AN EXPERIMENT IN THE AUTOMATIC CREATION OF MUSIC WHICH HAS SPECIFIC EMOTIONAL CONTENT

James Rutherford[1] *and Geraint A. Wiggins*[2]

[1] Department of Artificial Intelligence, University of Edinburgh; *jamesrutherford@btinternet.com*
[2]Department of Computing, City University, London; *geraint@city.ac.uk*

ABSTRACT

We present an experiment with HERMAN [9], a real-time music generator. HERMAN was designed to enrich the sensory environment of GhostWriter, an immersive educational tool. The aim of thus enhancing GhostWriter's virtual environment, was to offer a more convincing sense of presence, to motivate and stimulate users. HERMAN supplies continuous, *manipulable* music. It has an input parameter which, it is claimed, determines the *scariness* of the music generated. Here, we evaluate HERMAN's output; the results suggest a correlation between the input parameter and the scariness of the music as perceived by experimental subjects.

1 BACKGROUND

1.1 HERMAN and GhostWriter

GhostWriter [9] is an computer-based educational tool, intended to support primary schoolchildren writing ghost stories. It presents them with a three-dimensional virtual haunted house. Children assume different identities within the environment and communicate *via* a textual interface. A teacher directs the exploration of the environment by cueing particular characters to speak, or by assuming an identity in the story. Text dialogue is stored during the interaction and, later, the children use it as a basis for their own stories [9]. This approach system is believed to encourage both creative writing skills and co-operation in the group.

HERMAN is a real time computer-based music generator [3,14], developed to accompany GhostWriter. It has one input, the *scariness level*, varying between 0 and 99, which is adjusted via a graphical slider control. HERMAN's music is intended to complement the visually stimulating 3-d haunted house environment. The combination gives the children a richer virtual environment to build on. The teacher can manipulate the scariness level and this dynamic input can be used to provide direction in the story.

1.2 The Implementation of HERMAN

The HERMAN system is implemented in Java. The scariness parameter is passed through a mapping function, which converts its single value to a set of control parameters for the music generator. Using these control parameters to guide its rules, the generator creates intermediate music control data, which is rendered into MIDI format [10]. The detail of the system's operation is explained by Stapleford [14] and de Quincey [3].

To provide some degree of continuity in the music, HERMAN stores *phrases* of music. These phrases may be repeated, identically or transformed to fit the current progression in the music. Phrases incompatible with the current scariness level cannot be used and new phrases are generated for this purpose [14]. The ordering of the phrases is constrained to one of eight forms, though the system is able to deviate slightly from these guidelines.

Music is output on two MIDI channels, melody and accompaniment. A further planned channel for percussion was omitted due to time constraints on HERMANs development.

1.3 Defining Scariness

To extract a notion of what creates scariness, HERMAN's creators studied soundtracks from adults' and children's thrillers and horror films. Three people watched each film and independently made notes about "interesting structures" [3] evident in the music. A discussion about each film followed, to consolidate ideas. The films chosen were *Alien* [13], *Vertigo* [5], *North by Northwest* [6], *The Princess Bride* [8] and *Beetlejuice* [2]. The group concluded that these films covered a range of different scary styles, from horror and thriller through to comedy.

Stapleford and de Quincey isolated several scariness-building aspects of film music, including: a loud theme followed by silence (Alien); heartbeat-like rhythms, with increasing speed or a sudden halt (Alien); disjointed, 'bursty' rhythms (Alien, Beetlejuice); fast sequences of high-pitched strings (Alien, The Princess Bride); *etc.*

Bernard Herrmann provided music for some of Hitchcock's most suspenseful films, including *Vertigo*, *North by Northwest* and *Psycho*. When designing HERMAN, Stapleford and de Quincey studied Herrmann's work in more detail, *via* the writing of Brown [1].

To create scariness (which is a kind of dramatic tension), Herrmann's harmonic progressions break the rules that are supposed to govern Western tonal music [12], in particular, controlled ways. By not fulfilling a listener's expectations [7] or by playing more or less discordant notes simultaneously, the music is able to deviate from the expected norm to something more uncomfortable. Texts on musical theory [12] were used to provide guidelines for the basic structure upon which tension rules were overlaid [14].

2 AIMS

2.1 Verifying HERMAN

Stapleford [14] and de Quincey [3] performed a preliminary evaluation of a small number of human subjects' responses to HERMAN's output, but it was not the primary focus of their work. They acknowledge that their method was flawed, and so the results, though suggestive, are not conclusive. We aim to investigate whether HERMAN's output conveys a level of scariness dependent upon the scariness parameter, to the listener.

Our work aimed to measure human response to HERMANs output in comparison with human-composed pieces of music, to give a more reliable measure of its ability to convey affective content. Using human-composed music in the experiment means that pieces with strong affective content for Western listeners may be introduced, and stringent comparisons may be made with them.

The experiments presented HERMAN at 3 fixed levels of scariness, supposing that positive distinctions thus perceived will be heightened on hearing the system in its intended, dynamic usage.

We aim to evaluate whether the different affective outputs of HERMAN at different scariness settings lie in order in relation to each other. The outputs are also compared with other pieces of music.

2.2 Hypotheses

2.2.1 Not Scary/Scary

HERMANs purpose is to provide music at various levels of *scariness*. The comparative levels of scariness may be tested *via* the following hypotheses.

H1: HERMAN on medium scariness setting will be rated more scary than HERMAN on low tension setting, when compared on the *Not scary/Scary* continuum.

H2: HERMAN on high scariness setting will be rated more scary than HERMAN on medium tension setting, when compared on the *Not scary/Scary* continuum

H3: HERMAN on high scariness setting will be rated more scary than HERMAN on low tension setting, when compared on the *Not scary/Scary* continuum

2.2.2 Relaxed/Tense

HERMAN's scariness is derived mainly from harmonic tension, and also some rhythmic tension. The levels of tension provided by HERMAN can be compared *via* the following hypotheses.

H4: HERMAN on medium scariness setting will be rated more tense than HERMAN on low scariness setting, when compared on the *Relaxed/Tense* continuum.

H5: HERMAN on high scariness setting will be rated more tense than HERMAN on medium scariness setting, when compared on the *Relaxed/Tense* continuum.

H6: HERMAN on high scariness setting will be rated more tense than HERMAN on low scariness setting, when compared on the *Relaxed/Tense* continuum.

3 METHOD

3.1 Approach and Justification

Subjects were asked to rate their feelings about pieces of music in terms of bipolar scales. This approach was taken because both ends of a bipolar scale would be equally suggestive, so, if the suggestive methodology was driving the experiment, the result would be ratings exactly in the centre of the scale, interpreted as no more one of the terms than the other.

Verbal scales were used, with many distractors, allowing us to obscure which were most significant in the experimental design, so subjects could not tailor their answers according to their perception of "correctness".

3.2 Subjects

24 Subjects, 15 male and 9 female, took part in the main experiment, each in one of three different time slots. Average age of the group was 18.6 years (s.d.$= 1.28$) The age band is narrow, but, from the wide variety of musical skill, exposure and taste claimed by the subjects, we believe that there is enough diversity.

Results were collected individually and anonymously, to reduce social or peer pressure, and were normalised to reduce the difference between more and less expressive subjects.

3.3 Materials

Subjects were required to perform two distinct tasks in the experiment: a *Music Appreciation Task* and a *Number Frequency Task*.

3.3.1 Music Appreciation Task

Musical Stimuli Three clips of HERMAN generated music were provided by its creators, verified as a competent representation of its output. The clips were taken at three scariness settings, low (0), medium (50) and high (99). Each was 14 minutes long.

Four other pieces of music were selected, aiming to cover a broad spectrum of styles, and suggest a variety of affect to a UK-enculturated listener. Pieces with common associations were avoided, as their suggestions may come from outside the music alone. (*Night on a Bald Mountain* is used in Disney's *Fantasia* [4], but we believe that this film does not cause a lasting horror effect, and so does not cause significant associations.) The pieces chosen were as follows: *Night on a Bald Mountain* (Modest Mussorgsky, 1860); *Mars: The Bringer of War* (Gustav Holst, 1914); *Le Piccadilly* (Erik Satie, 1904); *Pavane* (Gabriel Fauré,1887). To minimise variation between HERMAN output and the human-composed pieces, other than the affective content which was being measured, all the music was rendered on a synthesiser.

60 seconds of each piece were presented: enough to convey the piece's intent, but not too much for subjects to cope with, and keeping the total duration sensible. The 60 seconds from the four human-composed pieces was chosen subjectively to represent the whole of the piece's mood and as consistent in terms of affect.

Since HERMAN's output at each level is stochastic and not structured to change mood, its minutes were randomly chosen from its output; a minute is enough to reduce the significance of minor stochastic anomalies. A different clip was used for each level of each experimental session, to minimise larger random anomalies.

Ratings Sheet Twenty bipolar scales were used, allowing detailed evaluation of the musical pieces [11]. The word pairs describing the scales were chosen to be as unambiguous as possible. A wide variety of scales was used, including visual sensations (*e.g., red/green*), physical descriptions (*e.g., full/empty*), performance descriptions (*e.g., dignified/ undignified*) and feelings (*e.g., angry/not angry*). Scales were chosen so that the subjects could describe the music with some depth in many different dimensions.

The scales were presented on vertical bars tall enough to encourage a wide range of ratings. There were two different *Music Appreciation* sheets, presenting the bars in two different orders, with poles in both orientations, so that order effects did not bias the results. A box was provided for subjects to record other adjectives suggested to them by the piece. Subjects were asked to indicate if a piece was familiar, and if so, whether they could name it.

3.3.2 Distractor Task: Number Frequency

A distractor task was used to reduce auditory memory transfer between the music conditions and to prevent subjects carrying over any feelings between pieces. The task needed to be emotionally neutral, musically irrelevant, cognitively engaging and long enough to provide a real cleansing effect. Subjects were asked to count the even occurrences in a sequence of 30 two-digit numbers.

3.3.3 Presentation of Materials

Audio Stimuli To reduce priming, the order of the seven clips was randomised between the three sessions, and the order of the HERMAN clips was changed between sessions (Table 1). Audio material was taped in sequence for uniform presentation.

	Session 1	Session 2	Session 3
Piece 1	Pavane	HERM M	Piccadilly
Piece 2	HERM L	Piccadilly	HERM H
Piece 3	Piccadilly	Night	Pavane
Piece 4	Mars	HERM L	HERM L
Piece 5	HERM H	Pavane	Night
Piece 6	Night	HERM H	Mars
Piece 7	HERM M	Mars	HERM M

Table 1: Sequences of stimuli reordered to minimise priming. Obvious abbreviations are used for the named music; HERM H/M/L denotes HERMAN at High/Medium/Low scariness.

Experimental Running Order Subjects were told that the experiment was anonymous, would take no more than 30 minutes, would be explained afterwards if requested, and that they might leave early if they wished. Questions about the subject's gender, age and musical experience were asked, and instructions on the experimental tasks were given.

Before each musical condition, a distractor task was presented. Then, subjects were played each excerpt once and then asked to draw a single horizontal bar at the appropriate place on each of the scales as the excerpt was played a second time.

For each session, the experimental duration was slightly under 30 minutes. Subjects sat face forward at benches and, from observation, were kept busy enough to prevent comparison of answers with neighbours.

3.4 Data Analysis

Scales were evaluated using a 7 point, evenly spaced overlay (-3 to +3). Scales with more than one mark, no marks or with adjustments to the labels were judged malformed and excluded; 19 of 3360 responses were so judged.

To ensure that responses of less expressive subjects were comparable with those of more expressive, responses were scaled so each contributed the same total of points. We suppose that each subject can recognise one type of feeling in music to the same degree that s/he can recognise another, so the ratings may be scaled up or down across the whole range for a particular subset.

4 RESULTS

A t-test analysis was performed between groups of adjusted ratings to establish significant differences between the feelings evoked by the music. Each t-test was ideally a comparison between two samples of 24 values, though in the case of malformed scores it compared fewer, the smallest sample being of 22. Significance was reported as ($p < 0.05$), ($p < 0.01$), ($p < 0.001$) or *ns* (for non-significant). Due to normalisation (§3.4), a sample of 24 values averages between +1.34 and -1.34. Our graphs show mean adjusted ratings, with 1 standard deviation above and below for each of the 7 stimuli for the 2 scales relevant to our hypotheses – Rutherford [11] reports the full results.

4.1 Scale 7: Not Scary/Scary

4.1.1 Related Hypotheses

Hypotheses H1, H2 and H3 (§2.2.1) define the expectation that HERMAN will be rated as scarier, the higher the input setting. The data from this continuum are shown in Figure 1.

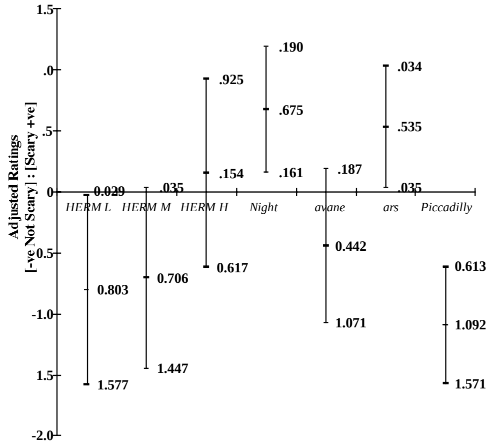

Figure 1: Results from the *Not scary/Scary* continuum. Abbreviations as before.

Hypothesis H1 was not supported by the t-test ($t[46] = 0.444$, *ns*). HERMAN's medium-scariness music was not significantly more scary than its low-scariness music. However, *HERMAN medium* was not rated less scary than *HERMAN low* either.

Hypothesis H2 was supported by the t-test ($t[46] = 3.939$, $p < 0.001$). HERMAN high was found to be significantly more scary than HERMAN medium, with a high degree of confidence.

Hypothesis H3 was supported by the t-test ($t[46] = 4.292$, $p < 0.001$), with a high degree of confidence. HERMAN was significantly more scary at the high setting than at the low setting.

4.1.2 Exploratory Analysis (2-tailed)

HERMAN high was significantly scarier than *Piccadilly* ($t[46] = 6.743$, $p < 0.001$) and *Pavane* ($t[46] = 2.935$, $p < 0.01$). It was significantly less scary than *Night on a Bald Mountain* ($t[46] = 2.754$, $p < 0.01$) and *Mars* ($t[46] = 2.023$, $p < 0.05$).

4.2 Scale 14: Relaxed/Tense

4.2.1 Related Hypotheses

Hypotheses H4, H5 and H6 (§2.2.2) define the expectation that HERMAN will be rated as more tense, the higher the input setting. The data from this continuum are presented in Figure 2.

Hypothesis H4 was not supported. *HERMAN medium* was not significantly different in perceived tension from *HERMAN low* ($t[46] = 0.812$, *ns*).

Hypothesis H5 was supported. *HERMAN high* was rated as more tense than*HERMAN medium* ($t[46] = 2.875$, $p < 0.01$).

Hypothesis H6 was supported. *HERMAN high* was rated by the listeners as producing more tense music than *HERMAN low* ($t[46] = 2.309$, $p < 0.05$).

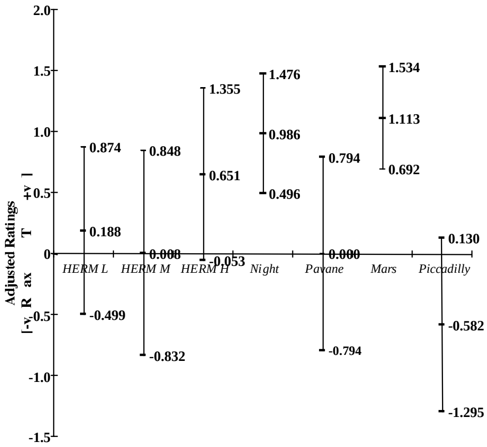

Figure 2: Results from the *Relaxed/Tense* continuum. Abbreviations as before.

4.2.2 Exploratory Analysis (2-tailed)

HERMAN high was significantly more tense than both *Piccadilly* ($t[46] = 6.033$, $p < 0.001$) and *Pavane* ($t[46] = 3.005$, $p < 0.01$). It was significantly less tense than *Mars* ($t[46] = 2.760$, $p < 0.01$), but not significantly different from *Night on a Bald Mountain* ($t[46] = 1.911$, ns).

4.3 Correlation between Scales 7 and 14

When examined across all the data (7 pieces of music for 24 subjects: $n = 168$), *Relaxed/Tense* was found to correlate strongly with *Not Scary/Scary* using Pearson's product-moment correlation ($r = 0.65$). However, when the data was split by musical sample ($n = 24$ for each), the correlations were lower: HERMAN low ($r = 0.53$), HERMAN medium ($r = 0.36$), HERMAN high ($r = 0.43$), *Night on a Bald Mountain* ($r = 0.52$), *Pavane* ($r = 0.24$), *Mars* ($r = 0.13$) and *Piccadilly* ($r = 0.46$).

5 CONCLUSIONS

In summary, our results demonstrate that there is a progression of increasing scariness as the input parameter increases, but that that increase is not uniform, as was intended.

HERMAN high was rated scarier than *Pavane* and *Piccadilly*, so *HERMAN high* is perceived as scary to some degree.

HERMAN high was found to be significantly less scary than both *Night on a Bald Mountain* and *Mars*. However, it is perhaps not surprising that it is rated as less scary than these two expertly composed pieces of music: both pieces are xemplars of scary music, deliberately designed to cause conflict with a listener's expectations – and thus make the music feel unsettling.

Similarly, there is a correlation between *HERMAN*'s scariness level and the tension perceived in the music. Again, however, the relation is not as uniform as intended.

Increasing the scariness input does increase the tension in the output but, again, the relation is not as uniform as desired.

Relaxed/Tense correlated very highly with *Not scary/scary* for the whole data set. However, when the data was split by music sample (testing the correlation of the two scales for the subjects within each piece of music), the correlations were less convincing. No real conclusions can be drawn without more detailed analysis.

Overal HERMAN attains some success as a music generator, and the scariness dimension seems to be noticed by human listeners. However, our results suggest that further work is needed to achieve a more uniform scale. Such proposals, however, are outside the scope of the current paper.

6 TOPICS

Computational models; Emotion in music; Aesthetic perception and response; Creativity in music; Music and communication.

7 REFERENCES

1. Brown, R. S. (1994). *Overtones and Undertones*. University of California Press, Berkeley and Los Angeles, CA.

2. Burton, T. (1988). *Beetlejuice*. Warner Bros./Geffen Pictures.

3. de Quincey, A. (1998). *HERMAN*. Master's thesis, Department of Artificial Intelligence, University of Edinburgh.

4. Disney, W. (1940). *Fantasia*. Disney Corporation.

5. Hitchcock, A. (1958). *Vertigo*. Paramount Pictures.

6. Hitchcock, A. (1959). *North by Northwest*. Paramount Pictures.

7. Narmour, E. (1990). *The Analysis and Cognition of Basic Melodic Structures: The Implication-Realistation Model*. The University of Chicago Press.

8. Reiner, R. (1987). *The Princess Bride*. Act III Communications.

9. Robertson, J., de Quincey, A., Stapleford, T., and Wiggins, G. A. (1998). Real-time music generation for a virtual environment. In F. Nack, editor, *Proceedings of the ECAI'98 Workshop on AI/Alife and Entertainment*, Brighton, England.

10. Rothstein, J. (1992). *MIDI : a comprehensive introduction*. Oxford University Press, Oxford.

11. Rutherford, J. (1999). An evaluation of the HERMAN automated music generation system. Undergraduate dissertation, Department of Artificial Intelligence, University of Edinburgh.

12. Schoenberg, A. (1983). *A Theory of Harmony*. Faber & Faber, London. Translated from the original German by Roy E. Carter.

13. Scott, R. (1979). *Alien*. Brandywine Productions Ltd/20th Century Fox.

14. Stapleford, T. (1998). *The Harmony, Melody and Form of HERMAN, a Real Time Music Generation System*. Master's thesis, Department of Artificial Intelligence, University of Edinburgh.