RC 22126 (99116) 27 July 2001 **Computer Science**

IBM Research Report

Affect Computing in Film through Sound Energy Dynamics

Simon Moncrieff, Svetha Venkatesh
Department of Computer Science
Curtin University of Technology
GPO Box U1987, Perth 6845,
W, Australia

Chitra Dorai
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

Affect Computing in Film through Sound Energy Dynamics

Simon Moncrieff
Department of Computer Science
Curtin University of Technology
GPO Box U1987, Perth, 6845, W. Australia
simonm@cs.curtin.edu.au

Chitra Dorai
IBM T. J. Watson Research Center
P.O. Box 704, Yorktown Heights
New York 10598
USA
dorai@watson.ibm.com

Svetha Venkatesh
Department of Computer Science
Curtin University of Technology
GPO Box U1987, Perth, 6845, W. Australia
svetha@cs.curtin.edu.au

ABSTRACT

We develop an algorithm for the detection and classification of *affective* sound events underscored by specific patterns of sound energy dynamics. We relate the portrayal of these events to proposed high level *affect* or emotional coloring of the events. In this paper, four possible characteristic sound energy events are identified that convey well established meanings through their dynamics to portray and deliver certain affect, sentiment related to the horror film genre. Our algorithm is developed with the ultimate aim of automatically structuring sections of films that contain distinct shades of emotion related to horror themes for nonlinear media access and navigation. An average of 82% of the energy events, obtained from the analysis of the audio tracks of sections of four sample films corresponded correctly to the proposed affect. While the discrimination between certain sound energy event types was low, the algorithm correctly detected 71% of the occurrences of the sound energy events within audio tracks of the films analyzed, and thus forms a useful basis for determining affective scenes characteristic of horror in movies.

Categories and Subject Descriptors

H.3.1 **[Information Systems]**: Information Storage and Retrieval—*Multimedia, Content Analysis and Indexing*

1. INTRODUCTION

The manipulation of the sound track of a film has a direct influence on the aesthetic impact of the accompanying visual component of the film. In this paper we examine certain characteristics in sound energy dynamics that are manipulated within the sound tracks of film, and examine the distinctive use of these characteristics in story telling and their impact on viewing experience. We term the occurrences of these characteristics within the film audio, *sound energy events*. We analyze changes in sound energy of the non-literal components of the audio tracks of films, namely sound effects and music including silence. We confine our study of the characteristics of sound energy dynamics at this time to horror films. Four types of sound energy events commonly used in horror films are identified. We hypothesise that the sound energy events correspond to specific *affect* and certain symbolic meanings within the film. The four proposed affects pertaining to examined sound energy events are (i) surprise or alarm, (ii) apprehension, or the emphasis of a significant event, (iii) surprise followed by a sustained alarm, and (iv) building apprehension up to a climax. We have developed a method to detect the four affective sound energy events based on the energy dynamics of the sound tracks of film. Related research in the area of film audio analysis studies the detection of violent sounds in audio [4] and the determination of scene transitions within film using features of the audio track [5].

This work was developed with a similar aim to our previous work [3] of detecting and reliably classifying film scenes based on their audio tracks. In particular, with the goal of automatically determining whether or not the sound energy events are concentrated in those scenes from horror movies that evoke in general horrific feelings among the movie viewers, thus enabling the automatic identification of these scenes within the film. The main contributions of this work lie in the automatic detection of well established audio cues or sound patterns in horror films to enable the subsequent analysis of the relation of occurrences of these events to scene content as a possible means of scene classification. Our work would also allow comparative evaluation of the prevalent and successful use of audio techniques in films.

2. SOUND ENERGY EVENTS

The four sound energy events, which are characterised by specific dynamics of the sound energy in the film audio, form the sound event set used for the detection of the affect conveyed by the film audio. The sound energy events are composed of varying periods of *attack*, *sustain*, and *decay* of sound. The sound energy events are depicted in Figure 1.

Sound Energy Event 1: This is characterised by a large increase in sound energy after a period of low sound energy, with a brief attack time. The attack is followed by a brief or no sustained period and subsequently by a decay in sound energy. This sound dynamic typically evokes surprise or alarm in a scene. The initial low sound energy is used to develop tension, which is subsequently followed by surprise, startling the audience using a sharp increase in sound energy. Either sound effects or music are used to create this dynamic in sound energy.

Sound Energy Event 2: An increase in sound energy over a period of time, characterised by a period of gradual sustained attack, for

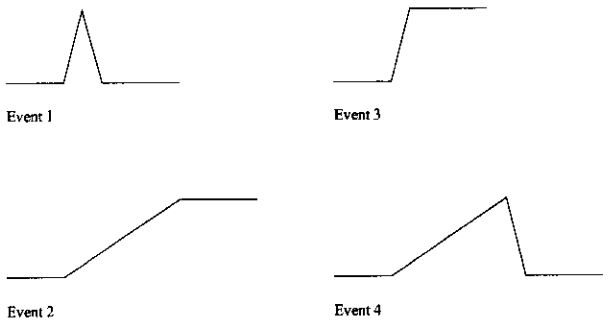

Figure 1: The four proposed sound energy events.

example, a gradual rising of the sound energy of the musical score. The attack is followed by an extended period of sustained sound energy. This sound dynamic is used to convey apprehension, or to emphasise a significant happening in the scene.

Sound Energy Event 3: A large increase in sound energy after a period of low sound energy, with a brief attack time. The attack is followed by an extended sustained period of sound energy. This sound event is similar to event 1, and is used to frighten the audience initially. However, the period of sustained sound energy leads to a different sound dynamic with the initial surprise followed by a period of sustained alarm.

Sound Energy Event 4: An increase in sound energy over a period of time, characterised by a period of gradual sustained attack, for example, a gradual rising of the sound energy of the musical score. The attack is followed by a decay period. This sound dynamic is used to build apprehension up to a climax, at which point the sound energy decays.

The symbolic meaning proposed for the sound events are summarised in Table 1. The essence of the events is characterised mainly by the attack phase of the sound. Energy events 1 and 3 are characterised by a sharp increase in sound energy, while events 2 and 4 are characterised by a more gradual, but sustained, increase in sound energy.

Table 1: Associated affect and symbolic meanings for the four sound energy events examined.

Sound energy event	Affect conveyed/symbolic meaning
Type 1	Startle, surprise, alarm
Type 2	Apprehension or the emphasis of an event
Type 3	Surprise, followed by sustained alarm
Type 4	Builds apprehension to a climax

3. DETECTION OF SOUND ENERGY EVENTS

Audio tracks from mainstream horror movies were used in our study.

3.1 Data

Data was extracted from sections of four films, *The Mummy*, *Pitch Black*, *Aliens*, and *Scream*. Sections were manually selected from each of the four movies, concentrating on horror scenes, which were then concatenated to form continuous segments of audio for analysis of the movies. The concatenated audio from *The Mummy* and *Pitch Black* were split into two separate tracks. The duration of the audio segments for *The Mummy* were 36:55 minutes and 9:01 minutes, 56:16 minutes and 24:45 minutes for *Pitch Black*, 38:22 minutes for *Scream* and 29:25 minutes for *Aliens*. The ground truth was manually determined for each film sound segment by noting starting and ending time indices for all the sound energy events occurring in the segment that corresponded to various affect experienced by both watching the section and listening to the sound. All sound segments were in 44.1kHz 16 bit mono wave sound format. The Speech Filing System [2] was used to calculate a smoothed energy envelope for each signal, the film audio segments, using a hamming window of size $36ms$. Due to the temporal coarseness of the sound energy events to be detected the energy envelope was averaged over 0.1s windows with no overlap in order to summarise the signal energy. The averaged energy envelope for each 0.1s window was then used as the sound energy for the input signal for the determination of the sound energy dynamics.

3.2 Sound Energy Event Detection Algorithm

The algorithm to detect sound energy events consists of two stages. Initially energy components of interest, such as sharp attack and decay, are detected in the sound energy envelope in order to summarise the *sound energy dynamics*. The summary of the dynamical characteristics is then searched for occurrences of the sound energy events of interest.

3.2.1 Computing Sound Dynamics

The dynamics of the sound energy are summarised by assigning one of 7 component energy dynamic labels to each value, or sample, of the sound energy according to the energy dynamics of the section. A section contains varying numbers of samples, 1 or more, with the same sound dynamic component label, indicating the energy dynamics of the signal over a region. For step edges the label is assigned to a single sample. For the remaining dynamical component labels a window of samples was used to determine the sound dynamic behaviour over the window, assigning a label according to dynamics of the window. A tolerance is used to reject the classification of a window of samples if less than a certain % of samples in the window meet the class label's prerequisite conditions. The window size and tolerance value used in the algorithm were determined through experimentation as discussed in Section 3.2.3. The dynamical component labels in order of decreasing precedence are:

Step edge attack: Detected by using first derivative edge detector to detect positive edges in energy [1].
Step edge decay: Detected by using first derivative edge detector to detect negative edges in energy [1].
Slope attack: The sound energy is monotonically increasing over a given window of the input energy envelope, with a degree of tolerance.
Slope decay: The sound energy is monotonically decreasing over a given window of the input envelope, with a degree of tolerance.
Low sound energy: The sound energy over a given window is below a threshold, with a degree of tolerance.
Sustained energy: The variance of the sound energy about the average energy of a given window is below a given threshold.
General: Does not correspond to the above conditions.

3.2.2 Detecting Types of Sound Energy Events

Types of sound energy events are determined by searching the energy dynamics for the sound dynamic components that comprise the events. For the detection of sound energy events, a window used and is characterised by the dominant components in the window. The dominant component is either the step edge attack or step edge decay, determined by which occurs first. If neither step edge is present, the dominant component is determined among the remaining components as one occurring as a majority.

Sound Energy Event 1: A period of low sound energy, followed by a step edge attack or a single window slope attack, subsequently followed by a window of slope decay or step edge decay, within n windows of the attack.
Sound Energy Event 2: Slope attack for a minimum of n windows, followed by a period of sustained or general energy.
Sound Energy Event 3: A period of low sound energy, followed by a step edge attack or a single window slope attack, subsequently followed by a period of n windows of sustained sound energy.
Sound Energy Event 4: Slope attack for a minimum of n windows, followed by a window of slope decay or step edge decay, within n windows of the attack.

3.2.3 *Determining Window Size*

The optimum neighborhood (window) size for the labeling of segment dynamics and for the sound energy event detection was determined by running the detection algorithm on a sample audio track taken from the *The Mummy* and comparing the results with the manually determined ground truth. The best performance was achieved using a window size of 9, corresponding to a window of 0.9s, with a tolerance of 80%, and using a minimum of 2 windows to detect the slope attack for sound energy events 2 and 3.

3.3 Sound Energy Event Detection Results

The algorithm was used to detect the sound energy events for the audio segments detailed in Section 3.1. For each film, the true positives, false positives, and false negatives were determined. The true positives (TP) are the sound energy events detected by the algorithm that were confirmed by the ground truth also to correspond to a symbolic meaning. The false negatives (FN) or misses correspond to the ground truth symbolic energy events that were not detected by the algorithm. The false positives (FP) detected by the algorithm are split into two sets: The FP events detected by the algorithm that display the sound dynamics of the events, yet do not correctly exhibit the affect or the meaning that corresponds to the energy event, are termed as *negative affect* events. The "other" category of the false positives corresponds to sections of the film audio that are erroneously detected as energy events.

Table 2 displays results for the automatic detection of the sound energy events for each of the audio tracks analyzed. While the detection rates of the sound events varied significantly, from 61% for track 1 of *Pitch Black* to 83% for the track 1 of *The Mummy*, an average of 71% of the sound energy events determined in the ground truth were automatically correctly detected.

Table 2: Results of automatic detection of sound energy events.

Movie	TP	FN	FP: *negative affect events*	FP: Other
The Mummy Track1	10	2	5	2
Track2	13	7	2	0
Pitch Black Track1	20	13	11	2
Track2	15	4	1	7
Scream	24	9	5	2
Aliens	13	6	10	12

Table 3 displays the accuracy in labeling the detected true positive sound energy events as one of the four types with a confusion matrix. The labeling accuracy of events 2 and 3 is low. However, the majority of misclassification occurs between events for types 1 and 3, and events for types 2 and 4. While the detection of the attack for each type is accurate, the misclassifications occur due to discrepancies in determining the decay of the sound energy. A further source of error lies in determining the ground truth, especially with borderline cases where there is no absolute distinction, particularly due to the subjective nature of the affect associated with the sound energy events. This is predominantly the case with types 1 and 3.

Table 3: Classification accuracy for the event type detection.

Sound Energy Event	1	2	3	4
1	24		1	
2	2	13		13
3	20	2	13	
4				7

Table 4 details the data support for the 4 sound energy event types and their intended affect. Positive support is the sum of the true positives and false negatives and negative support embodies the number of negative events detected. Overall 82% of the sound energy events support the associated affect and the proposed symbolic meaning. The high percentage of positive support for events, with the exception of sound energy event 1, lends credence to to the hypothesis that the sound energy events indicate affectivity. The positive results for the first event are above 50%, which suggests that the sound energy event may remain useful in detecting concentrated regions of the sound energy events that indicate horrific affect in the film audio tracks.

Table 4: Data support for the sound energy events and their affect.

Sound Event	Positive Support	Negative Support	Positive Support (%)
1	36	17	68
2	54	7	89
3	41	3	93
4	8	2	80

4. CONCLUSION

This paper proposes a method to automatically search the audio track of film for affective sound events characterised by the dynamics of the sound energy of the audio. Four sound energy events are proposed that correspond to specific affect and symbolic meaning in films from the horror genre. While the sound dynamics that characterise the sound energy events do not solely occur when the proposed affect is present in the film, our experimental results from film audio tracks show that the majority of the occurrences of the sound energy events do coincide with the affect evoked.

5. REFERENCES

[1] R. Deriche. Recursively implementing the Gaussian and it's derivatives. In *ICIP'92, Proc. 2nd Singapore Int. Conf. on Image Processing*, pages 263–267, 1992.

[2] M. Huckvale. Speech filing system. Url: www.phon.ucl.ac.uk/resources/sfs/.

[3] S. Moncrieff, C. Dorai, and S. Venkatesh. Detecting indexical signs in film audio for scene interpretation. To appear, ICME 2001.

[4] S. Pfeiffer, F. S., and W. Effelsberg. Automatic audio content analysis. Technical Report TR-96-008, University of Mannheim, April 1996. ftp://pi4.informatik.uni-mannheim.de/ pub/techreports/1996TR-96-008.ps.gz.

[5] H. Sundaram and S.-F. Chang. Determining computable scenes in films and their structures using audio visual memory models. In *ACM Multimedia*, Los Angeles, CA, October 2000.