DESIGN AND EVALUATION OF SEMANTIC MOOD MODELS FOR MUSIC RECOMMENDATION

Mathieu Barthet[1], David Marston[2], Chris Baume[2], György Fazekas[1], Mark Sandler[1]
[1]Centre for Digital Music, Queen Mary University of London
{firstname.lastname}@eecs.qmul.ac.uk
[2]BBC R&D London, {firstname.lastname}@bbc.co.uk

ABSTRACT

In this paper we present and evaluate two semantic music mood models relying on metadata extracted from over 180,000 production music tracks sourced from I Like Music (ILM)'s collection. We performed non-metric multidimensional scaling (MDS) analyses of mood stem dissimilarity matrices (1 to 13 dimensions) and devised five different mood tag summarisation methods to map tracks in the dimensional mood spaces. We then conducted a listening test to assess the ability of the proposed models to match tracks by mood in a recommendation task. The models were compared against a classic audio content-based similarity model relying on Mel Frequency Cepstral Coefficients (MFCCs). The best performance (60% of correct match, on average) was yielded by coupling the five-dimensional MDS model with the term-frequency weighted tag centroid method to map tracks in the mood space.

1. INTRODUCTION

Research on music and emotions is a burgeoning field in the Music Information Retrieval community, generating an ever-increasing number of studies [1]. Although the nature of emotional responses to music is still a misunderstood and controversial topic [6], several analyses of music consumer needs and behaviors have highlighted the usefulness of developing search engines which can organise tracks according to the moods they express (see e.g. [8]). Even before the advent of online music technologies and the concurrent removal of physical media for music such as the CD, production music labels [1] created track compilations according to suggested moods; this music delivery model persists today (see e.g. the "Mood for Love" compilation from West One Music Group [2]). Production music has, by tradition, been called "mood music" as it is often composed to convey mostly constant emotional responses facilitating its association with a narrative, as opposed to commercial music which often presents larger changes of perceived emotions over time. Creative producers searching for background music for television, film, or radio often use keywords which evoke emotions (e.g. "happy music for kids"), or emotion-eliciting situations (e.g. "basketball music"). Unfortunately, due to the complex and subjective nature of music emotions there is no consensual agreement between production music libraries on the genesis and organisation of mood-related metadata. This hinders the recommendation of tracks based on mood, especially on music search engines which aggregate several production music catalogues. In this paper we propose two different methods to represent the relationships between mood tags (semantic mood models) based on analysis of mood-related metadata extracted from 183,176 production music tracks from I Like Music (ILM)'s aggregated catalogue (29 different production music libraries). We then assess how well the various models perform in a task of music recommendation when combined with five different tag summarisation techniques. In a companion study, the mood model providing the best accuracy (a five-dimensional model derived from multidimensional scaling analysis of mood tag co-occurences) was used for audio content-based music emotion recognition using support vector regression (SVR).

2. RELATED WORK

Other data-driven semantic mood models have previously been proposed. In [9], latent semantic analysis (LSA) techniques were applied to 105 emotion words occurring in tags associated with 8,872 Last.fm tracks. The resulting low-dimensional representation obtained after MDS analysis was in line with the traditional arousal and valence dimensions from Russell's circumplex [12], and a third dimension correlated with the *spiritual* or *meditative* component of musical experience. The conclusions from [5] also found mood spaces in agreement with the AV plane after applying MDS to co-occurrence counts of artists from the Last.fm dataset whose styles were described by 146 mood tags. The model from [9] was extended in [14] in which the affective circumplex transform (ACT) was applied to LSA-MDS spaces in order to infer explicit configurations matching Russell's circumplex. The use of ACT provided

[1] Production music is recorded music for use in film, television, radio and other media.

[2] http://www.westonemusic.com/album?catno=WOM_RFM_0018

better accuracy than the baseline LSA-MDS technique according to human judgements. The semantic mood models proposed in this paper (Section 3) differ from previous approaches in several ways. First, to the best of our knowledge, no formal assessment of such models has previously been conducted in relation to music recommendation (Section 4). Second, the models are derived from mood annotations curated by music experts (production music libraries) rather than social tags which are more noisy and idiosyncratic by essence. The analysed mood tags come from the large-scale ILM dataset including a high number of unique mood-related tags (2,060). The comparative study [13] showed that mood models derived from this set of curated editorial mood tags significantly outperformed mood models derived from Last.fm social tags, based on human judgements of arousal, valence and tension. Third, the models which obtained the best performance rely on a higher number of dimensions (5-D and 10-D) than the classic 3-D emotion model.

3. DESIGN OF SEMANTIC MOOD MODELS

3.1 Data-driven Mood Model (MDS)

In the same way that a measure of similarity between tracks can be derived from tag co-occurrence counts, a measure of similarity between tags can be derived from their co-occurrence over tracks [10].

3.1.1 Processing of Mood Tags

In the case of curated editorial metadata, tracks are associated with a list of unique tags judged to be the most appropriate by professional music experts. Hence, a given tag is only attributed once to a track, unlike for social tags for which a large number of users apply tags to tracks. Initially the mood tags were cleaned by correcting misspellings (100 errors out of 2,398 mood tags), removing duplicates (338 duplicates yielding 2,060 unique tags), and stripping white spaces and punctuation marks (e.g. '.', '!'). Instead of following a bag-of-words approach, in which the meaning of certain tags consisting of a series of words can be lost (e.g. "guilty pleasure"), we collated multiple words together to further process them as single entities (using a hyphen between the words). The vocabulary used in the ILM editorial annotations is composed of conventional words and does not suffer from the idiosyncrasies of social tags which often include sentences, informal expressions (e.g. "good for dancing to in a goth bar", cited as an example in [10]), or artists' names. For this reason, we did not have to tokenise the tags with a stop-list (for instance, removing words such as "it", "and", "the"). However, we used a stemmer algorithm [3] to detect tags with similar base parts (e.g. "joyful" and "joy"), as these refer to identical emotional concepts. 1,873 mood-related stems were obtained out of the 2,060 unique mood tags. In order to reduce the size of the stem vector while maintaining the richness of the track descriptions, we only kept stems which were associated with at least 100 tracks in further analyses. This stem filtering process yielded a list of 453 stems which provided at least one mood tag for each of the 183,176 tracks from the ILM dataset. The associations between tracks and stems are provided in a document-term matrix $\mathbf{X} = \{x_{ij}\}$ where:

$$x_{ij} = \begin{cases} 1 & \text{if stem } j \text{ is associated with track } i \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

The stem pairwise co-occurrences over tracks c_{ij} are then given by:

$$c_{ij} = |\{x_{\bullet i}\} \cap \{x_{\bullet j}\}| \quad (2)$$

where $\{x_{\bullet i}\}$ is the set of tracks annotated with stem i and $|\ \ |$ is the cardinality operator. The measure of dissimilarity between stems s_{ij} is computed as follows:

$$s_{ij} = 1 - \frac{c_{ij}}{Max(c_{ij})} \quad (3)$$

where $Max(c_{ij})$ is the maximum of the pairwise stem co-occurrence in the ILM dataset (26,859).

3.1.2 Multidimensional Scaling Analysis

Non-metric multidimensional scaling (MDS) analyses were then applied to the stem dissimilarity matrix, $\mathbf{S} = \{s_{ij}\}$. Four outlier stems with a null or very small co-occurrence measure compared to all the other stems were discarded so as not to bias the MDS analysis (this yielded a list of 449 stems). Figure 1 shows the evolution of Kruskal's stress1 [7] as the number of dimensions (D) increases from 1 to 13. Following a rule of thumb for MDS [3], acceptable, good and excellent representations are obtained for D = 3 (stress < 0.2), D = 5 (stress < 0.1) and D = 11 (stress < 0.05). Interestingly, five dimensions yield a good representation (elbow of the scree plot). This result suggests that more than three dimensions are required to accurately categorise mood terms in the context of production music, which contrasts with the classic three-dimensional emotion model (arousal, valence and dominance) [11]. In further analyses, we mapped the mood stems back to mood tags to uncover the meaning of the dimensions. We inspected whether the organisation of the terms along each dimension of the MDS configurations was relevant according to definable emotion-related concepts. For mood tags common to the list from the Affective Norm for English Words (ANEW) [2] (123 common tags were found out of our list of 449), we computed the correlations (Pearson's r) between the coordinates of the mood tags along each dimension of the MDS configurations and the ANEW measures of arousal, valence and dominance (see results in Table 1 in the case where D=5). Interestingly, three out of the five MDS dimensions are significantly correlated with the arousal and/or valence and/or dominance dimensions, showing that the 5-D MDS configuration captures aspects of the core emotion dimensions. However, some of the MDS dimensions are concurrently correlated with the arousal, valence and dominance dimensions. This is

[3] The PorterStemmer algorithm from the Natural Language Toolkit (NLTK) package for Python was used.

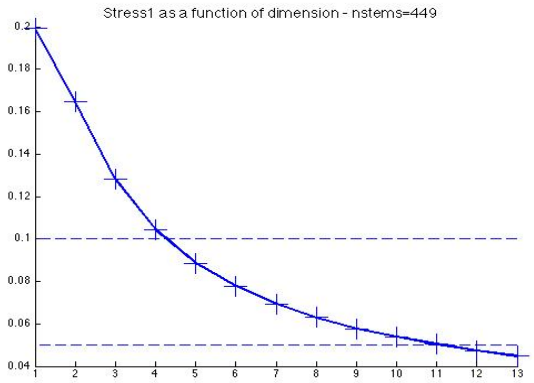

Figure 1. Kruskal's stress1 as a function of the number of dimensions in the MDS analysis of the mood stem dissimilarity matrix.

partly due to the fact that these dimensions can co-vary for certain emotions. For instance, the correlation between the ANEW valence and dominance dimensions is highly significant (r=.83, p<.001) which may explain why the second MDS dimension is correlated with both valence and dominance. A positive, although weak, correlation (r=.20, p<.05) was also found between the ANEW arousal and dominance dimensions. In contrast, no significant correlation was found between the ANEW arousal and valence dimensions. In order to disambiguate some of the MDS dimensions, and as MDS yields a solution which is invariant to rotation, we next applied a transformation to the MDS space to align it according to existing mood models.

3.1.3 Affective Circumplex Transform (ACT)

The affective circumplex transform (ACT) proposed in [14] was applied to the five-dimensional MDS configuration previously described. The goal of the ACT is to match the two first dimensions of the MDS configuration with the dimensions from Russell's arousal/valence (AV) model [12]. As in [13], we first determined mood terms common to both the MDS and Russell's AV spaces (37 terms were found in common). The ACT maintains the relative distances between mood terms in the initial MDS configuration, since it only allows translation, reflection, orthogonal rotation, and isotropic scaling. Measures of correlation between the transformed five-dimensional MDS configuration and the arousal, valence and dominance measures from the ANEW dataset are reported in Table 1. The first dimension of the MDS-ACT configuration is strongly correlated with valence (r=.56, p<.001) whereas before the ACT, no significant correlations with arousal, valence and dominance were found for that dimension. The strong relationship between valence and dominance can explain why the first MDS-ACT dimension is also correlated with dominance (r=.30, p<.001). The second and third dimensions are only correlated with arousal (r=.35, p<.001) and dominance (r=.34, p<.001), respectively. Hence, the ACT can infer an explicit representation according to the core emotion dimensions. Although no clear interpretations of the fourth and fifth dimensions have yet been found, the 5-D model yields a tag configuration which can be used to compute track distance measures, as described in the next two sections.

3.2 Tag Summarisation Methods

We devised several methods to summarise the tags of a track in a given multidimensional mood space. Let's denote $\mathbf{T} = \{t_{ij}\}$ as the tag matrix representing the coordinates of the tags i of a track across the dimensions j of the mood space. For the methods described in Sections 3.2.1 to 3.2.4, the tag summary matrix $\mathbf{Y} = \{y_{ij}\}$ is obtained by multiplying the tag matrix with a weight matrix $\mathbf{W} = \{w_i\}$.

3.2.1 Tag of Maximum Term Frequency (MTF)

This method assumes that a track is best represented by the tag from its set of tags which has the highest term frequency (TF) in the dataset. The weight w_i for the N tags of a track are as follows:

$$w_i = \begin{cases} 1 & if \quad TF(t_i) = Max_{i=(1...N)}[TF(t_i)] \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

3.2.2 Centroid (CEN)

This method summarises the tags of a track by their centroid or geometrical mean in the mood space. The tag weights are hence given by $w_i = \frac{1}{N}$.

3.2.3 Term-Frequency Weighted Centroid (TFW)

This method summarises the tags of a track by their centroid after attributing to each tag a weight proportional to its term frequency (TF): $w_i = \frac{TF(t_i)}{\sum(TF(t_i))}$. Hence the centroid is attracted by the tag with the highest term frequency.

3.2.4 Inverse Term-Frequency Weighted Centroid (ITF)

Conversely, this method attributes more weight to the tag with the lowest term frequency following the assumption that this tag may convey more specific information about the song: $w_i = \frac{1/TF(t_i)}{\sum(1/TF(t_i))}$.

3.2.5 Mean and Variance (MVA)

Rather than summarising the tags of a track by a point in the space, this method assumes that the tags can be represented by a Gaussian distribution. The tag summary matrix $\mathbf{Y}$ is given by the mean $\bar{\mathbf{T}}$ and variance $\sigma(\mathbf{T})$ of the tag matrix: $\mathbf{Y} = \{\bar{\mathbf{T}}; \sigma(\mathbf{T})\}$.

3.3 Model Derived from Mood Taxonomy (CLUST)

Popular mood key words were added to an initial selection provided by ILM to create a list of 355 mood words. Over 95% of the production music library contained at least one of these 355 words. Each of these words were placed in one of 27 categories, which became the starting point for a cluster-based model. Each category was treated as a cluster containing several mood words. Many

	Before ACT			After ACT		
Dimension	Arousal	Valence	Dominance	Arousal	Valence	Dominance
MDS Dim 1	-	-	-	-	.56***	.30***
MDS Dim 2	-	.31***	.36***	.35***	-	-
MDS Dim 3	-.33***	.47***	.19*	-	-	.34***
MDS Dim 4	-	-	-	-	-	-
MDS Dim 5	-.20 *	-.24**	-.31***	-	-	-

Table 1. Correlation (Pearson's r) between the dimensions of the five-dimensional MDS configuration and the arousal, valence and dominance dimensions, as characterised by the ANEW dataset. Only significant correlations are reported. *** $p < .001$, ** $p < .01$, * $p < .05$

of these clusters could be considered to overlap in their mood; some were clearly opposites while others had little in common. To convert these clusters into dimensions, the overlapping ones were combined into single dimensions; any opposite clusters were converted into negative (-ve) values of the dimension they were opposite to; and the non-overlapping clusters were treated as new dimensions. Using this method, the 27 clusters were converted to 10 dimensions, giving each of the 355 mood words 10 dimensional mood values. The choice of allocation of words to clusters and clusters to dimensions was performed based on only one person's opinion. The choice of 10 dimensions was a compromise between combining clusters that were too dissimilar and having too sparse a model. To illustrate the process, the first three dimensions represent the following mood clusters: 1) Confident (+ve scale), Cautious & Doubtful (-ve scale); 2) Sad & Serious (+ve scale), Happy & Optimistic (-ve scale); and 3) Exciting (+ve scale), Calm (-ve scale).

As each music track is associated with several mood tags which each mapped to 10 dimensional values, tags had to be combined. The most simple and obvious way would be to take the mean of all the mood values to generate a single 10-dimensional value for a track. However, it was felt that a music track can be represented by moods that differ significantly, so combining them into a single mood would be too crude. Therefore, a method (denoted PEA) to generate two mood values per track was devised. This method uses clustering of the 10-D scores where close scores are combined together. The means of the two most significant clusters are then calculated, resulting in two 10-D mood values for each track. A weight was assigned to each value according to the size of the cluster.

3.4 Track Distance Measures in Mood Space

For the purposes of searching a database of tracks with mood values assigned to them, a distance measurement is required to find which tracks most closely match each other. For the MDS-based models (Section 3.1), distances between tracks were obtained using either the Euclidean distance between tag summary vectors (methods MTF, CEN, TFW, ITF), or the Kullback-Leibler (KL) divergence between the Gaussian representations of the tags (method MVA). As the model described in Section 3.3 allocates two 10-D mood values per track (method PEA), a weighted Euclidean measure was used which exploited the weighting values associated with each of the two 10-D mood values. This is shown in Equation (5) where $m_s(i,k)$ is the mood of the seed track (where i is the value index, and k is the dimension index), $m_t(j,k)$ is the mood expressed by the test track (where j is the value index), $w_s(i)$ is the seed track weighting, and $w_t(j)$ is the test track weighting.

$$d = \sqrt{\sum_{i=0}^{1}\sum_{j=0}^{1}\sum_{k=0}^{9}((w_s(i)\cdot m_s(i,k)) - (w_t(j)\cdot m_t(j,k)))^2} \quad (5)$$

4. LISTENING EXPERIMENT

4.1 Corpus and Recommenders Tested

5,000 production music tracks were picked up randomly from the ILM dataset according to two constraints: (i) the durations of the tracks had to be at least 60 s (in order to discard short versions of the tracks), (ii) instrumental stems, i.e. individual tracks from multitrack recordings, were discarded. Six main genres were represented (jazz, dance, rock, electronic, folk and orchestral). 18 different recommenders were tested based on the two different mood models MDS and CLUST (Sections 3.1 and 3.3, respectively), the ACT transformation of the MDS model, MDS-ACT (Section 3.1.3), and a timbre-based model based on 20 MFFCs. Model MDS was tested with three different dimensions (3, 5 and 11) and the five different tag summarisation methods defined in Section 3.2 (MTF, CEN, TFW, ITF, MVA). Models CLUST, MFCC and MDS-ACT were tested with just one configuration each. As the ACT maintains the relative distances between mood terms in the MDS space, it doesn't affect track distances using the MTF, CEN, TFW, and ITF methods. The MDS-ACT model was hence only tested with the MVA method.

4.2 Procedure

To determine which mood model configuration used in a recommender gave the best matches according to human perception, a listening experiment was conducted. If a mood model is of any use it should ensure that a recommender generates tracks that closely match a seed track according to mood.

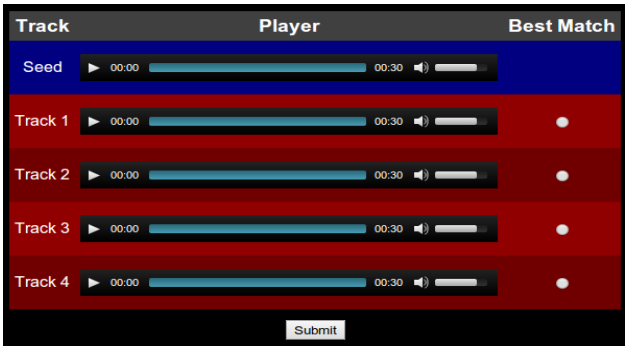

Figure 2. Survey interface.

The requirements of the test were that it should be simple to perform, not require specialist software or equipment, be accessible to enough people and not take too much time. To achieve these aims, a simple web-page was developed which made use of the audio tools in HTML5 (see Figure 2), so anyone could access the survey on an internet connection. The web-page presented the listener with a randomly selected seed track (from the 5,000 track dataset), plus four more tracks for assessment. In order to avoid potential causes of similarity between tracks due to the fact that they belong to the same album, recommended tracks were restricted to belong to a different album than that of the seed track. Of the four assessed tracks, one was from the recommender and the other three were randomly selected. The choice of recommender was also random. Participants were required to listen to each track at least once and select which one of the four tracks they felt most closely matched the seed track. They could repeat the process as many times as they wished, though they were encouraged to do at least 10 minutes' worth of testing.

4.3 Statistical Analysis

The system counted how many times the participants selected the recommenders' tracks. Therefore if a recommender generated recommendations no better than chance, the baseline score would be 25%. To determine which model gave the best performance, the confidence intervals for the scores had to demonstrate whether recommenders' scores were significantly different from each other. The score $s(m)$ for a particular recommender m is shown in the Equation 6, where $r(m,n)$ is 1 for a correct selection of the recommender m, 0 for a false selection, and n is the trial index. The number of times the recommender has been tested is $N(m)$.

$$s(m) = \frac{1}{N(m)} \sum_{n=1}^{N(m)} r(m,n) \qquad (6)$$

To calculate the confidence intervals, we decided to move away from the traditional parametric method which assumes a Gaussian distribution, and use a non-parametric bootstrapping method [4].

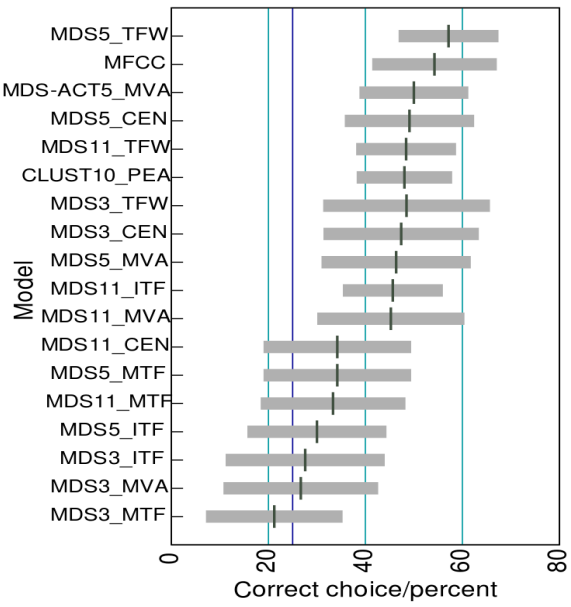

Figure 3. % of correct choice across recommenders. The recommenders are labelled using abbreviations in the format: 'ModelDimensions_TagSummarisationMethod'.

5. RESULTS AND DISCUSSION

The simple web-based survey design meant that the survey was easily accessible and simple to use, which allowed between 40-60 participants (estimated from email request list and activity) to perform the survey. There were 971 marks given in total in the survey, which was enough to determine which recommenders we should avoid using, but not enough to identify a clear leader. The scores for each model are shown in Figure 3, and are arranged in order with the best-scoring model at the top. The horizontal bars represent the confidence interval, with the fine vertical line representing the mean score. There is a line at 25% which corresponds to the random baseline score, so any scores that overlap that line can be considered to be no better than chance (given the sample size). To compare the performance of the different dimensions and of the MDS model, the scores from the versions of the model with the same dimensions were combined. To compare the methods, the scores from the same method were combined. Figure 4 shows how 3, 5 and 11 dimensions compare, and how the five tag summarisation methods perform. While the number of results per recommender ranged from 29 to 91 scores, the confidence intervals still remained quite large for making comparisons. While it would have been preferable to sample more participants to reduce the confidence intervals, the length of the experience had to be considered. While the confidence intervals do not necessarily allow a single recommender to be selected above all the others, they do still give us enough scope to eliminate the worst recommenders. The results show that the seven worst recommenders' confidence intervals overlap the 25% baseline score, although with some of those the overlap is small and a larger sample size could show they

are better than chance. For assessing the types of model, number of dimensions, method and distance measures, there was little evidence to conclude that any of those factors individually had a strong influence. However, the scores indicate the recommenders (for model MDS) with 5 dimensions could be stronger than 3 and 11 dimensions (though confidence intervals overlap). The MTF (tag with maximum term frequency) method yielded the smallest scores, which were significantly smaller than those of the strongest method, TFW (term-frequency weighted centroid).

The only recommender which does not use metadata relies on the timbre-based similarity measurement (MFCC). This recommender performed well, which can probably be explained by the fact that within the 5,000 tracks in the survey, there were still enough tracks that sounded very similar to each other (despite a certain amount of manual filtering of the full track list), and which would therefore indicate a similar mood. In practice, with a recommender searching over a million tracks, one that just returns very similar-sounding tracks may not fit the requirement of a mood-based recommender offering a more diverse selection.

Model CLUST performed as well as the best combinations of models MDS and MFCC. However, it was only tested with a single configuration, so it was not known whether the 10-dimensional two-maxima method was the best for it. It was also based on only one person's opinions of mood words, so could have benefited from a more extensive multi-person survey to refine the mood values.

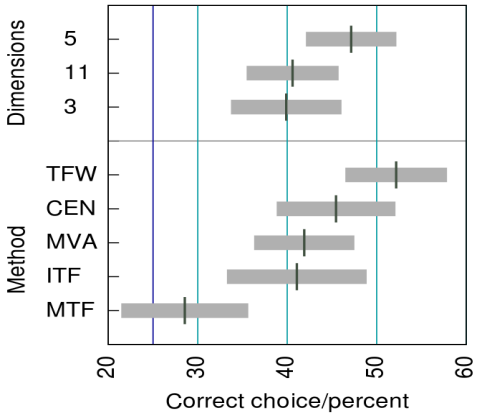

Figure 4. Dimension and method scores of model MDS.

6. SUMMARY AND CONCLUSIONS

The aim of the work was to design a mood model suitable for a recommender that could work over a very large database of music. It was felt that existing mood models with two [12] or three dimensions [11] were both too generalised for music analysis and lacking in dimensions to discriminate over large quantities of music. The survey conducted for this study assessed how various recommenders corresponded with human perception of music track matching based on perceived moods. The survey demonstrated that higher dimension models are effective in a recommender, and given the correct choice of summarisation methods and distance measures, can be tuned for better results. The results of the test showed that a five-dimensional model produced the best scores, although it was not statistically the clear leader. This work used production music associated with manually generated mood tags, and thus provided a source of information on emotional responses to music without analysis of the audio signal itself. Based on the results of the perceptual evaluation presented in this paper, a mood model (5-D MDS with term-frequency weighted centroid) was selected to develop an audio-content-based music emotion recognition (MER) system. We will follow up this study by testing to what extent this system can be used to assign mood tags in a robust manner to commercial music tracks which do not possess mood metadata.

7. ACKNOWLEDGEMENTS

This work was partly funded by the TSB project "Making Musical Mood Metadata" (TS/J002283/1). The authors would like to acknowledge the BBC Audio Research Partnership from which this collaboration was started.

8. REFERENCES

[1] M. Barthet, G. Fazekas, and M. Sandler. Multidisciplinary perspectives on music emotion recognition: Recommendations for content- and context-based models. In *Proc. of CMMR*, pages 492–507, 2012.

[2] M.M. Bradley and P.J. Lang. Affective norms for English words (ANEW): Instruction manual and affective ratings. Technical report c-2, University of Florida, Gainesville, FL, 2010.

[3] W. R. Dillon and M. Goldstein. *Multivariate Analysis*. Wiley series in probability and mathematical statistics. John Wiley & Sons, New York, 1984.

[4] T. C. Hesterberg, D. S. Moore, S. Monaghan, A. Clipson, and R. Epstein. *Bootstrap methods and permutation tests*. 2005.

[5] X. Hu. Music and mood: where theory and reality meet. In *Proc. of iConference*, 2010.

[6] P. Juslin and D. Västfäll. Emotional responses to music: The need to consider underlying mechanisms. *Behavioral and brain sciences*, 31:559–621, 2008.

[7] J.B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29:1–27, 1964.

[8] J. A. Lee and J. S. Downie. Survey of music information needs, uses, and seeking behaviors: preliminary findings. In *Proc. of ISMIR*, 2004.

[9] M. Levy. *Retrieval and Annotation of Music Using Latent Semantic Models*. PhD thesis, Queen Mary University of London, 2012.

[10] M. Levy and M. Sandler. A semantic space for music derived from social tags. In *Proc. ISMIR*, pages 411–416, 2007.

[11] C. E. Osgood, G. J. Suci, and P. H. Tannenbaum. *The Measurement of Meaning*. University of Illinois Press, 1957.

[12] J.A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161–1178, 1980.

[13] P. Saari, M. Barthet, G. Fazekas, T. Eerola, and M. Sandler. Semantic models of musical mood: comparison between crowd-sourced and curated editorial tags. In *Proc. of IEEE ICME (Affective Analysis in Multimedia workshop)*, San Jose, CA, 2013.

[14] P. Saari and T. Eerola. Semantic computing of moods based on tags in social media of music. *IEEE Transactions on Knowledge and Data Engineering*, manuscript submitted for publication available at http://arxiv.org/, 2013.