BBC

Research & Development

White Paper

WHP 228

May 2012

Musical Moods:
A Mass Participation Experiment for the Affective Classification of Music

Sam Davies (BBC)
Penelope Allen (BBC)
Mark Mann (BBC)
Trevor Cox (University of Salford)

BRITISH BROADCASTING CORPORATION

White Paper WHP 228

Musical Moods:
A Mass Participation Experiment
for the Affective Classification of Music

Sam Davies, Penelope Allen, Mark Mann BBC R&D & Trevor Cox University of Salford

Abstract

In this paper we present our mass participation experiment, Musical Moods. This experiment placed 144 theme tunes online, taken from TV and radio programmes from the last 60 years of the British Broadcasting Corporations (BBC) output. Members of the public were then invited to audition then rate these according to a set of semantic differentials based on the affective categories of evaluation, potency and activity. Participants were also asked to rate their familiarity with the theme tune and how much they liked the theme tune. A final question asked participants to identify the genre of the TV programme with which they associated the tune. The purpose of this is to aid in the affective classification of large-scale TV archives, such as those possessed by the BBC. We find correlations between evaluation and potency, potency and activity but none between activity and evaluation also no clear correlation between affect and genre. This paper presents our key findings from an analysis of the results along with our plans for further analysis. The initial results from this experiment are based on an analyses of over 51,000 answers from over 13,000 participants.

This document was originally presented at The 12th International Society for Music Information Retrieval Conference, Miami, Florida (USA) October 24–28, 2011

Additional key words: Psychology, Multimedia Classification, Music Information Retrieval

White Papers are distributed freely on request.

Authorisation of the Chief Scientist or General Manager is required for publication.

White Paper WHP 228

Musical Moods:
A Mass Participation Experiment
for the Affective Classification of Music

Sam Davies, Penelope Allen, Mark Mann BBC R&D & Trevor Cox University of Salford

1 INTRODUCTION

Music is an inherent part of nearly all broadcast programmes (TV and radio) and is often used to heighten the affective content of a scene or programme. Programme making teams have access to vast 'production music' libraries, which provide detail of not only the music tracks title and composer, but also keywords about the music which describes it's mood. This production music is used as backing music within a programme; providing an accompaniment to a scene. The British Broadcasting Corporation (BBC) provide internal access to a service called 'Desktop Jukebox' a production music library which contains over 38,000 production tracks along with a range of affective descriptors such as 'confident', 'bright' or 'sensitive'. This is an invaluable tool for helping programme producers to choose the right music as an accompaniment to a particular scene. Yet music is not just used as a background in productions. Most programmes also contain a theme tune – a piece of music designed to be recognizable and identifiable to introduce the programme. Generally especially commissioned, these pieces of music convey some idea of the affective content of the upcoming programme – a precursor to set the tone. For example, in preparation for the coverage of the 2010 UK General Election coverage, the BBC briefed the composer Blair-Oliphant that he should compose music that was "serious, important and classy" to reflect the fact that "this is likely to be a fairly historic election"[1].

In this paper we present our mass participation experiment Musical Moods that explored the link between theme tune and affect. Members of the public were asked to listen to theme tunes spanning 60 years of the BBCs output from 1950 and across 10 different genres, and rate each one on an affective differential scale. In the experiment, 144 theme tunes from 135 different programmes were made available as some long running programmes had multiple theme tunes. The breakdown of programmes, theme tunes and responses by genre is shown in table 1.

Each participant listened to a maximum of five theme tunes, chosen at random per experiment. After one month of the experiment being live, 13,183 participants had auditioned and ranked 51, 374 themes.

The aim of this experiment is to help develop automatic systems to classify the BBC archive using affective metadata. Currently all BBC programmes that are likely to be reused (either through re-broadcast or as clips) have manually created metadata, contents of which range from brief synopses to detailed shot listings. This results in London Classification (LonClass) database entry. LonClass, a Universal Decimal Classification extension developed by the BBC, is designed specifically to give factual information about a programme such as genre, shot type or recording location. Some programmes also have more in-depth analyses consisting of a full transcription and shot listing. However, this is a time and resource expensive process; a detailed analysis of a 30 minute programme can take a professional archivist 8 to 9 hours.

The purpose of this manually generated metadata is to allow for professional reuse. Frame accurate metadata is designed to allow users such as producers, and researchers to find stock shots such as landscapes or people, key interviews or other clips. However as the BBC open up their archives, this level of detail or type of metadata may not be best suited for non-professional users: viewers.

Genre	Number of Programmes (percentage of total)	Number of theme tunes (percentage of total)	Number of results (percentage of total)
Children's	16 (11.8%)	19 (13.2%)	6836 (13.3 %)
Comedy	33 (24.4%)	33 (23.0%)	11786 (22.9%)
Drama	38 (28.1 %)	40 (27.8%)	14204 (27.6%)
Entertainment	21 (15.6%)	22 (12.3%)	7867 (15.3%)
Factual	7 (5.2%)	8 (5.6%)	2769 (5.4%)
Lifestyle	8 (5.9%)	8 (5.6%)	2882 (5.6%)
News	7 (5.2%)	8 (5.6%)	2912 (5.7%)
Soaps	3 (2.2%)	3 (2.1%)	1047 (2.0%)
Sports	2 (1.5%)	3(2.1%)	1071 (2.1%)

Table 1. Breakdown of theme tunes, genres and results received.

BBC Information and Archives (BBC I&A), the section of the BBC that archive programmes and create associated metadata, periodically release digitised collections of programmes online to the UK viewing public [2]. These collections are grouped by theme and have semantic metadata; programme title, original transmission date, contributors and a brief synopsis; metadata similar to that in LonClass. This allows a user to accurately find what factual information is contained within a programme. This method of indexing may not be suitable when a viewer is looking for a programme for entertainment, not information. Thus, some form of semantic or affective metadata is required.

BBC Research and Development (R&D) are currently investigating automatic classification techniques [3]. These aim to create semantic and affective metadata from an archived programme through an analysis of the available audio and video or a programme. Current analysis techniques focus on non-music audio – speech and sound effects. The purpose of collecting metadata about theme tunes is to extend this to begin investigating how well music can aid automatic classification. Theme tunes are used to identify a programme; making it recognizable to the audience and setting up affective expectations. By affectively analysing the theme tune, we hope to be able to aid in an affective analysis of an entire programme.

In this paper we present our work as follows. We present an overview of the experiment and our methodology is given in section two and an initial analysis of our results in section three. We discuss these results in section four and conclude and present our plans for the future in section five.

2 METHODOLOGY

Musical Moods was an online experiment, accessible from the URL www.musicalmoods.org.uk. The experiment was launched as part of the British Science Association's National Science and Engineering Week, a yearly event in the UK with the aim of promoting participation in and the understanding of science in the UK. The experiment was also featured on the BBC television show, Bang Goes The Theory, a weekly science show with an average audience of around 2.5 million.

One of the key factors in this experiment was ease of use for participants. As such, each participant on hearing a clip of a theme tune was asked to rate a theme tune on one of the possible six semantic differentials. The clip was 15-20 seconds long, with the participants able to re-audition the theme tune if required. These clips were edited to ensure they contained the main musical themes of the theme tune, and did not contain any lyrics which alluded to the TV shows content. Participants were played five randomly chosen theme tunes in each experiment. Participants could take part in the experiment as many times as possible and no record was made of how many times a participant took part. Upon hearing a clip, the participants were asked to rate each semantic differential on a discrete scale of one to five, with each scale extreme labelled with an opposing pair of adjectives.

a. Semantic Differentials

One of the key issues found with previous music and affect research is the lack of a standard definition for semantic scales. Hevner was one of the first to attempt to define a taxonomy for music and affect [4]. This creates eight groups of adjectives arranged diametrically on a circle. Another approach is to use a valence/arousal space similar to that defined in [5]. The semantic differentials in this experiment were based upon Osgood's dimensional space; a three dimensional space incorporating Evaluation, Potency and Activity (EPA) [6]. Each of these allow mappings between different adjectives that have similar affective meaning allowing for affective unification of them. Evaluation relates to positive or negative feelings like happy or sad and accounts for around 50% of affective meaning. Potency relates to size or power, such as heavy or light and activity relates to amount of action. Potency and activity count for approximately the other 50% of semantic meaning, although there are 4 more minor categories. The affective adjectives used in this experiment were taken from mapping adjectives in [4] to the affective space in [6]. The semantic differentials were happy/sad, playful/serious for evaluation, masculine/feminine and heavy/light for potency and dramatic/relaxing, exciting/calm for activity. A score of one to five was used for these, with five relating to a maximum value of happy, playful, masculine, heavy, dramatic and exciting and one relating to a maximum score for the opposing adjectives although this numbering was not shown to the user. The adjective scales used for each EPA differential were found to correlate to each other, as is shown in table1. Here, a Pearson Correlation matrix shows the correlations between the different semantic differentials. Ideally there would be high correlation between semantic differentials in the same affective vector space (i.e. dramatic correlating highly with exciting. Whilst this is true for evaluation and activity, there is a weaker correlation for potency (heavy and masculine), with masculine correlating more with dramatic then heavy.

	Dramatic	Happy	Heavy	Masculine	Playful	Exciting
Dramatic	1	-0.061	0.697	0.673	-0.350	0.812
Happy	-0.061	1	-0.669	-0.060	0.902	0.404
Heavy	0.697	-0.669	1	0.620	-0.836	0.297
Masculine	0.674	-0.060	0.610	1	-0.261	0.574
Playful	-0.346	0.902	-0.836	-0.261	1	0.127
Exciting	0.812	0.404	0.297	0.574	0.127	1

Table 2. Pearson Correlation for Affective Results

b. Genre Identification

Participants were then asked to identify with which genre they associated the music clip. The options for genres were taken from an amended list to that on the BBC iPlayer service, an online programme catch-up facility available in the UK [7]. These were Children's, Comedy, Drama, Entertainment, Factual, Lifestyle, Soaps, News and Sport. The purpose of this was to look at the link between theme tune and genre – looking to identify if genres were readily identifiable from theme tunes and also if there was any link between the affective score and perceived genre.

c. Liked or Familiar

Participants were also asked either if they liked a theme tune, on a scale of 'Yes', 'No' and 'No Opinion' or how familiar they were with a theme tune, on a scale of 'Very', 'Not Very' and 'Sort of'. A very familiar or liked tune scored one and an unfamiliar or not very liked tune scoring three. The purpose of asking these was to ascertain if there was any link between familiarity and liking a theme tune, with the affective score given. This was not done to rank theme tunes by popularity or find out which were the best known theme tunes.

3 RESULTS

After over 51,000 results had been gathered a preliminary analysis was performed. Results were calculated by collating all the scores for each of the 144 programmes. The average and standard deviation for each answer was then calculated. To calculate the EPA values, each semantic differential for each of evaluation, potency and activity was combined, with equal weighting. The average and standard deviation of each of these was then calculated.

a. Participation

Of the 13,183 participants who took part in the experiment, 54% of participants identified themselves as female and 46% male. Age band results are shown in table 3.

Age Band	Percentage
< 16	44%
16 – 24	18%
25-39	18%
40-54	14%
55-69	5%
> 70	2%

Table 3. Age band breakdown of participants

b. Affective Scores

These results looked at how different theme tunes were classified according to the semantic differentials of evaluation, potency and activity. These are shown in figures 2, 3 and 4 respectively.

Figure 2 shows the average evaluation score against the average potency. Here, it can be quite clearly seen that across all theme tune genres there is a slight negative correlation of -0.2 between the potency of a theme tune and the evaluation (shown as a line), meaning tunes classified as happy are also broadly classified as light. Also, whilst genres do spread over the range of the scale, there is a tendency for theme tunes associated with children's ('+') and comedy ('o') programmes to rate higher on evaluation and for those associated with dramatic ('*') programmes to rate higher on the potency scale.

In Figure 3, the average potency against the average activity is shown for each programme. A positive correlation of 0.6 is shown between activity and potency meaning theme tunes perceived as heavier are also perceived as more exciting. Also, whilst clustering is less visible than in figure 2, there is some, with theme tunes associated with drama ('*') tending to rate higher on potency and activity (i.e. more dramatic and heavy) with theme tunes accompanying children's and comedy programmes ('+' and 'o' respectively) tending to be less so (i.e. slightly calmer and lighter), though children's programming does tend towards being more exciting.

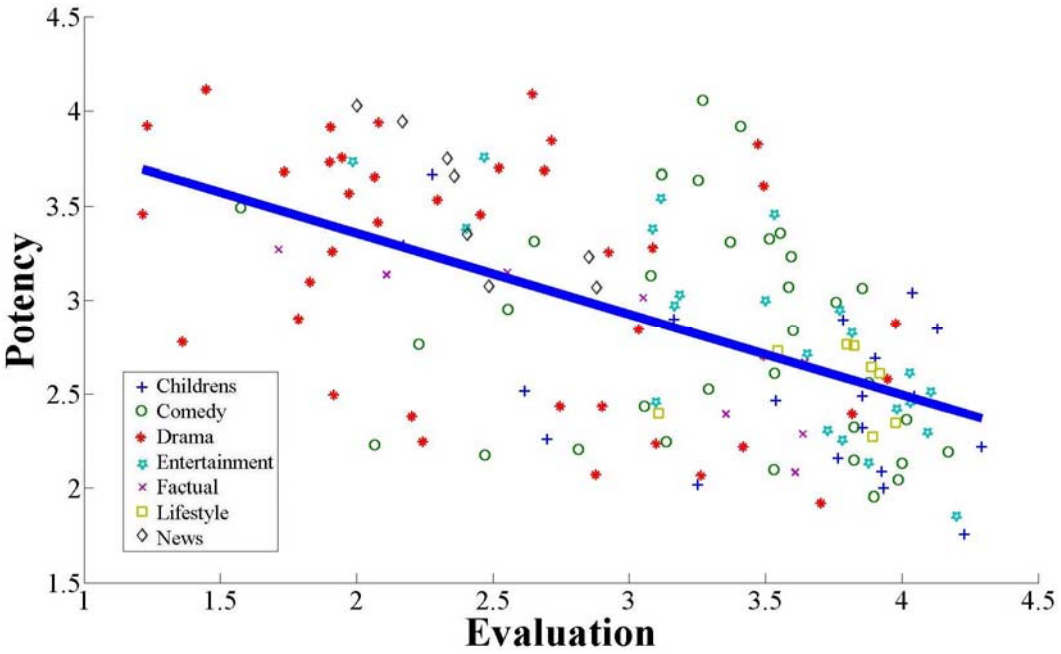

Figure 2. Potency and evaluation average classifications

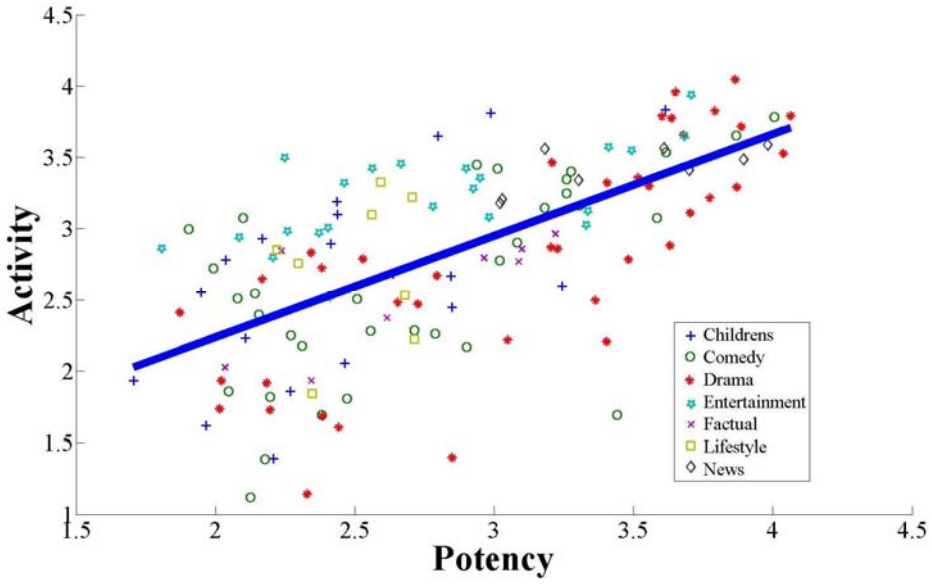

Figure 3. Activity and potency average classifications

Figure 4 shows no real correlation overall, but does show some clearer grouping. One of the most interesting groupings shown is that for theme tunes accompanying news and current affairs programmes (circled on figure 4). These can be seen to cluster around the centre scale for evaluation, but with a marked increase in activity meaning that they are classified as being neither happy nor sad but more dramatic.

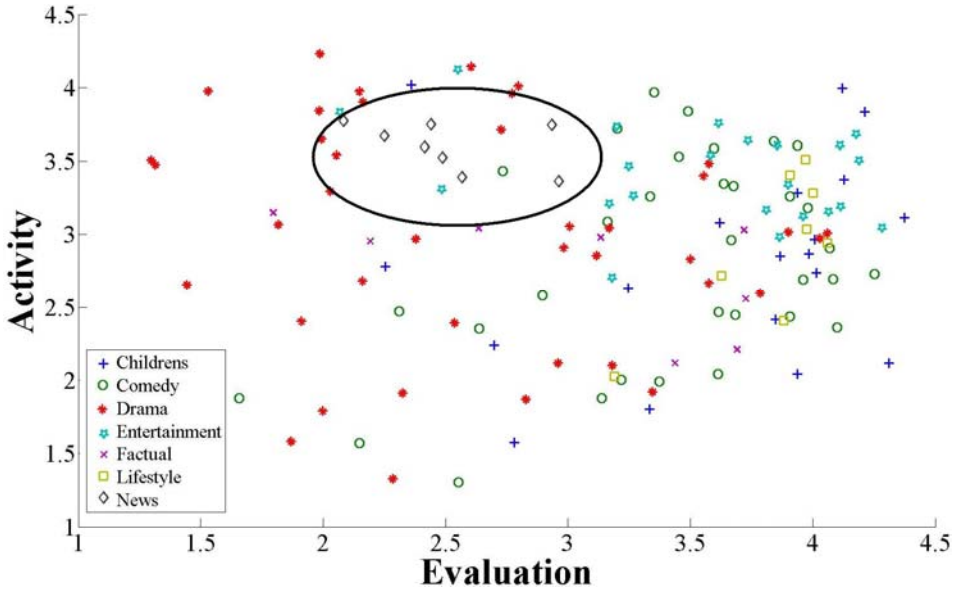

Figure 4. Activity and evaluation average classifications

c. Standard Deviation

In all instances, it was found that the standard deviation was significantly lower at the extremes of each semantic scale. This is shown in figure 5.

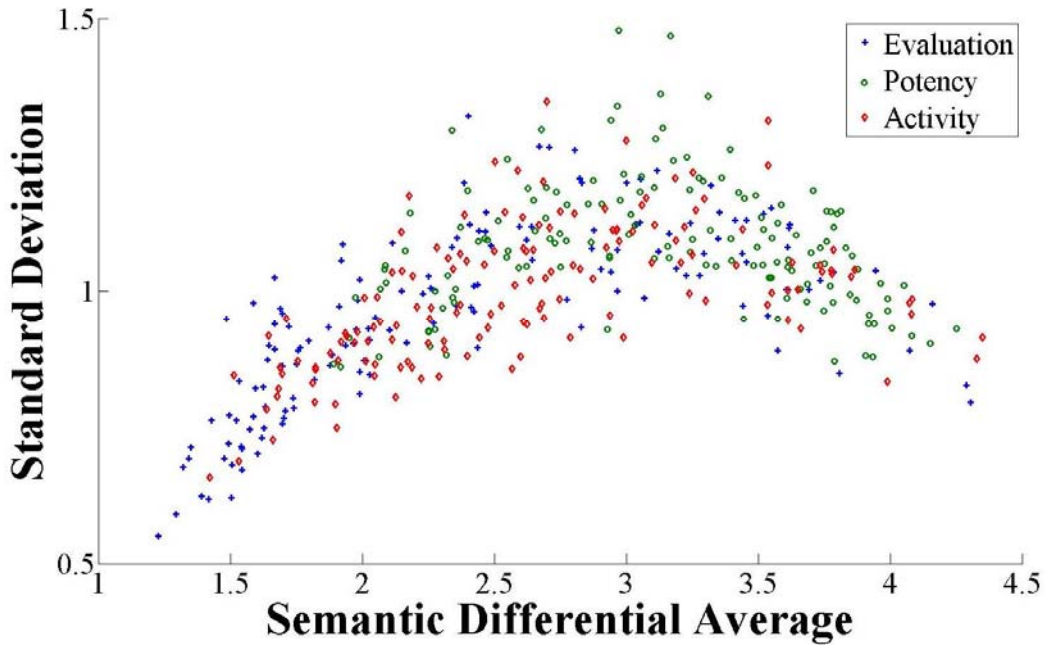

Figure 5. Standard deviation for semantic differentials

d. Genre, liking and familiarity

These looked at the results from the familiarity, how much participants liked a theme tune and genre identification questions.

Unsurprisingly, when participants were more familiar with the theme tune, they were generally able to identify the accompanying programmes genre. The results for correctly identified programmes are shown in figure 6.

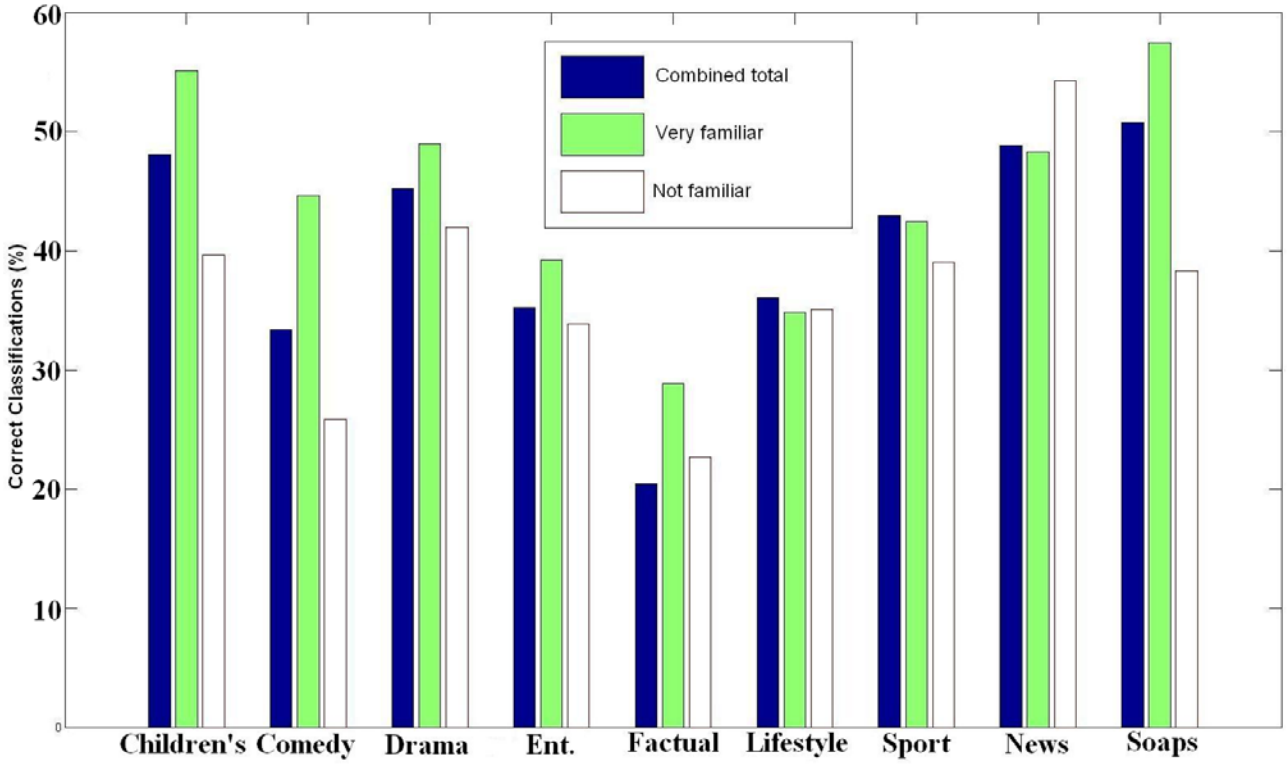

Figure 6. Genre classifications.

An interesting result from this genre classification is that participants did not seem to be able to correctly identify which genre a theme tune accompanied with only soaps being correctly identified more than 50% of the time. Participants also had most trouble identifying lifestyle programmes where only 35% of participants who stated they knew the theme tune being correct and 22% of those who stated they knew the theme tune incorrectly choosing entertainment. A further error was in factual genre identification where 23% of participants who did not know the theme tune identified the programme as factual and 24% as drama.

As can be seen from figure 7, theme tunes which were more familiar to people were also rated as being more liked, with one key exception, the theme tune to the programme the *Weakest Link*, where the theme tune is less familiar to participants, but liked. Familiarity and a value for liking was found to have a correlation coefficient of 0.69. However, familiarity was not seen to have a marked difference on the affective scores, with only a 7% difference noted between the scores of familiar and unfamiliar theme tunes.

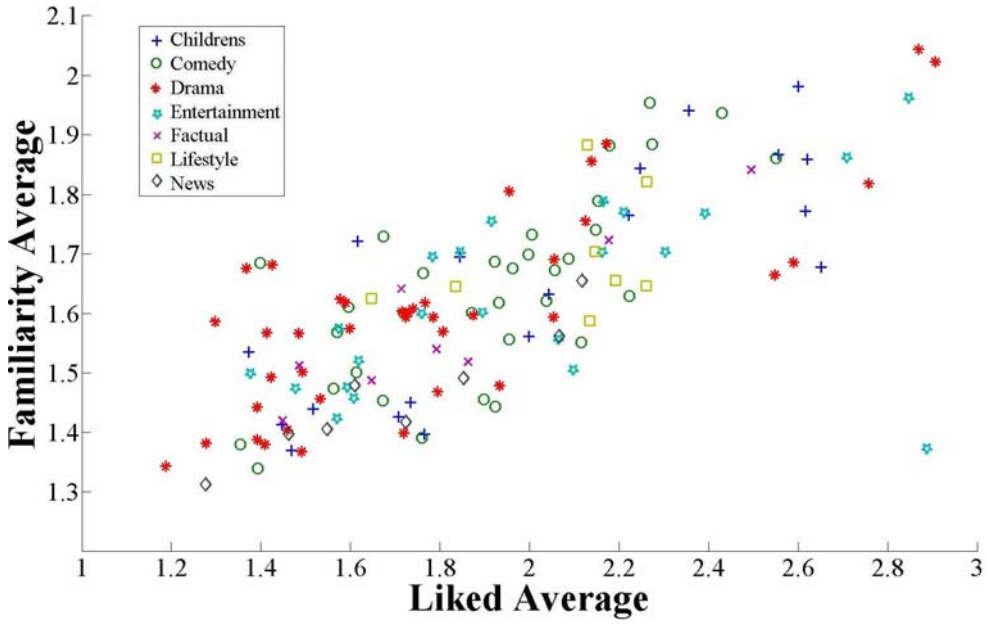

Figure 7.Familiarity against Likedness for Programme averages

All results and music will be available for download from www.musicalmoods.org .

4 DISCUSSION

From the results analysed so far some clear trends are visible. As figure 2 shows, there is a negative correlation between potency and evaluation. This suggests that tunes classified as happy or playful were also classified as light or feminine. When the genres of the programmes associated with the theme tunes is taken into account, this shows that these are generally genres which one would imagine as being happy or light – mainly children's, comedy and lifestyle programmes. This genre generally includes day-time TV programmes about home improvements or gardening and so again this is expected. However, no large scale affective analysis of the programmes has so far been conducted and so no strong link can as yet be drawn. At the other end of the scale, where theme tunes are classified as being heavier and sadder, mainly dramas are found. This would fit in with our understanding of drama programmes themselves – that they generally feature 'heavier' and less happy storylines. A brief analysis of the programmes by the authors whose theme tunes were found to have the lowest evaluation and highest potency scores found the overall affect of the programmes (the detective shows *Silent Witness* and *Ashes to Ashes*) found they also fitted into the same affective space. Conversely the two programmes which were found to have the highest evaluation scores and the lowest potency scores were found to be

both children's TV programmes; *Blue Peter*, a children's magazine programme and the *Teletubbies*, a show aimed at pre-school children. From the authors analysis of these programmes, it can be seen that programmes such as these which have theme tunes which score at extremes of the affective scales themselves would score highly at these extremes too.

In figure 3 a clear correlation between activity and potency is visible, meaning that theme tunes classified as being more dramatic and exciting (activity) are also classified as being heavier and more masculine (potency). However in this case much less grouping is observed. Whereas in figure 2 it was possible to see groupings in the genres children's and comedy and then a further group for dramatic programmes, here both children's, comedy and drama are spread along the scales. Theme tunes to the genres comedy and drama have classifications at both extremes of the scale. This is unsurprising when considering the affective nature of programmes. Genre's such as comedy and drama rely far more on affect for their programme substance than fact based programmes such as factual or news which rely more on their informative content. From this, it follows that the theme tunes to children's, comedy and drama programmes should have more affective spread. The correlation between activity and potency follows the findings in [6], which found that there was a high degree interchangeability between them.

Looking at the results shown in figure 4 there is no correlation shown between activity (dramatic or calm) and evaluation (playful or serious) shown in TV theme tunes. Therefore no link was found between theme tunes being happy or playful, and how dramatic or exciting they were found to be. However it is possible to see groupings of the genres of the programmes associated with the theme tunes more clearly. For example, children's, comedy and lifestyle and entertainment all tend to group towards the higher end of the evaluation scale. Conversely, theme tunes associated with drama tend towards the lower end of the evaluation scale, indicating a more sad or serious classification. One interesting cluster is that of theme tunes associated with programmes of the genre news, circled on figure 4. These cluster around the centre of the evaluation axis, suggesting that in terms of happy or sad they are neutral. However, these have the largest standard deviation (with an average of 1.2) so it could be that with classifying these types of theme tunes participants had the most trouble.

What is most interesting in looking at the genres associated with the theme tunes is that whilst some clear grouping occurs, individual genres spread out over semantic differential scales. From this, it is possible to conclude that using these affective values are not an accurate method for classifying genre i.e. children's programmes do not all classify as happy. This is further backed up by looking at the results shown in figure 6. Here, it can be seen that, in general, participants were not that accurate in associating genres and theme tunes. Even when participants stated they were very familiar with the theme tune, the highest percentage of correct genre identification was only 57% (for soaps). The lowest score was for factual at 29% and the average correct identification rate was only 44%. When the results for those who were not familiar with the theme tune are taken into account, correct identification falls even further with the most correct classifications being for the News genre, with 54% correct classifications and the lowest for Factual with 23%. The average was 37%. This again suggests that the genre of a programme and its associated theme tunes affective value do not show a strong link. This would have important implications in the design of any classification and recommendation system that looks to use genre and affect as a basis.

One of the problems with this could be the choice of genres that were made available. These were based on those offered by the BBC iPlayer service. These are very broad categories, with some ambiguity as to which genres some programmes belong too (for example between programmes in the entertainment and lifestyle genre) and with many programmes placed in multiple categories (for example children's news programmes). Further research is required to identify what genre classification system best maps to affect.

In looking at figure 6, it is clear that there is a positive correlation between the average value for liking a tune and familiarity with it. This suggests that the more familiar theme tunes were also more liked by participants. Whilst this in itself does not give any insight into how either liking a theme tune or being familiar with one has an effect on how participants affectively perceive a theme tune, it does suggest that when evaluating retrieval systems for programme archives, this

correlation should be noted. One interesting outlier in this set is the theme tune to the entertainment programme *The Weakest Link* – a quiz show where the familiarity is slightly below average but the tune has a high value for being liked. Further analysis is required to determine possible causes for this.

5 CONCLUSION AND FURTHER WORK

In this paper we have presented our experiment Musical Moods and our first analysis of the results. We have found that whilst some the genres of the TV programmes with which theme tunes are associated show some grouping on the affective scales, there is no real link between a programmes genre and the perceived affective value of its theme tune. We have also found that theme tune alone is not an accurate indication of programme genre. A further finding is there is strong correlation between familiarity and liking a theme tune.

Further work is also proposed to musically analyse the theme tunes and correlate these with the experiment results. It is planned to perform a full musical analysis on each track used – looking at features including key, harmonic progression, instrumentation and orchestration. These would then be analysed against the affective scores for the theme tunes. This would then be used as a ground truth dataset, looking to use automated musical analysis and machine learning techniques to identify the affective content of other programme theme tunes.

Another area of work is to affectively analyse the programme themselves, and look at any correlation between the affective classification of the video and audio content with the theme tune. This is planned through more large scale user evaluations.

We are also looking to get more participants to increase the validity of our results.

6 ACKNOWLEDGMENT

This project was launched as part of the British Science Association's National Science and Engineering Week and the BBC R&Ds Multimedia Classification Project.

7 REFERENCES

[1] K. Young, "TV theme tunes set tone for general election night," in *BBC News*, ed, 2010.

[2] BBC. (2010, 17th July). *BBC Archive Collections*. Available: **http://www.bbc.co.uk/archive/collections.shtml**

[3] -, "A Framework for Automatic Mood Classification of TV Programmes," in *5th International Conference on Semantic and Digital Media Technologies* Saarbrucken, Germany, 2010.

[4] K. Hevner, "Experimental Studies of the elements of expression in music," *American Journal of Psychology,* vol. 48, p. 246:268, 1936.

[5] J. Russell, "A Circumplex model of Affect," *Journal of Personality and Social Psychology,* vol. 39, pp. 1161-1178, 1980.

[6] C. E. Osgood, G. Suci, and P. Tannenbaum, *The measurement of meaning*. Urbana, USA: University of Illinois Press, 1957.

[7] BBC. (2010, 17th July). *BBC iPlayer*. Available: **http://www.bbc.co.uk/iplayer/**