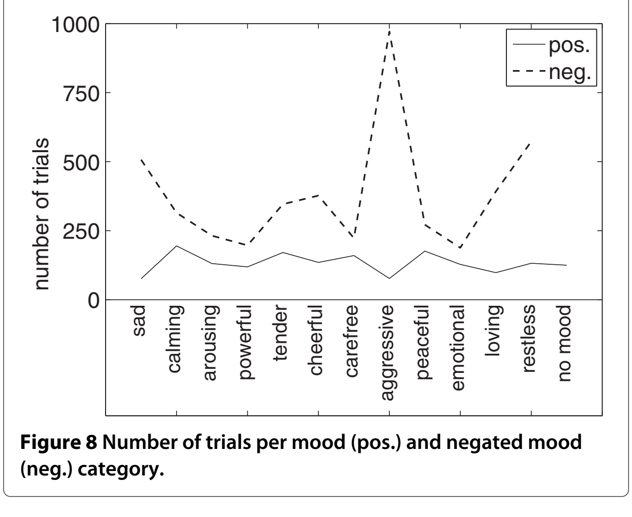

Figure 8 Number of trials per mood (pos.) and negated mood (neg.) category.

where r is the rating $(0, 1, \ldots, 6)$ of the (subject- and song-dependent) valence or arousal. Next this range is mapped to the full real axis by a logarithmic operator according to

$$\rho = \log_{10}\frac{1-z}{1+z}. \tag{7}$$

The means (over subjects) of ρ_v and ρ_a were determined where ρ_v and ρ_a are the mapped valence and arousal rating, respectively. The associated variances per song were determined as well. We note that all subsequent qualitative conclusions are independent of the mapping, i.e., we can draw the same conclusions without the mapping, but in that case interpretation of means and covariance matrices as a Gaussian blob is essentially not permitted.

Locations

In Figure 9, we have plotted ovals to represent the locations for each of the moods A to L (see Table 1) in the valence–arousal plane. The ovals are centered at the mean values for each specific mood and reflect the σ contour defined by the covariance matrix. We also included the no-mood category.

From this figure, we see that none of the categories has its centroid located in the lower-left quadrant. We also see that the ovals are large and overlapping. Some ovals are almost completely on top of each other, e.g., moods B, E, I, K. Lastly, these moods cover, roughly speaking, the outer range of the plane having either a positive valence or a positive arousal (or both).

In Figure 10, we have plotted ovals for each of the negated moods. Again, the ovals are centered at the mean values for each specific negated mood and reflect the σ contour defined by the covariance matrix. The union of these negated categories is an oval roughly stretching from top-left to bottom-right in the VA plane.

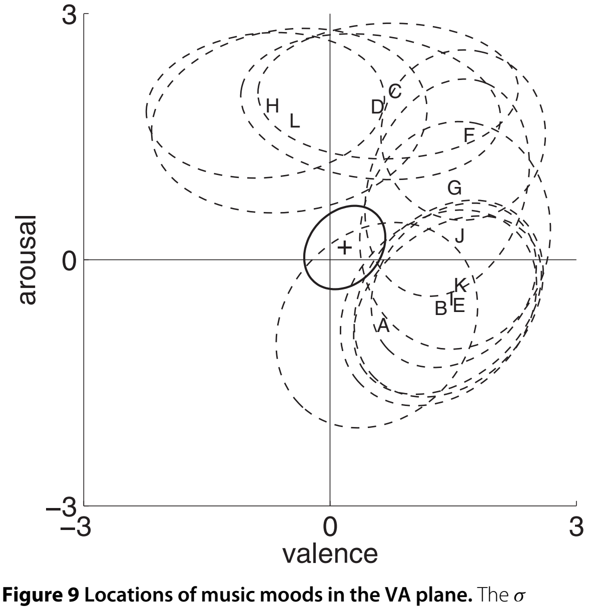

Figure 9 Locations of music moods in the VA plane. The σ contours of the different moods are indicated by dashed lines. The solid line indicates the *no mood* category.

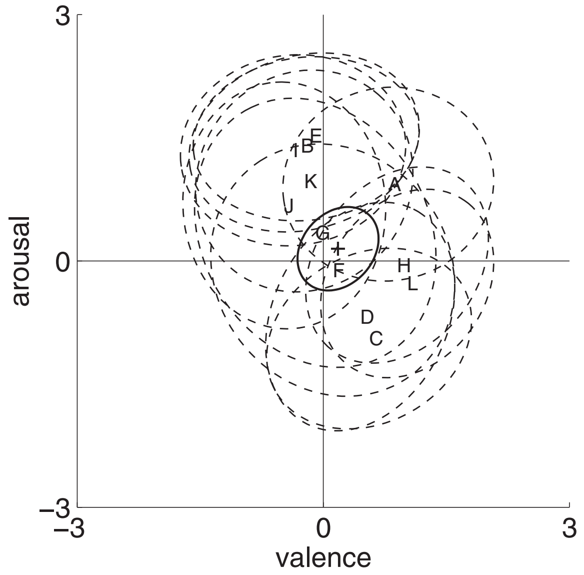

Figure 10 Locations of negated mood categories in the VA plane. The σ contours of the different negated moods are indicated by dashed lines. The solid line indicates the *no mood* category.

Comparison with affect word scaling

The circumplex model of affect is the dominant model for emotions which asserts that emotions are governed by two underlying variables: valence and arousal [16,17]. Emotions (affect words) have been scaled in this model and show a specific ordering of these words roughly on a circle in this plane. Since the music mood categories are dominantly determined by the same two variables, it is possible to compare the location of music mood categories characterized by the labels with those of affect words. From the discussion in the "Introduction" section concerning the difference between mood and emotions and the fact that our mood category locations are derived from mood, valence, and arousal ratings in music, it is not *a priori* clear what the correspondence or difference would be.

We can now compare the ordering of our music mood category labels with the ordering of the affect words on the circle. This is done in Table 6. Traveling clockwise from the upper-left corner in the VA plane we encounter the mood labels as given in the left column of Table 6. A similar exercise was done for the affect words using Figure two from Russell [16], using those affect words which have an intuitive correspondence with our labels. We infer that the ordering of the music moods in the VA plane obtained

Table 6 Ordering of music mood labels and affect words

Music mood	Affect word
Angry/Furious/Aggressive	Angry
Restless/Jittery/Nervous	Tense, Alarmed
Powerful/Strong	–
Arousing/Awakening	Aroused, Astonished
Cheerful/Festive	Excited
Carefree/Lighthearted/Light/Playful	Delighted
Emotional/Passionate/Touching/Moving	Happy
Loving/Romantic	Pleased, Glad
Peaceful	Serene
Tender/Soft	At ease
Calming/Soothing	Calm, Relaxed
Sad	Sad

in our experiment agrees well with the ordering of affect words in Russell's circumplex model.

Though there is a good agreement between the ordering of the music mood labels and the affect words, the actual positions are not always the same. Especially, the music mood category *sad* (mood A in Figure 9) has a small positive valence (a finding corroborated by Eerola et al. [12] for short excerpts) whereas the affect word scaling for *sad* shows a negative valence. Also the music mood category *calming/soothing* (mood B in Figure 9) appears to have a more positive arousal than that given for the affect word *calming*.

Overall, given the positions of the centers of music mood and negated mood categories (see Figures 9 and 10), we argue that the whole circle is distorted to roughly a semi-circle. This is also in line with our initial observations on the locations of the individual songs (Figure 7) where no songs with a large negative value for both arousal and valence were observed. It also agrees with the notion developed when collecting songs for this particular experiment: though we tried to include songs having all possible valence/arousal combinations, it was impossible to find a song in our database which had both an unambiguously negative valence and arousal.

In absence of any further research, we can only speculate why this is so. Putting aside the already noted distinction between emotion and (music) mood, we note that we considered mainly western popular music and used song-based annotations. Popular music is associated with entertainment, so one could argue that no negative valence and arousal is to be expected. If at all, one might have such instances in small parts of the song but not as an overall mood rating.

Another line of reasoning is that when emotion is put into a song, it has to be mapped to a musical structure or expression. Intuitively, a musical expression always tends to be positively valued in either valence or arousal at least if the musical expression is familiar to and recognized by the listener. Thus, the emotion expressed in music distorts the valence–arousal plane. In that interpretation, the lower-left corner of the VA plane would be associated with non-musical sounds, unfamiliar music, or non-recognized musical expressions.

Another element could be that some emotions are difficult to maintain its pure form when expressing them in a song. Consider a sad emotion. Translating that into a song presumably implies coping with that emotion which might involve a change in character from, e.g., a pure sad emotion to, e.g., a more melancholic or angry mood.

As an overall conclusion, we state that our music mood ordering in the valence and arousal coordinates agrees well with affect word scaling data but not their actual positions.

Distinctive mood labels

Using the data from the VA ratings, we can now reconsider some of the starting points of the experiment. These were the following (see Section "Mood categories").

- Twelve mood categories are relatively consistent between subjects and easy-to-use as well;
- Mood categories are non-exclusive categories;
- Moods should be ranked as 'belonging in this class' or 'not-belonging to this class' (as opposed to working with antagonistic labels);
- Proper wording of the labels is important.

The current findings concerning the VA ratings suggest the following. In view of the fact that several music moods have the same position in the VA plane, we argue that not all categories are easy-to-use even though they may be consistent. In view of coverage of the VA space, it is more convenient to use a more limited number of mood categories. Also, some of the categories can be used in an antagonistic manner.

In Figure 11, we have plotted a heuristically reduced set of moods showing that with five moods and one negated mood category, we are essentially covering the full space. Furthermore, the categories are roughly equal-sized and with similar amounts of overlap to neighboring categories. Presumably, this would serve as a more convenient subset of moods from which to build applications for a consumer-style usage.

Valence and arousal prediction

In this section, we will consider whether music features allow prediction of the valence and arousal ratings. To this end we used four sets of music features that were

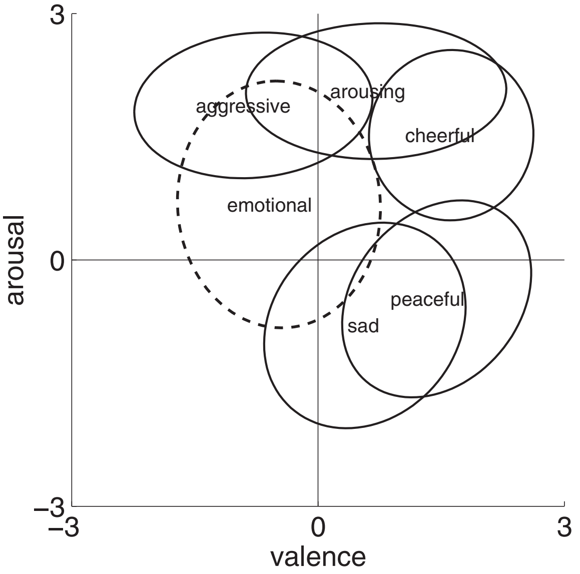

Figure 11 Sub-selection of mood categories. Five mood categories (solid lines) and one negated mood category (dashed line).

developed earlier. Table 7 gives an overview of the features extracted from the audio. The first category consists of spectral features (MFCCs, which comprises information concerning loudness and spectral tilt) and temporal behavior (modulations), see [19]. The set of tonality features is based on the calculation of the chromagram as described in [20,21] and includes features like key, major/minor classification, chroma flatness, consonant strength, chroma eccentricity in a circle of fifths and harmonic strangeness (correlation between long- and short-term chroma). The rhythm feature set [22,23] is based on the resonator filter bank approach [24], and contains features which represent statistics (e.g., means and standard deviations) concerning onset synchrony, inter-onset intervals and tempo. The percussion features are based on estimates of the signal envelopes of band-limited signals between two consecutive onsets. The envelopes are modeled in terms of attack, decay, sustain, and release phases. Next, a classifier rates these parts as percussive or non-percussive. The classifier ratings over the whole song are condensed into a set of statistics and form the basic percussiveness features [25,26].

Table 7 Feature categories, number per category and examples

Feature class	Number	Description
Spectral	22	MFCC and modulations
Tonality	26	Chroma, key, consonants, dissonants, harmonic strangeness chroma eccentricity
Rhythm	19	Tempos (fast and slow), onsets, inter-onset intervals
Percussiveness	21	Characterization and classification of onsets per band

We use the following terminology. The song index is called k with $1 \leq k \leq K$, where K is the total number of songs in the test, i.e., $K = 288$. Per song we have a mean rating for valence and arousal (mapped according to Equations 7 and 6) denoted as $\rho_v(k)$ and $\rho_a(k)$, respectively. The mean is established as the mean over the subjects that rated that particular song. From the individual ratings, we can estimate the variance which we denote as $s_v^2(k)$ and $s_a^2(k)$. In the remainder, we often drop the subscripts $_a$ and $_v$ since the treatment of the data is identical for both cases. That means that where we introduce new variables, these may reappear with the subscripts indicating that we consider specifically one or the other rating.

The features are called $f(k,i)$, k denoting the song index and i denoting the feature index. We have 79 different features so $i = 1, 2, \ldots, 79$. Suppose that I is a subset out of these 79 features. The model that we use is a linear model, i.e.,

$$R(k) = A_0 + \sum_{i \in I} A_i f(k,i) \tag{8}$$

where R is the model prediction and the set $\{A_i\}$ constitutes the model parameters. The prediction error $\epsilon(k)$ is defined as

$$\epsilon(k) = \rho(k) - R(k). \tag{9}$$

We optimize the parameters according to a mean-squared error criterion,

$$\arg \min_{A_i} \sum_{k=1}^{K} w(k) \epsilon^2(k), \tag{10}$$

and do this for valence and arousal ratings, separately. We used the weighting w to counteract the effect that we have a high density around the mean values (i.e., we emphasize the outlying valence and arousal samples slightly). The effect of the weighting is minor.

Since the set contains a number of highly correlated features we first reduced the set by removing 13 features such that high correlations between features were prevented. Next we used a greedy ordering method to get insight into the number of relevant features required to get a good prediction. For that purpose, we started with a prediction using the full set and next reduced this set by removing the feature which attributed least to the prediction. This procedure was repeated until we were left with the offset A_0. This procedure gives a different ordering per attribute (valence and arousal).

The result in terms of the ratio of adjusted squared error deviation σ^2 and measured variance of the mean

q^2 is plotted in Figure 12 as a function of the number of acoustic features where

$$\sigma^2 = \sum_{k=1}^{K} \frac{\epsilon^2(k)}{D} \tag{11}$$

with D the number of degree of freedom (i.e., the number of data points minus the number of parameters used in the fit). We see that for both valence and arousal a minimum is reached around 31. We use these subsets in the remainder of this article. The subsets contain elements of all four feature subcategories: MFCC, percussiveness, tonality, and rhythm. In total the least-squares fits uses $N + 1 = 32$ free parameters since a constant offset (A_0) is also used as a free variable.

In Figures 13 and 14, we have plotted the ratings $\rho(k)$ and the model prediction $R(k)$ for valence and arousal, respectively. These plots suggest that a linear model provides a reasonably accurate description of the subject-averaged ratings.

In order to assess this in a quantitative way, we applied the same method as in Section "Validation of axis interpretation", i.e., we considered whether the sum of squared errors divided by the estimated variance of the mean is within a 95% confidence interval of the probability density function defined by a χ^2 distribution with D degrees of freedom [18]. The number of parameters in the fit is $N+1 = 32$, the number of degrees of freedom D is therefore $D = K - N - 1 = 256$. In Table 8, we tabulated the weighted squared error S defined as

$$S = \sum_{k=1}^{K} \frac{\epsilon^2(k)}{q^2} \tag{12}$$

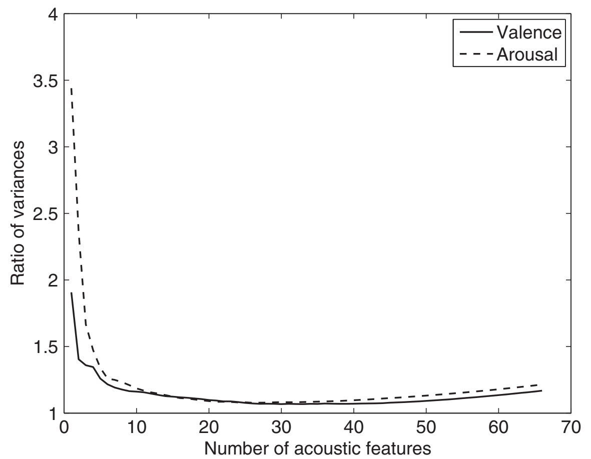

Figure 12 Number of acoustic features. Ratio of adjusted squared error deviation (σ^2) and measured variance of the mean (q^2) as a function of the number of acoustic features.

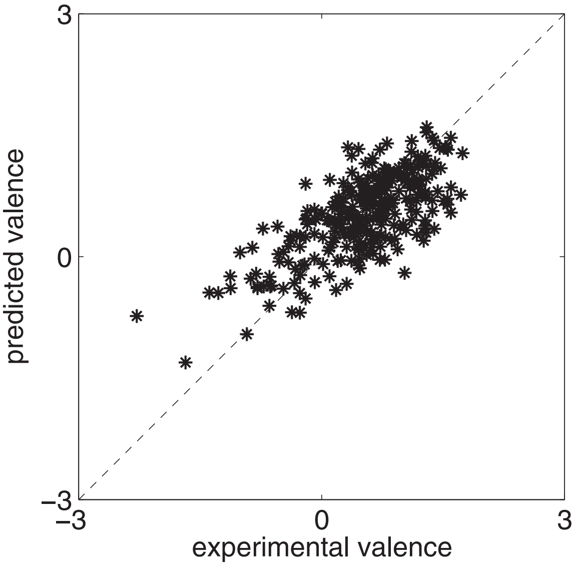

Figure 13 Valence prediction. Mean mapped valence ratings per song $\rho_v(k)$ and prediction $R_v(k)$ from features.

as well as the 2.5 and 97.5% points of the expected distribution. From the table, it is clear that the average error nicely agrees with the expected value based on the measurement noise since the error S lies in the 95% confidence interval. These results show that subject-averaged valence and arousal ratings can adequately be predicted using features automatically extracted from the music. For completeness, we have also included in the table the standard

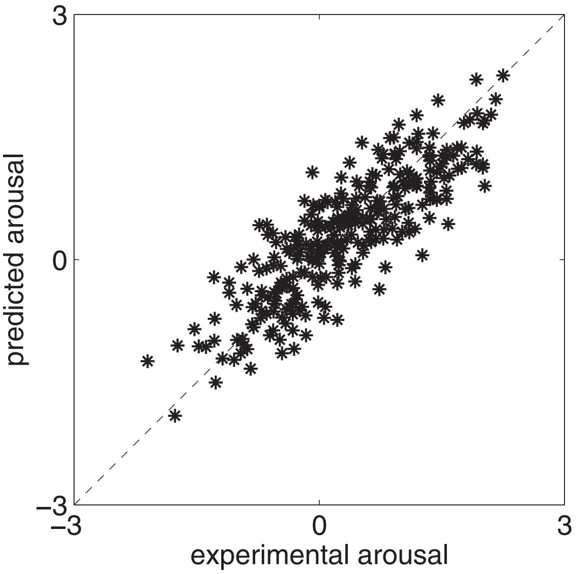

Figure 14 Arousal prediction. Mean mapped arousal ratings per song $\rho_a(k)$ and prediction $R_a(k)$ from features.

Table 8 Goodness-of-fit evaluation

Rating	S	D	2.5%	97.5%	q	σ	c
Valence	274	256	213	302	0.47	0.49	0.71
Arousal	277	256	213	302	0.47	0.48	0.85

Per rating, the weighted squared error sum S, the degrees of freedom D and the 2.5 and 97.5% confidence points of the χ^2 distribution are given. Also given are q, σ and c (for an explanation see text).

Table 9 Important features in the prediction of valence and arousal

Valence	Arousal
Chroma	Slow tempo
Percussiveness variability across bands	Loudness
Measure on ratio fast and slow tempos	Chroma eccentricity
Modulation spectrum	Fast tempo
Harmonic strangeness	Spectral tilt

deviation of the mean q, the standard deviation associated with the modeling error σ and the correlation coefficient c between the measurement ρ and the prediction R.

The goodness-of-fit tests indicate that the linear model neither overfits nor underfits the data: the mean valence and arousal ratings are on average predicted with an accuracy comparable to the measurement noise. There are no clear outliers: deviations larger than $3.5q$ do not occur. There are 13 songs with a deviation larger than $2q_v$ for valence and 10 songs with a deviation larger than $2q_a$ for arousal. These two sets of songs do not overlap. For a Gaussian distribution (i.e., the underlying assumption in a least-squares fit), one would expect about 5% of the data to be beyond the $2q$ boundary. Five percent of 288 amounts to 13 songs, i.e., in line with what we find. Lastly, we note that the two sets of songs beyond the $2q$ range were not concentrated in a specific area in the VA plane.

Given the predictability of valence and arousal, an obvious question that springs to mind is what are actually the dominant features, i.e., which features are essential in the fit. Before discussing this, we have several cautioning remarks. First, a slightly different criterion for the optimization may well result in a substantially different ordering. Second, we removed several features upfront. Third, the ordering process for the features is a greedy process due to the non-orthogonality of the features. Fourth, at the start of the ordering process there is a high number of features relative to the number of observations; a known source of problems for reliable linear regression results and ordering [12]. With this in mind and in view of the roll-off in Figure 12, we consider only the first five features according to the ordering process. These are given in Table 9.

A number of studies [11,27-29] considered correlations between music features and valence and arousal. A comparison is not straightforward due to differences in experimental conditions, in feature sets and in operational feature definition. Nevertheless, a comparison of the five features with highest correlation (in absolute sense) from these studies suggests that event densities, onsets, and spectral flux are important determinants for arousal. This is in line with the fact that tempo features rank high in our results. For valence, these studies suggest that modality measures are among the dominant factors. This has a counterpart in the top ranking of chroma and harmonic strangeness features in our case.

Discussion

In this section, we give evidence corroborating the particular shape of the distribution of the valence and arousal ratings, and we compare our results with earlier studies.

First, we use the model from Section "Valence and arousal prediction" to consolidate the finding that part of the VA plane does not naturally occur in western popular music. Since the emptiness of this part of the plane may stem from the particular songs that were used in the test (i.e., the test set is not representative for general popular music) we applied the VA prediction model to two larger databases also containing western popular music. The result for the largest database ($\approx$ 5000 songs) is shown in Figure 15, the result for the other database ($\approx$ 1700 songs) is similar. We again see that the lower-left quadrant is nearly empty which, in view of the linear prediction model, suggests that the songs in the test are not associated with an atypical set of song features. Thus, it supports the finding that lower-left quadrant of the VA plane for popular music is actually rather empty and indirectly refutes the notion that in our experiment this is due to a biased song selection.

From the earlier cited studies on mood in music, the studies of [6,9,11,12] are comparable as the first includes a PCA on the moods to arrive at the fundamental mood dimensions and the latter three contain data on direct VA rating. Our results concerning the VA ratings differ substantially from that in [9], where the coverage of the VA plane was essentially an oval with main axis at 45°. As noted in the "Introduction" section, this may be caused by many different factors in the experimental set up. Our experimental results are corroborated by those in [6] in all major aspects. Their analysis using linear discriminant analysis and PCA analysis also showed the boomerang-shaped 2D plane coverage that we observe and, as in line with our analysis in Section "Dimensions in music mood", we assume that their 2D PCA plane is also essentially the VA plane. Also their finding of a better inter-subject consistency for arousal than for valence is supported by our study. We showed that six (about 50% overlapping) mood categories cover

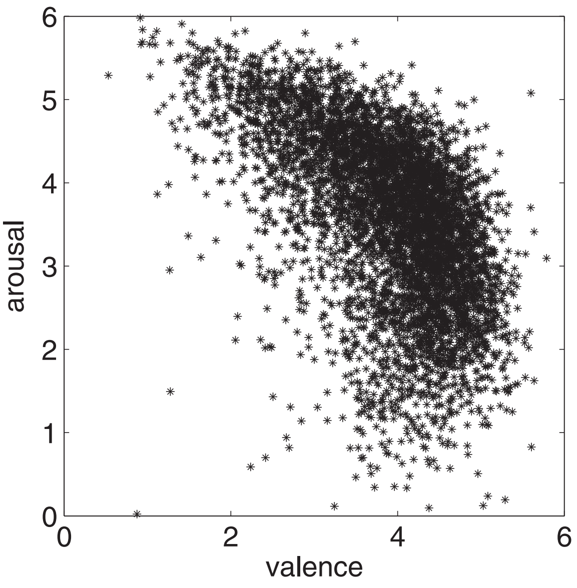

Figure 15 VA prediction. Prediction of the VA ratings for songs from the artist database.

the pertinent VA plane. This roughly translates to three non-overlapping categories as used in [6]. If we would reduce our categories to *aggressive*, *cheerful*, and *sad* only, we have three non-overlapping categories covering the major part of the VA plane where song ratings occur (see Figure 11). These three non-overlapping categories correspond well with their labels: *aggressive*, *happy*, and *melancholy*.

In [11], energy (arousal) and valence ratings for a set of popular ring tones is considered. A higher mean intersubject correlation is reported for energy than for valence, in line with our results and [6]. Results on prediction of valence and arousal from music features are reported as well, although these were considered preliminary outcomes. The performance of the prediction is given nevertheless in terms of amount of adjusted explained variance, with actual numbers of 0.68 and 0.50 for energy and valence, respectively. The adjusted explained variance for our data and feature set is in line with these results yet better: 0.75 for arousal and 0.59 for valence.

In [12], valence and arousal ratings are presented for short musical excerpts (about 15 s) covering, according to their terminology, five different emotions. The emotion categories are *happy*, *sad*, *tender*, *scary*, and *angry* and test excerpts adhering to these categories were selected by expert listeners. Valence and arousal ratings were collected and predicted from acoustic features using various modeling approaches. Depending on the approach, the explained variance for valence ranged between 0.42 and 0.72, that for arousal between 0.73 and 0.85. Leaving aside the difference between explained variance and adjusted explained variance (as in our case), these numbers agree well with ours. We also note that at least four out of five of their emotion adjectives have a strong correlate with our mood label adjectives.

Conclusions

A music mood web experiment was successfully organized and executed. The results indicate that with a careful set-up, the subjectivity of mood aspect can be controlled such as to generate meaningful subject-averaged ratings. Furthermore, the results largely confirm our assumptions with respect to the number of moods and non-antagonistic labeling. Nevertheless, the results also suggest that part of the labels (directions in the mood space) can be more properly condensed into a more limited number of dimensions including antagonistic labeling for some dimensions. Our study demonstrates how the VA plane can be used as an effective intermediate representation for finding a minimum number of mood categories.

The mood results were analyzed for basic dimensions underlying the mood judgment. An eigenvalue decomposition showed that there are at most three relevant directions in music mood judgments, a result in line with literature. The two main directions are valence and arousal.

The mood ratings were used to identify areas in the valence–arousal plane corresponding to different moods. The ordering of the moods in the valence and arousal plane is in line with the circumplex affect model. However, the actual positions of the mood centers (or the outer mood boundaries) do not constitute a full circle. Thus, we have shown that the music mood space for western popular music differs from the typical VA space associated with affect words.

We applied a linear model to predict the mean valence and arousal ratings. It is shown that this yields an accurate model for these dimensions. It implies that the mood (or moods) of a song can be estimated since the moods are determined by the position in the valence–arousal plane.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Philips Research, High Tech Campus 36, NL-5656 AE Eindhoven, The Netherlands. [2]Telekom Innovation Laboratories, Technical University Berlin, Ernst-Reuter-Platz 7, D-10587 Berlin, Germany.

Received: 21 February 2012 Accepted: 12 September 2012
Published: 3 October 2012

References

1. YE Kim, EM Schmidt, R Migneco, BG Morton, P Richardson, J Scott, JA Speck, D Turnbull, in *Proc. ISMIR 2010; 11th Int. Soc. Music Inf. Retrieval Conf.* Music mood recognition: a state of the art review. (Utrecht, The Netherlands, 2010), pp. 255–262
2. A Gabrielsson, PN Juslin, in *Handbook of Affective Sciences*, ed. by RJ Davidson, KR Scherer, HH Goldsmith Emotional expression in music. (Oxford University Press, Oxford, 2009), pp. 503–534

3. Y Feng, Y Zhuang, Y Pan, in *Proc. 26th Int. ACM SIGIR Conf. on R&D in Information Retrieval* Popular music retrieval by detecting mood. (Toronto, Canada, 2003), pp. 375–376
4. T Li, M Ogihara, in *Proc. ISMIR 2003; 4th Int. Symp. Music Information Retrieval* Detecting emotion in music. (Baltimore, MD, USA, 2003), pp. 239–240
5. D Liu, L Lu, HJ Zhang, in *Proc. ISMIR 2003; 4th Int. Symp. Music Information Retrieval* Automatic mood detection from acoustic mood data. (Baltimore, MD, USA, 2003), pp. 13–17
6. M Tolos, R Tato, T Kemp, in *Proc. CCNC'05; 2nd IEEE Consumer Communications and Networking Conference*. Mood-based navigation through large collections of musical data. (Las Vegas, NV, USA, 2005), pp. 71–75
7. A Friberg, E Schoonderwaldt, PN Juslin, CUEX: an algorithm for extracting expressive tone variables from audio recordings, Acta Acustica united with Acustica. **93**, 411–420 (2007)
8. K Trohidis, G Tsoumakas, G Kalliris, I Vlahavas, in *Proc. ISMIR 2008; 9th Int. Symp. Music Information Retrieval* Multi-label classification of music into emotions. (Philadelphia, PA, USA, 2008), pp. 325–330
9. B Schuller, J Dorfner, G Rigoll, Determination of nonprototypical valence and arousal in popular music: features and performances, EURASIP J. Audio Speech Music Process. **2010**, 735854 (2010). doi:10.1155/2010/735854
10. R Panda, RP Paiva, in *130th AES Convention* Using support vector machines for automatic mood tracking in audio music. (Conv Paper 8378, London, UK, 2011)
11. A Friberg, A Hedblad, in *Proc. 8th Sound and Music Computing Conference* A comparison of perceptual ratings and computed audio features. (Padova, Italy, 2011), pp. 122–127
12. T Eerola, O Lartillot, P Toiviainen, in *Proc. ISMIR 2009; 10th Int. Symp. Music Information Retrieval* Prediction from multidimensional emotional ratings in music from audio using multivariate regression models. (Kobe, Japan, 2009), pp. 621–626
13. J Skowronek, MF McKinney, in *Proc. 2006 ISCA Tutorial and Research Workshop on Perceptual Quality of Systems* Quality of music classification systems: how to build the reference? (Berlin, Germany, 2006), pp. 48–54
14. J Skowronek, MF McKinney, S van de Par, in *Proc. ISMIR 2006; 7th Int. Conf. Music Retrieval Information* Ground truth for automatic music mood classification. (Victoria, Canada, 2006), pp. 395–396
15. J Skowronek, MF McKinney, S van de Par, in *Proc. ISMIR 2007; 8th Int. Conf. Music Retrieval Information* A demonstrator for automatic music mood estimation. (Vienna, Austria, 2007), pp. 345–346
16. JA Russell, A circumplex model of affect, J. Personal. Soc. Psychol. **39**, 1161–1178 (1980)
17. J Posner, JA Russell, BS Peterson, The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology, Dev. Psychopathol. **17**(3), 715–734 (2005)
18. JR Wolberg, *Prediction Analysis* (Van Nostrand, New York, 1967)
19. MF McKinney, J Breebaart, in *Proc. 4th Int. Conf. Music Information Retrieval (ISMIR)* Features for audio and music classification. (Baltimore, USA, 2003), pp. 151–158
20. S van de Par, M McKinney, A Redert, in *Proc. 7th Int. Conf. Music Information Retrieval* Musical key extraction from audio using profile training. (Victoria, Canada, 2006), pp. 328–329
21. S Pauws, in *Intelligent Algorithms in Ambient and Biomedical Computing,* ed. by W Verhaegh, E Aarts, J Korst Extracting the key from music. (Springer, Dordrecht, 2006), pp. 119–132
22. MF McKinney, D Moelants, Ambiguity in tempo perception: what draws listeners to different metrical levels? Music Perception. **24**(2), 155–166 (2006)
23. MF McKinney, D Moelants, in *Proc. 5th Int. Conf. on Music Info. Retrieval* Extracting the perceptual tempo from music. (Barcelona, Spain, 2004)
24. ED Scheirer, Tempo and beat analysis of acoustic musical signals, J. Acoust. Soc. Am. **104**, 588–601 (1998)
25. J Skowronek, MF McKinney, Method and electronic device for determining a characteristic of a content item. US Patent US7718881B2, 18 May 2010
26. J Skowronek, M McKinney, in *Intelligent Algorithms in Ambient and Biomedical Computing,* ed. by W Verhaegh, E Aarts, J Korst Features for audio classification: percussiveness of sounds. (Springer, Dordrecht, 2006), pp. 119–132
27. J Fornari, T Eerola, in *Proc. CCMR 2008* The pursuit of happiness in music: retrieving valence with contextual music descriptors. (Copenhagen, Denmark, 2008), pp. 119–133
28. AP Oliveira, A Cardoso, in *10 Encontro de Engenharia de Áudio da AES Portugal* Emotionally-controlled music synthesis. (Lisbon, Portugal, 2008)
29. I Wallis, T Ingalls, E Campana, J Goodman, in *Proc. 8th International Sound and Music Computing (SMC) Conf.* A rule-based generative music system controlled by desired valence and arousal. (Padova, Italy, 2011)

doi:10.1186/1687-4722-2012-24
Cite this article as: den Brinker *et al.*: Expressed music mood classification compared with valence and arousal ratings. *EURASIP Journal on Audio, Speech, and Music Processing* 2012 **2012**:24.